Competitive Caching of Contents in 5G Edge Cloud Networks

Francesco De Pellegrini[◇], Antonio Massaro[◇], Leonardo Goratti[◇], and Rachid El-Azouzi[⋆]

***Abstract*—The surge of mobile data traffic forces network operators to cope with capacity shortage. The deployment of small cells in 5G networks shall increase radio access capacity. Mobile edge computing technologies can be used to manage dedicated cache memory at the edge of mobile networks. As a result, data traffic can be confined within the radio access network thus reducing latency, round-trip time and backhaul congestion. Such technique can be used to offer content providers premium connectivity services to enhance the quality of experience of their customers on the move.**

In this context, cache memory in the mobile edge network becomes a shared resource. We study a competitive caching scheme where contents are stored at a given price set by the mobile network operator.

We first formulate a resource allocation problem for a tagged content provider seeking to minimize the expected missed cache rate. The optimal caching policy is derived accounting for popularity of contents, spatial distribution of small cells, and caching strategies of competing content providers.

Next, we study a game among content providers in the form of a generalized non-smooth Kelly mechanism with bounded strategy sets and heterogeneous players. Existence and uniqueness of the Nash equilibrium are proved. Finally, numerical results validate and characterize the performance of the system.

***Index Terms*—Mobile Edge Computing, Caching, Convex Optimization, Kelly Mechanism, Nash Equilibrium**

I. Introduction

The recent boom of mobile data traffic is causing unprecedented stress over mobile networks. In fact, the global figures for such traffic reached 3.7 exabytes per month at the end of 2015. They are mostly ascribed to content providers (CP), e.g., video providers such as Vimeo, YouTube and NetFlix. Forecasts predict that the world's mobile data traffic will reach 30.6 monthly exabytes by 2020 [1].

As a consequence, capacity shortage has become a real threat for mobile network operators (MNOs), at the risk of reduced service quality. Solutions involving the deployment of small cell (SC) base stations [2] have been receiving large consensus from industry and academia for next LTE-based 5G systems. SCs are low power secondary base stations with limited coverage, to which user equipments (UEs) in radio range can connect, hence increasing spatial reuse and network capacity.

However, SCs are connected to a mobile operator's core network via backhaul technologies such as, e.g., DSL, Ethernet or flexible millimeter-wave links.

[◇]Fondazione Bruno Kessler, via Sommarive, 18 I-38123 Povo, Trento, Italy; [⋆]CERI/LIA, University of Avignon, 339, Chemin des Meinajaries, Avignon, France. This research received funding from the European Union's H2020 Research and Innovation Action under Grant Agreement No.671596 (SESAME project).

Mobile edge caching diverts traffic connections from remote servers using local proxies within the Radio Access Network (RAN). It can thus overcome limited backhaul connection of SCs [3]. Moreover, this strategy performs content localization, i.e., it brings the content close to content consumers: by increasing proximity to mobile users, it is possible to drastically reduce the round trip time, perform adaptation to radio access conditions and thus improve users' experience [4]. Finally, it confines content access within the MNO infrastructure. Insulation from public Internet network conditions, in turn, permits the MNO to exploit effectively the QoS mechanisms built into LTE standards to guarantee premium-grade mobile connectivity.

From the network management standpoint, in order to handle a large number of SCs and associated memory caches, MNOs will rely on the emerging mobile edge computing (MEC) [5] 5G technology. MEC platforms are designed to enable services to run inside the mobile RAN.

Ultimately, CPs will be able to leverage the MEC caching service offered by 5G MNOs. Contents such as, e.g., videos, music files, or online gaming data, can be directly replicated on lightweight server facilities embedded in the radio access network. In this context, the design of effective mobile edge caching policies requires to factor in popularity, number of contents, cache memory size as well as spatial density of small cells to which UEs may associate to. Indeed, due to storage limitations, allocation of contents on mobile edge caches has become an important optimization problem [6], [7], [8], [9], [10], [11], [12].

We consider a scheme in which CPs can reserve mobile edge cache memory from a MNO. The MNO can provide a multi-tenant environment where contents can be stored at given price and will assign the available caching resources to different content providers. In turn, this engenders competition of CPs for cache utilization.

