Efficient File Replication in Large Wireless Networks with Dynamic Popularity

Savvas Gitzenis
Information Technologies Institute
CERTH, Greece
E-mail: sgitz@iti.gr

Stavros Toumpis
Department of Informatics
Athens Univ. of Economics and Business, Greece
E-mail: toumpis@aueb.gr

Leandros Tassiulas
Dept. of Electrical & Computer Eng.
University of Thessaly, Greece
E-mail:leandros@uth.gr

***Abstract*—We investigate the problem of replication in large wireless networks that employ caching in the case of a single file whose popularity varies with time. As opposed to the case of static popularity, in this case for the network resources to be efficiently allocated the replication should vary with time. In this study, we first outline the low-level operations of wireless networks with caching, which involve decisions of combinatorial complexity, such as about the contents of all network caches. To overcome this complexity, we approximate the network optimization with a formulation based on the frequency of file replication across the network—a high-level perspective, amenable to mathematical analysis. We present a solution that is based on looking ahead into the future and has a simple graphical representation.**

I. Introduction

Large wireless networks have been the subject of extensive research in the last decades, due to the proliferation of wireless services and standards, and the advent of affordable mobile devices. The vision of joining large numbers of nodes into wireless networks that operate without any infrastructure support toward providing ubiquitous access has spurred many research efforts to characterize the properties of these networks (notably their scalability) and optimize their operation [1]–[6].

On the other hand, caching is a major technique in computing that lowers the average access delay and the system load by storing replicas of selected data across the system. In networking, caching has been widely employed, albeit in rudimentary ways, such as individually at each node or between node pairs. Key is the role of caching in the novel paradigm of Information Centric Networking (ICN) which aspires to replace the old primitives, based on node addresses, with new ones based on named content. The replication of the content in a comprehensive manner across the network [7]–[10] is a crucial element of any ICN architecture [11]–[13]. As a result, the ICN paradigm has stimulated investigations on the advantages of caching in large wireless networks [14]–[20].

Common in these works is the formulation of an analytic model of the network operations which is then applied to optimize performance. These models are often built at the *microscopic* level of the actual network, considering details, such as the precise routes of packets among the nodes, that capture the operation in the network with high accuracy, as in [14]–[16]. However, the resulting model is usually quite complex; more than often, suitable approximations and heuristics are required to analyze and optimize the network.

The alternative is a model relying on *macroscopic* quantities describing the wireless network and its operation. In this approach, the network is represented by an abstraction that involves only high-level quantities, such as—in the case of caching—how densely data are cached across the network as opposed to the contents of each cache. The merit of this approach is that the formulation becomes easier to handle with standard mathematical tools. The disadvantage is the additional step of translating macroscopic quantities into microscopic ones; this is often approximate and suboptimal, as important details missing from the abstract model must be filled in, which is sometimes not a straightforward task [19].

In this study, we investigate caching in large wireless networks. We go briefly through the microscopic operation, as our study is based mainly on the macroscopic model, as in [18]–[20]. These works consider flat wireless networks where requests are placed according to a given content popularity uniformly across the nodes; requests are identically distributed, independently from each other both temporally and spatially. There, the problem of replication is cast as minimizing the traffic over the network with the optimization variables being the replication densities of all data and constraints on the capacity of the individual nodes. The optimization is carried out using the tools of Lagrange multipliers and Karush Kuhn Tucker (KKT) conditions. Due to the translation between the macroscopic and microscopic models, the link load is estimated only up to a multiple of its actual value; hence, the performance of the proposed solution is within a multiplicative constant of the optimal. However, this *order-optimal* optimization (i) can be quite useful in comparing cache replacement policies—the goal of [18], and (ii) provides the exact order of the link traffic (with respect to the network size), which enables us to characterize the sustainability of large expanding wireless networks with caching—the object of [19], [20].

In [18]–[20], due to the assumption that content popularity is static, there is no need for the optimal replication to vary with time. Relaxing to time-varying popularity means that the replication should adjust to its fluctuations to be optimal. Our formulation captures this via an additional cost term related to the traffic generated by the adjustment of cache contents in response to the changes in the file popularity. In this work, we limit our study to the replication of a single file only.

The rest of the work is organized as follows: Section II

QSHINE 2014, August 18-20, RHODES, Greece

DOI 10.4108/icst.qshine.2014.256567

provides a microscopic wireless network model and its operations, which Section III abstracts away in a macroscopic model. We then optimize the file replication and present its solution. Section IV discusses next steps in this thread of research.

