EURECOM
Department of X
Campus SophiaTech
CS 50193
06904 Sophia Antipolis cedex
FRANCE

Research Report RR-14-297

Offloading on the Edge: Analysis and Optimization of Local Data Storage and Offloading in HetNets

December 1st, 2014
Last update December 1st, 2014

Pavlos Sermpezis, Luigi Vigneri, Thrasyvoulos Spyropoulos

Tel : (+33) 4 93 00 81 00
Fax : (+33) 4 93 00 82 00
Email : {pavlos.sermpezis,luigi.vigneri,thrasyvoulos.spyropoulos}@eurecom.fr

[1] EURECOM's research is partially supported by its industrial members: BMW Group Research and Technology, IABG, Monaco Telecom, Orange, Principaut de Monaco, SAP, SFR, ST Microelectronics, Symantec.

Offloading on the Edge: Analysis and Optimization of Local Data Storage and Offloading in HetNets

Pavlos Sermpezis, Luigi Vigneri, Thrasyvoulos Spyropoulos

Abstract

The rapid increase in data traffic demand has overloaded existing cellular networks. Planned upgrades in the communication architecture (e.g. LTE), while helpful, are not expected to suffice to keep up with demand. As a result, extensive densification through small cells, caching content closer to or even at the device, and device-to-device (D2D) communications are seen as necessary components for future heterogeneous cellular networks to withstand the data crunch. Nevertheless, these options imply new CAPEX and OPEX costs, extensive backhaul support, contract plan incentives for D2D, and a number of interesting tradeoffs arise for the operator. In this paper, we propose an analytical model to explore how much local storage and communication through "edge" nodes could help offload traffic in various heterogeneous network (HetNet) setups and levels of user tolerance to delays. We then use this model to optimize the storage allocation and access mode of different contents as a tradeoff between user satisfaction and cost to the operator. Finally, we validate our findings through realistic simulations and show that considerable amounts of traffic can be offloaded even under moderate densification levels.

Index Terms

Mobile Data Offloading; Device-to-Device Communications; Heterogeneous Cellular Networks; Caching; Performance Analysis

Contents

List of Figures

1 Introduction

The growth in the number of “smart” mobile devices and connection speeds has led to a high volume of mobile data traffic. Cellular networks are currently overloaded and, despite a lot of planned improvements on the physical layer technologies, they are not expected to be able to keep up with the rapidly increasing user data demand [1]. Radically reducing the communication distance by deploying, and *offloading* traffic to, many “small cells” (e.g. femto, pico, or even WiFi) is seen as the only viable solution [2–4]. Nevertheless, this requires a large investment in upgrading the backhaul network, increasingly based on wireless links, which will often be the new performance bottleneck [5]. Caching popular content at the “edge”, i.e. on storage devices installed at small cell base stations could alleviate backhaul congestion [5, 6], and is supported by a number of real data studies suggesting a high amount of demand overlap between user requests [7–9].

Reducing the communication distance is taken yet a step further with the newly proposed paradigm of device-to-device (D2D) communication [10, 11]. A device can store a (popular) content after consuming it, and give it directly to other neighboring devices also interested in it, offloading these requests from the main network. The connection between the two devices could be in-band (cellular frequencies) or out of band (e.g. Bluetooth, WiFi Direct). While D2D-based offloading normally assumes a content request will either be served *immediately* from a device currently in range or the cellular network, some recent works have suggested the use of *opportunistic offloading* through D2D: a device requesting some content might wait for some amount of time until it *encounters* another device sharing the content [12–14], and go back to the main network if not found before some set deadline.

Hence, more data could be offloaded from the main network through such D2D communication, perhaps at the expense of increased delay for some requests. Such increased delays could sometimes be acceptable (e.g. asynchronous requests, longer start-up or buffering delays easily amortized when considering large content). Yet, in many cases, the operator will need to provide appropriate incentives to these users, either in the form of instantaneous price reductions [15] or low(er) priced plans. What is more, operators will probably need to also provide incentives to the devices storing the content and acting as local *relays* on their behalf, as this raises important battery consumption, storage, as well as privacy and security issues.

The provision of these incentives constitutes another important form of cost for the operator, together with the costs of directly serving the content from the main (mostly macro-cell based) network, and that of installing, maintaining, and supporting with ample backhaul capacity, new small cells with large enough caches. It thus becomes increasingly important for an operator of such a future Heterogeneous Network (HetNet) with caching and D2D capabilities to be able to answer questions like: *”How much content can be offloaded by a given setup as a function of content demand patterns?”*, *”Is it worth investing in additional cell densifica-*

tion, or would it be more cost-efficient to provide incentives for D2D opportunistic offloading?”.

To this end, in this paper we propose an analytical model that can be used to study the problem of ”offloading on the edge” in a HetNet. Although capturing all the fine details of possible setups and technologies would be a rather daunting task, we assume two main mechanisms being employed in the considered network, namely (i) caching on small cells and mobile devices, collectively referred to as ”edge nodes”, and (ii) offloading requests through local, short range communications (e.g. D2D or low power communication to local femto or pico base stations). We describe the ”offloading on the edge” mechanism and propose a generic model that allows us to analytically study it (Section 2). We proceed by deriving useful results for the performance of content delivery through this mechanism and the incurred costs, as a function of key system parameters (Section 3). Then, we study the total offloading cost and provide insights for content placement and dissemination strategies that minimize this cost (Section 4). Finally, we validate our results through realistic simulations (Section 5) and discuss related work (Section 6).

Summarizing, the main contributions of our work are:

- To our best knowledge, this is the first work jointly and analytically studying offloading through small cells, opportunistic D2D, and caching at both.
- We provide closed-form analytical approximations applicable to a number of performance metrics and network setups.
- We provide initial insights into the various design tradeoffs involved, as well as the efficient allocation of storage space among different contents.

2 Offloading on the Edge

2.1 Network Setup

We consider a Heterogeneous Cellular Network (HetNet) [3], composed of 3 sets of nodes:

Macro-cell Base Stations ($\mathcal{BS}$): They provide full coverage to subscribed mobile nodes (MNs), but we assume their radio resources are congested.

Small Cells ($\mathcal{SC}$): These are shorter range, low power base stations (e.g. femto and pico-cells, or even WiFi access points) dispersed in the area of coverage. They provide ample capacity to the few MNs within range, and their communication cost to/from a MN is smaller [16]. Hence, they can be used to offload some traffic from BSs. However, the backhaul connection for these cells will often be wireless (either to a BS or to an aggregation point) and underprovisioned [5]. This makes a backhaul transmission to a small cell costly. To this end, each small cell is equipped with some storage capacity, as in [5, 6], where (popular) content could be cached to avoid duplicate backhaul accesses.

Mobile Nodes ($\mathcal{MN}$): These include smartphones, tablets, netbooks, etc. MNs can communicate with BSs, SCs (if in range), and even other MNs directly, if D2D communication is allowed. D2D communication potentially offers higher rates at lower interference levels [10]. Yet, appropriate incentives from the operator might be needed. Without loss of generality, we assume out-of-band communication (e.g. WiFi Direct or Bluetooth) for D2D. We also assume that each MN also has some storage capacity (normally less than that of a small cell) for caching (popular) content.

The number of nodes in each set is

$$N_{BS} = |\mathcal{BS}| \ , \ N_{SC} = |\mathcal{SC}| \ , \ N_{MN} = |\mathcal{MN}|$$

where $|\cdot|$ denotes the cardinality of a set.

2.2 Offloading Mechanism

Content Requests. We assume that each MN is interested in different contents over time (e.g. videos, web pages, software updates, etc.), and that the same content may be of interest to multiple MNs. This interest overlap is supported by recent studies (e.g. [7–9], to name a few), where the popularity distribution of contents is shown to be highly skewed. In the remainder, we will be assuming that the number of nodes interested in a content, the content popularity, is known in advance or can be estimated. For a number of applications, like *push services* [13], this information can be known in advance by the cellular network. Users are subscribed to a push service they are interested in (e.g. news, series episodes, trending videos, etc.), and when a content (of this service) is created or published, the content provider starts distributing (*pushing*) it to them[1]. Similarly, users might subscribe to certain categories of contents, such as personalized Internet radio stations like Pandora and Jango[2]. The content of these pseudo-random streams of songs can be decided in advance, and thus the popularity of songs belonging to different streams can be estimated.

Content Delivery. An operator can deliver a content to an interested MNs in one of the following ways: (i) *Direct transmission* from a BS; (ii) *Offloading through SCs and/or MNs*, where the operator transmits the content to some SCs over the backhaul and stores it there, or instructs some MNs to store a content for some time (e.g. keeping in their cache a content they consumed). Then, the operator can ask an interested MN within range of a SC or MN caching that content to retrieve it directly.

Moreover, an operator can ask an MN interested in a content θ, but not *currently* within range of an SC or MN with content θ in its cache, to wait for an amount of time, let TTL, until it *moves* within range of such an SC or MN. If this

[1] We assume that the content provider may be the cellular network operator itself or in cooperation with it (like the Akamai and Swisscom example [17]).

[2] `www.pandora.com`, `www.jango.com`

time expires, then the operator is obliged to deliver the content directly through the closest macro BS. While this *delay-tolerant* approach is in contrast to the usual ones considered for small cell and D2D based offloading [5, 6, 11], it is likely that the small cell and (D2D enabled) mobile node density will not always be enough to offload enough traffic. Hence, it is a valuable (and complementary) alternative, with potential benefits (increased offloading) and costs (reduced QoE and potential monetary incentives)[3].

2.3 Cost Model

The goal of an offloading mechanism is to minimize the cost of delivering a set of contents over time to different nodes. Hence, we need first to define a model for the costs involved in each phase of the ”offloading on the edge” mechanism.
– ***Initial Placement Costs:*** C_{BH}***,*** C_{BS}***.***
The content provider, at time $t_0 = 0$, places the content to some edge nodes (SCs and/or MNs). A content is placed to a SC through a backhaul (wired or wireless) transmission, and we denote this per placement cost as C_{BH}. A (possible) content placement to some MNs takes place through a macro-cell BS transmission. We denote this transmission cost, which mainly depends on the load/congestion of the BSs, as C_{BS}.
– ***Opportunistic Offloading Costs:*** C_{SC}***,*** C_{D2D}***.***
During time $t \in (0, TTL]$, the holders (which are either SCs or MNs) deliver the content to any requester they meet. We consider different costs for a SC-MN and a MN-MN (or D2D) transmission: C_{SC} and C_{D2D}. The former cost depends on the operating cost (transmission, energy consumption) of an SC, whereas the latter might exist if a compensation (or reward) is given by operator to MNs for each content they offload.
– ***Delayed Delivery Cost:*** $C_{BS}^{(TTL)}$***.***
At time TTL, the cellular network sends through macro-cell BSs the content to every non-served requester. This cost relates both to the load of BS (as C_{BS}) and to a (possible) compensation to the MNs for a delayed delivery. We denote this (per transmission) cost as $C_{BS}^{(TTL)}$.

2.4 Content Dissemination Model and Assumptions

Let us assume a content item (e.g. a popular video file) and a set of MNs interested in it. The content provider, at time $t_0 = 0$, places the content to the caches of some SCs and/or MNs. If by an expiry time TTL (if any), some of the

[3] Clearly, such delays might not be acceptable for all applications. However, many applications are inherently delay-tolerant, e.g. software updates, file downloads, one way streaming (e.g. YouTube or Netflix). Moreover, users might be willing to accept small or larger delays, if appropriate incentives are provided, and delayed content delivery has already been considered in a number of contexts, e.g [15, 18] .

interested MNs have not met any edge node (SC or MN) with the content, they are served by a macro-cell BS[4].

For the ease of reference, we define the following sets of ”edge nodes” that are involved in the offloading process:

Definition 1. *A* requester *of a content is a mobile node (MN) that (a) is interested in the content and (b) has not received it yet. We denote the set of requesters at time t as $\mathcal{R}(t)$.*

Definition 2. *A* holder *of a content is an edge node (SC or MN) that stores the content and will forward it to its requesters. We denote the set of holders at time t as $\mathcal{H}(t)$.*

We further denote the number of requesters and holders as:

$$R(t) = |\mathcal{R}(t)| \text{ and } H(t) = |\mathcal{H}(t)|$$

where $\mathcal{H}(t) = \mathcal{H}_{SC}(t) \cup \mathcal{H}_{MN}(t)$ and $H(t) = H_{SC}(t) + H_{MN}(t)$

To model the level of participation of MNs in the offloading mechanism, we make the following assumption.

