A Learning-Based Approach to Caching in Heterogenous Small Cell Networks

B. N. Bharath, K. G. Nagananda and H. Vincent Poor, *Fellow, IEEE**

Abstract—A heterogenous network with base stations (BSs), small base stations (SBSs) and users distributed according to independent Poisson point processes is considered. SBS nodes are assumed to possess high storage capacity and to form a distributed caching network. Popular files are stored in local cache of SBS, so that a user can download the desired files from one of the SBSs in its vicinity. The offloading-loss is captured via a cost function that depends on the random caching strategy we propose. The popularity profile of cached content is unknown and estimated using instantaneous demands from users within a specified time interval. An estimate of the cost function is obtained from which an optimal random caching strategy is devised. The training time to achieve an $\epsilon > 0$ difference between the achieved and optimal costs is finite provided the user density is greater than a predefined threshold, and scales as N^2, where N is the support of the popularity profile. A transfer learning-based approach to improve this estimate is proposed. The training time is reduced when the popularity profile is modeled using a parametric family of distributions; the delay is independent of N and scales linearly with the dimension of the distribution parameter.

Index Terms—Caching; small cell networks; popularity profile; transfer learning.

I. INTRODUCTION

The advent of multimedia-capable devices at economical costs has triggered the growth of wireless data traffic at an unprecedented rate. This trend is likely to continue, demanding wireless service providers to reevaluate design strategies for the next generation wireless infrastructure [1]. A promising approach to address this problem is to deploy small cells which can offload a significant amount of data from the macro BS [2]. Doing so, it is expected to lead to cost-effective integration of the existing WiFi and cellular technologies with improved performance of peak data traffic steering policies [3]. However, a potential shortcoming of the small cell infrastructure is that, during peak traffic hours, the backhaul link-capacity requirement to support data traffic is enormously high [4]. Also, the cost incurred in designing high capacity backbone network for small cells is quite discouraging especially from an economic standpoint. Therefore, in the grand scheme of things, small cell-based solutions alone will not suffice to efficiently solve the quality of service requirements associated with peak traffic demands.

A noteworthy development in this direction is to improve the accessibility of data content to users by storing the most popular data files in the *local caches* (intermediate servers such as gateways, routers, *etc.*) of small cell BSs, with the objective of reducing the peak traffic rates. This is commonly referred to as "caching" and has attracted significant attention [5] - [8]. In the next subsection we list a few references, which although by no means exhaustive, fairly indicate the scope and trend of research on caching.

A. Literature review on caching

Caching has received considerable attention in the wireless communications literature. In [9], a two-layer hierarchical strategy termed New Snoop was proposed to cache the unacknowledged packets from mobiles and the BSs to significantly enhance the TCP performance. In [10], a technique based on the concept of content-centric networking was devised for caching in 5G networks, while in [11] caching of video files was proposed by exploiting the redundancy of user requests and storage capacity of mobile devices with *a priori* knowledge of the locations of devices. In [12], the effects of cache size and cached-data popularity on a data access scheme were studied to mitigate the traffic load over the wireless channel. In [13], inner and outer bounds were proposed for the joint routing and caching problem in small cell networks, while in [14] in-network caching was proposed for information-centric networking architecture for faster content distribution in an energy-efficient manner. In-network caching was employed in [15] for content-centric networks using a tool called "contrace" for monitoring and operating the network. The tradeoff between the performance gain of coded caching and delivery delay in video streaming was characterized in [16]. A polynomial-time heuristic solution was proposed in [17] to address the NP-hard optimization problem of maximizing the caching utility of mobile users.

Caching has also made advances in device-to-device (D2D) communications. In [18], a practical method was devised for data caching and content distribution in D2D networks to enhance assisted communications between proximate nodes. In [19], the outage-throughput tradeoff was characterized for D2D nodes, which obtained the desired file from nodes which had that file in its cache. In [20], the conflict between collaboration-distance and interference was identified among D2D nodes to maximize frequency reuse by exploiting distributed storage of cached content. In [21], coded caching was shown to achieve multicast gain in a D2D network, where users had access to linear combinations of packets from cached files. In [22], the throughput scaling laws of random caching, where users with pre-cached information made arbitrary requests for cached files, were studied. New

*B. N. Bharath is with People's Education Society Institute of Technology, Bangalore South Campus, INDIA, E-mail: bharathbn@pes.edu. K. G. Nagananda is with People's Education Society University, INDIA, E-mail: kgnagananda@pes.edu. H. V. Poor is with Princeton University, Princeton, NJ 08544, USA, E-mail: poor@princeton.edu

caching mechanisms by modeling the network as independent Poisson point processes (PPPs) with full knowledge of the popularity profile can be found in [23] - [26], while the most recent results on caching in D2D networks and video content delivery are reported in [27], [28].

Caching has been addressed from an information-theoretic viewpoint as well. In [29], it was shown that when cached-content demand was uniformly distributed, joint optimization of caching and coded multicast delivery significantly improved the gains; this setup was extended to the case of nonuniform distribution on demand and to decentralized setting in [30] and [31], respectively. In [32], coded caching was achieved for content delivery networks with two layers of caches.

B. Main contributions of this paper

In the aforementioned references, the popularity profile of data files was assumed to be known perfectly. In practice, such assumptions cannot be reasonably justified; this was clearly highlighted in [33] - [37], where various learning-based approaches were employed to estimate the popularity profile. On the other hand, estimation procedures result in computational overheads especially in data-intensive realtime multimedia applications. Therefore, given the increasing demand for improving the quality of service for the end users, establishing the theoretical underpinnings of learning-based caching strategies is a topical research problem, and is the main subject of this paper.

In this work, we relax the assumption on *a priori* knowledge of the popularity profile to devise a caching strategy. We consider a heterogenous network where the users, BS and SBSs are assumed to be distributed according to PPPs. Each SBS is assumed to employ a random caching strategy with no caching at the user terminal (see [19]). A protocol model for communications is proposed using which a cost, that captures backhaul link overhead which depends on the popularity profile, is derived. Assuming a Poisson request model, a centralized approach is presented in which the BS computes an estimate of the popularity profile based on the requests observed during the time interval $[0, \tau]$; this estimate is then used in the cost function to optimize the caching probability. Thus, the actual cost incurred differs from the optimal cost, and this difference depends on the number of samples used to estimate the popularity profile. Further, the number of samples collected at the BS depends on the density of the Poisson arrival process and the training time during which the samples are collected. A lower bound on this training time is derived that guarantees a cost that is within $\epsilon > 0$ of the optimal cost. The results are improved using a transfer learning (TL)-based approach wherein samples from other domains, such as those obtained from a social network, are used to improve the estimation accuracy; the minimum number of source domain samples required to achieve a better performance is derived. Finally, we model the popularity profile using a parametric family of distributions (specifically, the Zipf distribution [38]) to analyze the benefits offered.

The following are the main findings of our study:

(i) The training time τ is finite, provided the user density is greater than a predefined threshold.

(ii) τ scales as $N^2 \log N$, where N is the total number of cached data files in the system.

(iii) Employing the TL-based approach, a finite training time can be achieved for *all* user density. In this case, the training time is a function of the "distance" between the probability distribution of the files requested and that of the source domain samples (the notion of distance will be made precise in the proof of Theorem 3).

(iv) When the popularity profile is modeled using a parametric family of distributions, the bound on the training time is independent of N, and scales only linearly with the dimension of the distribution parameter leading to a significant improvement in the performance compared to its nonparametric counterpart.

The problem of periodic caching without the knowledge of the popularity profile, but with access to the demand history, was addressed in [33] and [34], however, the model and objective function considered in our work are different from those presented therein. Learning-based approaches to estimate the popularity profile for devising caching mechanisms have also been reported in [35] - [37]; while caching in femtocell networks without prior knowledge of the popularity distribution was considered in [39], where it was shown that distributed caching was NP-hard and approximation algorithms were proposed for video content delivery. We would like to emphasize that the central focus of this paper is not on deriving new caching mechanisms. Our main contribution is the theoretical analysis of the implications of learning the popularity profile on the training time to achieve an offloading loss which is $\epsilon > 0$ close to the optimal policy. To the best of our knowledge, this is the first instance where an analytical treatment of training time and its relation to the probability distribution function of source domain samples have been reported in the literature on caching. Some preliminary aspects of this work can be found in [40].

In Section II, we present the system model followed by the main engineering problem addressed in the paper. The two methods for estimating the popularity profile and its corresponding training time analysis are developed in Section III. The training time analysis when the popularity profile is modeled as a parametric family of distributions is presented in Section IV. Experimental results are reported in Section V. Concluding remarks are provided in Section VI. The proofs of the theorems are relegated to appendices.

II. SYSTEM MODEL AND PROBLEM STATEMENT

In this section, we present the system model followed by the main problem addressed in the paper. The notation used in the rest of the paper is as follows: Φ_u, Φ_s and Φ_b (λ_u, λ_s and λ_b) denote the points (densities) corresponding to the user, SBSs and BS, respectively; k_x denotes the number of requests in $[0, \tau]$ by the user at x; $X_x^{(l)}$ denotes l^{th} request of the user x; λ_r is the average number of requests per unit time. A heterogenous cellular network is considered where the set $\Phi_u \subseteq \mathbb{R}^2$ of users, the set $\Phi_b \subseteq \mathbb{R}^2$ of BSs, and the set $\Phi_s \subseteq \mathbb{R}^2$ of SBSs are distributed according to independent PPP with density λ_u, λ_b and λ_s, respectively, in the two-dimensional space [41]. Each user independently requests a

data-file of size B bits from the set $\mathcal{F} \triangleq \{f_1, f_2, \ldots, f_N\}$; the popularity of data files is specified by the distribution $\mathcal{P} \triangleq \{p_1, \ldots, p_N\}$, where $\sum_{i=1}^N p_i = 1$ and is assumed to be stationary across time. In a typical heterogenous cellular network, the BS fetches a file using its backhaul link to serve a user. During peak data traffic hours, this results in an information-bottleneck both at the BS as well as in its backhaul link. To alleviate this problem, caching the most popular files (either at the user nodes or at the SBSs) is proposed. The requested file will be served directly by one of the neighboring SBSs depending on the availability of the file in its local cache. The performance of caching depends on the density of SBS nodes, cache size, users' request rate, and the caching strategy. It is assumed that the SBS can cache up to M files, each of length B bits. Each SBS in Φ_s caches its content in an independent and identically distributed (i.i.d.) fashion by generating M indices distributed according to $\Pi \triangleq \{\pi_i : f_i \in \mathcal{F}, \ i = 1, 2, \ldots, N\}$, $\sum_{i=1}^N \pi_i = 1$ (see [19]). One way of generating this is to roll an N sided dice M times in an i.i.d. fashion, where the outcomes correspond to the index of the file to be cached. Although this approach is suboptimal, it is mathematically tractable and the corresponding time complexity serves as a lower bound, albeit pessimistic, for optimal strategies.