II. Network Model—Microscopic Perspective

A. Wireless Network and Topology

Consider a set $\mathcal{N}$ of identical peers arranged on the plane in a square grid and let $N \triangleq |\mathcal{N}|$. As in [19], each node is connected to its four neighboring nodes at the same row or column, with non-interfering links. Unlike the random topology of [2], [18] and other works, this graph has a deterministic, regular structure. As explained in [19], it is a good model for a planar wireless networks as it bypasses the complexity of the PHY and MAC layers induced by interference, preserving at the same time the essential features of multihop wireless networks, i.e., (i) short links (many short links are better than a few long links due to better spatial reuse [2]) and (ii) the network diameter scales as $\sqrt{N}$—the key reason behind the $1/\sqrt{N}$ law of [2] on the per-flow throughput.

B. Content, User Requests and Caches

Users located in the nodes generate requests to access files. In this work, we focus on a single file and consider only the associated requests. Its popularity can be described by the arrival rate of the user requests for it at each node, $\lambda(t) > 0$, a function of time t (we let time be continuous). Note that the request arrival rate is common to all nodes; this symmetry is an essential simplifying element of our study. Moreover, all requests across both time (i.e., past and future requests) and space (i.e., requests at different nodes) are independent of each other. Hence, the requests at all nodes form independent Poisson processes of a time-varying rate.

The file requests are served by caches attached to every node. In this study, we assume that caching comes at a price $\gamma(t)$ per node and unit of time (i.e., the storage of the file in any node w for a short duration Δt costs $\gamma(t)\Delta t$). The rate $\gamma(t)$ is common to all nodes, but may vary with time (e.g., due to competition for storage from other files). Depending on its popularity, the file is stored in the set $\mathcal{W}(t)$ of nodes, a subset of $\mathcal{N}$. If a user request is placed at time t on a node n that belongs to $\mathcal{W}(t)$, then it is immediately served without engaging the network. Otherwise, node n has to place a request over the network to some node w of $\mathcal{W}(t)$.

In this work, we aim to optimize the traffic generated by user requests for the file against the cost of the file storage and the traffic load due to the updates of the set $\mathcal{W}(t)$. The trade-off is evident: a dense replication minimizes the traffic due to the user requests but is costly due to the high number of nodes required in $\mathcal{W}(t)$; moreover, the addition of nodes in $\mathcal{W}(t)$ induces load on the network. Given that popularity $\lambda(t)$ varies with time, the set $\mathcal{W}(t)$ should be accordingly adjusted. Intuitively, nodes should join or leave $\mathcal{W}(t)$ according to the current value of the popularity $\lambda(t)$ as well as its future evolution.

C. Delivery of User Requests and Cache Updates

In the static popularity context of [19], to serve a request at node n, one has to specify a delivery/routing path $\mathcal{R}_n$ of adjacent nodes from node n to some node $w \in \mathcal{W}$; no cache ever needs to get updated, hence $\mathcal{W}$ and $\mathcal{R}_n$ are constant in time. In the dynamic problem, as $\mathcal{W}$ and $\mathcal{R}_n$ vary with time, a mechanism to update them should be provided.

In order to minimize traffic, the network should support the delivery of the file to many nodes with a single operation through *multicast trees*. Then, to keep any considerations about the timing of the deliveries out of the model, we employ the Zero Download Delay (ZDD) assumption of [10]. According to it, a user file request is immediately served and any update on the contents of a cache takes place instantaneously when decided. As in the context of Poisson arrivals multiple user requests cannot happen at the same time, ZDD *precludes* serving multiple user requests in a single multicast delivery. Summing up, each delivery operation is multicast and serves up to one user request and any number of cache updates.

Due to space constraints, we refrain from specifying the delivery variables in more detail.

D. Optimization Formulation

Given the above model, we can precisely express the traffic over the network and optimize it jointly with the storage cost over an interval $[0, T]$ into their sum $J^{\mu}(T)$. The time horizon T should be chosen high enough so as to minimize the boundary effects at the interval ends. As initial condition, we specify a single replica at $t = 0$ at an arbitrary node.