Assumption 1 (Cooperation)**.** *A requester acts as a holder for a content it has received with probability $p_c \in [0, 1]$. The probability p_c is equal among all nodes and contents.*

The probability p_c captures either the chance a node to forward the content (e.g. it has enough resources at the time) or the percentage of nodes who are ”contracted” to help[5].

Finally, since edge nodes can exchange data only when they come within transmission range, the offloading is heavily affected by these *meeting events* between nodes. We assume the following class of node mobility.

Assumption 2 (Mobility)**.**
– The meeting events between two nodes $\{i, j\}$, $i \in \mathcal{MN}$ and $j \in \mathcal{MN} \cup \mathcal{SC}$, are given by a Poisson process with rate λ_{ij}.
– The meeting rates λ_{ij} are drawn from an (arbitrary) probability distribution $f_\lambda(\lambda)$ with mean value μ_λ.
– Meeting duration is negligible compared to the time intervals between nodes, but long enough for a content exchange.

[4] In the mechanism we consider, the content is cached only at the initial time, $t_0 = 0$, and macro-cell BSs deliver it only at its expiry time , $t = TTL$. Although one could place contents during time $t \in (0, TTL)$ as well, it has been shown (for similar settings) that placing contents at times $t \in (0, TTL)$ leads to a sub-optimal performance [13, 14].

[5] Here, we need to stress that the above assumption implies that MNs will never become holders of a content they are not interested in. Although there exist studies that assume that even not interested MNs might be willing to act as holders [13, 14, 19, 20], we believe that incentive mechanisms for these cases are difficult to implement (e.g. a user easier accepts to forward a content it already has stored, than to retrieve, cache and forward a content it will never use). Nevertheless, our framework could be easily extended also for such cases.

Assumption 2 is a tradeoff between realism (heterogeneous λ_{ij}) and tractability (Poisson process). Heterogeneous meeting rates are motivated by analysis of real mobility traces [21, 22], where not all people meet each other with the same frequency, and by the different communication ranges (SC-MN and MN-MN). Similar assumptions are common in related works [12–14, 20, 23] and have been shown to not be far from real mobility [21, 22]. Yet, in Section 5, we test our results against realistic scenarios where node mobility departs from our assumptions and involves much more complexity.

3 Analysis

An operator, in order to optimize the offloading performance and cost, has to weigh its options and take decisions about: *how to deliver* each content (directly or through offloading), *how many copies* of a content should be placed to different edge nodes, *which contents to store* in the SC and/or MN caches when their capacity is limited, etc. To this end, in this section, we provide the analytical results that are needed when trying to answer these questions. Specifically, we provide simple, closed form expressions for the performance of the ”offloading on the edge” mechanism (Section 3.1), and the costs it incurs (Section 3.2).

3.1 Content Dissemination

The performance of the “offloading on the edge” mechanism depends on how much traffic it can offload and/or how fast are contents delivered. To answer these questions, we calculate the two main (and most common) performance metrics, namely the *content delivery probability*, and *content delivery delay*.

First, we state the following Lemma, in which we use a mean field approximation and a resulting system of ODEs to approximate the number of holders and requesters over time.

Lemma 1. *The fluid-limit deterministic approximation for the expected number of holders ($H(t)$) and requesters ($R(t)$) at time t, is*

$$H(t) = H_0 \cdot \frac{(p_c \cdot R_0 + H_0) \cdot e^{\mu_\lambda \cdot (p_c \cdot R_0 + H_0) \cdot t}}{p_c \cdot R_0 + H_0 \cdot e^{\mu_\lambda \cdot (p_c \cdot R_0 + H_0) \cdot t}}$$

$$R(t) = R_0 \cdot \frac{p_c \cdot R_0 + H_0}{p_c \cdot R_0 + H_0 \cdot e^{\mu_\lambda \cdot (p_c \cdot R_0 + H_0) \cdot t}}$$

where $H_0 = H(0^+)$ and $R_0 = R(0^+)$.

Proof. Having assumed Poisson meeting processes, we can model the dissemination of a content with a continuous Markov Chain, whose states correspond to the different sets of holders and requesters $\{\mathcal{H}, \mathcal{R}\}$. Fig. 1 shows a segment of this Markov Chain; we present the different states with equal number of holders ($|\mathcal{H}|$) and requesters ($|\mathcal{R}|$) under the same group, which can be described by the tuples

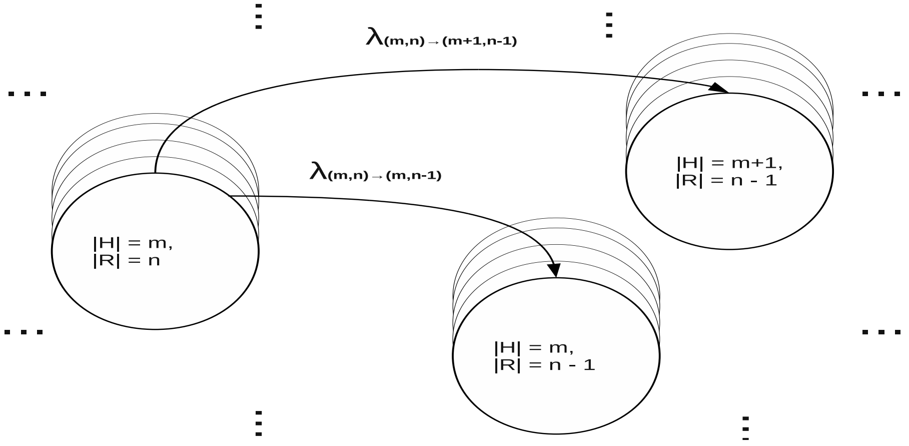


Figure 1: Content dissemination modeled by a Markov Chain.

$\{|\mathcal{H}|, |\mathcal{R}|\}$. To transition between states a *content delivery*, which takes place when a holder $i \in \mathcal{H}$ and a requester $j \in \mathcal{R}$ meet, is needed: (i) *Content delivery to cooperative node.* The next state is $\{|\mathcal{H}| = m + 1, |\mathcal{R}| = n - 1\}$ and the transition rate

$$\lambda_{(m,n)\to(m+1,n-1)} = p_c \cdot \textstyle\sum_{i\in\mathcal{H}} \sum_{j\in\mathcal{R}} \lambda_{ij} \qquad (1)$$

(ii) *Content delivery to non-cooperative node.* The next state is $\{|\mathcal{H}| = m, |\mathcal{R}| = n - 1\}$ and the transition rate

$$\lambda_{(m,n)\to(m,n-1)} = (1 - p_c) \cdot \textstyle\sum_{i\in\mathcal{H}} \sum_{j\in\mathcal{R}} \lambda_{ij} \qquad (2)$$

Statistics for the content dissemination process over time (e.g. distribution of $|\mathcal{H}(t)|$ or $|\mathcal{R}(t)|$), can be computed using the transition matrix of the Markov Chain of Fig. 1. However, this would render the problem analytically (and numerically, for large networks) intractable. To this end, we approach the problem with a mean field approximation of stochastic reaction models [24].

We first form the *drift equation* [24, Theorem 1.4.1] for the expected number of holders, $E\left[|\mathcal{H}(t)|\right] \equiv E\left[H(t)\right]$, as:

$$\frac{dE\left[H(t)\right]}{dt} = E\left[\lambda_{(m,n)\to(m+1,n-1)}\right] = p_c \cdot E\left[\sum_{i\in\mathcal{H}}\sum_{j\in\mathcal{R}} \lambda_{ij}\right]$$

The expectation in the right side of the drift equation is difficult to compute, as it requires the computation of the probabilities over the whole state space $\{\mathcal{H}, \mathcal{R}\}$. To this end, one can approximate $E[H(t)]$ with its deterministic equivalent $h(t)$. This approximation comes after neglecting the variability of $H(t)$ around its mean value and becomes more accurate for larger systems [24, Section 1.5].

Based on the deterministic approximation and since (a) the rates λ_{ij} are drawn independently from a distribution $f_\lambda(\lambda)$ with mean value μ_λ ($E[\lambda_{ij}] = \mu_\lambda$), and (b) the sum $\sum_{i\in\mathcal{H}}\sum_{j\in\mathcal{R}}\lambda_{ij}$ consists of $|\mathcal{H}|\cdot|\mathcal{R}|$ terms, we can write

$$E\left[\sum_{i\in\mathcal{H}}\sum_{j\in\mathcal{R}}\lambda_{ij}\right] \approx h(t)\cdot r(t)\cdot\mu_\lambda \tag{3}$$

The higher the number of terms in the above sum, and the less the heterogeneity of the meeting rates (i.e. the variance of $f_\lambda(\lambda)$), the more accurate the approximation in Eq. (3) is.

Substituting Eq. (3) in the drift equation (where $H(t)\rightarrow h(t)$), gives the ordinary differential equation (ODE) for $h(t)$[6]

$$\frac{dh(t)}{dt} = p_c\cdot h(t)\cdot r(t)\cdot\mu_\lambda \tag{4}$$

Proceeding similarly, the ODE for the deterministic approximation of the number of requesters ($R(t)\rightarrow r(t)$), is

$$\frac{dr(t)}{dt} = -h(t)\cdot r(t)\cdot\mu_\lambda \tag{5}$$

Finally, solving the system of the ODEs of Eq. (4) and Eq. (5), gives the expressions of Lemma 1. □

Based on Lemma 1 we, now, proceed to the calculation of the performance metrics. Let us consider a requester $i\in\mathcal{R}(0^+)$, and denote as T_i the time it receives the content. The probability this (random) requester to receive the content by a time t, i.e. $P\{T_i\leq t\}$, is equal to the *percentage of offloaded contents* by time t. Hence, we can write

$$P\{T_i\leq t\} = \frac{R_0 - R(t)}{R_0} = 1 - \frac{R(t)}{R_0} \tag{6}$$

Substituting the expression of Lemma 1 in Eq. (6), gives the following Result for the content delivery probability

Result 1 (Delivery Probability). *The probability a content to be delivered to a requester by time t is given by*

$$P\{T_d\leq t\} = 1 - \frac{p_c\cdot R_0 + H_0}{p_c\cdot R_0 + H_0\cdot e^{\mu_\lambda\cdot(p_c\cdot R_0+H_0)\cdot t}}$$

where $H_0 = H(0^+)$ and $R_0 = R(0^+)$.

[6] Note the differences between $H(t)$ and $h(t)$: (a) $H(t)$ is integer, whereas $h(t)$ is a real number; (b) the drift equation for $H(t)$ contains expectations, while the respective ODE for $h(t)$ does not.

With respect to the average delay a requester experiences till it receives the content, we state the following Result (the proof is technical and can be found in Appendix 8.1). We derive expressions for two cases: (a) the content does not expire (i.e. $TTL \to \infty$), and (b) a macro-cell BS serves undelivered contents at time $t = TTL$. .

Result 2 (**Delivery Delay**). *The expected content delivery delay, under an expiry time* $TTL \in [0, \infty)$*, is given by*
– for $p_c > 0$*:*

$$E[T_d|TTL] = \frac{\ln\left(1 + \frac{p_c \cdot R_0 - e^{-\mu_\lambda \cdot (p_c \cdot R_0 + H_0) \cdot TTL}}{H_0 + p_c \cdot R_0 \cdot e^{-\mu_\lambda \cdot (p_c \cdot R_0 + H_0) \cdot TTL}}\right)}{\mu_\lambda \cdot p_c \cdot R_0}$$

– for $p_c = 0$*:*

$$E[T_d|TTL] = \frac{1 - e^{-\mu_\lambda \cdot H_0 \cdot TTL}}{\mu_\lambda \cdot H_0}$$

where $H_0 = H(0^+)$ *and* $R_0 = R(0^+)$*.*

3.2 Content Delivery Cost

Incorporating the offloading costs (Section 2.3) in our content dissemination model, and using the analytical results of Section 3.1, we calculate the cost of a single content delivery in Result 3. The expression we derive, gives the cost as a (simple) function of the system parameters (e.g. R_0, μ_λ) and the operator selected parameters (e.g. $H_{SC}(0)$, $H_{MN}(0)$), providing, thus, the necessary information for the evaluation and tuning of the "offloading on the edge" mechanism.

