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

Pavlos Sermpezis
Institute of Computer Science
FORTH, Greece
sermpezis@ics.forth.gr

Thrasyvoulos Spyropoulos
Dept. Mobile Communications
EURECOM, France
spyropou@eurecom.fr

***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 keep up with demand. As a result, extensive densification through small cells, caching content closer to or even at the device, device-to-device (D2D) communications, and delayed content delivery 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, and contract plan incentives for D2D. A number of interesting tradeoffs, relating to performance and costs, arise thus for the operator. In this paper, we analytically investigate the extent to which local storage and communication through "edge" nodes could help offload traffic in a heterogeneous network (HetNet). We propose a model that can capture generic HetNet setups (comprising small cells, D2D communication, delayed delivery schemes, transmission costs, etc.). We analyse (a) the offloading performance and (b) the costs involved for the operator, and derive simple closed-form expressions as a function of the network parameters. Our results can be useful in performance evaluation and optimization of offloading and caching strategies, network dimensioning, pricing policies, etc.**

I. 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 data demand [1]. Radically reducing the communication distance by deploying many "small cells" (e.g. femto, pico, WiFi), and offloading traffic to them, 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 are predicted to become 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]. This 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]. A device can store content after consuming it, and give it directly to other neighboring devices also interested in it, offloading these requests from the main network. While D2D-based offloading normally assumes a content request will be served *immediately* either 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 [11]–[13], and go back to the main network if the requested content is 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 [14] 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, 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 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 densification, or would it be more cost-efficient to provide incentives for D2D opportunistic offloading?"*.

To this end, in this paper we analytically 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: (i) caching on small cells and mobile devices, collectively referred to as "edge nodes", and (ii) offloading requests through local, short range communications. After describing the main characteristics of

ISBN 978-3-901882-88-3

such an "offloading on the edge" mechanism (Section II), we propose a generic model that allows us to analytically study it (Section III). We proceed by deriving closed-form results for the performance of content delivery through this mechanism (Section IV) and the incurred costs (Section V), as a function of key system parameters. Our results can be used to study the performance gains and costs of offloading, and optimize content placement and dissemination strategies. Finally, we validate our results through realistic simulations (Section VI), propose a number of possible extensions for our framework (Section VII), and discuss related work (Section VIII).

Summarizing, the main contributions of our work are:

- We model offloading through small cells, opportunistic D2D, and caching at both. This unified approach allows us to study the joint effects of the many different parts that compose a HetNet.
- We provide closed-form expressions applicable to a number of performance metrics and network setups.
- We propose a generic model for the (direct and indirect) costs involved for the operator, and calculate the total offloading cost; our results allow to design strategies that minimize the cost, under performance guarantees.

II. Offloading on the Edge

A. 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 [15]. Hence, they can be used to offload some traffic from BSs. However, the backhaul connection for these cells is often 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

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

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

B. Offloading Mechanism

Content Requests. We assume that each MN is interested in different contents over time (e.g. videos, web pages, software updates), 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, which we will refer to as *content popularity*, is known. For a number of applications, the cellular networks can know this information in advance (e.g., for *push services* [12], [16]), or predict it (by applying various methods based, e.g., on past statistics, early demand of a content, social dynamics, etc.) [17].

Content Delivery. An operator can deliver a content to an interested MN 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 a MN requesting the content to wait and retrieve it when it *moves* within range of an SC or MN with the content in its cache. If a MN has been waiting for an amount of time, let TTL, without moving within range of a SC or MN with the content, then the operator is obliged to deliver the content directly through the closest macro BS.

While usual approaches of small-cell or D2D based offloading, e.g. [5], [6], do not allow a delayed delivery (i.e. $TTL \to 0$), it is likely that the small cell and (D2D enabled) mobile node density will not always be enough to offload enough traffic in these cases. Therefore, this *delay-tolerant* approach is a valuable (and complementary) alternative, with potential benefits (increased offloading) and costs (reduced QoE and potential monetary incentives).