The optimization decisions regard (i) the value of $\mathcal{W}(t)$ at each time t, i.e., about where the file is replicated over the network, and (ii) what delivery operations are taken to update $\mathcal{W}(t)$ and serve user file requests. Cache updates can happen either in response to changes in the content popularity alone or along with the user requests, as explained before. Hence, the microscopic problem is essentially a joint optimization on replication and delivery of a high complexity due to the exponential number of possibilities for the cache contents. The macroscopic formulation problem described next comes particularly useful in designing an efficient suboptimal solution for this microscopic problem (however, this is deferred to an upcoming extended version of this work).

Our optimization formulation is *centralized* and *non-causal*. The former implies the existence of a central controller that has instant and perfect knowledge about all the workings in the network and optimizes the operations globally. The latter means that the controller knows about the future evolution of popularity $\lambda(t)$ and price $\gamma(t)$ and manages the network in the most optimal way. Although these conditions would hardly hold in reality, they are useful in providing a bound on the network performance under the most favorable circumstances.

III. Macroscopic Replication Problem

A. Macroscopic Problem Formulation

The above joint replication-delivery problem can be abstracted to an optimization on replication only, that involves the macroscopic density of files across the network. In [19], the *planar* density d was defined as the fraction of caches that store the file under consideration. Here, we use the *linear* density, a related quantity whose inverse approximates the distance between nodes storing the file replicas:

$$\rho(t) \triangleq \sqrt{\frac{|\mathcal{W}(t)|}{N}} = \sqrt{d(t)}. \tag{1}$$

Compared to the static formulation of [19]–[21], the macroscopic problem must allow for changes in the density $\rho(t)$ in response to changes in the popularity $\lambda(t)$ and price rate $\gamma(t)$. This means that the optimization target of [19]–[21] should be accordingly modified. Toward this end, we express the density $\rho(t)$, a function of time t, as a difference of two weakly increasing functions $\rho_{\mathrm{i}}(t)$ and $\rho_{\mathrm{d}}(t)$: $\rho(t) = \rho_{\mathrm{i}}(t) - \rho_{\mathrm{d}}(t)$. This arrangement ensures that increases and decreases in $\rho(t)$ are realized through increases in $\rho_{\mathrm{i}}(t)$ and $\rho_{\mathrm{d}}(t)$ respectively; as a result, it enables us to express in a simple way the network load that arises from the changes in the file replication.

PROBLEM 1 [MACROSCOPIC]:
Minimize $J^{\mathrm{M}}(T, \rho_{\mathrm{i}}, \rho_{\mathrm{d}})$ *over* ρ_{i} *and* ρ_{d} *where*

$$J^{\mathrm{M}}(t, \rho_{\mathrm{i}}, \rho_{\mathrm{d}}) \triangleq \rho_{\mathrm{i}}(t) - \rho_{\mathrm{i}}(0) + \int_0^t \left[\frac{\lambda(\tau)}{\rho(\tau)} + \gamma(\tau)\left(\rho(\tau)\right)^2 \right] \mathrm{d}\tau, \tag{2}$$

and $\rho(t) \triangleq \rho_{\mathrm{i}}(t) - \rho_{\mathrm{d}}(t)$, *subject to*

1) *for any* t, $1/\sqrt{N} \le \rho(t) \le 1$,
2) ρ_{i} *and* ρ_{d} *are weakly increasing functions,*
3) $\rho_{\mathrm{d}}(0) = 0$ *and* $\rho_{\mathrm{i}}(0) = 1/\sqrt{N}$.

The objective function (2) is a sum of the expected volume of traffic carried per network link in the interval $[0, t]$ plus the associated cache usage cost per node; in fact, $J^{\mathrm{M}}(t)$ can be shown to be to a constant factor of the microscopic $J^{\mathrm{\mu}}(t)$.

Indeed, the inverse of the linear density $\rho(t)$ approximates the average number of hops needed to reach the file from a random node at time t; scaled by the user arrival rate $\lambda(t)$, it expresses the instantaneous network load related to serving user requests. The product $\gamma(t)(\rho(t))^2$ corresponds to the replication cost per node: $(\rho(t))^2$ is the planar density which represents the fraction of nodes that replicate the file under consideration, and is scaled by cache price $\gamma(t)$. These two terms are integrated over the interval $[0, T]$ to find the aggregate traffic due to user requests and total replication cost.