We now present a simple communications protocol to determine the set of neighboring SBS nodes for any user in Φ_u. Essentially, we let each SBS at location $y \in \Phi_s$ communicate with a user at location $x \in \Phi_u$ if $\|y - x\| < \gamma$, $(\gamma > 0)$; this condition determines the communication radius. In this protocol, we have ignored the interference constraint. The set of neighbors of the user at location x is denoted

$$\mathcal{N}_x \triangleq \{y \in \Phi_s : \|y - x\| < \gamma\}. \tag{1}$$

A. The main problem addressed in this paper

The user located at $x \in \Phi_u$ requests for a data-file from the set $\mathcal{F}$, with the popularity profile chosen from the probability distribution function $\mathcal{P}$. The requested file will be served directly by one of the neighboring SBS at location $y \in \Phi_s$ depending on the availability of the file in its local cache, and following the protocol described in the previous paragraph. The problem of caching involves minimize the time overhead incurred due to the unavailability of the requested file. Without loss of generality and for ease of analysis, we focus on the performance of a typical user located at the origin, denoted $o \in \Phi_u$. The unavailability of the requested file from a user located at o is given by

$$\mathcal{T}(\Pi, \mathcal{P}) \triangleq \frac{B}{R_0} \mathbb{E} \sum_{i=1}^N [\mathbf{1}\{f_i \notin \mathcal{N}_o\}] \, \mathbf{1}\{f_i \text{ requested}\}, \tag{2}$$

where $\mathcal{N}_o$ is as defined in (1), R_0 is the rate supported by BS to the user, and $\frac{B}{R_0}$ is the time overhead incurred in transmitting the file from BS to the user. Further, we use $f_i \notin \mathcal{N}_o$ to denote the event that the file f_i is not stored in any of the SBSs in $\mathcal{N}_o$. The expectation is with respect to Φ_u, Φ_s and $\mathcal{P}$. The indicator function $\mathbf{1}\{A\}$ is equal to one if event A occurs,

and zero otherwise. We refer to $\mathcal{T}(\Pi, \mathcal{P})$ as the "offloading loss", which we seek to minimize:

$$\min_{\Pi \succeq 0} \quad \mathcal{T}(\Pi, \mathcal{P}) \tag{3}$$

$$\text{subject to } \sum_{i=1}^N \pi_i = 1,$$

where $\pi_i \geq 0$, for $i = 1, \ldots, N$. To solve the optimization setup (3), we need an analytical expression for $\mathcal{T}(\Pi, \mathcal{P})$ which is provided in the following theorem.

Theorem 1: For the caching strategy proposed in this paper, the average offloading loss is given by

$$\mathcal{T}(\Pi, \mathcal{P}) = \frac{B}{R_0} \left[\sum_{i=1}^N \exp\{-\lambda_s \pi \gamma^2 \left[1 - (1 - \pi_i)^M\right]\} p_i \right]. \tag{4}$$

Proof: See Appendix A. ∎

We note that, solving the optimization problem posed in (3) is not the main focus of this paper. We assume that there exists a method to solve the problem posed in (3), and instead focus on analyzing the training time required to obtain a good estimate of the popularity profile that results in an offloading loss which is within ϵ of the optimal offloading loss. Interestingly, although the problem in (3) is non-convex, since it is separable, a bound on the duality gap can be obtained with respect to the solution derived using the Karush-Kuhn-Tucker conditions.

In practice, the popularity profile $\mathcal{P}$ is generally unknown and has to be estimated. Denoting the estimated popularity profile by $\hat{\mathcal{P}} \triangleq \{\hat{p}_1, \ldots, \hat{p}_N\}$, and the corresponding offloading loss by $\mathcal{T}(\Pi, \hat{\mathcal{P}})$, (3) becomes

$$\min_{\Pi \succeq 0} \quad \mathcal{T}(\Pi, \hat{\mathcal{P}}) \tag{5}$$

$$\text{subject to } \sum_{i=1}^N \pi_i = 1,$$

with $\pi_i \geq 0$, for $i = 1, \ldots, N$. Naturally, the solution to (5) differs from that of the original problem (3). Let $\hat{\Pi}^*$ and Π^* denote the optimal solutions to the problems in (3) and (5), respectively, and let the throughput achieved using $\hat{\Pi}^*$ be denoted $\hat{\mathcal{T}}^* \triangleq \mathcal{T}(\hat{\Pi}^*, \mathcal{P})$. **The central theme of this paper is the analysis of the offloading loss difference, *i.e.*, $\hat{\mathcal{T}}^* - \mathcal{T}^*$, where $\mathcal{T}^* \triangleq \mathcal{T}(\Pi^*, \mathcal{P})$ is the minimum offloading loss incurred with perfect knowledge of the popularity profile $\mathcal{P}$.** Theorems 2 - 5 are devoted to this analysis.

III. ESTIMATING THE POPULARITY PROFILE

In this section, we present two methods for estimating the popularity profile and provide the corresponding training time analyses. The efficiency of the estimate $\hat{\mathcal{P}}$ of the popularity profile depends on the number of available data samples, which in turn is related to the number of requests made by the users. We first obtain an expression for the estimate of the popularity profile. We then study, in Section III-A, the minimum training time in obtaining the samples to achieve a desired estimation accuracy $\epsilon > 0$. Finally, in Section III-B, we employ the TL-based approach to improve the bound on

the training time. We begin with the definition of the request model.

Definition 1: (*Request Model*) Each user requests for a file $f \in \mathcal{F}$ at a random time $t \in [0, \infty]$ following an independent Poisson arrival process with density $\lambda_r > 0$.

For notational convenience, the same density is assumed across all the users. The following centralized scheme is used where BS collects the requests from all the users in its coverage area in a time interval $[0, \tau]$ to estimate the popularity profile of the requested files: Let the number of users in the coverage area of BS $z \in \Phi_b$ of radius $R > 0$ be n_R, which is distributed according to a PPP with density λ_u. Let the number of requests made by the user at the location $x \in \{\Phi_u \bigcap \mathbb{B}(0, R)\}$ in the time interval $[0, \tau]$ be k_x, where $\mathbb{B}(0, R)$ is a two-dimensional ball of radius R centered at 0. We assume that requests across the users are known at BS. The requests from the user x is denoted $\mathcal{X}_x \triangleq \{X_x^{(1)}, \ldots, X_x^{(k_x)}\}$, where $X_x^{(l)} \in \{1, \ldots, N\}$ denotes the indices of the files in $\mathcal{F}$, $l = 0, \ldots, k_x$. After receiving $\mathcal{X}_x$, $x \in \{\Phi_u \bigcap \mathbb{B}(0, R)\}$, in the time interval $[0, \tau]$, BS computes an estimate of the popularity profile as follows:

$$\hat{p}_i = \frac{1}{\sum_{x \in \{\mathbb{B}(0,R) \bigcap \Phi_u\}} k_x} \sum_{x \in \{\mathbb{B}(0,R) \bigcap \Phi_u\}} \sum_{l=0}^{k_x} \mathbf{1}\{X_x^{(l)} = i\}, \quad (6)$$

$i = 1, \ldots, N$. Given the number n_R of users in the coverage area of BS, the expression $\sum_{x \in \{\mathbb{B}(0,R) \bigcap \Phi_u\}} k_x$ is a PPP with density $n_R \lambda_r$. Also, $\mathbb{E}\{\hat{p}_i | |\{\Phi_u \bigcap \mathbb{B}(0, R)\}| = n_R\} = p_i$, which leads us to conclude that $\hat{p}_i$ is an unbiased estimator. The estimated popularity profile $\hat{p}_i$ given by (6) is shared with every SBS in the coverage area of BS, and then used in (5) to find the optimal caching probability.

The proposed estimator can be improved by using samples from other related domains, for example, a social network. The term "target domain" is used when samples are obtained only from users in the coverage of BS. In the next subsection we derive the minimum training time τ, corresponding to the estimator in (6), required to achieve the desired estimation accuracy $\epsilon > 0$.

A. *A lower bound on the training time τ*

Theorem 2: For any $\epsilon > 0$, with a probability of at least $1 - \delta$, a throughput of $\hat{\mathcal{T}}^* \leq \mathcal{T}^* + \epsilon$ can be achieved using the estimate in (6) provided

$$\tau \geq \begin{cases} \left\{ \frac{1}{\lambda_r g^*} \log\left(\frac{1}{1 - \frac{1}{\lambda_u \pi R^2} \log \frac{2N}{\delta}}\right) \right\}^+ & \text{if } \lambda_u > \mathcal{L}, \\ \infty & \text{otherwise,} \end{cases} \quad (7)$$

where $\{x\}^+ \triangleq \max\{x, 0\}$, $g^* \triangleq (1 - \exp\{-2\bar{\epsilon}^2\})$, $\mathcal{L} \triangleq \frac{1}{\pi R^2} \log\left(\frac{2N}{\delta}\right)$ and

$$\bar{\epsilon} \triangleq \frac{R_0 \epsilon}{2B \sup_\Pi \sum_{i=1}^N g(\pi_i)}, \quad (8)$$

with $g(\pi_i) \triangleq \exp\{-\lambda_s \pi \gamma^2 \left[1 - (1 - \pi_i)^M\right]\}$.

Proof: See Appendix B. ∎

To achieve a finite training time that results in an estimation accuracy $\epsilon > 0$, the user density λ_u has to be greater than a

threshold. Further insights into (7) are obtained by making the following approximation: $1 - x \leq e^{-x}$ for all $x \geq 0$. This is combined with $\sup_{\Pi : \Pi \succeq 0, \mathbf{1}^T \Pi = 1} \sum_{i=1}^N g(\pi_i) \leq N$ yielding the following lower bound on the training time τ:

$$\tau \geq \frac{2B^2}{\pi R^2 \lambda_u \lambda_r R_0^2 \epsilon^2} N^2 \log\left(\frac{2N}{\delta}\right). \quad (9)$$

The lower bound (9) enables us to make the following observations:

(i) The training time τ to achieve an ϵ-offloading loss difference scales as N^2,
(ii) τ is inversely proportional to (λ_u, λ_r), and
(iii) as the coverage radius increases, the delay decreases as $1/R^2$, and
(iv) as the data-file size B increases, the training time scales as B^2.

The bound in (9) is a lower bound on the training time per request per user, since the offloading loss is derived for a given request per user. There are on an average λ_r requests per unit time per user. Thus, to obtain the training time per user, the offloading loss has to be multiplied by λ_r. This amounts to replacing ϵ by ϵ/λ_r. Therefore, (9) becomes

$$\tau \geq \frac{2B^2 \lambda_r}{\pi R^2 \lambda_u R_0^2 \epsilon^2} N^2 \log\left(\frac{2N}{\delta}\right). \quad (10)$$

It is seen that the training time scales linearly with λ_r. Although the training time per user per request tends to zero as $\lambda_r \to \infty$, the training time per user tends to ∞. This is because the number of requests per unit time approaches ∞, and thus, a small fraction of errors results in an infinite difference in offloading loss leading to an infinite training time. With the increasing demand to provide higher quality of service for the end user, the question of whether it is possible to improve (*i.e.* decrease) the training time τ to achieve the desired estimation accuracy ϵ deserves attention. In the next subsection we show that the lower bound on the training time can indeed be improved by employing a TL-based approach.