Result 3. *The cost of "offloading on the edge" a content is given by*

$$\begin{aligned} C =& C_{BH} \cdot H_{SC}(0) + C_{BS} \cdot H_{MN}(0) \\ &+ (C_{SC} \cdot q + C_{D2D} \cdot (1-q)) \cdot R_0 \cdot P\{T_d \leq TTL\} \\ &+ C_{BS}^{(TTL)} \cdot R_0 \cdot (1 - P\{T_d \leq TTL\}) \end{aligned}$$

where $q = \frac{H_{SC}(0) \cdot \ln\left(\frac{H(TTL)}{H_0}\right)}{p_c \cdot (R_0 - R(TTL))}$*, and* $P\{T_d \leq\ TTL\}$*,* $H(TTL)$ *and* $R(TTL)$ *are given in Lemma 1 and Result 1.*

Proof.
– Initial Placement. The first two terms correspond to the initial placement phase: The cellular network operator, at time $t = 0$, places the content to $H_{SC}(0)$ SCs and $H_{MN}(0)$ MNs; in total ($H_0 = H_{SC}(0) + H_{MN}(0)$) holders. The costs per content placement are C_{BH} and C_{BS}, respectively.

– *Opportunistic Offloading.* During the opportunistic offloading phase, i.e. $t \in (0, TTL)$, the average number of requesters that receive the content by an edge node is $R_0 \cdot P\{T_d \leq TTL\}$. If we denote with q the percentage of requesters that receive the content by a SC, it is easy to see that the costs due to SC-MN and MN-MN content deliveries are

$$C_{SC} \cdot q \cdot R_0 \cdot P\{T_d \leq TTL\} \tag{7}$$

$$C_{D2D} \cdot (1 - q) \cdot R_0 \cdot P\{T_d \leq TTL\} \tag{8}$$

respectively.

To calculate the percentage q we proceed as following:

At first, the total number of requesters that receive the content by time TTL is

$$\#R_{tot} = R_0 - R(t) \tag{9}$$

Second, the total number of requesters that receive the content in the interval $(t, t + dt]$, $t \in (0, TTL)$ is

$$R(t) - R(t, t + dt) = -dR(t) \tag{10}$$

The probability that a content delivery that takes place in the interval in the interval $(t, t + dt]$ is due to a SC is equal to

$$\frac{H_{SC}(0)}{H(t)} \in [0, 1] \tag{11}$$

where $H_{SC}(0)$ is the number of SC holders (which does not change over time), and $H(t)$ the total number of holders at time t.

Therefore, the number of requesters that receive the content by an SC holder in the interval $(t, t + dt]$ is given by $-dR(t) \cdot \frac{H_{SC}(0)}{H(t)}$, and the total number of requesters that receive the content by an SC holder by time TTL is

$$\begin{aligned}\#R_{SC} &= \textstyle\int_0^{TTL} -dR(t) \cdot \frac{H_{SC}(0)}{H(t)} = \int_0^{TTL} -\frac{dR(t)}{dt} \cdot \frac{H_{SC}(0)}{H(t)} \cdot dt \\ &\overset{\text{Eq. (5)}}{=} \textstyle\int_0^{TTL} H(t) \cdot R(t) \cdot \mu_\lambda \cdot \frac{H_{SC}(0)}{H(t)} \cdot dt \\ &= \mu_\lambda \cdot H_{SC}(0) \textstyle\int_0^{TTL} R(t) \cdot dt \end{aligned} \tag{12}$$

Using the expression of Lemma 1 for $R(t)$ to calculate the above integral, we get

$$\begin{aligned}\#R_{SC} &= \frac{H_{SC}(0)}{p_c} \cdot \ln\left(\frac{(p_c \cdot R_0 + H_0) \cdot e^{\mu_\lambda \cdot (p_c \cdot R_0 + H_0) \cdot TTL}}{p_c \cdot R_0 + H_0 \cdot e^{\mu_\lambda \cdot (p_c \cdot R_0 + H_0) \cdot TTL}}\right) \\ &= \frac{H_{SC}(0)}{p_c} \cdot \ln\left(\frac{H(TTL)}{H_0}\right) \end{aligned} \tag{13}$$

where the last equality follows from the expression for $H(t)$ given in Lemma 1.

Now, q easily follows from Eq. (9) and Eq. (13)

$$q = \frac{\#R_{SC}}{\#R_{tot}} = \frac{H_{SC}(0)}{p_c} \cdot \frac{\ln\left(\frac{H(TTL)}{H_0}\right)}{R_0 - R(TTL)} \tag{14}$$

$-$ *Delayed Delivery.* Finally, the average number of requesters that do not receive the content before its expiry time, is given by $R_0 \cdot (1 - P\{T_d \leq TTL\})$. Since, the cost of each content transmission at time $t = TTL$ is $C_{BS}^{(TTL)}$, the total cost of delayed delivery phase (last line of the expression in Lemma 3) follows easily. □

4 Applications: Cost Optimization

In a real scenario, the network operator would have to offload simultaneusly many different contents. Using the results of the previous section, the average performance or the total cost over all the contents can be calculated easily, by evaluating them for each content separately and then averaging or summing them, respectively. However, since some of the system parameters are controlled by the operator (e.g. H_0), they can be selected such that they lead to optimal performance. To this end, in this section, as an application of our analytical results, we study how offloading and caching can be designed in order to minimize the total cost.

Let us assume that the content provider has to deliver $M \geq 1$ contents to their requesters. We denote the set of the contents as $\mathcal{M}$ ($M = |\mathcal{M}|$). Since in a real scenario, not all contents are expected to be equally popular [7–9], nor tolerate equal delays, we denote the popularity (i.e. the number of initial requesters) and the expiry time of each content $\theta \in \mathcal{M}$ as R_0^θ and TTL^θ, respectively.

Under a given setting (i.e. with certain mobility, cooperation, traffic, etc., characteristics), what the cellular network can select, is the initial placement (*caching*) for each content $\theta \in \mathcal{M}$; namely, the number of SC and MN initial holders, $H_{SC}^\theta(0)$ and $H_{MN}^\theta(0)$, respectively (note that $H_0^\theta(0) \equiv H_{SC}^\theta(0) + H_{MN}^\theta(0)$). Additionally, it might be possible that the delay-tolerance of each content, TTL^θ, can be selected as well.

Therefore, if we denote as C^θ is the delivery cost of a content $\theta \in \mathcal{M}$ (which is given by Result 3), we can express the *total* cost optimization problem as

Problem 1.

$$\min_{\overline{H}_{SC},\, \overline{H}_{MN},\, \overline{TTL}} \ \left\{\textstyle\sum_{\theta\in\mathcal{M}} C^\theta\right\}$$

$$\begin{aligned} s.t.\ \forall\theta \in \mathcal{M}:\ & 0 \leq H_{SC}^\theta(0) \leq N_{SC} \\ & 0 \leq H_{MN}^\theta(0) \leq R^\theta(0) \\ & T_{min} \leq TTL^\theta \leq T_{max} \\ \textit{and} \qquad & \sum_{\theta\in\mathcal{M}} H_{SC}^\theta(0) \leq \sum_{i\in\mathcal{SC}} Q(i) \end{aligned}$$

where $\overline{H_{SC}}$, $\overline{H_{MN}}$ and $\overline{TTL}$ denote the vectors with components $H^{\theta}_{SC}(0)$, $H^{\theta}_{MN}(0)$ and TTL^{θ} ($\theta \in \mathcal{M}$), respectively, and $Q(i)$ is the caching capacity (in number of contents) of a SC node i.

Remark: Since MNs cache only contents in which they are interested in, we assume that their storage capacity is enough for all the contents of interest. Hence, storage capacity constraints for MN are not considered in Problem 1.

Since the costs C^{θ} are expressed as a function of the optimization variables (Result 3), well known numerical methods can be employed to solve Problem 1. Under certain scenarios, analytical solutions for Problem 1 can be found as well. In the remainder, we focus on two characteristics cases, which are analytically solvable, and provide useful insights for the system.

4.1 Offloading through SCs

We first consider the case where contents are offloaded only through SCs (i.e. when $p_c = 0$ and $H^{\theta}_{MN}(0) = 0$, or equivalently, $H^{\theta}_{0} = H^{\theta}_{SC}(0)$). This is the most common and feasible scenario considered in previous literature, since MNs are not required to share their contents, and thus incentive mechanisms are easier to implement. In this case and for[7] $C_{SC} < C^{(TTL)}_{BS}$ it can be proved that Problem 1 is convex and we compute the analytical solution in Result 4. For notation simplicity, we consider equal expiry times $TTL^{\theta} = TTL$, $\forall \theta \in \mathcal{M}$, and cache sizes $Q(i) = Q$, $\forall i \in \mathcal{SC}$. However, Result 4 can be easily modified for different[8] TTL^{θ} and $Q(i)$ values.

Result 4. *Under a base scenario ($p_c = 0$, $H_{MN}(0) = 0$), the initial allocation $\overline{H_{SC}}$ that minimizes the total cost, is given by*

$$H^{\theta}_{SC}(0) = \begin{cases} N_{SC} & , R^{\theta}(0) > U \\ \frac{1}{\gamma} \cdot \ln\left(\frac{1}{L} \cdot R^{\theta}(0)\right) & , L \leq R^{\theta}(0) \leq U \\ 0 & , R^{\theta}(0) < L \end{cases}$$

with $\gamma = \mu_{\lambda} \cdot TTL$, $L = \frac{1}{\gamma \cdot \Phi} \cdot \left(1 + \frac{\lambda_0}{C_{BH}}\right)$, $U = L \cdot e^{\gamma \cdot N_{SC}}$, $\Phi = \frac{C^{(TTL)}_{BS} - C_{SC}}{C_{BH}}$, and

$$\lambda_0 = \inf\left\{\lambda_0 \geq 0 : \sum_{\theta \in \mathcal{M}} H^{\theta}_{SC}(0) \leq \sum_{i \in \mathcal{SC}} Q(i)\right\}$$

Proof. Applying the method of Lagrange multipliers [25] to Problem 1, gives (for brevity we use the notation $H^{\theta}_{0} \equiv H^{\theta}_{SC}(0^{+}) = H^{\theta}_{SC}(0)$ and $R^{\theta}_{0} \equiv R^{\theta}(0^{+}) =$

[7] The “offloading on the edge” mechanism is meaningful if $C_{SC} < C^{(TTL)}_{BS}$, as in the opposite case, offloading would cost more than directly delivering from the macro-cell BSs.

[8] In particular, one has to substitute γ with $\gamma^{\theta} = \mu_{\lambda} \cdot TTL^{\theta}$ for each content. The expressions for $H^{\theta}_{SC}(0)$ remain the same, and only the expressions of L and U need to be modified.

$R^{\theta}(0)$):

$$\nabla\left(\sum_{\theta\in\mathcal{M}} C^{\theta}\right) = \nabla\lambda_0\left(\sum_{i\in\mathcal{SC}} Q(i) - \sum_{\theta\in\mathcal{M}} H_0^{\theta}\right) + \nabla\sum_{\theta\in\mathcal{M}}\lambda_{\theta}\cdot H_0^{\theta} + \nabla\sum_{\theta\in\mathcal{M}}\mu_{\theta}\cdot(N_{SC} - H_0^{\theta}) \quad (15)$$

where $\lambda_0 \geq 0$ and $\lambda_{\theta}, \mu_{\theta} \geq 0, \forall\theta \in \mathcal{M}$ are the langrangian multipliers.