First, we study the single CP optimization problem: under a given spatial distribution of SCs, the CP decides the optimal cache memory share to be reserved to different classes of contents, thus minimizing missed cache rate as a function of the purchased memory.

Finally, the competition among CPs is formulated using a new generalized Kelly mechanism with bounded strategy set. CPs trade off the cost for caching contents in the radio access network versus the expected missed cache rate. The game is showed to have a unique Nash equilibrium. Further properties of the game, including convergence and the optimal revenue of the MNO, are investigated numerically.

978-3-901882-90-6/17 ©2017 IFIP

II. Related Works and main contribution

The authors of [6] consider a device-to-device (D2D) network and derive throughput scaling laws under cache coding and spatial reuse. Content delay is optimized in [13] by performing joint routing and caching, whereas in [7] a distributed matching scheme based on the deferred acceptance algorithm provides association of users to SC base stations based on latency figures. In [8] contents to be cached minimize a cost which depends on the expected number of missed cache hits. In [9] a model for caching contents for D2D networks is proposed. A convex optimization problem is obtained and solved using a dual optimization algorithm.

In [10] a coded caching strategy is developed to optimize contents' placement based on SC association patterns. In [14] a Stackelberg game is investigated to study a caching system consisting of a content provider and multiple network providers: CPs lease their videos to the network providers to gain profit and network providers aim to save the backhaul costs by caching popular videos. Finally, [12] proposes proactive caching in order to take advantage of contents' popularity. The scheme we develop in this work can also be applied to proactive caching. Recent results [15] show that by online estimation of the contents' popularity it is possible to minimize the missed cache rate. We leave the online estimation of the contents' demand rates as part of future works.

Main results. The main contributions of this work are the following: 1) we introduce a model accounting for contents' popularity, spatial distribution of small cells, the price for cache memory reservation and the effect of competing content providers under multi-tenancy; 2) the optimal caching policy is found to possess a waterfilling-type of structure 3) a competitive game is formulated where the price for cache memory reservation is fixed by the network provider; it results in a Kelly mechanism with bounded strategy set which admits a unique Nash equilibrium.

Our uniqueness results for the Nash equilibrium are new. Existing results, e.g., [16], [17], require the cost function to be twice continuously differentiable. Even in the smooth case, uniqueness for bounded strategy sets proved in a recent work [18] only applies to linear costs. Our result applies as well to non linear yet convex cost functions. The complete technical discussion is reported in [19] for space's sake.

III. System Model

We consider a MNO serving a set $\mathcal{C}$ of content providers, where $|\mathcal{C}| = C$. Each CP c serves his customers leveraging the MNO network.

Each content provider is assumed to host up to M different content classes, e.g., videos, music files, online gaming data, etc. The i-th content class features N_c^i contents. For the sake of analysis, we assume that files within each class are equally popular and contents of class i generate $\tilde{g}_c^i$ requests per day. A class of contents will hence generate $N_c^i\,\tilde{g}_c^i$ requests per day.

We assume that each SC is attached to a local edge caching server, briefly *cache*. Multiple caches are aggregated by connecting them through the MNO backhaul and managed using a local MEC orchestrator, thus forming a seamless local edge cache unit as in Fig. 1. N *caching slots* represent the available memory on such local edge cache unit; the total cache space across the whole deployment is hence $N_0 = K \cdot N$ where K is the number of local edge cache units. For the sake of simplicity, each content is assumed to occupy one caching slot; since we assume $N_0, N \gg 1$, we rely on fluid approximations to describe the dynamics of cache occupation.