III. Model

The generic "offloading on the edge" setup of Section II is characterised by a number of parameters, like, number of nodes, location of SCs and mobility of MNs, willingness of users to act as relays, etc. In this section, we propose an analytic model to describe the main system parameters and their interplay. Since capturing all these aspects in detail would not be analytically tractable, we model here the main characteristics of them; then, in Section VII we show how this model can be refined and extended towards various dimensions.

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 to some MNs, which are interested in it[1]. If by an expiry time

[1]Here, we need to stress that, under the "offloading on the edge" mechanism, MNs will never become holders of a content they are not interested in. Although some previous studies assume that MNs might act as holders for every content [12], [13], [18], [19], we believe that incentive mechanisms for these cases are difficult to implement (e.g. a user easier accepts to forward a content it has already consumed/stored). Nevertheless, our framework could be easily extended also for such cases.

TTL (if any), some of the interested MNs have not met and received the content by any edge node (SC or MN), they are served by a macro-cell BS.

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:

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

with $\mathcal{H}(t) = \mathcal{H}_{SC}(t) \cup \mathcal{H}_{MN}(t)$ and $H(t) = H_{SC}(t) + H_{MN}(t)$, where the subscripts denote the subset of holders belonging to each type, SC or MN.

The number of requesters and holders are important quantities for the content dissemination/delivery (and necessary for the performance analysis in the following sections). The number of requesters, $R(t)$, indicates how many users still need to be served at a given time (or, equivalently how many have been served), and thus it shows to what extent the offloading process has been completed. The number of holders, $H(t)$, denotes the amount of resources used for serving user requests, and thus it relates to how fast a content can be delivered.

With respect to the holders of a content, we assume that SCs store their cached contents until they expire, and during this time interval SCs always deliver them to encountered requesters (i.e. $\mathcal{H}_{SC}(t) = \mathcal{H}_{SC}(0), \forall t \in (0, TTL]$). This is a reasonable assumption, since SCs are under the control of the mobile operator, which knows the operating state of each SC, and thus content discards (e.g. due to cache overloads) can be avoided. On the other hand, MNs cannot be entirely controlled by the mobile operator. Despite the possible incentives given to MNs for forwarding cached contents to their requesters, some users might still not be willing to contribute to the offloading mechanism (temporarily or always). Some examples could be the cases where (a) some users might not accept becoming holders (ever) for privacy or power consumption reasons, or (b) some users that have accepted to become holders (e.g. by signing a contract with the operator that compensates them for each content they offload) might be temporarily reluctant to do so due to low battery levels or memory space, e.g. by turning off a "content forwarding option" in their device.

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

Assumption 1 (Cooperation)**.** *The probability a requester of a content to act as a holder for it, after its reception, is $p_c \in [0, 1]$. The probability p_c is equal among all nodes and contents.*

In the above assumption, we consider, for simplicity, a uniform node behavior captured by a single parameter p_c. However, our model and analysis can be extended and capture also heterogeneous patterns of node cooperation, e.g. based on social characteristics of users [20]. We provide a discussion of such issues in Section VII-A.

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.

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 [11]–[13], [19] and have been shown not to be far from real mobility [21], [22]. Yet, in Section VI, we test our results against realistic scenarios where node mobility departs from our assumptions and involves much more complexity.

In Table I we summarize the important notation used in this and the following sections.

TABLE I: Important notation

N_{BS}, N_{SC}, N_{MN}	number of BSs, SCs, and MNs
$\mathcal{R}(t), \mathcal{H}(t)$	set of *requesters* and *holders* at time t
$R(t) = \lvert\mathcal{R}(t)\rvert$ $H(t) = \lvert\mathcal{H}(t)\rvert$	number of *requesters* and *holders* at time t
$H_{SC}(t), H_{MN}(t)$	number of *holders* of type SC and MN at time t
p_c	probability for node participation to offloading
λ_{ij}	meeting rate between nodes i and j
μ_λ	mean value of meeting rates
$R_0 = R(0^+)$, $H_0 = H(0^+)$	number of *requesters* and *holders* at time $t = 0^+$ (just after the initial placement of the content)

IV. Performance 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 derive expressions for the performance of the "offloading on the edge" mechanism, which are useful when trying to answer the above questions.