Using the expression of Result 1 for the delivery probability, the offloading cost (Result 3) of a content θ, in a base scenario, can be written as

$$C^{\theta} = C_{BH}\cdot H_0^{\theta} + C_{SC}\cdot R_0^{\theta} + (C_{BS}^{(TTL)} - C_{SC})\cdot R_0^{\theta}\cdot e^{-\mu_{\lambda}\cdot H_0^{\theta}\cdot TTL} \quad (16)$$

Substituting C^{θ} from Eq. (16) to Eq. (15), the differentiation over H_0^{θ} gives

$$H_0^{\theta} = \frac{1}{\gamma}\cdot\left[\ln\left(\Phi\cdot\gamma\cdot R_0^{\theta}\right) - \ln\left(1 + \frac{\lambda_0 - \lambda_{\theta} + \mu_{\theta}}{C_{BH}}\right)\right] \quad (17)$$

The conditions for the lagrangian multipliers, i.e.

$$\lambda_{\theta}\cdot H_0^{\theta} = 0, \quad \text{and} \quad \mu_{\theta}\cdot(N_{SC} - H_0^{\theta}) = 0 \quad ,\forall\theta\in\mathcal{M}$$

imply that H_0^{θ} either

(a) is given by Eq. (17) and $\lambda_{\theta} = \mu_{\theta} = 0$, or

(b) is equal to N_{SC} and $\lambda_{\theta} = 0, \mu_{\theta} > 0$, or

(c) is equal to 0 and $\lambda_{\theta} > 0, \mu_{\theta} = 0$

From condition (a), we calculate the limits of the interval within which the optimal H_0^{θ} is given by Eq. (17). To find the lower limit, L, we set H_0^{θ} (Eq. (17) with $\lambda_{\theta} = \mu_{\theta} = 0$) equal to 0 and for the upper limit, U, equal to N_{SC}, which give

$$L = \frac{1}{\gamma\cdot\Phi}\cdot\left(1 + \frac{\lambda_0}{C_{BH}}\right) \quad (18a)$$

$$U = \frac{1}{\gamma\cdot\Phi}\cdot e^{\gamma\cdot N_{SC}}\cdot\left(1 + \frac{\lambda_0}{C_{BH}}\right) = L\cdot e^{\gamma\cdot N_{SC}} \quad (18b)$$

Combining Eq. (17) and Eqs. (18), we can express the optimal placement as

$$H_0^{\theta\,*} = \begin{cases} N_{SC} & , R_0^{\theta} > U \\ \frac{\ln\left(\gamma\cdot\Phi\cdot R_0^{\theta}\right) - \ln\left(1 + \frac{\lambda_0}{C_{BH}}\right)}{\gamma} & , L \leq R_0^{\theta} \leq U \\ 0 & , R_0^{\theta} < L \end{cases} \quad (19)$$

The only unknown parameter in Eq. (19) is λ_0 (since we expressed L and U as functions of λ_0). Lemma 2, which we state and prove in Appendix 8.2, suggests that the total cost, $\sum_{\theta \in \mathcal{M}} C^{\theta}$, is monotonically increasing with λ_0. Therefore, the optimal placement policy corresponds to the smaller *non-negative* value of λ_0 that satisfies the storage constraint, $\sum_{\theta \in \mathcal{M}} H_0^{\theta} \leq \sum_{i \in \mathcal{SC}} Q(i)$, and this proves the Result. □

In general, the value of the parameter λ_0 can be found (within some precision) with e.g. a binary search. Nevertheless, for a large number of contents, and given their popularity distribution, its value can be directly calculated using the Corollary 1, which follows after substituting the expression of Result 4 and the popularity density function in the storage constraint $\sum_{\theta \in \mathcal{M}} H_{SC}^{\theta}(0) = \sum_{i \in \mathcal{SC}} Q(i)$.

Corollary 1. *Under a content popularity distribution $\rho(x)$, the parameter λ_0 in Result 4 is given by $\lambda_0 = \max\left\{0, \hat{\lambda_0}\right\}$, where $\hat{\lambda_0}$ is the (minimum) solution of*

$$\int_L^U \ln\left(\gamma \cdot \Phi \cdot x\right) \cdot \rho(x)dx - \ln\left(1 + \frac{\lambda_0}{C_{BH}}\right) \cdot \int_L^U \rho(x)dx + \gamma \cdot N_{SC} \cdot \int_U^{\infty} \rho(x)dx = \frac{\gamma \cdot N_{SC} \cdot Q}{M}$$

Result 4 reveals how resources should be allocated: (i) The optimal allocation is logarithmic in popularity, with either large or small caches. (ii) When capacity is limited, an extra factor (λ_0) is introduced, so that the *relative* allocation remains logarithmic, but the absolute allocation is reduced (normalized) as the number of contents increase, or total capacity decreases. (iii) Some low popularity contents might get no allocation, either because it does not help the offloading cost, or because there is not enough capacity for them.

Practical Example: Assume an urban area covered by $N_{BS} = 4$ macro-cell BSs and $N_{SC} = 100$ SCs. On average, in this area reside $N_{MN} = 10000$ users[9] with an average meeting rate $\mu_{\lambda} = 3.3 \cdot 10^{-5}$ meetings/sec (equal to this of the real mobility trace [26]). The cellular network has to deliver M contents (e.g. YouTube video files of an average size $10MB$ [7]) with expiry time $TTL \approx 5min$ and popularity given by a bounded Pareto distribution in the interval $R_0 \in [10, 1000]$ with shape parameter $\alpha = 0.5$ [7]. The costs are[10] $C_{BS}^{(TTL)} = 10 \cdot C_{BH}$ and $C_{SC} \ll C_{BH}, C_{BS}^{(TTL)}$.

[9] Vodafone Germany reported an average number of 1700 users per cell (`http://mobilesociety.typepad.com/mobile_life/2009/06/base-station-numbers.html`). In an urban environment, users density is expected to be higher.

[10] In general, the offloading costs incurred in each phase, might differ between areas, time periods and operators. Their absolute values are not available and/or are difficult to estimate. To this end, in this example, as well as in other numerical results, we use relative values inferred by some average values proposed in [16].

Substituting the given values, and taking the expectation over the popularity distribution, it follows that the necessary buffer size of a SC, $Q = \frac{E[H_0]}{N_{SC}} \cdot M \cdot L$, is approximately $1MB$ per content. This means that, even under very high traffic demand, the caching capacity of the SCs would be adequate such that the last constraint of Problem 1 is not violated; e.g. for $M = 100000$ (i.e. each user requests 10 videos per 5 minutes!), the needed capacity is $Q = 100GB$ (which is a feasible and relatively cheap investment).

4.2 Offloading through MNs

We now consider the case where offloading takes place only through MN-MN communication ($p_c > 0$) and *without* content storing on SCs (i.e. $H_{SC}(0) = 0$). A content is initially sent by the BSs to $H_{MN}(0)$ (out of $R(0)$) of its requesters, which start disseminating it to the other requesters. However, not all nodes might be willing to participate by acting as holders, which in our framework means that each node (including the initial nodes in which the content is placed) cooperates with probability p_c. Therefore, we can write

$$H_0 \equiv H_{MN}(0^+) = p_c \cdot H_{MN}(0)$$

Also, as defined in Lemma 1,

$$R_0 \equiv R(0^+) = R(0) - H_{MN}(0)$$

As in the previous case, we assume equal expiry times $TTL^\theta = TTL$, $\forall \theta \in \mathcal{M}$.

Result 5. *Under an opportunistic MN-MN scenario ($p_c > 0$, $H_{MN}(0) = 0$), the initial allocation $\overline{H_{MN}}$ that minimizes the total cost, is given by*

$$H^\theta_{MN}(0) = \begin{cases} R^\theta(0) & , R^\theta(0) \leq OPT^\theta \\ OPT^\theta & , 0 \leq OPT^\theta < R^\theta(0) \\ 0 & , OPT^\theta < 0 \end{cases}$$

where $OPT^\theta = \dfrac{R^\theta(0) \cdot \left(\sqrt{\Phi'} \cdot e^{\frac{1}{2}\gamma \cdot p_c \cdot R^\theta(0)} - 1\right)}{e^{\gamma \cdot p_c \cdot R^\theta(0)} - 1}$*, and* $\Phi' = \frac{C_{BS}^{(TTL)} - C_{D2D}}{C_{BS} - C_{D2D}}$ *and* $\gamma = \mu_\lambda \cdot TTL$*.*

Proof. The cost for offloading a content θ under an opportunistic MN-MN scenario, where $H_0^\theta = p_c \cdot H^\theta_{MN}(0)$ and $R^\theta_0 = R(0)^\theta - H^\theta_{MN}(0)$, is (see Result 3)

$$\begin{aligned} C^\theta = C_{BS} \cdot H^\theta_{MN}(0) & \\ + \left(C_{D2D} - C_{BS}^{(TTL)}\right) \cdot (R^\theta(0) - H^\theta_{MN}(0)) \cdot P\{T_d \leq TTL\} & \\ + C_{BS}^{(TTL)} \cdot (R^\theta(0) - H^\theta_{MN}(0)) & \end{aligned} \tag{20}$$

Similarly, for $H_0^\theta = p_c \cdot H_{MN}^\theta(0)$ and $R_0^\theta = R^\theta(0) - H_{MN}^\theta(0)$, the delivery probability $P\{T_d \leq TTL\}$ can be written as

$$P\{T_d \leq TTL\} = 1 - \frac{R^\theta(0)}{R^\theta(0)+H_{MN}^\theta(0)\cdot\left(e^{\gamma\cdot p_c\cdot R^\theta(0)}-1\right)} \tag{21}$$

where $\gamma = \mu_\lambda \cdot TTL$.

Substituting Eq. (21) in Eq. (20), and taking the derivative over the initial number of transmissions $H_{MN}^\theta(0)$, gives

$$\frac{dC^\theta}{dH_{MN}^\theta(0)} = (C_{BS}^{(TTL)} - C_{D2D}) + \frac{(C_{D2D} - C_{BS}) \cdot (R^\theta(0))^2 \cdot e^{\gamma\cdot p_c\cdot R^\theta(0)}}{\left(R^\theta(0) + H_{MN}^\theta(0) \cdot (e^{\gamma\cdot p_c\cdot R^\theta(0)} - 1)\right)^2} \tag{22}$$

From Eq. (22) it follows that

$$\frac{dC^\theta}{dH_{MN}^\theta(0)} = \begin{cases} < 0 & , H_{MN}^\theta(0) < OPT^\theta \\ > 0 & , H_{MN}^\theta(0) > OPT^\theta \end{cases}$$

where

$$OPT^\theta = \frac{R^\theta(0) \cdot \left(\sqrt{\Phi'} \cdot e^{\frac{1}{2}\gamma\cdot p_c\cdot R^\theta(0)} - 1\right)}{e^{\gamma\cdot p_c\cdot R^\theta(0)} - 1} \tag{23}$$

Therefore, when $OPT^\theta \in [0, R^\theta(0)]$, the minimum cost is achieved for $H_{MN}^\theta(0) = OPT^\theta$. Otherwise, for $OPT^\theta \notin [0, R^\theta(0)]$, and since it must hold that $H_{MN}^\theta(0) \in [0, R^\theta(0)]$, the minimum cost is achieved for the largest or lowest possible values of $H_{MN}^\theta(0)$. □

Result 5 reveals how content storage should be delivered when offloading only through MNs is considered. As it can be seen, the initial allocation is much different that in the offloading through SCs case (see Result 4), and this is mainly due to the fact that some of the requesters get the content at the beginning.

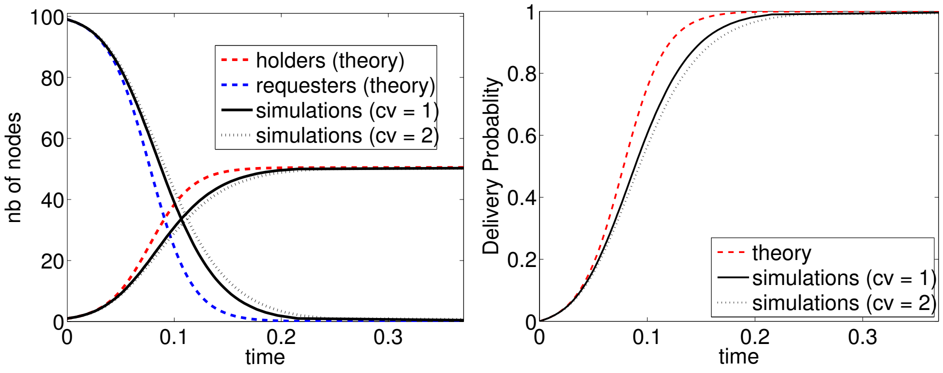
5 Simulation Results

To validate our analysis, we compare the theoretical predictions against Monte Carlo simulations (Section 5.1). Then, we evaluate the cost efficiency of ”offloading on the edge” in scenarios with realistic traffic demand patterns (Section 5.2).

5.1 Model Validation

5.1.1 Synthetic Scenarios

We first compare the theoretical results against Monte Carlo simulations on various synthetic scenarios. Synthetic simulations allow us to create a number of different scenarios with varying parameters.