TABLE I
Main notation used throughout the paper

Symbol	*Meaning*
M	number of content classes
Λ	intensity, i.e., spatial density of small-cells
$\mathcal{C}$	set of content providers, $\lvert\mathcal{C}\rvert = C$
r	covering radius of UEs
N	storage capacity of a local edge cache unit (number of caching slots)
N_0	total storage capacity of the deployment
N_c^i	number of contents of class i for content provider c
$\tilde{g}_c^i$	demand rate for contents of class i of content provider c
g_c^i	$g_c^i := w_c^i\, N_c^i\, \tilde{g}_c^i$
Λ_c^i	$\Lambda_c^i := \Lambda \pi r^2 w_c^i\, \tilde{g}_c^i$
b_c	caching rate of content provider c, $b_c \in [0, B_c]$
b	total caching rate $b = \sum_{c\in\mathcal{C}} b_c$
$b_{-c} = \sum_{v\neq c} b_v$	total caching rate of competing content providers;
δ	mobile network provider's own caching rate
$\mathbf{u}_c$	caching policy for content provider c, $\mathbf{u}_c = (u_c^1, \ldots, u_c^M)$, $\sum u_c^i = 1$
x_c	share of cache memory occupied by content provider c
x_c^i	share of cache memory for i-th class contents of content provider c
B_c	maximum caching rate for content provider c
λ_c	price per caching slot for content provider c

Fetching a non cached content from the remote CP server beyond the backhaul comes at unitary cost; such cost may represent the content's access delay or the throughput to fetch the content from the remote server. Conversely, such cost is negligible if the user associates to a small cell storing a cached copy of the content. However, such cache should be reached by connecting to a SC within the UE radio range $r > 0$. SCs are distributed according to a spatial Poisson point process with intensity Λ.

The following assumptions characterize the caching process:
i. each CP c can purchase edge-caching service from the MNO and issue b_c caching slot requests per day; we call b_c the *caching rate*, where $0 \leq b_c \leq B_c$;
ii. MNO reserves $\delta > 0$ caching slots per day for her own purposes;
iii. reserved slots expire at rate $\eta > 0$, i.e., after $1/\eta$ days;
iv. in order to attain b_c caching slots per day, CP c bids $\tilde{b}_c \in [0, 1]$, and the MNO grants $b_c = b_0\, \tilde{b}_c$ caching slots per day, where b_0 is such that $\sum b_c + \delta \leq N_0$. In our analysis we assume $b_0 = 1$ for the sake of simplicity[1].
v. CPs are charged based on the caching rate b_c;
vi. demand rates g_c^i per content class are uniform across the MNO's network.

The MNO accommodates X_c memory slots for CP c according to

$$\dot{X}_c = b_c - \eta\, X_c, \tag{1}$$

[1]We refer to [20] for a connection between linear bids, efficiency and fair share of resources of the type studied in this paper.

so that the whole cache memory occupation is ruled by

$$\dot{X} = b - \eta X, \tag{2}$$

where $b := \sum_c b_c + \delta$ is the total caching rate. Let $X(0) = 0$: the corresponding dynamics for the reserved cache memory writes

$$X(t) = \min\left\{N_0, \frac{b}{\eta}\left(1 - e^{-\eta t}\right)\right\}$$

The MNO will ensure full memory utilization ($X(\infty) = N_0$) by choosing η such that $b/\eta \geq N_0$. It follows from a simple calculation that, in steady state, the fraction of the caching space for CP c is

$$x_c(t) = \frac{b_c}{b_c + b_{-c} + \delta} \tag{3}$$

Because contents' requests are uniform across the MNO's network, the same fraction of cache space is occupied by CP c in each local edge cache unit.

In particular, CP c will split his reserved memory among content classes according to a proportional share allocation with weighting coefficients u_c^i, $i = 1, \ldots, M$, where $\sum_{i=1}^M u_c^i = 1$. Vector $\mathbf{u}_c := (u_c^1, \ldots, u_c^M)$ defines the *caching policy* of CP c.

Then, the fraction of local edge cache memory occupied by contents of class i from CP c is

$$x_c^i = \frac{b_c}{\sum_{v \in \mathcal{C}} b_v + \delta} u_c^i \tag{4}$$

Finally, a tagged content of class i of CP c is found in the memory of a local edge cache with probability $P_c^i = \min\{\frac{N}{N_c^i} x_c^i, 1\}$. In the rest of the paper, we will assume $N < N_c^i$ for the sake of simplicity.

Now, we want to quantify the probability for a given requested content not to be found in the local edge cache memory, i.e., the *missed cache* probability.