B. *Transfer learning to improve the training time*

In practice, the minimum training time required to achieve an estimation accuracy $\epsilon > 0$ can be expected to be very large. An approach to overcome this drawback is to utilize the knowledge obtained from users' interactions with a social community (termed "source domain"). Specifically, by cleverly combining samples from the source domain and users' request pattern (target domain), one can potentially reduce the training time. In fact, the estimation accuracy is indicative of the dependence between the source and target domains. These techniques are commonly referred to as TL-based approaches, and have implications on the training time to achieve a given estimation accuracy. The TL-based approaches were also employed in [36], [37] to negotiate over-fitting problems in estimating the content popularity profile matrix. However, unlike in [36], [37], in this paper we are interested in deriving the minimum training time to achieve a desired performance accuracy. Furthermore, the model we consider is quite different from those considered in [36], [37].

The TL-based approach considered here comprises two sources, namely, source domain and target domain, from which the samples are acquired. An estimate of the popularity profile is obtained in a stepwise manner as follows:

(i) Using target domain samples, the following parameter is computed at BS

$$\hat{S}_i^{(tar)} \triangleq \sum_{x \in \mathbb{B}(0,R) \cap \Phi_u} \sum_{l=0}^{k_x} \mathbf{1}\{X_x^{(l)} = i\}, i = 1, \ldots, N. \quad (11)$$

Recall that k_x is the number of requests made by the user at the location x. The corresponding l^{th} request by the user at the location x in the time interval $[0, \tau]$ is denoted $X_x^{(l)}$, $l = 1, 2, \ldots, k_x$.

(ii) The source domain samples $\mathcal{X}^s \triangleq \{X_1^s, \ldots, X_m^s\}$ are drawn i.i.d. from a distribution $\mathcal{Q}$, where $X_l^s = i$ ($i = 1, \ldots, N$) denotes that the user corresponding to the lth sample has requested the file f_i. The nature of the distribution will be made precise in Proposition 1. Using this, BS computes

$$\hat{S}_i^s \triangleq \sum_{k=1}^{m} \mathbf{1}\{X_k^s = i\}, \; i = 1, 2, \ldots, N. \quad (12)$$

(iii) BS uses (11) and (12) to compute an estimate of $\hat{p}_i^{(tl)}$ (the superscript tl indicates transfer learning) given by

$$\hat{p}_i^{(tl)} = \frac{\hat{S}_i^{(tar)} + \hat{S}_i^s}{\sum_{x \in \{\mathbb{B}(0,R) \cap \Phi_u\}} k_x + m}. \quad (13)$$

Using the estimate given by (13), a lower bound on the training time is obtained as stated in the next theorem.

Theorem 3: Let $g(\pi_i) \triangleq \exp\{-\lambda_s \pi \gamma^2 [1 - (1 - \pi_i)^M]\}$. Then, for any accuracy

$$\epsilon > \frac{2B \sup_\Pi \left\{ \sum_{i=1}^N g(\pi_i) \right\}}{R_0} \|\mathcal{P} - \mathcal{Q}\|_\infty, \quad (14)$$

with a probability of at least $1 - \delta$, a throughput of $\hat{\mathcal{T}}^* \leq \mathcal{T}^* + \epsilon$ can be achieved using the estimate in (13) provided the training time τ satisfies the following condition:

$$\tau \geq \begin{cases} \left\{ \frac{1}{\lambda_r(1 - e^{-2\epsilon_{pq}^2})} \log\left(\frac{1}{1-\Lambda}\right) \right\}^+, & \text{if } \lambda_u > \rho, \\ \infty, & \text{otherwise,} \end{cases} \quad (15)$$

where $\rho \triangleq \frac{1}{\pi R^2} \left(\log \frac{2N}{\delta} - 2\epsilon_{pq}^2 m \right)$, $\epsilon_{pq} \triangleq \bar{\epsilon} - \|\mathcal{P} - \mathcal{Q}\|_\infty$, $\Lambda \triangleq \frac{1}{\lambda_u \pi R^2} \left(\log \frac{2N}{\delta} - 2\epsilon_{pq}^2 m \right)$, and $\bar{\epsilon} \triangleq \frac{R_0 \epsilon}{2B \sup_\Pi \{ \sum_{i=1}^N g(\pi_i) \}}$.

Proof: See Appendix C. ∎

From Theorem 3, we see that under suitable conditions the TL-based approach performs better than the source domain sample-based agnostic approach. The following inferences are drawn:

(1) The minimum user density to achieve a finite delay is reduced by a positive offset $2\epsilon_{pq}^2 m$. In fact, for $m > \frac{\log\left(\frac{2N}{\delta}\right)}{2(\bar{\epsilon} - \|\mathcal{P} - \mathcal{Q}\|_\infty)^2}$, a finite delay can be achieved for all user density which provides a significant advantage.

(2) The finite delay achieved is smaller compared to source domain sample-based agnostic approach for large enough

source samples, and the distributions are "close." This is made precise in the following proposition, and a detailed discussion is provided in Section V.

Proposition 1: For any $\epsilon > 0$ and $\delta \in [0, 1]$, the TL-based approach performs better than the source sample-based agnostic approach provided the number m of source samples satisfies $m \geq \frac{1}{2\epsilon_{pq}^2} \left[\log\left(\frac{2N}{\delta}\right) - F \right]^+$, and the distributions satisfy the following condition:

$$\|\mathcal{P} - \mathcal{Q}\|_\infty < \frac{\epsilon R_0}{2B \lambda_u \pi \gamma^2 N}, \quad (16)$$

where $F \triangleq \lambda_u \pi R^2 \left(1 - \exp\left\{ \frac{1 - e^{-2\epsilon_{pq}^2}}{1 - e^{-2\bar{\epsilon}^2}} \log(1 - \mathcal{L}) \right\} \right)$ and $\mathcal{L} \triangleq \frac{1}{\lambda_u \pi R^2} \log\left(\frac{2N}{\delta}\right)$.

In fact, (16) provides the guiding principle to decide if the samples drawn from the distribution $\mathcal{Q}$ should be used to estimate the distribution $\mathcal{P}$. In general, the distance between the distributions has to be estimated from the available samples (relative to the distribution on $\mathcal{P}$).

An estimate of the popularity profile can also be obtained by linearly combining its estimates obtained from the source domain and target domain samples. In particular, we have

$$\hat{p}_i = \alpha \hat{p}_i^{(s)} + (1 - \alpha) \hat{p}_i^{(t)}, \quad (17)$$

where $\hat{p}_i^{(s)}$ and $\hat{p}_i^{(t)}$ are the estimates of the popularity profile obtained from the source domain samples and the target domain samples, respectively. The estimates are given by

$$\hat{p}_i^{(t)} = \frac{\hat{S}_i^{(tar)}}{\sum_{u \in \{\mathbb{B}(0,R) \cap \Phi_u\}} k_u}, \quad (18)$$

$$\hat{p}_i^{(s)} = \frac{\hat{S}_i^s}{m}. \quad (19)$$

Note that, in this case the coefficients are independent of the realization of the network. For the estimate proposed in (17), we have the following result:

Theorem 4: For any accuracy

$$\epsilon > \frac{2B \sup_\Pi \left\{ \sum_{i=1}^N g(\pi_i) \right\}}{R_0} \|\mathcal{P} - \mathcal{Q}\|_\infty, \quad (20)$$

with a probability of at least $1 - \delta$, a throughput of $\hat{\mathcal{T}}^* \leq \mathcal{T}^* + \epsilon$ can be achieved using the estimate in (17) provided the training time τ satisfies the condition specified by (21) at the top of the next page, where

$$\rho_{\text{thresh}} \triangleq \frac{1}{\pi R^2} \left(\log\left(\frac{2N}{\delta}\right) + \log\left\{ \frac{1}{1 - \left(\frac{2N}{\delta}\right) \exp\{-2\bar{\omega}^2 m\}} \right\} \right),$$

$\bar{\epsilon} \triangleq \frac{R_0 \epsilon}{2B \sup_\Pi \{ \sum_{i=1}^N g(\pi_i) \}}$, $g_t^* \triangleq \left(1 - \exp\{-2\eta^2\} \right)$, $g(\pi_i) \triangleq \exp\{-\lambda_s \pi \gamma^2 [1 - (1 - \pi_i)^M]\}$ and $\omega \triangleq \frac{\bar{\epsilon} - (1 - \alpha)\eta}{\alpha} > 0$. This is valid for all $0 < \alpha < \min\left\{\frac{\bar{\epsilon}}{G}, 1\right\}$ and $0 \leq \eta < \frac{\bar{\epsilon} - \alpha G}{1 - \alpha}$, where $G \triangleq \|\mathcal{P} - \mathcal{Q}\|_2 + \sqrt{\frac{1}{2m} \log \frac{2N}{\delta}}$.

Proof: See Appendix D. ∎

$$\tau \geq \begin{cases} \frac{1}{\lambda_r g_t^*} \log \left[\cfrac{1}{1 - \frac{1}{\lambda_u \pi R^2} \left(\log\left(\frac{2N}{\delta}\right) + \log\left\{ \cfrac{1}{1 - \left(\frac{2N}{\delta}\right) \exp\{-2\bar{\omega}^2 m\}} \right\} \right)} \right], & \text{if } \lambda_u > \rho_{\text{thres}}, \\ \infty, & \text{otherwise,} \end{cases} \tag{21}$$

IV. PARAMETERIZED FAMILY OF POPULARITY PROFILE

In the previous sections, no structure was imposed on the popularity profile. In practice, the popularity profile is modeled using a ==parametric family of distributions== such as the Zipf distribution [38], which, with a one-dimensional parameter $\Theta \in \mathbb{R}$, is specified by $p_{\Theta,i} = \frac{1/i^\Theta}{\sum_{j=1}^N 1/j^\Theta}, i = 1, 2, \ldots, N$. To obtain an estimate of the Zipf distribution it suffices to find the parameter Θ; estimating a single parameter requires fewer samples which can potentially reduce the training time. We now derive bounds on the training time when the popularity profile belongs to a parametric family of distributions. We begin with the following assumption:

Assumption 1: Let the family of parameterized popularity distributions be defined by $\mathcal{P} = \{\mathcal{P}_\Theta : \Theta \subseteq [a,b]^d, a < b\}$. Further, for all $\Theta \subseteq [a,b]^d$, $\mathcal{P}_\Theta$ satisfies $\sum_{i=1}^N \|\partial_\Theta p_{\Theta,i}\|_2 < C$, where $C < \infty$ is independent of N, and $\partial_\Theta p_{\Theta,i} \in \mathbb{R}^d$ denotes the sub-differential of $p_{\Theta,i}$.[1] For example, the Zipf distribution $p_{\Theta,i} = \frac{1/i^\Theta}{\sum_{j=1}^N 1/j^\Theta}, i = 1, 2, \ldots, N$ satisfies this property.