(a) $H_0 = 1, p_c = 0.5$

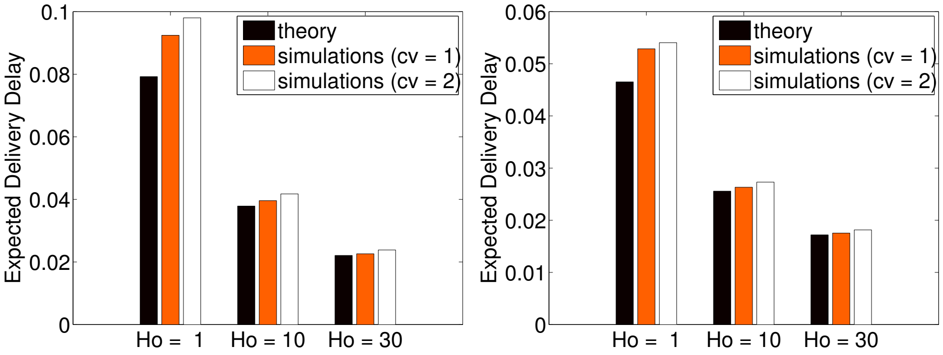
(b) $H_0 = 1, p_c = 0.5$

Figure 2: (a) Expected number of holders, $H(t)$, and requesters, $R(t)$, over time for generic scenarios with $R_0 = 100$, $H_{SC} = 0$; (b) shows the corresponding results for the delivery probability, i.e. $P\{T_d \leq TTL\}$, where TTL is the x-axis variable.

We generate synthetic networks, conforming to the model of Section 3, as following:
(i) We choose a probability distribution $f_\lambda(\lambda)$ and for each pair $\{i, j\}$ we draw randomly a meeting rate λ_{ij}.
(ii) We create a sequence of contact events for every pair in the network with rate (Poisson processes with rates λ_{ij}).
(iii) We select the content traffic parameters (R_0, H_0, p_c, $H_{SC}(0)$, $H_{MN}(0)$, N_{SC}), and we simulate a large number of content disseminations, choosing randomly each time the set of requesters and the set of holders (note, however, that the set of holders depends also on the parameters $H_{SC}(0)$, $H_{MN}(0)$ and N_{SC}).

We have created many scenarios with different combinations of mobility ($f_\lambda(\lambda)$) and traffic (R_0, H_0, p_c, $H_{SC}(0)$, $H_{MN}(0)$, N_{SC}) characteristics. We present here a representative subset of them, which allow us demonstrate the accuracy of our predictions and their sensitivity when varying certain parameters. In the presented scenarios we create nodes mobility according to a gamma distribution $f_\lambda(\lambda)$ with mean value $\mu_\lambda = 1$ (i.e. normalized value) and variance σ_λ^2 (or, equivalently, coefficient of variation $CV_\lambda = \frac{\sigma_\lambda}{\mu_\lambda}$) [27]. Gamma distributions allow us to capture different levels of mobility heterogeneity by varying the value of CV_λ.

Content Dissemination. In Fig. 2 we compare simulation results (average values over the different runs) of expected number of holders ($H(t)$) / requesters ($R(t)$) and content delivery probability $P\{T_d \leq TTL\}$ with the respective theoretical predictions (Lemma 1 and Result 1, respectively). Considering the same content traffic parameters, we simulated scenarios with moderate ($CV_\lambda = 1$) and high ($CV_\lambda = 2$) mobility variance, in order to show how mobility heterogeneity affects the accuracy of our predictions. It can be seen that our predictions become



(a) $p_c = 0.5$

(b) $p_c = 1$

Figure 3: Expected delivery delay, $E[T_d]$, for various generic scenarios with $R_0 = 100$, $H_{SC} = 0$ and (a) $p_c = 0.5$, (b) $p_c = 1$.

more accurate for lower mobility heterogeneity ($CV_\lambda = 1$). This is due to the mean field approximation of the transitions rates we used in the analysis (see Section 3.1). For scenarios with even lower mobility heterogeneity (e.g. $CV_\lambda = 0.5$ - not shown in the plots) the accuracy is even better. Additionally, we need to highlight that these results correspond to an initial allocation of only one holder ($H_0 = 1$), which is the *worst case* scenario (i.e. lowest accuracy of the mean field approximation, and, thus our predictions) among the ones with the given mobility and traffic (other than H_0) characteristics. In the same scenarios, when considering a few more initial holders, e.g. $H_0 = 10$, theoretical results achieve an almost exact prediction.

Similar observations can be made in Fig. 3, where we compare the theoretically predicted delivery delays with the respective simulation results. The results in Fig. 3 are in accordance with the above observations, i.e. the predictions' accuracy increases for (a) lower CV_λ, and (b) higher number of initial holders H_0.

Offloading Cost. We finally present results that validate the cost optimization analysis of Section 4. Fig. 4 shows the incurred cost for the cellular network (y-axis) under different number of initial holders H_0 (x-axis) for various generic traffic scenarios. Different cooperation policies (top plots: $p_c = 1$, middle plots: $p_c = 0.5$, and bottom plots: $p_c = 0$) and expiry times TTL (or, equvalently, $\gamma = \mu_\lambda \cdot TTL$) are considered. It can be seen that our results accurately predict the content dissemination cost.

Some remarkable observations about the optimal initial allocation of holders that can be made in Fig. 4 (as well as in other scenarios we investigated) are the following: (i) In many cases, offloading on the edge can signifantly reduce the cost of a content dissemination. For instance, in the scenario shown in Fig. 4 (bottom plot - bottom curve / black color), even without node cooperation ($p_c = 0$), offloading on the edge can reduce the cost 10 times, compared to the corresponding scenario without offloading (i.e. $C = 100$). (ii) An optimal initial allocation

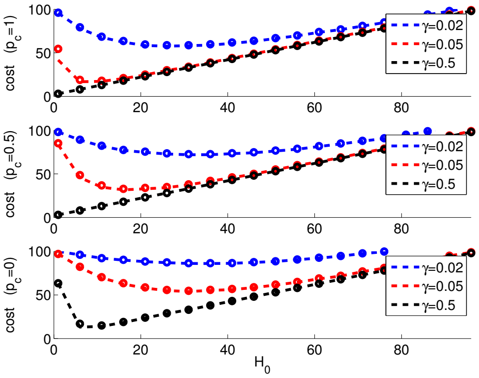


Figure 4: Single content offloading cost C (Lemma 3) under different number of initial holders (H_0, x-axis) for a synthetic mobility scenario with $R_0 = 100$, $H_{SC} = H_0$, and $C_{BH} = C_{BS}^{(TTL)} = 50 \cdot C_{SC}$. Dashed lines correspond to theoretical predictions and markers to simulation results. We denote $\gamma = \mu_\lambda \cdot TTL$.

requires only a small number of (initial) storage resources, which in most of the cases we present is equal or less than 20% of the content requesters. (iii) The higher the allowed delay (i.e. expiry time TTL or parameter γ) is, the larger the gain the cellular network can have is. For example, consider the red line ($\gamma = 0.05$) in the bottom plot. Increasing $\times 10$ the value of TTL (black line - $\gamma = 0.5$) can reduce the cost (e.g. for $H_0 = 5$ which is close to the optimal allocation) almost 8 times.

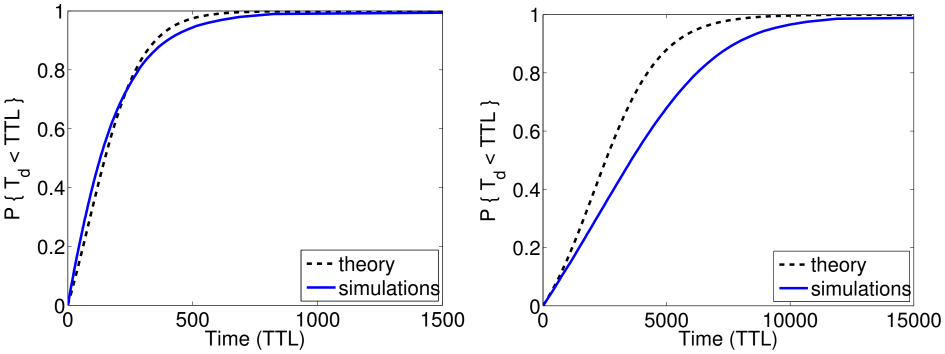
5.1.2 Mobility Traces

Results of synthetic simulations demonstrate a significant accuracy of our predictions and verify the arguments used in the derivation of our results. In this section, we present results in more challenging scenarios, where node mobility characteristics depart from our model assumptions.

Specifically, we use the TVCM [28] and SLAW [29] mobility models, which have been shown to capture well real mobility patterns, like power-law flights [29], community structure [28], etc. The generated scenarios we present are

TVCM scenario: Mobile nodes move in a square area $1000m \times 1000m$, which contains three areas of interest (communities). Nodes move mainly inside their community (60% of the time) and leave it for a few short periods. Macro-cell BSs provide full coverage of the whole area, while 25 non-overlapping (placed on a grid) small-cell base stations (SCs), with a communication range of $100m$, provide further connectivity. Mobile nodes are equipped with *D2D* communication interfaces, for which we assume a range of $30m$.

SLAW scenario: A square area of edge length $2000m$ is simulated, where mobile nodes either move or remain static for a maximum time of $20min$ (the other mobility parameters are set as in the source code provided by [29]). Macro-cell BSs



(a) TVCM: $H(0) = 10$, $R(0) = 50$

(b) SLAW: $H(0) = 5$, $R(0) = 50$

Figure 5: Delivery probability $P\{T_d \leq TTL\}$ over time TTL (x-axis), for the (a) TVCM and (b) SLAW scenarios with $p_c = 0.5$ and $H_{SC}(0) = H_{MN}(0) = \frac{H(0)}{2}$.

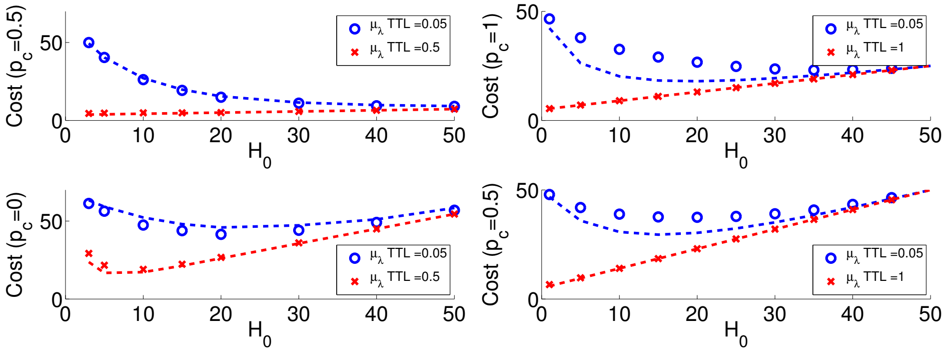
cover the whole area and coexist with 100 non-overlapping small-cells. Communication ranges are set as above.

In Fig. 5 we present the delivery probability $P\{T_d \leq TTL\}$, along with the theoretical prediction, for two content traffic scenarios in the TVCM (Fig. 5(a)) and SLAW (Fig. 5(b)) traces. Contents with popularity $R(0) = 50$ are initially cached to $H(0)$ edge nodes (half of which are MNs). The MNs' participation in offloading is set to $p_c = 0.5$. In the TVCM trace (Fig. 5(a)) it can be seen that the accuracy of our results is significant, despite the community structure of the network (which cannot be captured explicitly by our mobility Assumption 2). In the SLAW scenario (Fig. 5(b)), our results overestimate the delivery probability. However, note here that the number of holders in the SLAW scenario is smaller, and, thus, our approximation is expected to be less accurate. For scenarios with more initial holders the accuracy of the predictions increase (see e.g. Fig. 6(b), where the accuracy is higher for higher H_0 values).

Although in some points the theoretical performance metrics deviate considerably from simulations (e.g. 20%), the accuracy of the cost metrics (Lemma 3) is less affected. Fig. 6 shows the incurred cost for delivering a content to $R(0) = 30$ requesters (y-axis) under different number of initial holders H_0 (x-axis). Different initial placement policies ($H_{SC}(0), H_{MN}(0)$), levels of MNs participation (p_c), and expiry times TTL are considered. In the majority of scenarios our results accurately predict the offloading cost. Yet, even in the case where the predictions are less accurate (e.g. in Fig. 6(b) for $\mu_\lambda \cdot TTL = 0.05$), they can still capture the actual optimal initial allocation regimes.