Under the Poisson assumption, the probability for a tagged UE not to find any SC within distance r is $e^{-\pi r^2 \Lambda}$. Applying a thinning argument, the probability not to find a content of class i of CP c within distance r is $e^{-\pi r^2 \Lambda P_c^i}$.

The cost function of CP c is the weighted sum of the missed cache probabilities, namely the *weighted missed cache rate* (WMCR)

$$U_c(b_c, b_{-c}, \mathbf{u}_c) = \sum_i g_c^i \, e^{-\pi r^2 \Lambda \frac{N}{N_c^i} \frac{b_c}{b_c + b_{-c} + \delta} u_c^i} \tag{5}$$

where we have defined $g_c^i := w_c^i N_c^i \tilde{g}_c^i$, and w_c^i weights the relative importance of class i at CP c. The cost function depends on caching rate b_c and on caching policy $\mathbf{u}_c$. Also, $b_{-c} := \sum_{v \neq c} b_v$ accounts for the fact that other CPs share the same cache space.

Next, we shall describe the optimal caching policy $\mathbf{u}_c^*$ attained when CP c aims at minimizing (5), for a fixed value b_c of the caching rate.

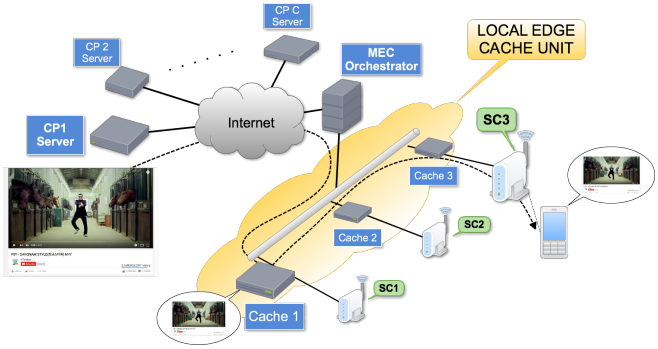

Fig. 1. Local edge cache unit providing N memory slots.

IV. Optimal Caching Policy

In order to analyze the model introduced before, we need to characterize the CPs' response to competitors' actions, i.e., b_{-c}. Hence, we assume that each CP aims at minimizing his own WMCR, and that the network provider guarantees full information to CPs, i.e., storage capacity, spatial density of SCs and total caching rate. We defer the study of the system under partial information at the CP's side to later works. We hence consider the following resources allocation problem for the single CP:

Definition 1 (Optimal Caching Policy)**.** *Given opponents' strategy profile* $\mathbf{b}_{-c} = (b_1, \ldots, b_{c-1}, b_{c+1}, \ldots, b_C)$ *the optimal caching policy of* $c \in \mathcal{C}$ *is the solution of*

$$\mathbf{u}_c^* := \arg \min_{u_c^1, \ldots, u_c^M} U_c(b_c, b_{-c}, \mathbf{u}_c) \tag{6}$$

subject to the following constraints:

$$u_c^i \geq 0, \quad \sum_i u_c^i = 1 \tag{7}$$

It is immediate to observe that $U_c(b_c, b_{-c}, \mathbf{u}_c)$ is a strictly convex function in the single CP control $\mathbf{u}_c$, so that a unique solution exists [21]. In order to solve the constrained minimization problem in equations (6) and (7) we can write the Lagrangian for player $c \in \mathcal{C}$ as follows

$$L_c(\mathbf{u}_c, \mu, \nu) = \sum_i g_c^i e^{-\Lambda_c^i \frac{b_c}{b+\delta} u_c^i} - \sum_i \mu_i u_c^i + \nu \left(\sum_i u_c^i - 1\right)$$

For notation's sake, we have defined $\Lambda_c^i = \pi r^2 \Lambda \frac{N}{N_c^i}$. Furthermore, since constraints are affine, the Karush Kuhn Tucker (KKT) conditions solve the problem [21]:

$$\begin{aligned} &\nabla_u L_c(\mathbf{u}_c, \mu, \nu) = \mathbf{0} \\ &\sum_i u_c^i - 1 = 0 \\ &u_c^i \geq 0, \quad \mu_i \geq 0, \quad \nu \geq 0, \quad \mu_i u_c^i = 0 \end{aligned}$$

Using a standard argument [21], by complementary slackness, $u_c^i > 0$ implies $\mu_i = 0$; let us define index set $I := \{i \in \{1, 2, \ldots, M\} \,|\, u_c^i > 0\}$.