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

As discussed earlier, the number of requesters and holders define and determine, respectively, the offloading performance. Hence, we first state the following Lemma (proved in [23]), in which we use a mean field approximation and a resulting system of ordinary differential equation (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 $R_0 = R(0^+)$.

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 that this (random) requester receives 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

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

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

Result 1 (Delivery Probability). *The probability that a content is delivered by an edge node 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 $R_0 = R(0^+)$.

With respect to the average delay a requester $i \in \mathcal{R}(0^+)$ experiences until it receives the content, we state the following Result (the proof can be found in [23]).

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

$$E[T_d|TTL] = \begin{cases} \dfrac{1}{\mu_\lambda \cdot p_c \cdot R_0} \cdot \ln\left(\dfrac{H(TTL)}{H_0}\right) & , p_c > 0 \\ \dfrac{1}{\mu_\lambda \cdot H_0} \cdot \left(1 - e^{-\mu_\lambda \cdot H_0 \cdot TTL}\right) & , p_c = 0 \end{cases}$$

where $H_0 = H(0^+)$ and $R_0 = R(0^+)$, and $H(TTL)$ is given by the expression of Lemma 1 for $H(t)$ with $t = TTL$, i.e.

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

V. Cost Analysis

In many cases, the cost of content delivery might be of higher interest than the offloading performance for the network operator . The "cost" of offloading can correspond to monetary expenses, resources consumption (e.g., bandwidth, energy), etc. It is useful for an operator to know the cost incurred by different offloading options, in order to select how to deliver each content, how to set the TTLs, or how many copies to place in MNs or SCs.

In this section, we use the performance predictions of Section IV to derive closed-form expressions for the total cost of "offloading on the edge" as a function of network parameters and operator costs.

A. Cost Model

Previous works usually define the offloading cost as the number of transmissions from BSs. However, content delivery through SCs, participation of MNs in the dissemination process, or tolerance of the users to delayed content, might also involve extra costs for the operator. To this end, we propose a generic model that captures the costs involved in each phase of the "offloading on the edge" mechanism.

– ***Initial Placement Costs:*** C_{BH}, C_{BS}.
Initially the content provider 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 as C_{BS}. This BS to MN transmission cost might depend on the operating cost of a BS, as well as on a number of other parameters, like the employed transmission technology, the load of the BS, the area, etc. Also, possible congestions of the backhaul or the cellular wireless interfaces, might affect the relative difference between the costs C_{BS} and C_{BH} as well.

– ***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 mainly 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 $C_{BS}^{(TTL)}$.

Remark: In the remainder we assume these costs as constant and independent of other network parameters. However, in general, the costs may change, e.g., at different times of day. In this case, our results can be applied on a per time window basis. Moreover, in Section VII-B, we provide a relevant discussion on how the model can be modified to capture dependencies between costs and parameters like TTL, p_c, etc.

B. Content Delivery Cost

We, now, calculate the cost incurred for the mobile network operator, when "offloading on the edge" is used. The knowl-

edge of this cost can play a crucial role towards designing or optimizing the offloading/caching policy.

Incorporating the offloading costs (Section V-A) in our content dissemination model, and using the analytical results of Section IV, we calculate the cost of a single content delivery in Result 3 (we provide the proof in Appendix A). 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. the number of initial content placements in SCs and MNs, $H_{SC}(0)$ and $H_{MN}(0)$, respectively), providing, thus, the necessary information for the evaluation and tuning of the "offloading on the edge" mechanism.