5.2 Performance Evaluation

After validating our analysis, we now investigate the cost efficiency of the ”offloading on the edge” mechanism in a realistic traffic scenario. We present results



(a) TVCM ($H_{SC}(0) = H_{MN}(0)$)

(b) SLAW ($H_{SC}(0) = 0,$)

Figure 6: Offloading cost (y-axis) vs number of initial holders (H_0, x-axis). Dashed lines correspond to theoretical predictions and markers to simulation results. Transmission costs are: (a) $C_{BS}^{(TTL)} = 10 \cdot C_{BH} = 10 \cdot C_{BS} = 20 \cdot C_{SC} = 20 \cdot C_{D2D}$ (top plot) and $C_{BS}^{(TTL)} = C_{BH} = C_{BS} = 10 \cdot C_{SC}$ (bottom plot); (b) $C_{BS}^{(TTL)} = 2 \cdot C_{BS} = 10 \cdot C_{D2D}$.

that demonstrate the effect of different system factors, and provide useful conclusions for cellular network operators.

The parameters of the scenario we consider are the following:

– *Popularity:* Content popularity has been shown to follow a power-law distribution [7–9]. Thus, we draw the popularity of each content from a bounded-Pareto distribution ($R_0 \in [1, 100]$) with shape parameter $\alpha = 0.5$ [7].

– *Traffic Intensity:* Mobile operators do not release real mobile traffic data. To this end, and since traffic demand is directly related to the number of mobile users that reside in an area, we infer traffic patterns from an available dataset of the Gowalla *location-based social network*. The Gowalla dataset [30] contains information (logs of position and time) of user checkins (through their mobile devices) in different venues. In the scenarios we present, we create different number of contents during a $24h$ time interval. The number of contents M is proportional to the number of mobile users that checked-in a certain area (we selected the most popular venue) at the same time. The maximum number of concurrent contents is $M = 200$.

– *Delay Tolerance:* We set equal expiry times TTL for each content, and we consider different sets of scenarios with low ($TTL = 5min$), moderate ($TTL = 25min$), and high ($TTL = 25min$) delay tolerance.

– *Costs:* The relative costs are set $C_{BS} = C_{BS}^{(TTL)} = 10 \cdot C_{BH} = 20 \cdot C_{SC} = 20 \cdot C_{D2D}$, values selected based on some data presented in [16].

– *Node Mobility:* We use the TVCM mobility scenario presented in the previous section.

Offloading through SCs

We first consider the case of offloading through SCs. We simulate two sets of scenarios with small ($Q = 5$) or large ($Q = 200$) caches. We choose the optimal initial caching policy of Result 4.

In Fig. 7 we present the total offloading cost (marked lines) incurred for the cellular network operator over different times of the day. The gray area shows the intensity of mobile users that reside in the considered area. The dashed line denotes traffic demand over time, or equivalently, the cost when content delivery *without* offloading is considered.

Some interesting observations that follow from Fig. 7 are:

(i) Under the optimal caching policy, ”offloading on the edge” can significantly reduce the cost of content delivery, up to an order of magnitude, or even more in some cases.

(ii) The ”offloading on the edge” cost changes over time much smoother than traffic demand. In particular, for large caches (cross/red line), the offloading cost curve is almost flat, despite the large peaks in traffic demand. In cellular networks, such temporal variations of the traffic intensity is an important issue, since operators are required to over-provision the network capacity (high CAPEX costs) [15]. As we show, ”offloading on the edge” can amortize these costs. Even under higher transmission costs C_{BH}, C_{SC} than these we assumed, although the operating cost (OPEX) increases, the cost curve remains smooth, reducing thus a need for over-provisioning.

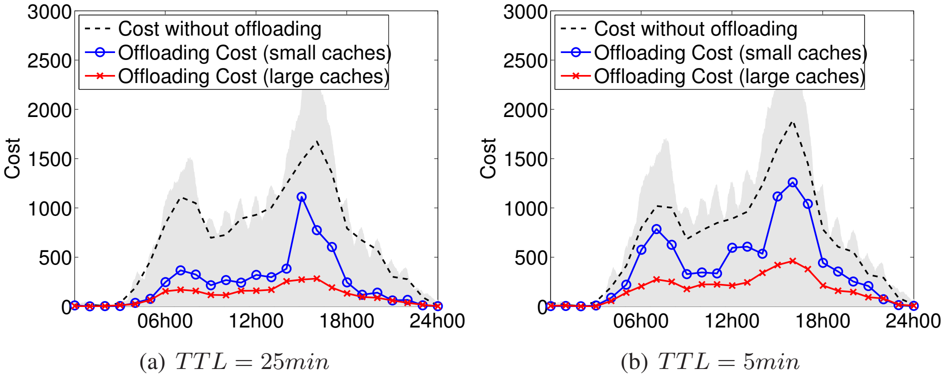
(iii) Large caching capacity has as a result a smoother cost curve (cross/red vs circle/blue curves). This is a positive message for operators, because to equip SCs with large enough caches is both feasible and inexpensive, as discussed in the example scenario of Section 4.1.

(iv) Comparing Fig. 7(a) and Fig. 7(b), we see that the tolerated delay has also a significant effect on the smoothness of the cost curve (higher TTL values lead to smaller variations). This implies that an alternative way of avoiding the over-provision cost (CAPEX), is to give incentives (OPEX) to users for accepting delayed content. Such solutions have been previously considered, e.g. [15], however, our framework allows an easy investigation of their effects (due to the closed-form results) and an analytic approach of pricing policies, etc.

Offloading through MNs

Now, we evaluate the performance of offloading through MNs. We simulate scenarios with different levels of node cooperation p_c. We choose the optimal initial content placement policy of Result 5.

In Fig. 8(a) we present the total offloading cost (marked lines) incurred for the cellular network operator over different times of the day. We simulate three scenarios with low, moderate and high delay tolerance ($TTL = 5, 25, 60min$), and 10% of user cooperation in offloading ($p_c = 0.1$). Similarly to the offloading



(a) $TTL = 25min$ (b) $TTL = 5min$

Figure 7: Traffic demand and offloading cost over a $24h$ period.

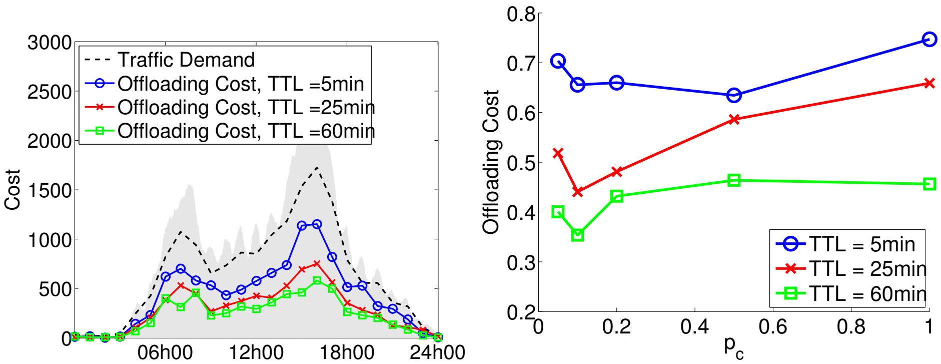
through SCs case (see e.g. Fig. 7), for higher TTL values, the cost decreases and its variations are smoother. However, it can be seen that improvement between the scenarios with $TTL = 25min$ and $TTL = 60min$ is not significant. This has an important implication for the system: Although increasing the delay tolerance is beneficial for the operator, after a point or gradually (depending on the scenario), the effects of this improvement become negligible. Bearing in mind that user satisfaction decreases with TTL indicates that there is a tradeoff, which should be carefully assessed by the system designer or considered for further optimization.

In Fig. 8(a) we show how the total offloading cost over a day period (normalized to the respective cost without offloading) changes with p_c. It is evident that varying user cooperation does not have the same effects for different scenarios, and that the minimum total cost is achieved at different values of p_c. This introduces one extra dimension, which can be used for system optimization as well. Such optimization options (with respect to TTL, p_c, etc.) could lead to interesting conclusions, we intend to consider them in future research.

6 Related Work

In this section we discuss works that are closer to ours, rather than studies which do not consider caching and/or delay tolerant delivery, and which are mainly based on pure infrastructure architectures, e.g. with WiFi access points [4] or small-cell base stations [2, 3], or on the D2D paradigm [10].

Mobile data offloading through opportunistic communications and epidemic content dissemination is studied in [13, 14, 19, 20]. In the setting of [19], copies of a content are distributed through the infrastructure to a subset of mobile nodes, which then start propagating them epidemically. The performance of different content "pushing" techniques (e.g. slow/fast start) is investigated through simulations on a real vehicular mobility trace. Analytical approaches for pushing techniques can be found in [13, 14], which study the optimal selection of the number of initial



(a) Offloading Cost vs. Time

(b) Offloading Cost vs. p_c

Figure 8: (a) Traffic demand and offloading cost over a $24h$ period. User cooperation is 10%. (b) Total offloading cost over a $24h$ period, normalized to the total cost without offloading.

and final content pushes. [14] models the content dissemination as a control system and proposes an adaptive algorithm, *HYPE*, which aims to minimize the load of the cellular network by using real time measurements. On the other hand, [13] uses a fluid limit approximation and focuses on the cost optimization problem. Finally, [20] takes into account fairness among different contents/nodes, and derives schedulers that maximize the throughput, under given mobility and wireless channel conditions. These studies, in contrast to our framework, assume that *every* user is willing to offload contents, even if they are *not of her interest*. Difficulties in devising incentive mechanisms or limitations of device capabilities, might render such settings unrealistic.

To this end, [12, 31] consider a limited number of (designated) holders. [12] proposes centralized algorithms for selecting the best set of available holders, in order to minimize the traffic load served by the infrastructure. In a different approach, [31] focuses on the effects of *popularity* (number of requesters) and *availability* (number of holders) on the performance of content delivery. Our paper extends these works, by introducing generic offloading costs and policies, and deriving insightful, closed-form results for the optimal caching.

Finally, [32] proposes caching in femto-cells and user devices, in a different setting than ours, where users communicate with several holders simultaneously. D2D communication is controlled by a macro-cell BS, which is aware of the status of caches, location of users, and channel state information between them. The objective of the paper is to decide which files should be stored and on which helper node, a problem that is shown to be *NP-hard*. This problem is formally presented, studied in more detail, and extended for coded contents in [5].

7 Conclusion

In this work we studied “offloading on the edge”, a mechanism that employs edge nodes (SCs and/or MNs) to opportunistically offload popular content. We built a model that can capture heterogeneous traffic demand, user cooperation and mobility characteristics, and describe generic caching and offloading policies. Based on our model, we derived closed-form expressions for predicting the offloading performance. These allowed us to analytically study the cost optimization problem, and provide results that shed light on how caching policies should be designed. Realistic simulations verified the insights that stem from our analysis, and led to useful conclusions.

Our closed-form expressions reveal how and to what extent each system parameter affects performance and cost. Thus, they could be easily applied to sensitivity analysis, network planning and dimensioning, or design of pricing strategies; issues that have recently attracted a lot of attention from network operators, who seek novel solutions to alleviate the effects of the rapidly growing traffic demand.

References

[1] CISCO, “Cisco visual networking index: Global mobile data traffic forecast update, 2013–2018,” Tech. Rep., 2014.

[2] V. Chandrasekhar, J. Andrews, and A. Gatherer, “Femtocell networks: a survey,” *Communications Magazine, IEEE*, vol. 46, no. 9, pp. 59–67, September 2008.

[3] J. Andrews, “Seven ways that hetnets are a cellular paradigm shift,” *Communications Magazine, IEEE*, vol. 51, no. 3, pp. 136–144, March 2013.

[4] K. Lee, J. Lee, Y. Yi, I. Rhee, and S. Chong, “Mobile data offloading: How much can wifi deliver?” in *Proc. ACM CoNEXT*, 2010.

[5] K. Shanmugam, N. Golrezaei, A. Dimakis, A. Molisch, and G. Caire, “Femtocaching: Wireless content delivery through distributed caching helpers,” *IEEE Transactions on Information Theory*, vol. 59, no. 12, Dec 2013.

[6] K. Poularakis, G. Iosifidis, A. Argyriou, and L. Tassiulas, “Video delivery over heterogeneous cellular networks: Optimizing cost and performance,” in *Proc. IEEE INFOCOM*, 2014.

[7] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, “Youtube traffic characterization: A view from the edge,” in *Proc. ACM IMC*, 2007.

[8] J. Erman, A. Gerber, K. K. Ramadrishnan, S. Sen, and O. Spatscheck, “Over the top video: The gorilla in cellular networks,” in *Proc. ACM IMC*, 2011.