Last, the term $\rho_{\mathrm{i}}(t)$ corresponds to the traffic required to increase density—this is given without proof (it is a key element of the derivation that links the microscopic to the macroscopic problem which will appear in an extended version of this work). The aggregate increases of $\rho(t)$ are reflected in $\rho_{\mathrm{i}}(T)$, which expresses the overall traffic carried for cache contents adjustment in the interval $[0, T)$, and, hence, in $J^{\mathsf{CD}}(T)$. In contrast, decreases in density are realized by increasing $\rho_{\mathrm{d}}(t)$. This does not induce any traffic: i.e., in the microscopic problem, no reallocation in the caches takes place—just a subset of the caches replicating the file are released. Last, as our modeling assumes that the first (primary) copy of a file appears at some node of the network at $t = 0$ spontaneously without incurring any traffic, we have to subtract the respective density of $1/\sqrt{N}$ from file's $\rho_{\mathrm{i}}(T)$ in $J^{\mathsf{CD}}(T)$. Note, however, that the subtraction of $1/\sqrt{N}$ does not affect the optimal solution $(\rho_{\mathrm{i}}^*(t), \rho_{\mathrm{d}}^*(t))$.

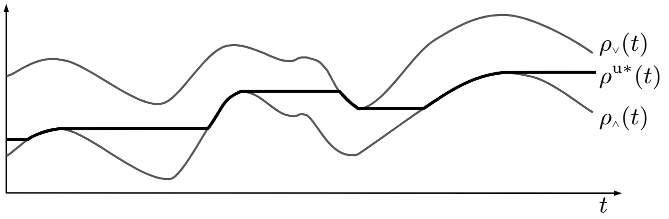


Fig. 1: Example of optimal density $\rho^{\mathrm{u}*}(t)$, shown in black line, against the bounds $\rho_\wedge(t)$ and $\rho_\vee(t)$, shown in gray lines.

Regarding the problem constraints, the ones on $\rho_{\mathrm{d}}(0)$ and $\rho_{\mathrm{i}}(0)$ come from the microscopic problem initial condition at $t = 0$ about the single replica. Last, the lower and upper bounds $1/\sqrt{N}$ and 1 on $\rho(t)$ express the fact that there should be at least one copy of the file in the network, and, at most, the file is replicated at all nodes, respectively.

Towards finding the solution, we relax the pair of constraints on the value of $\rho(t)$ into non-negative numbers as follows:

PROBLEM 2 [MACROSCOPIC UNCONSTRAINED]: *Minimize* $J^{\mathrm{M}}(T, \rho_{\mathrm{i}}, \rho_{\mathrm{d}})$ *over* ρ_{i} *and* ρ_{d}, *subject to*

1) ρ_{i} *and* ρ_{d} *are weakly increasing non-negative functions,*
2) $\rho_{\mathrm{d}}(0) = \rho_{\mathrm{i}}(0) = 0$.

The minimization of the objective function (2) can, in principle, be formulated as a problem of Calculus of Variations by expressing $\rho_{\mathrm{i}}(t)$ as an integral of its derivative leading to

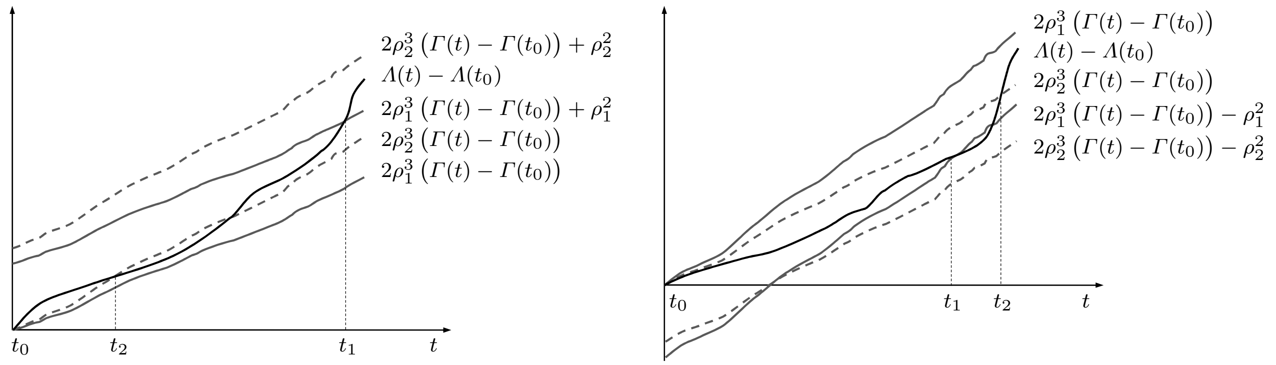
$$J^{\mathrm{M}}(t, \rho_{\mathrm{i}}, \rho_{\mathrm{d}}) = \int_0^t \left[\dot{\rho}_{\mathrm{i}}(\tau) + \frac{\lambda(\tau)}{\rho(\tau)} + \gamma(\tau)\left(\rho(\tau)\right)^2 \right] \mathrm{d}\tau,$$