A. Waterfilling solution

The solution to the KKT conditions can be formulated as a waterfilling-like solution [21]. In fact, from stationarity conditions, μ_r writes as

$$\frac{\partial L_c}{\partial u_c^i} = -\Lambda_c^i \frac{b_c}{b+\delta} g_c^i e^{-\Lambda_c^i \frac{b_c}{b+\delta} u_c^i} - \mu_i + \nu = 0$$
$$\mu_i = \nu - \Lambda_c^i \frac{b_c}{b+\delta} g_c^i e^{-\Lambda_c^i \frac{b_c}{b+\delta} u_c^i}$$

which can be specialized into the following two cases.
Case i: $\nu > \Lambda_c^i g_c^i \frac{b_c}{b+\delta}$. In this case $\mu_i > 0$ for any $u_c^i \geq 0$. Hence, by complementary slackness, $u_c^i = 0$.
Case ii: $\nu \leq \Lambda_c^i g_c^i \frac{b_c}{b+\delta}$. It is always possible to find $u_c^i > 0$ satisfying the stationarity condition and a μ_i that satisfies the complementary slackness condition: just set $\mu_i = 0$ and

$$u_c^i = \frac{b+\delta}{\Lambda_c^i b_c} \log\left(\frac{\Lambda_c^i g_c^i}{\nu} \frac{b_c}{b+\delta}\right)$$

Finally, let $\alpha_i := \frac{b+\delta}{\Lambda_c^i g_c^i b_c}$. For notation's sake, the solution writes

$$u_c^{*i} = \begin{cases} \frac{b+\delta}{\Lambda_c^i b_c} \left(\log(1/\nu) - \log(\alpha_i)\right) & \text{if } 1/\nu > \alpha_i \\ 0 & \text{if } 1/\nu \leq \alpha_i \end{cases} \quad (8)$$
$$\text{subject to:} \quad \sum_i u_c^{*i} = 1$$

It is immediate to recognize a waterfilling solution in logarithmic scale. Let $\alpha = \min_i \alpha_i$. Indeed $\sum_i u_c^{*i}$ is strictly increasing in $1/\nu, 1/\nu > \alpha$. Also, $\sum_i u_c^{*i}(1/\nu) = 0$ for $1/\nu \leq \alpha$, and $\lim_{1/\nu \to \infty} \sum_i u_c^{*i}(1/\nu) = \infty$. Thus, there exists a unique positive ν satisfying our problem.

Actually, the solution is determined in polynomial time $O(M)$. Now, we can sort class indexes in increasing order of the α_is. For every choice $\alpha_i \leq 1/\nu \leq \alpha_{i+1}$, we can determine a value of ν

$$\log(1/\nu) = \frac{\frac{b_c}{b_c+b_{-c}+\delta} + \sum_{r=1}^{k} \frac{\log \alpha_r}{\Lambda_c^r}}{\sum_{r=1}^{k} \frac{1}{\Lambda_c^r}}$$

for $k = 1, \ldots, M$. Then, consider the only $1/\nu$, compatible with (8). We observe that $\alpha_i \leq \alpha_{i+1}$ is equivalent to state

$$w_c^i \tilde{g}_c^i \geq w_c^{i+1} \tilde{g}_c^{i+1} \quad (9)$$

and if $u_c^i = 0$, indeed $u_c^{i+1} = 0$, so that we can derive the following

Corollary 1 (Threshold structure)**.** *There exists $1 \leq r_0 \leq M$ such that $u_c^{*s} > 0$ for $s \leq r_0$ and $u_c^{*s} = 0$ otherwise.*