[9] Z. Li, J. Lin, M.-I. Akodjenou, G. Xie, M. A. Kaafar, Y. Jin, and G. Peng, “Watching videos from everywhere: A study of the pptv mobile vod system,” in *Proc. ACM IMC*, 2012.

[10] A. Asadi, Q. Wang, and V. Mancuso, “A survey on device-to-device communication in cellular networks,” *Communications Surveys Tutorials, IEEE*, vol. PP, no. 99, pp. 1–1, 2014.

[11] M. Ji, G. Caire, and A. F. Molisch, “Fundamental limits of caching in wireless d2d networks,” arXiv preprint, http://arxiv.org/abs/1405.5336, 2014.

[12] Y. Li, M. Qian, D. Jin, P. Hui, Z. Wang, and S. Chen, “Multiple mobile data offloading through disruption tolerant networks,” *IEEE Transactions on Mobile Computing*, vol. 13, no. 7, pp. 1579–1596, 2014.

[13] X. Wang, M. Chen, Z. Han, T. Kwon, and Y. Choi, “Content dissemination by pushing and sharing in mobile cellular networks: An analytical study,” in *Proc. IEEE MASS*, 2012.

[14] V. Sciancalepore, D. Giustiniano, A. Banchs, and A. Picu, “Offloading cellular traffic through opportunistic communications: Analysis and optimization,” *arXiv*, vol. 1405.3548, 2014.

[15] S. Ha, S. Sen, C. Joe-Wong, Y. Im, and M. Chiang, “Tube: Time-dependent pricing for mobile data,” *ACM SIGCOMM Comput. Commun. Rev.*, vol. 42, no. 4, pp. 247–258, 2012.

[16] K. Johansson, “Cost effective deployment strategies for heterogenous wireless networks,” *Doctoral Thesis*, 2007.

[17] Akamai, “Swisscom and akamai enter into a strategic partnership,” Press Release, March 2013.

[18] K. Suh, C. Diot, J. Kurose, L. Massoulie, C. Neumann, D. Towsley, and M. Varvello, “Push-to-peer video-on-demand system: Design and evaluation,” *Selected Areas in Communications, IEEE Journal on*, vol. 25, no. 9, pp. 1706–1716, December 2007.

[19] J. Whitbeck, M. Amorim, Y. Lopez, J. Leguay, and V. Conan, “Relieving the wireless infrastructure: When opportunistic networks meet guaranteed delays,” in *Proc. IEEE WoWMoM*, 2011.

[20] H. Cai, I. Koprulu, and N. Shroff, “Exploiting double opportunities for deadline based content propagation in wireless networks,” in *Proc. IEEE INFOCOM*, 2013.

[21] W. Gao, Q. Li, B. Zhao, and G. Cao, “Multicasting in delay tolerant networks: a social network perspective,” in *Proc. ACM MobiHoc*, 2009.

[22] V. Conan, J. Leguay, and T. Friedman, “Characterizing pairwise inter-contact patterns in delay tolerant networks,” in *Proc. ACM Autonomics*, 2007.

[23] W. Gao and G. Cao, “User-centric data dissemination in disruption tolerant networks,” in *Proc. IEEE INFOCOM*, 2011.

[24] J.-Y. Le Boudec, “Modelling the Immune System Toolbox: Stochastic Reaction Models,” http://infoscience.epfl.ch/record/98734/files/toolbox.pdf, 2006.

[25] A. Antoniou and W.-S. Lu, *Practical Optimization: Algorithms and Engineering Applications*. Springer, 2007.

[26] P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft, and C. Diot, “Pocket switched networks and human mobility in conference environments,” in *Proc. ACM WDTN*, 2005.

[27] A. Passarella, R. I. Dunbar, M. Conti, and F. Pezzoni, “Ego network models for future internet social networking environments,” *Computer Communications*, vol. 35, no. 18, pp. 2201–2217, 2012.

[28] W.-J. Hsu, T. Spyropoulos, K. Psounis, and A. Helmy, “Modeling spatial and temporal dependencies of user mobility in wireless mobile networks,” *IEEE/ACM Trans. on Networking*, vol. 17, no. 5, 2009.

[29] K. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong, “Slaw: A new mobility model for human walks,” in *IEEE INFOCOM*, 2009.

[30] T. Hossmann, G. Nomikos, T. Spyropoulos, and F. Legendre, “Collection and analysis of multi-dimensional network data for opportunistic networking research,” *Comput. Communications*, vol. 35, no. 13, 2012.

[31] P. Sermpezis and T. Spyropoulos, “Not all content is created equal: Effect of popularity and availability for content-centric opportunistic networking,” in *Proc. ACM MOBIHOC*, 2014.

[32] N. Golrezaei, A. Molisch, A. Dimakis, and G. Caire, “Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution,” *IEEE Comm. Magazine*, vol. 51, no. 4, April 2013.

8 Appendix

8.1 Proof of Result 2

Proof. The probability a content to be delivered in the time interval $[t, t+dt)$ is given by

$$P\{T_d \in [t, t+dt)\} = \frac{dP\{T_d \leq t\}}{dt} \cdot dt \tag{24}$$

Since a requester gets the content at time $t = TTL$ from a BS, if it has not received it earlier, we can write for the expected delay

$$\begin{aligned} E[T_i|TTL] &= TTL \cdot (1 - P\{T_d \leq TTL\}) \\ &\qquad + \int_0^{TTL} t \cdot P\{T_d \in [t, t+dt)\} \\ &= TTL \cdot (1 - P\{T_d \leq TTL\}) + \int_0^{TTL} t \cdot \frac{dP\{T_d \leq t\}}{dt} \cdot dt \end{aligned} \tag{25}$$

where the last equality follows from Eq. (24).

Using the expression of Result 1, we first compute the derivative $\frac{dP\{T_d \leq t\}}{dt}$, and, then, the integral in Eq. (25), and we get

$$\begin{aligned} E[T_i|TTL] &= TTL \cdot (1 - P\{T_d \leq TTL\}) \\ &+ \frac{1}{p_c \cdot R_0} \cdot \left(\frac{TTL \cdot H_0 \cdot (p_c \cdot R_0 + H_0) \cdot e^{\mu_\lambda \cdot (p_c \cdot R_0 + H_0) \cdot TTL}}{p_c \cdot R_0 + H_0 \cdot e^{\mu_\lambda \cdot (p_c \cdot R_0 + H_0) \cdot TTL}} \right. \\ &\qquad \left. + \frac{1}{\mu_\lambda \cdot p_c \cdot R_0} \cdot \ln \left(\frac{p_c \cdot R_0 + H_0}{p_c \cdot R_0 + H_0 \cdot e^{\mu_\lambda \cdot (p_c \cdot R_0 + H_0) \cdot TTL}} \right) \right) \end{aligned}$$

Substituting the value of $P\{T_d \leq TTL\}$ from Result 1 in the above equation, after some algebraic manipulations, we can successively get

$$\begin{aligned} E[T_i|TTL] &= \frac{TTL \cdot (p_c \cdot R_0 + H_0)}{p_c \cdot R_0} \\ &\qquad + \frac{1}{\mu_\lambda \cdot p_c \cdot R_0} \cdot \ln \left(\frac{p_c \cdot R_0 + H_0}{p_c \cdot R_0 + H_0 \cdot e^{\mu_\lambda \cdot (p_c \cdot R_0 + H_0) \cdot TTL}} \right) \\ &= \frac{1}{\mu_\lambda \cdot p_c \cdot R_0} \cdot \ln \left(\frac{(p_c \cdot R_0 + H_0) \cdot e^{\mu_\lambda \cdot (p_c \cdot R_0 + H_0) \cdot TTL}}{p_c \cdot R_0 + H_0 \cdot e^{\mu_\lambda \cdot (p_c \cdot R_0 + H_0) \cdot TTL}} \right) \\ &= \frac{1}{\mu_\lambda \cdot p_c \cdot R_0} \cdot \ln \left(1 + \frac{p_c \cdot R_0 - e^{-\mu_\lambda \cdot (p_c \cdot R_0 + H_0) \cdot TTL}}{H_0 + p_c \cdot R_0 \cdot e^{-\mu_\lambda \cdot (p_c \cdot R_0 + H_0) \cdot TTL}} \right) \end{aligned}$$

which is the expression of Result 2 for $p_c > 0$. The expression for $p_c = 0$ follows after taking the limit ($p_c \rightarrow 0$) of the above expression. □

8.2 Lemma 2: Cost Monotonicity with λ_0

Lemma 2. *Under a content placement policy given by Eq.* (19)*, the derivative of the total cost, $\sum_{\theta \in \mathcal{M}} C^\theta$, with respect to λ_0 is*

$$\frac{d}{d\lambda_0} \left[\sum_{\theta \in \mathcal{M}} C^\theta \right] = \frac{1}{\gamma} \cdot \left(1 - \frac{1}{1 + \frac{\lambda_0}{\Phi_1}} \right) \cdot |A| \geq 0$$

where $\mathcal{A} = \{\theta \in \mathcal{M} : L \leq R_0^\theta \leq U\}$.

Proof. From the conditions (b) and (c) (see, proof of Result 4), and similarly to Eqs. (18), we can express the multipliers λ_θ and μ_θ as a function of λ_0, as

$$\lambda_\theta = \begin{cases} \lambda_0 + C_{BH}\left(1 - \gamma \cdot \Phi \cdot R_0^\theta\right) & , R_0^\theta < L \\ 0 & , R_0^\theta \geq L \end{cases} \tag{26a}$$

$$\mu_\theta = \begin{cases} -\lambda_0 - C_{BH}\left(1 - \gamma \cdot \Phi \cdot e^{-\gamma \cdot N_{SC}} R_0^\theta\right) & , R_0^\theta > U \\ 0 & , R_0^\theta \leq U \end{cases} \tag{26b}$$

The cost of a single content dissemination, Eq. (16), under the content placement policy of Eq. (19), can be written as

$$\begin{aligned} C^\theta &= \frac{\Phi_1}{\gamma} \cdot \left[\ln\left(\gamma \cdot \Phi \cdot R_0^\theta\right) - \ln\left(1 + \frac{\lambda_0 - \lambda_\theta + \mu_\theta}{\Phi_1}\right)\right] \\ &+ \Phi_2 \cdot R_0^\theta \\ &+ (\Phi_3 - \Phi_2) \cdot R_0^\theta \cdot \frac{1}{\gamma \cdot \Phi \cdot R_0^\theta} \cdot \left(1 + \frac{\lambda_0 - \lambda_\theta + \mu_\theta}{\Phi_1}\right) \\ &= \frac{\Phi_1}{\gamma} \cdot \left[\ln\left(\gamma \cdot \Phi \cdot R_0^\theta\right) - \ln\left(1 + \frac{\lambda_0 - \lambda_\theta + \mu_\theta}{\Phi_1}\right)\right] \\ &+ \Phi_2 \cdot R_0^\theta + \frac{\Phi_1}{\gamma} \cdot \left(1 + \frac{\lambda_0 - \lambda_\theta + \mu_\theta}{\Phi_1}\right) \end{aligned} \tag{27}$$

Taking its derivative, with respect to λ_0, gives

$$\begin{aligned} \frac{d}{d\lambda_0}\left[\sum_{\theta \in \mathcal{M}} C^\theta\right] &= -\frac{\Phi_1}{\gamma} \cdot \frac{d}{d\lambda_0}\left[\sum_{\theta \in \mathcal{M}} \ln\left(1 + \frac{\lambda_0 - \lambda_\theta + \mu_\theta}{\Phi_1}\right)\right] \\ &+ \frac{1}{\gamma} \cdot \frac{d}{d\lambda_0}\left[\sum_{\theta \in \mathcal{M}} (\lambda_0 - \lambda_\theta + \mu_\theta)\right] \end{aligned} \tag{28}$$

because the terms including only the scenario parameters (R_0^θ, γ, and costs) do not depend on the selected resource allocation and, thus, on the parameter λ_0.