Result 3. *The cost of offloading a content through the "offloading on the edge" mechanism, 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

$$q = \frac{H_{SC}(0) \cdot \ln\left(\frac{H(TTL)}{H_0}\right)}{p_c \cdot R_0 \cdot P\{T_d \leq TTL\}}$$

and $P\{T_d \leq TTL\}$ and $H(TTL)$ are given in Result 1 and Lemma 1, i.e.

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

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

C. Application: Cost Optimization

In this section, we briefly discuss how Result 3 can be used by the operator to design its content placement strategy and minimize the offloading cost. Since a detailed optimization study is out of the scope of this paper, we provide only some key points in the following discussion.

In a real scenario, the network operator would have to offload simultaneously $M \geq 1$ contents to their requesters. Let us denote the set of the contents as $\mathcal{M}$ ($M = |\mathcal{M}|$). Since not all contents are expected to be equally popular [7]–[9] or 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.), 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[2]. Therefore, if we denote as C^θ the delivery cost of a content $\theta \in \mathcal{M}$ (given by Result 3), we can define the following optimization problem, whose objective is to minimize the *total* cost among all offloaded contents.

Problem 1.

$$\begin{aligned} &\min_{\overrightarrow{\boldsymbol{H_{SC}}},\ \overrightarrow{\boldsymbol{H_{MN}}}} \ \left\{\textstyle\sum_{\theta \in \mathcal{M}} C^\theta\right\} \\ &s.t. \ \ \forall \theta \in \mathcal{M}: \ 0 \leq H_{SC}^\theta(0) \leq N_{SC} \\ &\qquad\qquad\qquad 0 \leq H_{MN}^\theta(0) \leq R^\theta(0) \\ &\qquad \textit{and} \qquad \sum_{\theta \in \mathcal{M}} H_{SC}^\theta(0) \leq \sum_{i \in \mathcal{SC}} Q(i) \end{aligned}$$

where $\overrightarrow{\boldsymbol{H_{SC}}}$ and $\overrightarrow{\boldsymbol{H_{MN}}}$ denote the vectors with components $H_{SC}^\theta(0)$ and $H_{MN}^\theta(0)$ ($\theta \in \mathcal{M}$), respectively, and $Q(i)$ is the caching capacity (in number of contents) of a SC node[3] i.

The costs C^θ in the objective function of Problem 1 can be expressed as a function of the optimization variables (Result 3). As a result, well known numerical or analytic methods (see, e.g., [23]) can be employed to solve Problem 1.

VI. Simulation Results

To validate our analysis, we compare the theoretical predictions against Monte Carlo simulations on various synthetic scenarios (with node meetings drawn from Poisson processes), and on traces generated by state-of-the-art mobility models.

Results of synthetic simulations demonstrate a significant accuracy of our predictions and verify the arguments used in our analysis. Due to space limitation we omit these synthetic simulations, which can be found in [23], and in the remainder we present results in the more challenging scenarios, where mobility characteristics depart from our model assumptions.

Specifically, we use the TVCM [24] and SLAW [25] mobility models, which have been shown to capture well real mobility patterns, like power-law flights [25], community structure [24], etc. The generated scenarios we present are:

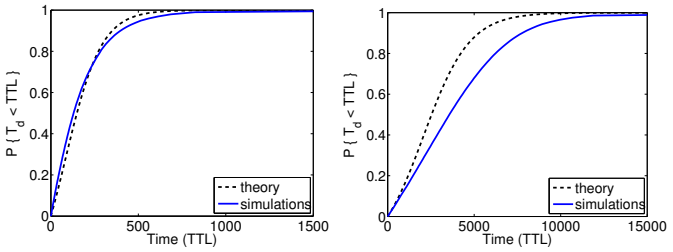
TVCM scenario: Mobile nodes move in a square area $1000m \times 1000m$, which contains three areas of interest (communities). Nodes move mainly inside their community (60% of the time) and leave it for a few short periods. Macro-cell BSs provide full coverage of the whole area, while 25 non-overlapping 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 [25]). Macro-cell BSs cover the whole area and coexist with 100 non-overlapping small-cells. Communication ranges are set as above.