Closed form. The stationarity conditions can be used to determine the structure of the optimal solution in closed form. Let $0 \leq i \leq r_0$, then $\mu_i = \mu_{r_0} = 0$, so that

$$\Lambda_c^{r_0} g_c^{r_0} e^{-\Lambda_c^{r_0} \frac{b_c}{b+\delta} u^{*r_0}_c} = \Lambda_c^i g_c^i e^{-\Lambda_c^i \frac{b_c}{b+\delta} u^{*i}_c}$$

and

$$u^{*i}_c = \frac{\Lambda_c^{r_0}}{\Lambda_c^i} u^{*r_0}_c - \frac{b+\delta}{\Lambda_c^i b_c} \log\left(\frac{g_c^{r_0} \Lambda_c^{r_0}}{g_c^i \Lambda_c^i}\right)$$

Finally, due to the constraint saturation

$$u^{*r_0}_c = \frac{1 + \frac{b+\delta}{b_c} \sum_{i=0}^{r_0} \frac{1}{\Lambda_c^i} \log\left(\frac{g_c^{r_0} \Lambda_c^{r_0}}{g_c^i \Lambda_c^i}\right)}{\sum_{i=0}^{r_0} \frac{\Lambda_c^{r_0}}{\Lambda_c^i}} \quad (10)$$

From Cor. 1, the optimal solution corresponds to the maximal r_0 such that $u^{*r_0}_c$ solving (10) lies in $[0, 1]$

Threshold structures in waterfilling-type solutions are expected: here, from (9), by increasing the cache space, classes in the cache will appear according to increasing α_is, i.e., decreasing values of $w_c^i \tilde{g}_c^i$. From now on, we assume content classes sorted according to Cor. 1. The weights w_is let CPs control such order, whereas, when $w_c^i = w_c^j$ for all i, j, the optimal caching policy depends on contents' demand rate only, as expected.

V. Optimal Cost Function

CPs who optimize contents to be cached, for a given value of b_c, minimize the expected WMCR $U_c(b_c, b_{-c}, \mathbf{u}_c)$ in the caching policy $\mathbf{u}_c$. In the game model presented in the next section we shall leverage the convexity properties of the optimal cost function $U_c : \mathbb{R}_+^2 \to \mathbb{R}$, defined as

$$U(b_c, b_{-c}) := \min_{\mathbf{u_c} \in \Pi} \left\{ \sum_i g_c^i e^{-\Lambda_c^i \frac{b_c}{b_c+b_{-c}+\delta} u_c^i} \right\} \quad (11)$$

where $\Pi = \{\mathbf{u} \in \mathbb{R}^M | \mathbf{u}_c \geq 0, \sum u_c^i = 1\}$. As already proved, the minimum in (11) is unique, hence $U_c(b_c, b_{-c})$ is well defined. Hereafter we demonstrate its convexity in b_c.

Actually, convexity can be derived for a class of functions wider than the posynomial expression appearing in (11). We need the following fact, whose proof is found in [19].

Lemma 1. *Let f be non increasing, with domain $\mathbb{R}_+$. Let $H(x) = x\, f(x)$ be convex. Then f is convex on $\mathbb{R}_+$.*

We can now derive the general conditions for the convexity of the optimal missed cache rate

Theorem 1. *Let $h : \mathbb{R}^M \to \mathbb{R}$, convex and decreasing in each variable x_i for $i = 1, \ldots, M$, then*

$$U_c(b_c, b_{-c}) := \min_{\mathbf{u_c} \in \Pi} h\Big(\frac{u_c^1 b_c}{b_c + b_{-c} + \delta}, \ldots, \frac{u_c^M b_c}{b_c + b_{-c} + \delta}\Big)$$

is convex and decreasing in b_c.

The case in (11) satisfies the assumptions by letting $h(\mathbf{x}) = \sum_i g_c^i e^{-\Lambda_c^i x_i}$.

For presentation's sake, in Sec. VI we shall identify $U(b_c, b_{-c}, \mathbf{u}_c^*) := U(b_c, b_{-c})$. There, we also need the following result, whose proof is found in [19].