Let the true underlying parameter be $\Theta \triangleq \{\Theta_1, \ldots, \Theta_d\}$. Note that $\Theta_j \in [a,b]$ for all $j = 1, 2, \ldots, d$. Let the BS observe n_p (number of requests) i.i.d samples $(X_{t,1}, X_{t,2}, \ldots, X_{t,n_p}) \in \mathcal{X}_t^{n_p}$ drawn from the distribution $\mathcal{P}_\Theta$.[2] Also, let $\hat{\Theta}_{n_p}^{(i)} \triangleq \left(\hat{\Theta}_{n_p,i,1}, \hat{\Theta}_{n_p,i,2}, \ldots, \hat{\Theta}_{n_p,i,d} \right) \in \mathbb{R}^d$, $i = 1, 2, \ldots, n_p$ denote the estimate of Θ, based on a single observation, i.e., $\hat{\Theta}_{n_p}^{(i)} = f(X_{t,i})$, $i = 1, 2, \ldots, n_p$, where $f : \mathcal{X}_t \to [a,b]^d$ is an unbiased estimator of Θ. In the above, n_p denotes the number of requests made by the users corresponding to the BS z in a time interval of $[0, \tau]$. Since $f(.)$ is an unbiased estimator of Θ, we have $\mathbb{E}\left\{ \hat{\Theta}_{n_p}^{(i)} | \Theta \right\} = \Theta$ for all $i = 1, 2, \ldots, n_p$. The estimate of Θ using n_p samples is obtained as follows:

$$\hat{\Theta}_{n_p} = \frac{1}{n_p} \sum_{j=1}^{n_p} \hat{\Theta}_{n_p}^{(j)}. \tag{22}$$

Note that $\hat{\Theta}_{n_p} \triangleq (\hat{\Theta}_{n_p,1}, \hat{\Theta}_{n_p,2}, \ldots, \hat{\Theta}_{n_p,d})$, is also an unbiased estimator of Θ, i.e., $\mathbb{E}\left\{ \hat{\Theta}_{n_p} | \Theta, n_p \right\} = \Theta$. The following theorem provides a bound on the time complexity for a family of parameterized popularity profile satisfying **Assumption 1**.

Theorem 5: For the family $\mathcal{P}_\Theta$ satisfying **Assumption 1**, and given the estimator $\hat{\Theta}_{n_p}$, $\hat{\mathcal{T}}^* \leq \mathcal{T}^* + \epsilon$ for every $\epsilon > 0$ with probability at least $1 - \delta$ if

$$\tau > \frac{1}{\lambda_r(1 - e^{-\sigma^2})} \log \left(\frac{1}{1 - \frac{1}{\lambda_u \pi R^2} \log \frac{2d}{\delta}} \right), \tag{23}$$

[1] Here, Θ and the sub-differential are assumed to be row vectors.
[2] Here, t is used to denote the target domain samples.

for $\lambda_u > \frac{1}{\pi R^2} \log \frac{2d}{\delta}$, otherwise $\tau = \infty$, where $\sigma^2 \triangleq \frac{2\Omega^2}{dC^2(b-a)^2}$ and $\Omega \triangleq \frac{R_0 \epsilon}{2B}$.

Proof: See Appendix E. ∎

From (23), we see that the bound on the training time is independent of N, and from a scaling perspective, the training time scales with d, λ_r and λ_u. This amounts to a significant improvement compared to the nonparametric model studied in the previous sections of this paper, where the training time is shown to scale as $N^2 \log N$. A natural extension is to utilize the knowledge obtained from user's interactions with a social community, namely, the source domain samples. In the next subsection, we analyze the time complexity bound employing the TL-based approach for popularity profiles modeled using a parametric family of distributions.

A. ==Transfer Learning== for Parametric Models

In this subsection, we derive a lower bound on the training time when the BS has access to the source domain samples along with the target domain samples. Let the source domain samples $(X_{s,1}, X_{s,2}, \ldots, X_{s,m}) \in \mathcal{X}_s^m$ drawn i.i.d from $\mathcal{P}_{\Theta_s}$, where $\Theta_s \in \mathbb{R}^d$. Further, as before, we assume that $\exists f : \mathcal{X}_s \to \mathbb{R}^d$, an unbiased estimate of Θ_s. As before, let the BS observe n_p i.i.d. target domain samples from $\mathcal{X}_t^{n_p}$ drawn from $\mathcal{P}_\Theta$. An estimate of Θ based on the available source and target domain samples is obtained as follows:

(i) Using the source domain samples an estimate of Θ_s, denoted $\hat{\Theta}_s$, is obtained in manner similar to that of target domain parameter Θ as explained earlier in this section.

(ii) Using the target domain samples, an estimate of Θ denoted $\hat{\Theta}_t$ is obtained as in (22).

(iii) The two estimates are fused to get an estimate of Θ as $\hat{\Theta}_{tl} \triangleq \lambda \hat{\Theta}_t + (1-\lambda)\hat{\Theta}_s$, where $\lambda \in [0,1]$ will be described shortly.

Theorem 6: For the family $\mathcal{P}_{\Theta'}$ satisfying **Assumption 1**, and given the estimator $\hat{\Theta}_{tl} \triangleq \lambda \hat{\Theta}_t + (1-\lambda)\hat{\Theta}_s$, we have $\hat{\mathcal{T}}^* \leq \mathcal{T}^* + \epsilon$ for every $\epsilon > 0$ with a probability of at least $1 - \delta$ if the condition specified by (24) at the top of the next page is satisfied, for $\lambda_u > \frac{1}{\pi R^2} \left(\log \frac{2d}{\delta} + \log \frac{1}{1 - \frac{2d}{\delta} \exp\left\{ -\frac{2m(\bar{\Omega} - \|\Theta - \Theta_s\|_2)^2}{(b-a)^2} \right\}} \right)$. This holds for all $D_t < \frac{\Omega}{C} - \lambda \bar{G}$ and $0 < \lambda < \min\left\{ \frac{\Omega}{CG}, 1 \right\}$. Here, $\bar{\Omega} \triangleq \frac{1}{\lambda} \left(\frac{\Omega}{C} - D_t \right)$, $\sigma_t^2 \triangleq \frac{2D_t^2}{d(1-\lambda)^2(b-a)^2}$, $\Omega \triangleq \frac{\tilde{\epsilon}}{\sup_i g(\pi_i)}$, $\tilde{\epsilon} \triangleq \frac{R_0 \epsilon}{2B}$, and $\bar{G} := \|\Theta - \Theta_s\|_2 + (b-a)\sqrt{\frac{1}{2m} \log \frac{2d}{\delta}}$.

Proof: See Appendix F. ∎

It is important to note that the aforementioned bound is independent of N. In the following section, we provide numerical results to get further insights into the expressions derived in the paper.

$$\tau \geq \frac{1}{\lambda_r(1-e^{-\sigma_t^2})} \log \left(\frac{1}{1 - \frac{1}{\lambda_u \pi R^2}\left(\log \frac{2d}{\delta} + \log \frac{1}{1 - \frac{2d}{\delta}\exp\left\{-\frac{2m(\bar{\Omega} - \|\Theta - \Theta_s\|_2)^2}{(b-a)^2}\right\}}\right)} \right) \tag{24}$$
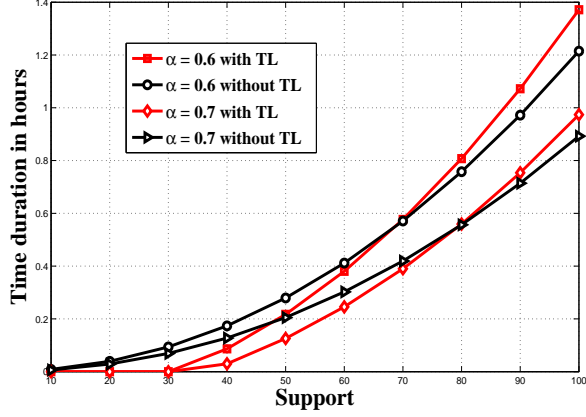

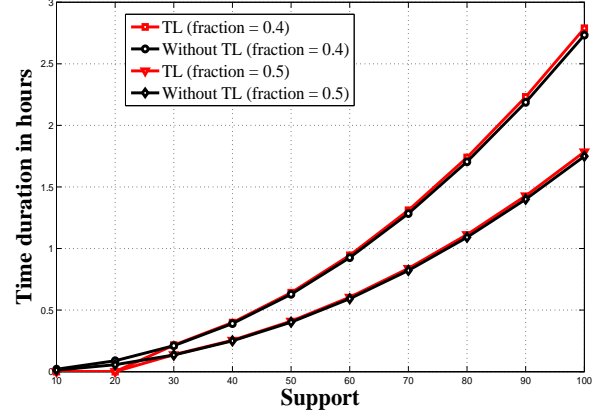Fig. 1: Training duration versus N, corresponding to Theorems 2 and 3.

Fig. 2: Training duration versus N, corresponding to Theorem 4.

V. NUMERICAL RESULTS

In this section, we perform experiments to provide insights into the results derived in the previous sections. The parameter values used for simulations are as follows: $B = 10^7$ bits, $R_0 = 10^6$ bits/s, $\gamma = 100$m, $\lambda_u = 0.001$ nodes/m^2, $\lambda_r = 1/360$ requests/s, $\lambda_s = 10^{-5}$ nodes/m^2, $\delta = 0.02$, $R = 2$ Km, $m = 10^5$ samples, $\beta_e = 0.6$ and $\beta_l = 0.2$. ϵ is chosen as a fraction of a lower bound on the offloading loss, *i.e.*, $\mathcal{T}(\Pi, \mathcal{P}) \geq \frac{B}{R_0}\exp\{-\lambda_s\pi\gamma^2\}$. In particular, $\epsilon = $ fraction $\times \frac{B}{R_0}\exp\{-\lambda_s\pi\gamma^2\}$. Further, $\|\mathcal{P} - \mathcal{Q}\|_\infty = 0.1\left(\frac{\epsilon R_0}{2BN}\right)$ which for the above parameters is of the order of $10/N$.

Fig. 1 shows a plot of the lower bounds on the training duration obtained in Theorems 2 and 3 as a function of the support N. It is seen that, for $N \leq 70$ the TL-based approach provides significant performance improvement. However, for $N > 70$, the performance of the TL-based approach degrades compared to the approach which used only the source domain samples (and, hence, can be called agnostic). This suggests that for larger values of N, the estimate of the popularity profile obtained using (17) performs poorly due to incorrect fusion of the estimates obtained from source and target domains. Fig. 2 shows the plots of the lower bound in Theorem 4 corresponding to the estimate obtained by a fixed linear combination of the source and target estimates (see (17)). As seen in the figure, it does not bring any performance improvement and in fact sometimes performs poorly compared to the source domain agnostic approach. This is because the fixed linear combination does not have the flexibility to adapt to different realizations of the network, proving the sub-optimality of the estimate in (17) compared to that in (13). It is also seen that the coefficients used in the estimate that
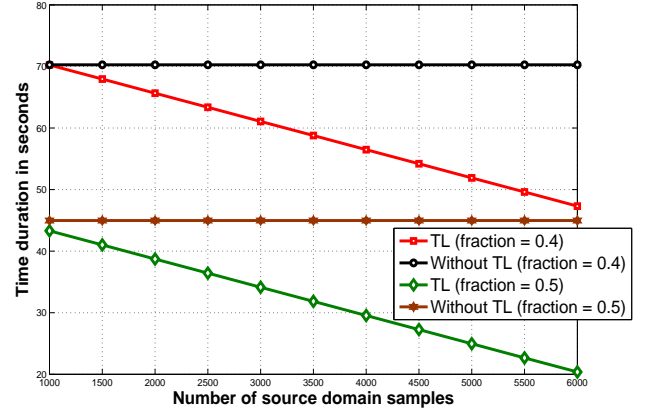

Fig. 3: Training duration versus m for a fixed $N(=10)$.

adapts to the varying realizations of the network as in (17) is beneficial.