In Fig. 1 we present the delivery probability $P\{T_d \leq TTL\}$, along with the theoretical prediction, for two content traffic scenarios in the TVCM (Fig. 1(a)) and SLAW (Fig. 1(b)) traces. Contents with popularity $R(0) = 50$ are initially cached to $H(0)$ edge nodes (half of which are MNs).

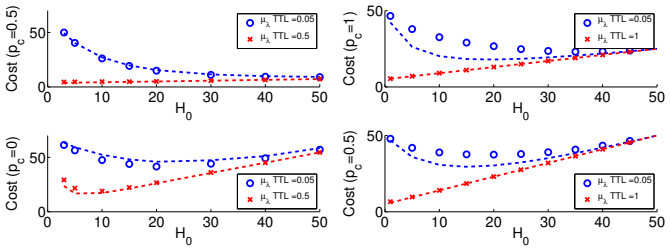
[2]Additionally, it might be possible that the delay-tolerance of each content, TTL^θ, can be selected as well. This case can be captured by our framework as well. However, for notation simplicity, in the remainder we consider only the initial placement parameters as the optimization variables.

[3]Since MNs cache *only* contents in which they are interested in, we assume that their storage capacity is enough for all the contents of their interest. Hence, storage capacity constraints for MNs are not considered in Problem 1. On the other hand, since SCs are expected to cache much more contents than a MN, it is essential to take into account their capacity.



(a) TVCM: $H(0) = 10$, $R(0) = 50$ (b) SLAW: $H(0) = 5$, $R(0) = 50$

Fig. 1: 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 $H_{SC}(0) = H_{MN}(0) = \frac{H(0)}{2}$.



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

Fig. 2: 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) $C_{BS}^{(TTL)} = 2 \cdot C_{BS} = 10 \cdot C_{D2D}$.

The MNs' participation in offloading is set to $p_c = 0.5$. In the TVCM trace (Fig. 1(a)) it can be seen that the accuracy of our results is significant, despite the community structure of the network (which cannot be captured by Assumption 2). In the SLAW scenario (Fig. 1(b)), our results overestimate the delivery probability. However, note here that the number of holders in the SLAW scenario is smaller, and, thus, the prediction of Result 1 is expected to be less accurate (since it is based on the expressions of Lemma 1, which become more accurate as the number of requesters and/or holders increase). For scenarios with more initial holders the accuracy of the predictions increases (see e.g. Fig. 2(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 (Result 3) is less affected. Fig. 2 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. 2(b) for $\mu_\lambda \cdot TTL = 0.05$), they can still capture the actual optimal initial allocation regimes.

VII. Discussion and Extensions

In this section we discuss how our study and base model can be modified and/or extended to make our framework applicable to more generic "offloading on the edge" setups.

A. Mobility Model and Cooperation/Caching Heterogeneity

The mobility model we use allows heterogeneous meeting rates λ_{ij} (Assumption 2) in order to account for various node mobility patterns and communication ranges. However, in Lemma 1 we apply an approximative method (see details in [23]), which leads to considering only the mean value of the meeting rates μ_λ in the analytic results. Although the same expressions could have been derived (easier) by using a homogeneous model, i.e. $\forall i, j : \lambda_{ij} = \mu_\lambda$, our main motivation for considering heterogeneous rates is the following. We can easily incorporate further heterogeneous characteristics (related to mobility patterns), like: (i) social selfishness, where node cooperation is related also to their social ties / mobility [20]; or (ii) smart, mobility-aware content placement algorithms, where the edge nodes that are encountered more frequently by the requesters are selected as holders [26]. These characteristics could *not* have been taken into account under a homogeneous mobility assumption.