To calculate the derivatives appearing in the right side of Eq. (28), we use the definition of a derivative, i.e.

$$\frac{df(\lambda_0)}{d\lambda_0} = \lim_{d\lambda_0 \to 0} \frac{f(\lambda_0 + d\lambda_0) - f(\lambda_0)}{d\lambda_0} \tag{29}$$

and proceed as following:

We first define the sets

$$\mathcal{A} = \{\theta \in \mathcal{M} : L \leq R_0^\theta \leq U\} \tag{30a}$$

$$\mathcal{B} = \{\theta \in \mathcal{M} : R_0^\theta > U\} \tag{30b}$$

$$\mathcal{C} = \{\theta \in \mathcal{M} : R_0^\theta < L\} \tag{30c}$$

and, respectively, for $\lambda_0 \to \lambda_0 + d\lambda_0$, the sets

$$\mathcal{A}^{'} = \{\theta \in \mathcal{M} : L + \Delta L \leq R_0^\theta \leq +\Delta U\} \tag{31a}$$

$$\mathcal{B}^{'} = \{\theta \in \mathcal{M} : R_0^\theta > U + \Delta U\} \tag{31b}$$

$$\mathcal{C}^{'} = \{\theta \in \mathcal{M} : R_0^\theta < L + \Delta L\} \tag{31c}$$

where we denoted

$$L + \Delta L = \frac{1}{\gamma \cdot \Phi} \cdot \left(1 + \frac{\lambda_0 + d\lambda_0}{C_{BH}}\right) = L + \frac{d\lambda_0}{\gamma \cdot C_{BH} \cdot \Phi} \tag{32a}$$

$$\begin{aligned} U + \Delta U &= \frac{1}{\gamma \cdot \Phi} \cdot e^{\gamma \cdot N_{SC}} \cdot \left(1 + \frac{\lambda_0 + d\lambda_0}{C_{BH}}\right) \\ &= U + \frac{d\lambda_0}{\gamma \cdot C_{BH} \cdot \Phi} \cdot e^{\gamma \cdot N_{SC}} = (L + \Delta L) \cdot e^{\gamma \cdot N_{SC}} \end{aligned} \tag{32b}$$

Regarding the first derivative term in Eq. (28), we proceed as following

$$\begin{aligned} &\frac{d}{d\lambda_0}\left[\sum_{\theta \in \mathcal{M}} \ln\left(1 + \frac{\lambda_0 - \lambda_\theta + \mu_\theta}{C_{BH}}\right)\right] \\ &\overset{\text{Eqs. (26)}}{=} \frac{d}{d\lambda_0}\left[\sum_{\theta \in \mathcal{A}} \ln\left(1 + \frac{\lambda_0}{C_{BH}}\right)\right] \\ &\qquad + \frac{d}{d\lambda_0}\left[\sum_{\theta \in \mathcal{B}} \left(\ln\left(\gamma \cdot \Phi \cdot R_0^\theta\right) - \gamma \cdot N_{SC}\right)\right] \\ &\qquad + \frac{d}{d\lambda_0}\left[\sum_{\theta \in \mathcal{C}} \ln\left(\gamma \cdot \Phi \cdot R_0^\theta\right)\right] \\ &= \frac{d}{d\lambda_0}\left[|\mathcal{A}| \ln\left(1 + \frac{\lambda_0}{C_{BH}}\right)\right] \\ &\qquad + \frac{d}{d\lambda_0}\left[\sum_{\theta \in \mathcal{B}} \ln\left(\gamma \cdot \Phi \cdot R_0^\theta\right)\right] - \gamma \cdot N_{SC} \cdot \frac{d|\mathcal{B}|}{d\lambda_0} \\ &\qquad + \frac{d}{d\lambda_0}\left[\sum_{\theta \in \mathcal{C}} \ln\left(\gamma \cdot \Phi \cdot R_0^\theta\right)\right] \\ &= |\mathcal{A}| \cdot \frac{1}{C_{BH}} \cdot \frac{1}{1 + \frac{\lambda_0}{C_{BH}}} + \ln\left(1 + \frac{\lambda_0}{C_{BH}}\right) \cdot \frac{d|\mathcal{A}|}{d\lambda_0} \\ &\qquad + \frac{d}{d\lambda_0}\left[\sum_{\theta \in \mathcal{B}} \ln\left(\gamma \cdot \Phi \cdot R_0^\theta\right)\right] - \gamma \cdot N_{SC} \cdot \frac{d|\mathcal{B}|}{d\lambda_0} \\ &\qquad + \frac{d}{d\lambda_0}\left[\sum_{\theta \in \mathcal{C}} \ln\left(\gamma \cdot \Phi \cdot R_0^\theta\right)\right] \end{aligned} \tag{33}$$

The derivatives in the above sum are calculated as following

$$\begin{aligned}
\frac{d|\mathcal{A}|}{d\lambda_0} &= \frac{|\mathcal{A}'| - |\mathcal{A}|}{d\lambda_0} \\
&= \frac{\int_{L+\Delta L}^{U+\Delta U} M \cdot \rho(x)dx - \int_{L}^{U} M \cdot \rho(x)dx}{d\lambda_0} \\
&= M \cdot \frac{\int_{U}^{U+\Delta U} \rho(x)dx - \int_{L}^{L+\Delta L} \rho(x)dx}{d\lambda_0} \\
&\approx M \cdot \frac{p(U) \cdot \Delta U - p(L) \cdot \Delta L}{d\lambda_0} \\
&\overset{\text{Eqs. (32)}}{=} M \cdot \frac{p(U) \cdot \Delta L \cdot e^{\gamma \cdot N_{SC}} - p(L) \cdot \Delta L}{d\lambda_0} \\
&= M \cdot \frac{\Delta L}{d\lambda_0} \cdot \big(p(U) \cdot e^{\gamma \cdot N_{SC}} - p(L)\big) \\
&\overset{\text{Eqs. (32)}}{=} \frac{M}{\gamma \cdot C_{BH} \cdot \Phi} \cdot \big(p(U) \cdot e^{\gamma \cdot N_{SC}} - p(L)\big) \qquad \text{(34a)}
\end{aligned}$$

and, similarly,

$$\frac{d|\mathcal{B}|}{d\lambda_0} \approx -M \cdot \frac{e^{\gamma \cdot N_{SC}}}{\gamma \cdot C_{BH} \cdot \Phi} \cdot p(U) \qquad \text{(34b)}$$

and

$$\begin{aligned}
&\frac{d}{d\lambda_0}\left[\sum_{\theta \in \mathcal{B}} \ln\left(\gamma \cdot \Phi \cdot R_0^{\theta}\right)\right] \\
&= \frac{\sum_{\theta \in \mathcal{B}'} \ln\left(\gamma \cdot \Phi \cdot R_0^{\theta}\right) - \sum_{\theta \in \mathcal{B}} \ln\left(\gamma \cdot \Phi \cdot R_0^{\theta}\right)}{d\lambda_0} \\
&= \frac{-\int_{U}^{U+\Delta U} \ln(\gamma \cdot \Phi \cdot x) \cdot M \cdot \rho(x)dx}{d\lambda_0} \\
&\approx -M \cdot \frac{\ln(\gamma \cdot \Phi \cdot U) \cdot p(U) \cdot \Delta U}{d\lambda_0} \\
&\overset{\text{Eqs. (32)}}{=} -M \cdot \frac{e^{\gamma \cdot N_{SC}}}{\gamma \cdot C_{BH} \cdot \Phi} \cdot \ln(\gamma \cdot \Phi \cdot U) \cdot p(U) \\
&\overset{\text{Eqs. (18)}}{=} -M \cdot \frac{e^{\gamma \cdot N_{SC}}}{\gamma \cdot C_{BH} \cdot \Phi} \cdot p(U) \cdot \left(\gamma \cdot N_{SC} + \left(1 + \frac{\lambda_0}{C_{BH}}\right)\right) \qquad \text{(34c)}
\end{aligned}$$

and, similarly,

$$\begin{aligned}
&\frac{d}{d\lambda_0}\left[\sum_{\theta \in \mathcal{C}} \ln\left(\gamma \cdot \Phi \cdot R_0^{\theta}\right)\right] \\
&\qquad \approx M \cdot \frac{1}{\gamma \cdot C_{BH} \cdot \Phi} \cdot p(L) \cdot \ln\left(1 + \frac{\lambda_0}{C_{BH}}\right) \qquad \text{(34d)}
\end{aligned}$$

Substituting Eqs. (34) in Eq. (33), gives

$$\frac{d}{d\lambda_0}\left[\sum_{\theta\in\mathcal{M}}\ln\left(1+\frac{\lambda_0-\lambda_\theta+\mu_\theta}{\Phi_1}\right)\right]=|\mathcal{A}|\cdot\frac{1}{\Phi_1}\cdot\frac{1}{1+\frac{\lambda_0}{\Phi_1}} \tag{35}$$

Regarding the second derivative term in Eq. (28), we proceed as following

$$\begin{aligned}
&\frac{d}{d\lambda_0}\left[\sum_{\theta\in\mathcal{M}}(\lambda_0-\lambda_\theta+\mu_\theta)\right]\\
\overset{\text{Eqs. (26)}}{=}&\frac{d}{d\lambda_0}\left[\sum_{\theta\in\mathcal{A}}\lambda_0\right]+\frac{d}{d\lambda_0}\left[\sum_{\theta\in\mathcal{B}}(\lambda_0+\mu_\theta)\right]+\frac{d}{d\lambda_0}\left[\sum_{\theta\in\mathcal{C}}(\lambda_0-\lambda_\theta)\right]\\
=&\frac{d}{d\lambda_0}\left[\lambda_0\cdot|\mathcal{A}|\right]\\
&+\frac{d}{d\lambda_0}\left[\sum_{\theta\in\mathcal{B}}\left(-C_{BH}+\gamma\cdot C_{BH}\cdot\Phi\cdot e^{-\gamma\cdot N_{SC}}\cdot R_0^\theta\right)\right]\\
&+\frac{d}{d\lambda_0}\left[\sum_{\theta\in\mathcal{C}}\left(-C_{BH}+\gamma\cdot C_{BH}\cdot\Phi\cdot R_0^\theta\right)\right]\\
=&|\mathcal{A}|+\lambda_0\cdot\frac{d|\mathcal{A}|}{d\lambda_0}\\
&-C_{BH}\cdot\frac{d|\mathcal{B}|}{d\lambda_0}+\gamma\cdot C_{BH}\cdot\Phi\cdot e^{-\gamma\cdot N_{SC}}\cdot\frac{d}{d\lambda_0}\left[\sum_{\theta\in\mathcal{B}}R_0^\theta\right]\\
&-C_{BH}\cdot\frac{d|\mathcal{C}|}{d\lambda_0}+\gamma\cdot C_{BH}\cdot\Phi\cdot\frac{d}{d\lambda_0}\left[\sum_{\theta\in\mathcal{C}}R_0^\theta\right]
\end{aligned} \tag{36}$$

Similarly as before, we get

$$\frac{d|\mathcal{C}|}{d\lambda_0}\approx M\cdot\frac{1}{\gamma\cdot C_{BH}\cdot\Phi}\cdot p(L) \tag{37a}$$

and

$$\begin{aligned}
\frac{d}{d\lambda_0}\left[\sum_{\theta\in\mathcal{B}}R_0^\theta\right]&=\frac{-\int_U^{U+\Delta U}x\cdot M\cdot\rho(x)dx}{d\lambda_0}\\
&\approx -M\cdot\frac{U\cdot p(U)\cdot\Delta U}{d\lambda_0}\\
&\overset{\text{Eqs. (32)}}{=}-M\cdot\frac{\Delta L}{d\lambda_0}\cdot L\cdot p(U)\cdot e^{2\cdot\gamma\cdot N_{SC}}\\
&\overset{\text{Eqs. (18)}}{=}-M\cdot\frac{e^{\gamma\cdot N_{SC}}}{\gamma\cdot C_{BH}\cdot\Phi}\cdot\frac{1}{\gamma\cdot\Phi}\cdot\left(1+\frac{\lambda_0}{C_{BH}}\right)p(U)
\end{aligned} \tag{37b}$$

and, similarly,

$$\frac{d}{d\lambda_0}\left[\sum_{\theta\in\mathcal{C}} R_0^\theta\right] \approx -M \cdot \frac{1}{\gamma \cdot C_{BH} \cdot \Phi} \cdot \frac{1}{\gamma \cdot \Phi} \cdot \left(1 + \frac{\lambda_0}{C_{BH}}\right) p(L) \tag{37c}$$

Substituting Eqs. (37) in Eq. (36), gives

$$\frac{d}{d\lambda_0}\left[\sum_{\theta\in\mathcal{M}} (\lambda_0 - \lambda_\theta + \mu_\theta)\right] = |\mathcal{A}| \tag{38}$$

Finally, substituting the expressions of Eq. (35) and Eq. (38) in Eq. (28), proves the Lemma. □