Fig. 3 shows a plot similar to that in Fig. 1 but with $N = 10$ and varying m. It can be seen that the TL-based approach performs better for all $m \geq 1000$ demonstrating its applicability in practice. As seen, the performance is better for higher values of the fraction which corroborates intuition. Fig. 4 also shows a plot of time duration versus m for a fixed $N = 10$. It can be seen that the estimate in (17) outperforms the agnostic approach; however, this is observed at very high values of source domain samples ($m = 10500$ and $m = 16000$ for fraction $= 0.5$ and fraction $= 0.4$, respectively). Thus, although the TL-based approach using the estimate (17) has some benefits, it is not desirable for practical applications.

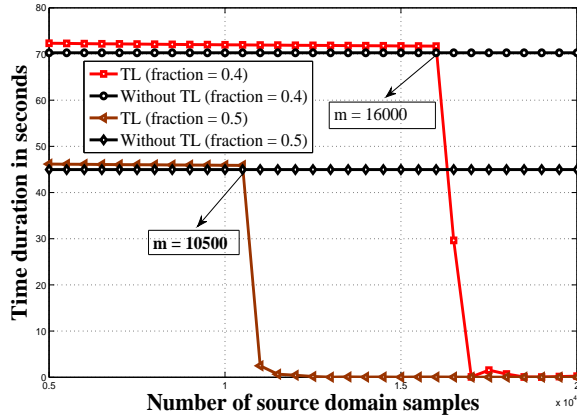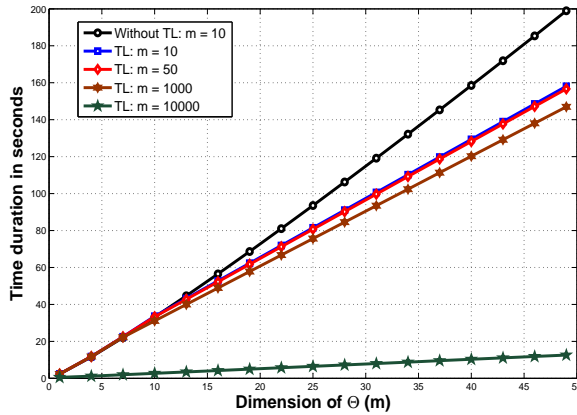Fig. 4: Training duration versus m for a fixed $N(=10)$. Comparing TL-based and agnostic approaches.

Fig. 5: Training duration versus Θ when the popularity profile is modeled using a parametric family of distributions. $C = 2$, fraction $= 0.6$, $\|\Theta - \Theta_s\|_2 = 0.1$, $(b-a) = 0.5$.

The main benefits of the TL-based approach are shown in Fig. 5 for the parametric family of popularity profile. It can be seen that the TL-based approach performs significantly better than source domain agnostic approach for values of m as low as 10. This is because the number of parameters to be estimated scales with the dimension of Θ rather than with the support. In particular, as d increases the training duration also increases, which is quite expected. However, the delay scales only linearly in d as compared to quadratic scaling experienced in the nonparametric method.

VI. CONCLUDING REMARKS

In this paper, we have considered a learning-based approach to caching in heterogenous small cell networks. In particular, we have studied situations in which the popularity profile for caching content is estimated at the BS using the available instantaneous demands from users in a time interval $[0, \tau]$. We have shown that a training time τ to achieve an $\epsilon > 0$ difference between the achieved cost and the optimal cost is finite, provided the user density is greater than a threshold;

τ was shown to scale as the square of the support of the popularity profile. A TL-based approach has been proposed to estimate the popularity profile, and a condition has been derived under which it performs better than the target domain sample only based approach. Although the TL-based approach performs better, the error that is achieved in (20) depends on $\|\mathcal{P} - \mathcal{Q}\|_\infty$, suggesting that the lower the distance between the two distributions the better the TL scheme performs. From Proposition 1, the benefits of using target domain samples can only be realized with the knowledge of the distance $\|\mathcal{P} - \mathcal{Q}\|_\infty$. The main benefit of the TL-based approach is recognized when the popularity profile is modeled using a parametric family of distributions. In this case, the delay is independent of N and scales only linearly with the dimension of the distribution parameter. In practice, caching depends on several factors such as the scheduling scheme used, which in turn depends on the channel conditions, QoS requirements, *etc*. We assume that if the requested file is present in one (or more) of the neighboring SBSs, the transmissions are scheduled within a tolerable time frame, which can be relaxed and can be thought of as an abstraction of the scheduling scheme employed. If the file is not present, regardless of the scheduling policy, the file cannot be served locally. Hence, the approach that we have leads to a lower bound, albeit pessimistic, on the training time. Thus, even under pessimistic situations, the training time scales as N^2 for achieving an offloading loss that is $\epsilon > 0$ away from the optimal offloading loss.

REFERENCES

[1] A. Furuskar, J. Charles, M. Frodigh, S. Jeux, M. Sayed Hassan, A. Saadani, A. Stidwell, J. Soder, and B. Timus, "Refined statistical analysis of evolution approaches for wireless networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 5, pp. 2700 – 2710, May 2015.

[2] S.-F. Chou, T.-C. Chiu, Y.-J. Yu, and A.-C. Pang, "Mobile small cell deployment for next generation cellular networks," in *Proc. IEEE Global Commun. Conf.*, Dec. 2014, pp. 4852–4857.

[3] M. Bennis, M. Simsek, A. Czylwik, W. Saad, S. Valentin, and M. Debbah, "When cellular meets WiFi in wireless small cell networks," *IEEE Commun. Magazine*, vol. 51, no. 6, pp. 44–50, Jun. 2013.

[4] J. Kim, C. Jeong, H. Yu, and J. Park, "Areal capacity limit on the growth of small cell density in heterogeneous networks," in *Proc. IEEE Global Commun. Conf.*, Dec. 2014, pp. 4263–4268.

[5] Y.-B. Lin, W.-R. Lai, and J.-J. Chen, "Effects of cache mechanism on wireless data access," *IEEE Trans. Wireless Commun.*, vol. 2, no. 6, pp. 1247–1258, Nov. 2003.

[6] E. Baştuğ, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Commun. Magazine*, vol. 52, no. 8, pp. 82–89, Aug. 2014.

[7] H. R. Sadjadpour, "A new design for Information Centric Networks," in *Proc. Conf. Inf. Sciences Syst.*, Mar. 2014, pp. 1–6.

[8] 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.

[9] J.-H. Hu, G. Feng, and K. Yeung, "Hierarchical cache design for enhancing TCP over heterogeneous networks with wired and wireless links," *IEEE Trans. Wireless Commun.*, vol. 2, no. 2, pp. 205–217, Mar. 2003.

[10] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. Leung, "Cache in the air: exploiting content caching and delivery techniques for 5G systems," *IEEE Trans. Wireless Commun.*, vol. 52, no. 2, pp. 131–139, Feb. 2014.

[11] N. Golrezaei, P. Mansourifard, 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.

[12] G. Lee, I. Jang, S. Pack, and X. Shen, "FW-DAS: Fast wireless data access scheme in mobile networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 8, pp. 4260–4272, Aug. 2014.

[13] K. Poularakis, G. Iosifidis, and L. Tassiulas, "Approximation algorithms for mobile data caching in small cell networks," *IEEE Trans. Commun.*, vol. 62, no. 10, pp. 3665–3677, Oct. 2014.

[14] C. Fang, F. Yu, T. Huang, J. Liu, and Y. Liu, "A survey of energy-efficient caching in information-centric networking," *IEEE Commun. Magazine*, vol. 52, no. 11, pp. 122–129, Nov. 2014.

[15] H. Asaeda, K. Matsuzono, and T. Turletti, "Contrace: a tool for measuring and tracing content-centric networks," *IEEE Commun. Magazine*, vol. 53, no. 3, pp. 182–188, Mar. 2015.

[16] R. Pedarsani, M. A. Maddah-Ali, and U. Niesen, "Online coded caching," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2014, pp. 1878–1883.

[17] Y. Guan, Y. Xiao, H. Feng, C.-C. Shen, and L. J. Cimini, "MobiCacher: Mobility-aware content caching in small-cell networks," in *Proc. IEEE Global Commun. Conf.*, Dec. 2014, pp. 4537–4542.

[18] A. Pyattaev, O. Galinina, S. Andreev, M. Katz, and Y. Koucheryavy, "Understanding practical limitations of network coding for assisted proximate communication," *IEEE J. Select. Areas Commun.*, vol. 33, no. 2, pp. 156–170, Feb. 2015.

[19] 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.

[20] N. Golrezaei, A. G. Dimakis, and A. F. Molisch, "Wireless device-to-device communication with distributed caching," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2012, pp. 2781–2785.

[21] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of distributed caching in D2D wireless networks," in *Proc. IEEE Inf. Theory Workshop*, Sep. 2013, pp. 1–5.

[22] ——, "Fundamental limits of caching in wireless D2D networks," *IEEE Trans. Inf. Theory*, vol. 62, no. 2, pp. 849–869, Jan. 2016.

[23] 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.

[24] C. Yang, Y. Yao, Z. Chen, and B. Xia, "Analysis on cache-enabled wireless heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 131–145, Jan. 2015.

[25] S. H. Chae, J. Y. Ryu, T. Q. S. Quek, and W. Choi, "Cooperative transmission via caching helpers," in *Proc. IEEE Global Commun. Conf.*, Dec. 2015.

[26] S. Tamoor-ul Hassan, M. Bennis, P. H. J. Nardelli, and M. Latva-Aho, "Modeling and analysis of content caching in wireless small cell networks," in *Proc. IEEE Int. Symp. Wireless Commun. Syst.*, Aug. 2015.

[27] L. Zhang, M. Xiao, G. Wu, and S. Li, "Efficient scheduling and power allocation for D2D-assisted wireless caching networks," 2015, IEEE Trans. Commun. submitted. [Online]. Available: http://arxiv.org/pdf/1509.06932v3.pdf

[28] J. Li, M. Xiao, W. Chen, and X. Liu, "Efficient video pricing and caching in heterogeneous networks," *IEEE Trans. Vehicular Tech.*, 2015, to appear.

[29] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.

[30] U. Niesen, C. Beilken, and M. A. Maddah-Ali, "Coded caching with nonuniform demands," in *Proc. IEEE Conf. Comp. Commun. Workshop*, Apr. 2014, pp. 221–226.

[31] M. Maddah-Ali and U. Niesen, "Decentralized coded caching attains order-optimal memory-rate tradeoff," *IEEE/ACM Trans. Networking*, vol. 23, no. 4, pp. 1029–1040, Aug. 2015.

[32] N. Karamchandani, U. Niesen, M. A. Maddah-Ali, and S. N. Diggavi, "Hierarchical coded caching," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2014, pp. 2142–2146.