To extend our results for the two above cases, it is just needed to modify the analytic expressions by substituting the average meeting rate μ_λ with the effective average meeting rate $\mu_\lambda^{(eff.)}$ (see [20, Lemmas 3.1 and 3.2] and [26, Result 4], respectively), which is given by

$$\mu_\lambda^{(eff.)} = E[\lambda \cdot p(\lambda)] \quad \text{and} \quad \mu_\lambda^{(eff.)} = \frac{E[\lambda \cdot \pi(\lambda)]}{E[\pi(\lambda)]} \qquad (2)$$

where the function $p(\lambda)$ describes the social selfishness and $\pi(\lambda)$ the mobility-aware content placement algorithm. The expectations in Eq. (2) are taken over the meeting rates distribution f_λ, and, as a result, the mobility heterogeneity is actively involved in the performance prediction expressions. On the contrary, using a homogeneous model ($\lambda_{ij} = \mu_\lambda, \forall i, j$), selfishness can be captured only with a single parameter (as the one we already use, p_c), while traffic heterogeneity cannot be captured at all:

$$\mu_\lambda^{(eff.)-HOM} \rightarrow \mu_\lambda \cdot p(\mu_\lambda) \rightarrow \mu_\lambda \cdot p_c$$
$$\mu_\lambda^{(eff.)-HOM} \rightarrow \frac{\mu_\lambda \cdot \pi(\mu_\lambda)}{\pi(\mu_\lambda)} = \mu_\lambda$$

B. Dependence of Costs on System Parameters

In our model we considered constant costs for the different transmission types. However, in different scenarios, there might exist some correlation between transmission costs and other system parameters. Some examples could be:

(i) The delayed content delivery cost $C_{BS}^{(TTL)}$ might be a function of TTL. If user impatience increases with time, the cost $C_{BS}^{(TTL)}$ might increase with TTL, e.g.,

$$C_{BS}^{(TTL)} = c_1 + c_2 \cdot e^{c_3 \cdot TTL}, \quad c_1, c_2, c_3 = const.$$

(ii) If multicasting is used for initial content placement to MNs, the transmission cost C_{BS} might not be *linearly* related to $H_{MN}(0)$, i.e., the cost of multicasting a content to $H_{MN}(0)$ nodes, might not be equal to $H_{MN}(0)$ times the cost of a unicast transmission.

(iii) The MN-MN transmission cost C_{D2D} might be correlated with the cooperation probability p_c. For instance, the willingness of nodes to participate in offloading (which is captured by p_c) might be higher when the reward for each offloaded content (which is captured by C_{D2D}) increases.

Our results predicting the content dissemination performance and cost remain the same, or need minor (and straightforward) modifications, when adopting such more generic cost models. For example, in the aforementioned *multicast* case, if the cost for multicasting a content to x users is given by a known function $g(x)$ (which may depend on transmission technology, density of users, size of content, etc.), we need only to replace the second term in Result 3 as follows

$$C_{BS} \cdot H_{MN}(0) \rightarrow g\left(H_{MN}(0)\right)$$

VIII. 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 [12], [13], [18], [19]. In the setting of [18], 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 [12], [13], which study the optimal selection of the number of initial and final content pushes. [13] 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, [12] uses a fluid limit approximation and focuses on the cost optimization problem. Finally, [19] 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, [11], [26] consider a limited number of (designated) holders. [11] 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, [26] 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. Moreover, our results could be incorporated to the framework of [26] to calculate the average offloading performance and cost (among all the contents delivered).

Finally, [27] 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].

IX. 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 and cost.

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, 2015–2020," Tech. Rep., 2016.

[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] 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.