However, the Euler-Lagrange equation, the standard solution in Calculus of Variations, is not applicable as it requires the optimal ρ_{i} and ρ_{d} to be twice differentiable functions of time—it turns out that this is not true in our case; in terms of Calculus of Variations, we have to find the points of non-differentiability and treat separately the optimal ρ_{i} and ρ_{d} at these points. In essence, the difficulty stems from the fact that the optimization has to look ahead into the future (possibly up to time T), as revealed next from the solution. Although the solution can be computed numerically with the tools of Dynamic Programming, here we present an alternative method that brings out the solution structural properties and is amenable to a graphical interpretation.

B. Problem Solution

Next, we introduce the solution $\rho^{\mathrm{u}*}(t)$ of Problem 2 first at a high level (see Fig. 1) and then move down to the details. The results are presented next without proof. The proof will appear in an extended version of this work.

Problem 2 has a solution described by two density bounds $\rho_\wedge(t)$ and $\rho_\vee(t)$, both functions of time. These define an interval $\mathcal{P}(t) \triangleq [\rho_\wedge(t), \rho_\vee(t)]$: for all t, the optimal density



(a) An example of $\rho_1 < \rho_\wedge(t_0) < \rho_2$

(b) An example of $\rho_2 < \rho_\vee(t_0) < \rho_1$

Fig. 2: The user requests $\mathit{\Lambda}(t)-\mathit{\Lambda}(t_0)$ (black line) are compared against $2\rho^3(\mathit{\Gamma}(t)-\mathit{\Gamma}(t_0))$ and $2\rho^3(\mathit{\Gamma}(t)-\mathit{\Gamma}(t_0)) \pm \rho^2$ (gray lines) to find $\rho_\wedge(t_0)$ and $\rho_\vee(t_0)$. Solid gray lines correspond to density $\rho=\rho_1$ that satisfy (3)–(4) or (5)–(6), respectively, while dashed lines correspond to $\rho=\rho_2$ with no solution:

(a) User requests $\mathit{\Lambda}(t)-\mathit{\Lambda}(t_0)$ are compared against $2\rho^3(\mathit{\Gamma}(t)-\mathit{\Gamma}(t_0))+\rho^2$ and $2\rho^3(\mathit{\Gamma}(t)-\mathit{\Gamma}(t_0))$ to determine $\rho_\wedge(t_0)$. For $\rho=\rho_1$, (3) is satisfied for $t_\wedge=t_1$, and (4) is true for $t\in[t_0,t_\wedge]$. Hence, $\rho_\wedge(t_0)$ is at least equal to ρ_1. However, for $\rho=\rho_2$, (3) is not satisfied for any $t_\wedge \le t_2$. Given the crossing at t_2, assuming a solution of $\rho=\rho_2$ for $t_\wedge > t_2$, (4) is violated at $t=t_2\in[t_0,t_\wedge]$. Hence, $\rho_\wedge(t_0)$ is strictly less than ρ_2.

(b) User requests $\mathit{\Lambda}(t)-\mathit{\Lambda}(t_0)$ are compared against $2\rho^3(\mathit{\Gamma}(t)-\mathit{\Gamma}(t_0))-\rho^2$ and $2\rho^3(\mathit{\Gamma}(t)-\mathit{\Gamma}(t_0))$ to determine $\rho_\wedge(t_0)$. For $\rho=\rho_1$, (5) is satisfied for $t_\vee=t_1$, and (6) is true for $t\in[t_0,t_\vee]$. Hence, $\rho_\vee(t_0)$ is at most equal to ρ_1. However, for $\rho=\rho_2$, (5) is not satisfied for any $t_\vee \le t_2$. Given the crossing at t_2, assuming a solution of $\rho=\rho_2$ for $t_\wedge > t_2$, (6) is violated at $t=t_2\in[t_0,t_\vee]$. Hence, $\rho_\vee(t_0)$ is strictly less than ρ_2.