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

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

[35] A. Sengupta, S. Amuru, R. Tandon, R. M. Buehrer, and T. C. Clancy, "Learning distributed caching strategies in small cell networks," in *Proc. IEEE Int. Symp. Wireless Commun. Syst.*, Aug. 2014, pp. 917–921.

[36] E. Baştuğ, M. Bennis, and M. Debbah, "Anticipatory caching in small cell networks: A transfer learning approach," in *Workshop Anticipatory Net.*, Germany, Sep. 2014.

[37] ——, "A transfer learning approach for cache-enabled wireless networks," in *Int. Symp. Model. Opt. Mobile, Ad Hoc Wireless Net. (WiOpt)*, May 2015, pp. 161–166.

[38] J. Llorca, A. M. Tulino, K. Guan, and D. C. Kilper, "Network-coded caching-aided multicast for efficient content delivery," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2013, pp. 3557–3562.

[39] N. Golrezaei, K. Shanmugam, A. Dimakis, A. Molisch, and G. Caire, "Femto caching: Wireless video content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

[40] B. N. Bharath and K. G. Nagananda, "Caching with unknown popularity profiles in small cell networks," in *Proc. IEEE Global Commun. Conf.*, Dec. 2015. [Online]. Available: http://arxiv.org/pdf/1504.03632v1.pdf

[41] 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.

[42] V. Vapnik and A. Chervonenkis, "Theory of pattern recognition," *Moscow (in Russian), German Translation: Theorie der Zeichenerkennung, Akademie Verlag, Berlin*, 1979.

[43] L. Devroye, L. Gyorfi, and G. Lugosi, *A Probability Theory of Pattern Recognition.* Springer, 2014.

APPENDIX A
PROOF OF THEOREM 1

The expression on the RHS in (2), $\mathbb{E}\sum_{i=1}^{N}\mathbf{1}\{f_i \notin \mathcal{N}_o\}\mathbf{1}\{f_i \text{ requested}\}$, can be written as

$$\mathbb{E}_{n_s}\sum_{i=1}^{N}\mathbb{E}\mathbf{1}\{f_i \text{ requested}\}\Pr\{f_i \notin \mathcal{N}_o|\ |\mathcal{N}_o| = n_s\}$$

$$\overset{(a)}{=} \mathbb{E}_{n_s}\sum_{i=1}^{N}[\Pr\{f_i \notin a, \text{ for } \mathbf{any}\ a \in \mathcal{N}_o\}]^{n_s}p_i$$

$$\overset{(b)}{=} \sum_{i=1}^{N}\mathbb{E}_{n_s}(1-\pi_i)^{n_s M}p_i$$

$$\overset{(c)}{=} \mathbb{E}\sum_{i=1}^{N}\sum_{j=0}^{\infty}(1-\pi_i)^{jM}e^{\{-\lambda_s\pi\gamma^2\}}\frac{(\lambda_s\pi\gamma^2)^j}{j!}$$

$$= \sum_{i=1}^{N}\exp\{-\mathcal{U}\}p_i,$$

where $\mathcal{U} \triangleq \lambda_s\pi\gamma^2\left[1-(1-\pi_i)^M\right]$. In the above exposition, (a) follows from the fact that the proposed random caching scheme is independent across users, (b) is due to the fixed cache size (M), and (c) follows since n_s is a PPP with mean $\lambda_s\pi\gamma^2$, where n_s is the number of SBSs in a circular area of radius γ. This completes the proof of Theorem 1. ∎

APPENDIX B
PROOF OF THEOREM 2

For any $\epsilon > 0$, the inequality $\Pr\{\hat{\mathcal{T}}^* \geq \mathcal{T}^* + \epsilon\} \leq \Pr\left\{2\sup_{\mathbf{1}\succeq\Pi\succeq\mathbf{0}:\mathbf{1}^T\Pi=1}|\Delta\mathcal{T}| > \epsilon\}\right\}$ is proved, where $\Delta\mathcal{T} \triangleq \mathcal{T}(\Pi,\hat{\mathcal{P}}) - \mathcal{T}(\Pi,\mathcal{P})$. And, $\hat{\mathcal{T}}^* - \mathcal{T}^*$ can be written as (see [42])

$$\hat{\mathcal{T}}^* - \inf_{\Pi}\mathcal{T}(\Pi,\mathcal{P}) \leq \hat{\mathcal{T}}^* - \hat{\mathcal{T}} + \sup_{\Pi}\left|\mathcal{T}(\Pi,\hat{\mathcal{P}}) - \mathcal{T}(\Pi,\mathcal{P})\right|$$

$$\leq 2\sup_{\Pi}\left|\mathcal{T}(\Pi,\hat{\mathcal{P}}) - \mathcal{T}(\Pi,\mathcal{P})\right|,$$

where $\hat{\mathcal{T}} \triangleq \mathcal{T}(\Pi, \hat{\mathcal{P}})$, thus proving the inequality. Substituting for $\mathcal{T}(\Pi, \hat{\mathcal{P}})$ and $\mathcal{T}(\Pi, \mathcal{P})$ from (4) we get $\Pr\left\{\sup_{\Pi} \left|\sum_{i=1}^{N} g(\pi_i)(\hat{p}_i - p_i)\right| > \tilde{\epsilon}\right\}$, which can be upper bounded as follows:

$$\Pr\left\{\sup_{\Pi} \sum_{i=1}^{N} g(\pi_i)\hat{\delta}_{p,i} > \tilde{\epsilon}\right\} \leq \Pr\left\{\max_{i=1,2,\ldots,N} \hat{\delta}_{p,i} > \bar{\epsilon}\right\}$$

$$\leq \sum_{i=1}^{N} \Pr\left\{\hat{\delta}_{p,i} > \bar{\epsilon}\right\}$$

$$\leq 2N\mathbb{E}\left[\exp\left\{-2\bar{\epsilon}^2 n_p\right\}\right] \quad (25)$$

where $\hat{\delta}_{p,i} \triangleq |\hat{p}_i - p_i|$, $\tilde{\epsilon} \triangleq \frac{R_0\epsilon}{2B}$, $\bar{\epsilon} \triangleq \frac{\tilde{\epsilon}}{\sup_{\Pi}\left\{\sum_{i=1}^{N} g(\pi_i)\right\}}$, and $g(\pi_i) \triangleq \exp\{-\lambda_s\pi\gamma^2\left[1 - (1-\pi_i)^M\right]\}$, and the last inequality follows by applying Hoeffdings inequality (see [43]) since the estimator $\hat{\mathcal{P}}$ is unbiased and $\pi_i, \pi \in [0,1]$. Note that, the expectation in (25) is with respect to n_p. Conditioned on the number n_R of users in the coverage area of BS, n_p is a Poisson distributed random variable with density $n_R\lambda_r\tau$. Therefore, $2N\mathbb{E}\sum_{n=0}^{\infty} \exp\{-\bar{g}\} \frac{(\lambda_r n_R\tau)^n}{n!} = 2N\mathbb{E}_{n_R} \exp\{-\lambda_r n_R\tau g^*\}$, where $\bar{g} \triangleq \left(2\bar{\epsilon}^2 k + \lambda_r n_R\tau\right)$ and $g^* \triangleq \left(1 - \exp\left\{-2\bar{\epsilon}^2\right\}\right)$ which can further be simplified as

$$2N\sum_{k=0}^{\infty} \exp\{-\lambda_r k\tau g^*\}\exp\{-\lambda_u\pi R^2\}\frac{(\lambda_u\pi R^2)^k}{k!}$$

$$= 2N\exp\{-\lambda_u\pi R^2\left(1 - \exp\left\{-\lambda_r\tau g^*\right\}\right). \quad (26)$$

We see that $\Pr\left\{\sup_{\Pi} |\Delta\mathcal{T}| > \frac{\epsilon}{2}\right\} \leq \delta$ if (26) is upper bounded by δ, resulting in

$$\tau \geq \frac{1}{\lambda_r g^*} \log\left(\frac{1}{1 - \frac{1}{\lambda_u\pi R^2}\log\frac{2N}{\delta}}\right),$$

provided $\lambda_u > \frac{1}{\pi R^2}\log\frac{2N}{\delta}$, otherwise $\tau = \infty$, proving Theorem 2. ∎

APPENDIX C
PROOF OF THEOREM 3

It is easy to see that $\Pr\{\hat{\mathcal{T}}^* \geq \mathcal{T}^* + \epsilon\} \leq \Pr\left\{\sup_{1\leq i\leq N} \left|\hat{p}_i^{(tl)} - p_i\right| > \bar{\epsilon}\right\}$, where $\bar{\epsilon} \triangleq \frac{R_0\epsilon}{2B\sup_{\Pi}\left\{\sum_{i=1}^{N} g(\pi_i)\right\}}$ and $g(\pi_i) \triangleq \exp\{-\lambda_s\pi\gamma^2\left[1 - (1-\pi_i)^M\right]\}$. Denote by n_p the total number of requests in the coverage area of BS. Conditioned on the number of users n_R in the coverage area of BS, n_p is a Poisson distributed random variable with density $n_R\lambda_r$. Further, $\mathbb{E}\left\{\hat{p}_i^{(tl)}|n_p\right\} = \frac{n_p}{n_p+m}p_i + \frac{m}{n_p+m}q_i$. Using this, we can write

$$\Pr\left\{\sup_{1\leq i\leq N} \left|\hat{p}_i^{(tl)} - \mathbb{E}\hat{p}_i^{(tl)} + \mathbb{E}\hat{p}_i^{(tl)} - p_i\right| > \bar{\epsilon}\right\}$$

$$\leq \Pr\left\{\sup_{1\leq i\leq N} \left|\hat{p}_i^{(tl)} - \mathbb{E}\hat{p}_i^{(tl)}\right| + \left|\mathbb{E}\hat{p}_i^{(tl)} - p_i\right| > \bar{\epsilon}\right\}$$

$$\leq \Pr\left\{\sup_{1\leq i\leq N} \left|\hat{p}_i^{(tl)} - \mathbb{E}\hat{p}_i^{(tl)}\right| > \bar{\epsilon} - \sup_{i\in[1,N]} \left|\mathbb{E}\hat{p}_i^{(tl)} - p_i\right|\right\}$$

$$\leq \Pr\left\{\sup_{1\leq i\leq N} \left|\hat{p}_i^{(tl)} - \mathbb{E}\hat{p}_i^{(tl)}\right| > \bar{\epsilon} - \frac{m}{n_p+m}\|\mathcal{P} - \mathcal{Q}\|_\infty\right\}$$

$$\leq \mathbb{E}_{n_p} \Pr\left\{\sup_{1\leq i\leq N} \left|\hat{p}_i^{(tl)} - \mathbb{E}\hat{p}_i^{(tl)}\right| > \bar{\epsilon} - \|\mathcal{P} - \mathcal{Q}\|_\infty \mid n_p\right\}$$

$$\leq N\mathbb{E}_{n_p} \Pr\left\{\left|\hat{p}_i^{(tl)} - \mathbb{E}\hat{p}_i^{(tl)}\right| > \bar{\epsilon} - \|\mathcal{P} - \mathcal{Q}\|_\infty \mid n_p\right\},$$