Lemma 2 (Limit solution for $b_c \to 0$)**.** *There exists $\varepsilon > 0$ such that, for any $b_c < \varepsilon$, $\mathbf{u}_c^* = (1, 0, \ldots, 0)$ and the optimal WMCR is*

$$U_c(b_c, b_{-c}, \mathbf{u}_c^*) = g_c^1 e^{-\Lambda_c^1 \frac{b_c}{b_c+b_{-c}+\delta}} + \sum_{i>1} g_c^i \quad (12)$$

A. The case $M = 2$

For two classes of contents, $M = 2$, the expression for $U_c(b_c, b_{-c})$ is derived in closed form. This sample case retains the main properties of the optimal policy and provides insight into the structure of the optimal WMCR. The optimal WMCR writes

$$U_c(b_c, b_{-c}) = \min_{0 \le u_c^1 \le b_c} g_c^1 e^{-\Lambda_c^1 \frac{b_c\, u_c^1}{b_c + b_{-c} + \delta}} + g_c^2 e^{-\Lambda_c^2 \frac{b_c (1 - u_c^1)}{b_c + b_{-c} + \delta}}$$

For notation's sake, we denote $\Gamma := \frac{w_c^2 \tilde{g}_c^2}{w_c^1 \tilde{g}_c^1}$. The (unconstrained) minimum of the right hand term is attained at

$$u^{*1}_c = \frac{\Lambda_c^2}{\Lambda_c^1 + \Lambda_c^2} - \frac{(b_c + b_{-c} + \delta)}{b_c(\Lambda_c^1 + \Lambda_c^2)} \log(\Gamma) \tag{13}$$

When $u^{*1}_c \in (0, 1)$, the utility function of $c \in \mathcal{C}$ is

$$U_c(b_c, b_{-c}) = K_c\, e^{-\frac{\Lambda_c^1 \Lambda_c^2}{\Lambda_c^1 + \Lambda_c^2} \frac{b_c}{b_c + b_{-c} + \delta}}$$

where the constant appearing on the first term is

$$K_c = g_c^1 \cdot \Gamma^{\frac{\Lambda_c^1}{\Lambda_c^1 + \Lambda_c^2}} + g_c^2 \cdot \Gamma^{-\frac{\Lambda_c^2}{\Lambda_c^1 + \Lambda_c^2}} \tag{14}$$

Incidentally, the convexity of $U_c(\cdot, b_{-c})$ for $M = 2$ can be verified directly from the convexity of $\exp(1/x)$ and by composition with an affine function, which preserves convexity.

We are interested in precisely characterizing the behavior of the expected WMCR as a function of b_c and of other system parameters.

Now, we can obtain the following result

Proposition 1. *i. Assume* $\Gamma < 1$. *Let* $\Lambda_c^1 > \log(1/\Gamma)$, *and define the threshold for content* 2

$$b_c^\star = (b_{-c} + \delta) \frac{\log(1/\Gamma)}{\Lambda_c^1 - \log(1/\Gamma)} \tag{15}$$

then it holds

$$U_c(b_c, b_{-c}) = \begin{cases} g_c^1 e^{-\Lambda_c^1 \frac{b_c}{b_c + b_{-c} + \delta}} + g_c^2 & \text{if} \quad 0 \le b_c < b_c^\star \\ K_c\, e^{-\frac{\Lambda_c^1 \Lambda_c^2}{\Lambda_c^1 + \Lambda_c^2} \frac{b_c}{b_c + b_{-c} + \delta}} & \text{if} \quad b_c \ge b_c^\star \end{cases} \tag{16}$$

where the corresponding optimal caching policy is $(1, 0)$ *in the first case,* $(u^{*1}_c, 1 - u^{*1}_c)$ *in the second case; constant* K_c *is defined in* (14)

ii. Let $\Lambda_c^1 \le \log(1/\Gamma)$, *then* $(1, 0)$ *case holds for any* $b_c > 0$ *with associated expected WMCR defined as in case i.*

iii. If $\Gamma > 1$, *both i. and ii. hold with role of content* 1 *and* 2 *reversed.*

The proof follows by inspection of (13) considering u^{*1}_c as an unconstrained minimizer. First, we observe that if $\Gamma < 1$, then $u^{*1}_c > 0$, i.e., the first content class is always cached. The other conditions follow by imposing $u^{*1}_c \ge 1$.