$\rho^{\mathrm{u}*}(t)$ has to lie in $\mathcal{P}(t)$. Assuming, moreover, a given density value at time t_0, the density should remain constant in $t>t_0$ as long as $\rho^{\mathrm{u}*}(t)\in\mathcal{P}(t)$. As $\rho_\wedge(t)$ and $\rho_\vee(t)$ vary with time, the density should change so as not to violate $\rho(t)\in\mathcal{P}(t)$. Hence, density increases come only from $\rho_\wedge(t)$, and similarly, density decreases come only from $\rho_\vee(t)$, as illustrated in Fig. 1.

To formally define the bounds, we introduce two quantities, (i) the expected volume $\mathit{\Lambda}(t)$ of user requests per node placed in interval $[0,t]$, and (ii) the cumulative cache cost $\mathit{\Gamma}(t)$ for using a unit of cache in the same interval:

$$\mathit{\Lambda}(t) \triangleq \int_0^t \lambda(\tau)\,\mathrm{d}\tau, \qquad \mathit{\Gamma}(t) \triangleq \int_0^t \gamma(\tau)\,\mathrm{d}\tau.$$

Now, consider an arbitrary non-negative value ρ as candidate for the optimal density at time t_0. We eliminate the possibility of too high or too low values ρ at t_0 as follows:

- Consider all $t_\wedge \in (t_0, T]$ and the following inequalities:

$$\frac{\mathit{\Lambda}(t_\wedge)-\mathit{\Lambda}(t_0)}{\mathit{\Gamma}(t_\wedge)-\mathit{\Gamma}(t_0)} \ge 2\rho^3 + \frac{\rho^2}{\mathit{\Gamma}(t_\wedge)-\mathit{\Gamma}(t_0)}, \text{ and} \tag{3}$$

$$\frac{\mathit{\Lambda}(t)-\mathit{\Lambda}(t_0)}{\mathit{\Gamma}(t)-\mathit{\Gamma}(t_0)} \ge 2\rho^3 \qquad \text{for all } t\in(t_0,t_\wedge]. \tag{4}$$

  If, for the assumed ρ, a $t_\wedge$ that satisfies the above pair of conditions exists, as we show next, the optimal density is bounded below by it: $\rho^{\mathrm{u}*}(t_0)\ge\rho$. We define then as $\rho_\wedge(t_0)$ the *supremum* of all such bounds, i.e., all ρ which satisfy the above pair of conditions for some $t_\wedge > t_0$.

- Consider all $t_\vee \in (t_0, T]$ and the following inequalities:

$$\frac{\mathit{\Lambda}(t_\vee)-\mathit{\Lambda}(t_0)}{\mathit{\Gamma}(t_\vee)-\mathit{\Gamma}(t_0)} \le 2\rho^3 - \frac{\rho^2}{\mathit{\Gamma}(t_\vee)-\mathit{\Gamma}(t_0)}, \text{ and} \tag{5}$$

$$\frac{\mathit{\Lambda}(t)-\mathit{\Lambda}(t_0)}{\mathit{\Gamma}(t)-\mathit{\Gamma}(t_0)} \le 2\rho^3 \qquad \text{for all } t\in(t_0,t_\vee]. \tag{6}$$

  If, for the assumed ρ, a $t_\vee$ that satisfies the above pair of conditions exists, as we show next, the optimal density is bounded above by it: $\rho^{\mathrm{u}*}(t_0)\le\rho$. We define then as $\rho_\vee(t_0)$ the *infimum* of all such bounds, i.e., all ρ which satisfy the above pair of conditions for some $t_\vee > t_0$.

Making use of the convexity properties of the Problem, we provide some intuition on conditions (3)–(6). Compare the case that in the interval $[t_0,t_\wedge]$ density $\rho(t)$ is held constant against increasing density by $\Delta\rho$ at t_0 and letting it remain constant till $t_\wedge$. If we multiply both sides of (3) by some $\Delta\rho>0$, the LHS approximates the decrease in link traffic that results from the increase in density by $\Delta\rho$ in this interval, while the RHS is the sum of two terms, (i) the increase in replication cost by the $\Delta\rho$ increase in density in this interval and (ii) the link traffic overhead to realize this $\Delta\rho$ increase in the replication density at t_0, the beginning of the interval. If the LHS is greater than the RHS, then the density increase is advantageous, and clearly we should consider greater values of $\rho'>\rho$ to minimize the total cost incurred as much as possible.