provided $\bar{\epsilon} > \|\mathcal{P} - \mathcal{Q}\|_\infty$, where $\|\mathcal{P} - \mathcal{Q}\|_\infty \triangleq \sup_{i\in[1,N]} |q_i - p_i|$. From Hoeffding's inequality,

$$2N\mathbb{E}_{n_p} \exp\left\{-2\epsilon_{pq}^2(n_p + m)\right\}$$

$$= 2N\mathbb{E}_{n_R} \exp\left\{-(2\epsilon_{pq}^2 m + \lambda_r n_R\tau)\right\} \sum_{k=0}^{\infty} \frac{a^k}{k!},$$

$$= 2N\exp\left\{-2\epsilon_{pq}^2 m\right\} \mathbb{E}_{n_R} \exp\{-\bar{g}_{pq}\}$$

$$= 2N\exp\left\{-2\epsilon_{pq}^2 m\right\} \exp\{-\lambda_u\pi R^2\} \times$$

$$\sum_{l=0}^{\infty} \exp\{-\lambda_r l\tau(1 - \exp\{-2\epsilon_{pq}^2\})\}\frac{(\lambda_u\pi R^2)^l}{l!}, (27)$$

where $a \triangleq \lambda_r n_R\tau \exp\left\{-2\epsilon_{pq}^2\right\}$, $\epsilon_{pq} \triangleq \bar{\epsilon} - \|\mathcal{P} - \mathcal{Q}\|_\infty$ and $\bar{g}_{pq} \triangleq \lambda_r n_R\tau(1 - \exp\{-2\epsilon_{pq}^2\})$. Therefore, $2N\mathbb{E}_{n_p} \exp\left\{-2\epsilon_{pq}^2(n_p + m)\right\} = 2\exp\left\{-2\epsilon_{pq}^2 m\right\} \exp\{-\lambda_u\pi R^2 t\}$, where $t \triangleq \left(1 - \exp\{-\lambda_r\tau\left(1 - \exp\{-2\epsilon_{pq}^2\}\right)\}\right)$, and is at most $\delta > 0$ if

$$\tau \geq \frac{1}{\lambda_r(1 - e^{-2\epsilon_{pq}^2})} \log\left(\frac{1}{1 - \frac{1}{\lambda_u\pi R^2}\left(\log\frac{2N}{\delta} - 2\epsilon_{pq}^2 m\right)}\right), (28)$$

provided $\lambda_u > \frac{1}{\pi R^2}\left(\log\frac{2N}{\delta} - 2\epsilon_{pq}^2 m\right)$, otherwise $\tau = \infty$, thus proving Theorem 3. ∎

APPENDIX D
PROOF OF THEOREM 4

We begin with $\Pr\{\hat{\mathcal{T}}^* \geq \mathcal{T}^* + \epsilon\} \leq \Pr\left\{\sup_{1\leq i\leq N} \left|\hat{p}_i^{(tl)} - p_i\right| > \bar{\epsilon}\right\} \leq \sum_{i=1}^{N} \Pr\left\{\left|\hat{p}_i^{(tl)} - p_i\right| > \bar{\epsilon}\right\}$, where $\bar{\epsilon} \triangleq \frac{R_0\epsilon}{2B\sup_{\Pi}\left\{\sum_{i=1}^{N} g(\pi_i)\right\}}$ and $g(\pi_i) \triangleq \exp\{-\lambda_s\pi\gamma^2\left[1 - (1-\pi_i)^M\right]\}$. Each term in the summation can be upper bounded as shown in (29) - (32) at the top of the next page, where $\omega \triangleq \frac{\bar{\epsilon}-(1-\alpha)\eta}{\alpha} > 0$. From (26), we have $\Pr\left\{\left|\hat{p}_i^{(t)} - p_i\right| > \eta\right\} = 2\exp\{-\lambda_u\pi R^2\left(1 - \exp\left\{-\lambda_r\tau g_t^*\right\}\right), where $g_t^* \triangleq \left(1 - \exp\left\{-2\eta^2\right\}\right)$ and the second term can be bounded as follows:

$$\Pr\left\{\left|\hat{p}_i^{(s)} - p_i\right| > \omega\right\} = \Pr\left\{\left|\hat{p}_i^{(s)} - q_i + q_i - p_i\right| > \omega\right\}$$

$$\overset{(a)}{\leq} \Pr\left\{\left|\hat{p}_i^{(s)} - q_i\right| > \omega - \|\mathcal{P} - \mathcal{Q}\|_\infty\right\},$$

where (a) follows from the triangular inequality and using $\|\mathcal{P} - \mathcal{Q}\|_\infty = \sup_i |p_i - q_i|$. Note that, the inequality (a) is valid if $\omega > \|\mathcal{P} - \mathcal{Q}\|_\infty$. Using $\mathbb{E}\hat{p}_i = p_i$ and Hoeffding's inequality, we have $\Pr\left\{\left|\hat{p}_i^{(s)} - q_i\right| > \omega - \|\mathcal{P} - \mathcal{Q}\|_\infty\right\} \leq 2\exp\left\{-2(\omega - \|\mathcal{P} - \mathcal{Q}\|_\infty)^2 m\right\}$. Therefore,

$$\sum_{i=1}^{N} \Pr\left\{\left|\hat{p}_i^{(tl)} - p_i\right| > \bar{\epsilon}\right\} \leq 2N[\exp\left\{-2\bar{\omega}^2 m\right\}$$

$$+ \exp\{-\lambda_u\pi R^2\left(1 - \exp\{-\lambda_r\tau g_t^*\}\right)], \quad (33)$$

$$\Pr\left\{\left|\hat{p}_i^{(tl)} - p_i\right| > \bar{\epsilon}\right\} = \Pr\left\{\left|\alpha\hat{p}_i^{(s)} + (1-\alpha)\hat{p}_i^{(t)}) - p_i\right| > \bar{\epsilon}\right\} \tag{29}$$

$$\leq \Pr\left\{\alpha\left|(\hat{p}_i^{(s)} - p_i)\right| + (1-\alpha)\left|(\hat{p}_i^{(t)} - p_i)\right| > \bar{\epsilon}\right\} \tag{30}$$

$$= \Pr\left\{\left[\alpha\left|\hat{p}_i^{(s)} - p_i\right| + (1-\alpha)\left|\hat{p}_i^{(t)} - p_i\right| > \bar{\epsilon}\right]\bigcap\left|\hat{p}_i^{(t)} - p_i\right| > \eta\right\} \tag{31}$$

$$+ \Pr\left\{\alpha\left|\hat{p}_i^{(s)} - p_i\right| + (1-\alpha)\left|\hat{p}_i^{(t)} - p_i\right| > \bar{\epsilon}\bigcap\left|\hat{p}_i^{(t)} - p_i\right| \leq \eta\right\}$$

$$\leq \Pr\left\{\left|\hat{p}_i^{(t)} - p_i\right| > \eta\right\} + \Pr\left\{\left|\hat{p}_i^{(s)} - p_i\right| > \omega\right\}, \tag{32}$$

---

where $\bar{\omega} = \omega - \|\mathcal{P} - \mathcal{Q}\|_\infty > 0$. Finally, it is clear that (33) can upper bounded by δ provided

$$\tau \geq \frac{1}{\lambda_r g_t^*} \log\left[\frac{1}{1 - \frac{1}{\lambda_u \pi R^2}\left[\log\left(A_1\right) + \log\left(B_1\right)\right]}\right], \tag{34}$$

where $A_1 = \frac{2N}{\delta}$, $B_1 = \frac{1}{1 - \left(\frac{2N}{\delta}\right)\exp\{-2\bar{\omega}^2 m\}}$, and $\lambda_u > \frac{1}{\pi R^2}\left(\log\left(\frac{2N}{\delta}\right) + \log\left\{\frac{1}{1 - \left(\frac{2N}{\delta}\right)\exp\{-2\bar{\omega}^2 m\}}\right\}\right)$, which is valid if $\left(\frac{2N}{\delta}\right)\exp\{-2\bar{\omega}^2 m\} < 1$. This along with $\omega - \|\mathcal{P} - \mathcal{Q}\|_\infty > 0$ leads to the constraints stated in Theorem 4. ∎

APPENDIX E
PROOF OF THEOREM 5

We begin with

$$\Pr\{\hat{\mathcal{T}}^* \geq \mathcal{T}^* + \epsilon\}$$

$$\leq \Pr\left\{\sup_{0 \leq \pi \leq 1} g(\pi)\sum_{i=1}^N\left|p_{\hat{\Theta}_{n_p},i} - p_{\Theta,i}\right| > \Omega\right\}$$

$$\overset{(a)}{\leq} \Pr\left\{\sum_{i=1}^N\left|p_{\hat{\Theta}_{n_p},i} - p_{\Theta,i}\right| > \Omega\right\},$$

where (a) follows from the fact that $\sup_{0 \leq \pi \leq 1} g(\pi) = 1$, $p_{\hat{\Theta}_{n_p},i}$ is the estimate of $p_{\Theta,i}$, $\Omega \triangleq \frac{R_0\epsilon}{2B}$ and $g(\pi) \triangleq \exp\{-\lambda_u \pi \gamma^2\left[1 - (1 - \pi_i)^M\right]\} \leq 1$. By using the remainder form of the Taylors series, $p_{\hat{\Theta}_{n_p},i} = p_{\Theta,i} + (\Theta - \hat{\Theta}_{n_p})^T \partial p_{\Theta^*,i}\big|_{\Theta^* \in [\Theta, \hat{\Theta}_{n_p}]}$, where $[\Theta, \hat{\Theta}_{n_p}]$ represents the line joining the points Θ and $\hat{\Theta}_{n_p}$, leading to (recall that the i-th component of $\hat{\Theta}_{n_p}$ is denoted by $\hat{\Theta}_{n_p,i}$, $i = 1, 2, \ldots, d$)

$$\Pr\left\{\sum_{i=1}^N\left|p_{\hat{\Theta}_{n_p},i} - p_{\Theta,i}\right| > \Omega\right\}$$

$$\overset{(a)}{\leq} \Pr\left\{\|\hat{\Theta}_{n_p} - \Theta\|_2 \sup_{\Theta^* \in [\Theta, \hat{\Theta}_{n_p}]}\sum_{i=1}^N\|\partial p_{\Theta^*,i}\|_2 > \Omega\right\}$$

$$\leq \Pr\left\{\|\hat{\Theta}_{n_p} - \Theta\|_2^2 > \frac{\Omega^2}{C^2}\right\}$$

$$\overset{(b)}{\leq} \Pr\left\{\sup_{1 \leq i \leq d}\left|\hat{\Theta}_{n_p,i} - \Theta_i\right|^2 > \frac{\Omega^2}{dC^2}\right\}$$

$$\leq d\Pr\left\{\left|\hat{\Theta}_{n_p,i} - \Theta_i\right|^2 > \frac{\Omega^2}{dC^2}\right\}, \tag{35}$$

where (a) follows from the Cauchy-Schwartz inequality and **Assumption 1** in Section IV, and (b) follows from

the fact that $\|\hat{\Theta}_{n_p} - \Theta\|_2^2 = \sum_{i=1}^d\left|\hat{\Theta}_{n_p,i} - \Theta_i\right|^2 \leq d\sup_{1 \leq i \leq d}\left|\hat{\Theta}_{n_p,i} - \Theta_i\right|^2$.