[12] 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.
[13] V. Sciancalepore, D. Giustiniano, A. Banchs, and A. Picu, "Offloading cellular traffic through opportunistic communications: Analysis and optimization," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 1, pp. 122–137, 2016.
[14] 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.
[15] K. Johansson, "Cost effective deployment strategies for heterogenous wireless networks," *Doctoral Thesis*, 2007.
[16] 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.
[17] A.-F. Tatar, "Predicting user-centric behavior: Content popularity and mobility," *Doctoral Thesis*, 2014.
[18] 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.
[19] H. Cai, I. Koprulu, and N. Shroff, "Exploiting double opportunities for deadline based content propagation in wireless networks," in *Proc. IEEE INFOCOM*, 2013.
[20] P. Sermpezis and T. Spyropoulos, "Understanding the effects of social selfishness on the performance of heterogeneous opportunistic networks," *Computer Communications, Elsevier, Volume 48*, 04 2014.
[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] P. Sermpezis, L. Vigneri, and T. Spyropoulos, "Offloading on the edge: Analysis and optimization of local data storage and offloading in HetNets," Tech. Rep. http://arxiv.org/abs/1503.00648, 2015.
[24] 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.
[25] K. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong, "Slaw: A new mobility model for human walks," in *IEEE INFOCOM*, 2009.
[26] 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.
[27] 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.

APPENDIX A
PROOF OF RESULT 3

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 (which as said are also requesters of it)[4]. 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 $R_0 \cdot P\{T_d \leq TTL\}$. Denoting with q the percentage of requesters that receive the content by a SC, we can express the costs due to SC-MN and MN-MN content deliveries as

$$C_{SC} \cdot q \cdot R_0 \cdot P\{T_d \leq TTL\} \quad (3)$$

$$C_{D2D} \cdot (1-q) \cdot R_0 \cdot P\{T_d \leq TTL\} \quad (4)$$

[4]In total the content is placed to $H_{SC}(0)+H_{MN}(0)$ edge nodes. However, since some MNs/requesters might not be willing to act as holders, the number of holders after the initial placement will be equal to $H_0 = H_{SC}(0) + H_{MN}(0^+)$, where $H_{MN}(0^+) \leq H_{MN}(0)$.

respectively.

To calculate the percentage q we proceed as follows:

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

$$\#R_{tot} = R_0 - R(TTL)$$

which can be also expressed as (see Eq. (1))

$$\#R_{tot} = R_0 \cdot P\{T_d \leq TTL\} \quad (5)$$

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

$$R(t) - R(t, t+dt) = -dR(t) \quad (6)$$

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

$$\frac{H_{SC}(0)}{H(t)} \in [0,1] \quad (7)$$

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

$$\begin{aligned}\#R_{SC} &= \int_0^{TTL} -dR(t) \cdot \tfrac{H_{SC}(0)}{H(t)} = \int_0^{TTL} -\tfrac{dR(t)}{dt} \cdot \tfrac{H_{SC}(0)}{H(t)} dt \\ &= \int_0^{TTL} H(t) \cdot R(t) \cdot \mu_\lambda \cdot \tfrac{H_{SC}(0)}{H(t)} \cdot dt \\ &= \mu_\lambda \cdot H_{SC}(0) \int_0^{TTL} R(t) \cdot dt \end{aligned} \quad (8)$$

where (in the second line) $\frac{dR(t)}{dt} = -H(t) \cdot R(t) \cdot \mu_\lambda$ follows from the expressions of Lemma 1. Now, using the expression of Lemma 1 for $R(t)$ to calculate the above integral, we get

$$\begin{aligned}\#R_{SC} &= \frac{H_{SC}(0)}{p_c} \cdot \ln\left(\frac{(p_c \cdot R_0 + H_0) \cdot e^{\mu_\lambda \cdot (p_c \cdot R_0 + H_0) \cdot TTL}}{p_c \cdot R_0 + H_0 \cdot e^{\mu_\lambda \cdot (p_c \cdot R_0 + H_0) \cdot TTL}}\right) \\ &= \frac{H_{SC}(0)}{p_c} \cdot \ln\left(\frac{H(TTL)}{H_0}\right) \end{aligned} \quad (9)$$

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

Now, q easily follows from Eq. (5) and Eq. (9)

$$q = \frac{\#R_{SC}}{\#R_{tot}} = \frac{H_{SC}(0)}{p_c} \cdot \frac{\ln\left(\frac{H(TTL)}{H_0}\right)}{R_0 \cdot P\{T_d \leq TTL\}} \quad (10)$$

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 Result 3) follows easily.