However, (4) must also hold for any $t\in(t_0,t_\wedge]$ as well. Indeed, multiplying by $\Delta\rho>0$ both its sides, the resulting LHS gives the decrease in link traffic by the density increase by $\Delta\rho$ in interval $[t_0,t]$, while the RHS expresses the associated increase in storage cost. If for some $t'\in(t_0,t_\wedge]$, the RHS outweighs the LHS, then such an increase in density is unfavorable for this subinterval, as it raises the total cost incurred in $[t_0,t']$. In simple words, this means that while (3) says that we gain by increasing the density at t_0 in the interval $[t_0,t_\wedge]$, (4) says that in the first part $[t_0,t']\subset[t_0,t_\wedge]$, we lose from such an increase; it is better to postpone the increase and make it happen at t'. Hence, the benefit of (3) for interval $[t_0,t_\wedge]$ is not diminished by the loss of $[t_0,t']$. Indeed, upon a moment of reflection, we can see that (3) will still hold true if we consider $t_0' \triangleq t'$ and the same $t_\wedge$ as before.

These considerations are depicted in Fig. 2a for graphically finding $\rho_\wedge(t_0)$: if there exists a $t_\wedge$ that satisfies (3)–(4) for an assumed ρ, a differential increase at t_0 in density above ρ is advantageous, thus the lower bound $\rho_\wedge$ is higher or equal to ρ. To determine $\rho_\wedge$, we have to try again to satisfy (3)–(4) with $\rho' > \rho$ up to the point that this is not possible. If, reversely, (3)–(4) are not satisfied for the assumed ρ, then $\rho_\wedge < \rho'$, and we should try with $\rho' < \rho$ to satisfy (3)–(4).

Reversely, decreases in ρ are decided by (5)–(6). The second term of the RHS of (5) corresponds to a cost of decreasing ρ. Although such a decrease cost does not explicitly appear in the J^{CDU}, it is related to the cost that we will have to pay later if we decide to reverse this decrease and restore density back to its value. Hence, decreases in ρ should happen if with respect to interval $[t_0, t_\vee]$, the increase in the total cost (LHS of (5)) plus the cost to increase replication density in the future (second negative term of RHS of (5)) is less than the reduction in the cache cost (first term of the RHS of (5)). Again, (6) safeguards against a premature density reduction (Fig. 2b).

The above are formally expressed as follows:

THEOREM 1 [DENSITY BOUNDS]: *The optimal density function $\rho^{\text{u}*}(t)$ of Problem 2 is bounded by $\rho_\wedge(t)$ and $\rho_\vee(t)$:*

$$\rho_\wedge(t) \leq \rho^{\text{u}*}(t) \leq \rho_\vee(t), \quad \textit{a.e.} \tag{7}$$

In (7), almost everywhere (a.e.) means that the integral of the set of violation of the inequalities is zero.

Last, if neither (3)–(4) or (5)–(6) justify a change in the density, the optimal strategy is to keep it constant:

THEOREM 2 [DENSITY FLUCTUATION]: *The optimal density function $\rho^{\text{u}*}(t)$ of Problem 2 stays constant as long as Theorem 1 does not dictate any change through* (7).

The above Theorem justifies the intervals of constant value in Fig. 1. Using convexity properties of Problem 1, we can use Problem 2 and restrict its solution $\rho^{\text{u}*}(t)$ to the allowed range of $[1/N, 1]$ to find the solution $\rho^*(t)$.

IV. Conclusions and Future Work

In this work, we studied the problem of optimizing the replication of a single file of time-varying popularity in a planar wireless network, assuming a given time-varying cache cost. Section II delineated the operations of the actual wireless network with caching, while Section III, the main contribution of this work, provided a macroscopic formulation based on the replication density of the file across the network and a graphical way to find the solution of the replication.

In an extended version of this work, we will show (i) the correspondence of the microscopic and the macroscopic problems of Sections II and III, and (ii) the proofs of the theorems that describe the solution of the macroscopic problem. The subjects of further research are (i) to consider the problem given a set of files of diverse time-varying popularities, and (ii) to investigate the sustainability of large wireless networks.

Acknowledgment

This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) – Research Funding Program: **THALES CROWN**. Investing in knowledge society through the European Social Fund.

References

[1] M. Franceschetti and R. Meester, *Random Networks for Communication*. New York, NY, USA: Cambridge University Press, Series: Cambridge Series in Statistical and Probabilistic Mathematics (No. 24), 2007.

[2] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Trans. Inf. Theory*, vol. 46, pp. 388–404, Mar. 2000.