First, note that for all i, $\hat{\Theta}_{n_p,i}$ is an unbiased estimate of Θ_i. Further, $a \leq \Theta_j \leq b$ for $j = 1, 2, \ldots, d$. Thus, by applying Hoeffding's inequality, we have

$$d\Pr\left\{\left|\hat{\Theta}_{n_p,i} - \Theta_i\right|^2 > \frac{\Omega^2}{dC^2}\right\} \leq 2d\mathbb{E}_{n_p}\exp\left\{-n_p\sigma^2\right\}, \tag{36}$$

where $\sigma^2 \triangleq \frac{2\Omega^2}{dC^2(b-a)^2}$. Conditioned on the number of users (denoted n_R) in a radius of R around the BS, n_p is PPP with density $\lambda_r \tau n_R$. Using this fact in (36), we can write

$$d\Pr\left\{\left|\hat{\Theta}_{n_p,i} - \Theta_i\right|^2 > \frac{\Omega^2}{dC^2}\right\}$$

$$\leq 2d\mathbb{E}_{n_R}\left[\mathbb{E}_{n_p}\left[\exp\left\{-n_p\sigma^2\right\}|n_p\right]\right]$$

$$= 2d\mathbb{E}_{n_R}\sum_{k=0}^\infty \exp\{-k\sigma^2\}e^{-n_R\lambda_r\tau}\frac{(n_R\lambda_r\tau)^k}{k!}$$

$$= 2d\mathbb{E}_{n_R}e^{-n_R\lambda_r\tau}\sum_{k=0}^\infty\frac{(n_R\lambda_r\tau e^{-\sigma^2})^k}{k!}$$

$$= 2d\mathbb{E}_{n_R}\exp\{-n_R\lambda_r\tau(1 + e^{-\sigma^2})\}$$

$$= 2d\sum_{k=0}^\infty\exp\{-k\lambda_r\tau(1 + e^{-\sigma^2})\}\frac{(\lambda_u\pi R^2)^k}{k!}e^{-\lambda_u\pi R^2}$$

$$= f_{\sigma^2}(\tau), \tag{37}$$

where $f_{\sigma^2}(\tau) \triangleq 2d\exp\left\{-\lambda_u\pi R^2\left(1 - \exp\{-\lambda_r\tau(1 - e^{-\sigma^2})\}\right)\right\}$ is a monotonically decreasing function of τ for all $\tau > 0$. Thus, $f_{\sigma^2}(\tau) \leq \delta$ if $\tau > \frac{1}{\lambda_r(1 - e^{-\sigma^2})}\log\left(\frac{1}{1 - \frac{1}{\lambda_u\pi R^2}\log\frac{2d}{\delta}}\right)$, for $\lambda_u > \frac{1}{\pi R^2}\log\frac{2d}{\delta}$, proving Theorem 5. ∎

APPENDIX F
PROOF OF THEOREM 6

We begin with

$$\Pr\{\hat{\mathcal{T}}^* \geq \mathcal{T}^* + \epsilon\} \leq \Pr\left\{\sup_{0 \leq \pi \leq 1} g(\pi)\sum_{i=1}^N\left|p_{\hat{\Theta}_{tl},i} - p_{\Theta,i}\right| > \tilde{\epsilon}\right\}$$

$$\leq \Pr\left\{\sum_{i=1}^N\left|p_{\hat{\Theta}_{tl},i} - p_{\Theta,i}\right| > \Omega\right\},$$

where $p_{\hat{\Theta}_{tl},i}$ is the estimate of $p_{\Theta,i}$ using TL-based approach described in Section IV-A, $\Omega \triangleq \frac{R_0\epsilon}{2B}$, and $g(\pi) \triangleq \exp\{-\lambda_u\pi\gamma^2\left[1 - (1 - \pi_i)^M\right]\}$. Note that, $\hat{\Theta}_{tl} \triangleq \lambda\hat{\Theta}_s + (1 -$

$$\Pr\{\hat{\mathcal{T}}^* \geq \mathcal{T}^* + \epsilon\} \quad \leq \quad \Pr\left\{\|\hat{\Theta}_{\mathrm{tl}} - \Theta\|_2 > \frac{\Omega}{C}\right\} \tag{38}$$

$$\overset{(a)}{\leq} \quad \Pr\left\{\lambda\|\hat{\Theta}_s - \Theta\|_2 + (1-\lambda)\|\hat{\Theta}_t - \Theta\|_2 > \frac{\Omega}{C}\right\} \tag{39}$$

$$= \quad \Pr\left\{\lambda\|\hat{\Theta}_s - \Theta\|_2 + (1-\lambda)\|\hat{\Theta}_t - \Theta\|_2 > \frac{\Omega}{C}\bigcap \mathcal{E}\right\}$$

$$+ \Pr\left\{\lambda\|\hat{\Theta}_s - \Theta\|_2 + (1-\lambda)\|\hat{\Theta}_t - \Theta\|_2 > \frac{\Omega}{C}\bigcap \mathcal{E}^c\right\} \tag{40}$$

$$\leq \quad \Pr\left\{\lambda\|\hat{\Theta}_s - \Theta\|_2 > \frac{\Omega}{C} - D_t\right\} + \Pr\left\{\|\hat{\Theta}_t - \Theta\|_2 > \frac{D_t}{1-\lambda}\right\}, \tag{41}$$

$\lambda)\hat{\Theta}_t$. Further, from the reminder form of the Taylor series around Θ (true parameter), we get $p_{\hat{\Theta}_{\mathrm{tl},i}} = p_{\Theta,i} + (\hat{\Theta}_{\mathrm{tl}} - \Theta)^T \partial p_\Theta|_{\Theta \in [\Theta_{\mathrm{u}},\Theta]}$, $i = 1, 2, \ldots, N$, which implies that

$$\sum_{i=1}^{N}\left|p_{\hat{\Theta}_{\mathrm{tl},i}} - p_{\Theta,i}\right| = \sum_{i=1}^{N}\left|(\hat{\Theta}_{\mathrm{tl}} - \Theta)^T \partial p_\Theta|_{\Theta \in [\Theta_{\mathrm{u}},\Theta]}\right|$$

$$\overset{(a)}{\leq} \|\hat{\Theta}_{\mathrm{tl}} - \Theta\|_2 \sum_{i=1}^{N}\|\partial p_\Theta|_{\Theta \in [\Theta_{\mathrm{u}},\Theta]}\|_2$$

$$\overset{(b)}{\leq} C\|\hat{\Theta}_{\mathrm{tl}} - \Theta\|_2,$$

where (a) follows from Cauchy-Schwartz inequality and (b) follows from **Assumption 1** in Section IV. Therefore we have the relations (38) - (41) at the top of this page, where (a) follows from using $\hat{\Theta}_{\mathrm{tl}} = \lambda\hat{\Theta}_s + (1-\lambda)\hat{\Theta}_t$ followed by the triangle inequality. Here, $\mathcal{E} \triangleq \{\|\hat{\Theta}_t - \Theta\|_2 < \frac{D_t}{1-\lambda}\}$, and we let $D_t < \Omega/C$. The first term can be expressed as follows:

$$\Pr\left\{\|\hat{\Theta}_s - \Theta\|_2 > \frac{1}{\lambda}\left(\frac{\Omega}{C} - D_t\right)\right\}$$

$$\leq \Pr\left\{\|\hat{\Theta}_s - \Theta_s\|_2 + \|\Theta_s - \Theta\|_2 > \bar{\Omega}^2\right\}$$

$$\leq \Pr\left\{\sup_{1 \leq i \leq d}\left|\hat{\Theta}_{s,i} - \Theta_{s,i}\right|^2 > (\bar{\Omega} - \|\Theta_s - \Theta\|_2)^2\right\}$$

$$\leq d\Pr\left\{\left|\hat{\Theta}_{s,i} - \Theta_{s,i}\right|^2 > (\bar{\Omega} - \|\Theta_s - \Theta\|_2)^2\right\}, \tag{42}$$

where $\bar{\Omega} \triangleq \frac{1}{\lambda}\left(\frac{\Omega}{C} - D_t\right) > \|\Theta_s - \Theta\|_2$. However, $\hat{\Theta}_{s,i} \in [a, b]$ is an unbiased estimator of $\Theta_{s,i}$. Therefore, by Hoeffding's inequality, we can write

$$d\Pr\left\{\left|\hat{\Theta}_{s,i} - \Theta_{s,i}\right|^2 > (\bar{\Omega} - \|\Theta_s - \Theta\|_2)^2\right\}$$

$$\leq 2d\exp\left\{-\frac{2m(\bar{\Omega} - \|\Theta_s - \Theta\|_2)^2}{(b-a)^2}\right\}. \tag{43}$$

Next, we have

$$\Pr\left\{\|\hat{\Theta}_t - \Theta\|_2 > \frac{D_t}{1-\lambda}\right\} \leq$$

$$2d\exp\left\{-\lambda_u\pi R^2\left(1 - \exp\{-\lambda_r\tau(1 - e^{-\sigma_t^2})\}\right)\right\}, \tag{44}$$

where $\sigma_t^2 \triangleq \frac{D_t^2}{2(b-a)^2(1-\lambda)^2}$, and the inequality follows from (37) by replacing Ω^2/dC^2 with $\frac{D_t^2}{(1-\lambda)^2}$.

Therefore, $\Pr\{\hat{\mathcal{T}}^* \geq \mathcal{T}^* + \epsilon\}$ will be upperbounded by

$$2d\exp\left\{-\frac{2m(\bar{\Omega} - \|\Theta_s - \Theta\|_2)^2}{(b-a)^2}\right\}$$

$$+2d\exp\left\{-\lambda_u\pi R^2\left(1 - e^{\{-\lambda_r\tau(1 - e^{-\sigma_t^2})\}}\right)\right\},$$

which is less than or equal to δ if

$$\tau \geq \frac{1}{\lambda_r(1 - e^{-\sigma_t^2})}\log\left(\frac{1}{1 - \frac{1}{\lambda_u\pi R^2}\left(\log\frac{2d}{\delta} + \log C_1\right)}\right),$$

for $\lambda_u > \frac{1}{\pi R^2}\left(\log\frac{2d}{\delta} + \log\frac{1}{1 - \frac{2d}{\delta}\exp\left\{-\frac{2m(\bar{\Omega} - \|\Theta - \Theta_s\|_2)^2}{(b-a)^2}\right\}}\right)$,

where

$$C_1 = \frac{1}{1 - \frac{2d}{\delta}\exp\left\{-\frac{2m(\bar{\Omega} - \|\Theta - \Theta_s\|_2)^2}{(b-a)^2}\right\}}$$

These together with the conditions $\bar{\Omega} > \|\Theta_s - \Theta\|_2$ and $\frac{2d}{\delta}\exp\left\{-\frac{2m(\bar{\Omega} - \|\Theta - \Theta_s\|_2)^2}{(b-a)^2}\right\} < 1$ proves Theorem 6. $\blacksquare$