[3] S. Toumpis, "Asymptotic capacity bounds for wireless networks with non-uniform traffic patterns," *IEEE Trans. Wireless Commun.*, vol. 7, pp. 2231–2242, Jun. 2008.

[4] A. Özgür, O. Lévêque, and D. Tse, "Hierarchical cooperation achieves optimal capacity scaling in ad hoc networks," *IEEE Trans. Inf. Theory*, vol. 53, pp. 3549–3572, Oct. 2007.

[5] M. Franceschetti, M. D. Migliore, and P. Minero, "The capacity of wireless networks: information-theoretic and physical limits," *IEEE Trans. Inf. Theory*, vol. 55, pp. 3413–3424, Aug. 2009.

[6] M. Franceschetti, O. Dousse, D. Tse, and P. Thiran, "Closing the gap in the capacity of wireless networks via percolation theory," *IEEE Trans. Inf. Theory*, vol. 53, pp. 1009–1018, Mar. 2007.

[7] I. Psaras, W. K. Chai, and G. Pavlou, "Probabilistic in-network caching for information-centric networks," in *Proceedings of the Second Edition of the ICN Workshop on Information-centric Networking*, ser. ICN '12. New York, NY, USA: ACM, 2012, pp. 55–60.

[8] K. Katsaros, G. Xylomenos, and G. C. Polyzos, "A hybrid overlay multicast and caching scheme for information-centric networking," in *INFOCOM IEEE Conference on Computer Communications Workshops, 2010*. IEEE, 2010, pp. 1–6.

[9] E. Cohen and S. Shenker, "Replication strategies in unstructured peer-to-peer networks," in *SIGCOMM Comput. Commun. Rev. 32, 4*, Pittsburgh, PA, USA, Oct. 2002, pp. 177–190.

[10] E. J. Rosensweig, D. S. Menasch, and J. Kurose, "On the steady-state of cache networks," in *Proc. of INFOCOM*, Torino, Italy, Apr. 2013.

[11] V. Jacobson, D. K. Smetters, J. D. Thornton, M. F. Plass, N. H. Briggs, and R. L. Braynard, "Networking named content," in *Proc. of the 5th int'l conference on Emerging networking experiments and technologies (CoNEXT '09)*, Rome, Italy, Dec. 2009, pp. 1–12.

[12] D. Trossen, M. Sarela, and K. Sollins, "Arguments for an information-centric internetworking architecture," *ACM SIGCOMM Computer Communication Review*, vol. 40, no. 2, pp. 26–33, 2010.

[13] I. Stoica, R. Morris, D. Karger, M. F. Kaashoek, and H. Balakrishnan, "Chord: A scalable peer-to-peer lookup service for internet applications," in *ACM SIGCOMM Computer Communication Review*, vol. 31, no. 4. ACM, 2001, pp. 149–160.

[14] J. Zhao, P. Zhang, G. Cao, and C. R. Das, "Cooperative caching in wireless P2P networks: Design, implementation, and evaluation," *IEEE Trans. Parallel Distrib. Syst.*, vol. 21, pp. 229–241, Feb. 2010.

[15] U. Niesen, D. Shah, and G. Wornell, "Caching in wireless networks," in *IEEE International Symposium on Information Theory*, Seoul, Korea, Jun. 2009, pp. 2111–2115.

[16] M. G. G. Alfano and E. Leonardi, "Content-centric wireless networks with limited buffers: when mobility hurts," in *Proc. of IEEE INFOCOM*, Torino, Italy, Apr. 2013.

[17] N. Golrezaei, A. G. Dimakis, and A. F. Molisch, "Wireless device-to-device communications with distributed caching," in *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*, 2012, pp. 2781–2785.

[18] S. Jin and L. Wang, "Content and service replication strategies in multi-hop wireless mesh networks," in *MSWiM '05: Proc. of the 8th ACM int'l symposium on Modeling, analysis and simulation of wireless and mobile systems*, Montréal, QC, Canada, Oct. 2005, pp. 79–86.

[19] S. Gitzenis, G. S. Paschos, and L. Tassiulas, "Asymptotic laws for joint content replication and delivery in wireless networks," *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 2760–2776, 2013.

[20] ——, "Enhancing wireless networks with caching: Asymptotic laws, sustainability & trade-offs," *Computer Networks*, vol. 64, pp. 353–368, 2014.

[21] ——, "Asymptotic laws for content replication and delivery in wireless networks," in *Proc. of INFOCOM*, Orlando, FL, USA, Mar. 2012, pp. 126–134.