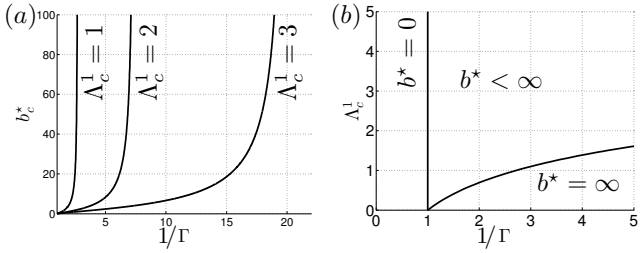

Fig. 2. Case $M = 2$: (a) Increasing value of $b_c^\star$ as a function of Γ, for $\Lambda_c^1 = 1, 2, 3$, $w_c^1 = w_c^2 = 1$ and $u_{-c} = \delta = 1$ (b) Region of switch on of content 2.

Discussion

Hereafter we draw insight from Prop. 1. First, as seen there, the optimal caching rate u^{*1}_c depends solely on a few system parameters, namely g_c^i and Λ_c^i for $i = 1, 2$. Actually, when $\Gamma < 1$ then $u_c^{1*} = 1 - u_c^{2*} > 0$: contents of type 1 are always cached because $w_c^2 \tilde{g}_c^2 < w_c^1 \tilde{g}_c^1$. The fact that contents of type 2 are cached depends on the sign of $\Lambda_c^1 - \log(1/\Gamma)$, which in turn determines the actual structure of the waterfilling solution.

From Prop. 1, Λ_c^1 determines whether contents of type 1 will be cached or not. In practice, when $\Lambda_c^1 > \log(1/\Gamma)$, there exists a critical value of the CP caching rate b_c, i.e., the threshold (15). Above such value, contents of type 2 are cached, below that they are not cached. For the sake of consistency, in the case when $\Lambda_c^1 \le \log(1/\Gamma)$, $b^\star = +\infty$ while for $\Gamma > 1$, $b^\star = 0$.

Furthermore, $b_c^\star$ increases linearly with both the MNO caching rate δ and the competitors' aggregate caching rate b_{-c}: competition tends to prevent caching of contents with smaller $g_c^i \Lambda_c^i$. Actually, under higher competition figures, optimal caching policies are of the type $u_c^{1\,*} = 1$, and $u^{*2}_c = 0$. We observe that, as detailed in case ii., not always there exists a caching rate b_c such that it is worth caching the least profitable content class.

We have provided a pictorial representation of the results of this section in Fig. 2 for the case $M = 2$. In Fig. 2(a) the value of Λ_c^1 has been fixed at different values and the corresponding behavior of the threshold value $b_c^\star$ has been reported as a function of Γ. For $\exp(\Lambda_c^1) \le 1/\Gamma$, it holds $b_c^\star = \infty$ since there is no switch-on value of b_c for class 2. Fig. 2(b) represents the region where the switch-on of the less popular content is possible as it can be derived from the expression (15) as a function of $1/\Gamma$ and Λ_c^1.

VI. Game Model for Content Providers

So far the caching rate b_c has been input for the CPs in order to decide how to optimize the caching policy $\mathbf{u}_c$. Let MNO propose to CPs costs λ_c per caching slot. The strategy of CP c in turn is the number b_c of caching slots he reserves per day, with convex and compact strategy set $[0, B_c]$. The best response b_c^* of CP c depends on his contents and his opponents' strategies. It minimizes *cost function* $U_c(b_c, b_{-c}, \mathbf{u}_c) + \lambda_c \cdot b_c$ by solving:

$$\min_{b_c} U_c(b_c, b_{-c}, \mathbf{u}_c) + \lambda_c \cdot b_c \tag{17}$$
$$0 \le b_c \le B_c$$

Note: all the following results can be generalized for the case of a general continuously differentiable convex cost $\lambda_c(b_c)$.

Here $b_{-c} = \sum_{v \neq c} b_{-v}$ and opponents' strategy profile writes $\mathbf{b}_{-c} = (b_1, \ldots, b_{c-1}, b_{c+1}, \ldots, b_C)$.

The $\mathbf{u}_c$ appearing in (17) is a general caching policy and we shall consider two cases.

Caching Rate Optimizers. In this case, the best response of CPs is decided for a *fixed caching policy* $\mathbf{u}_c$. I.e., each CP decides beforehand the caching policy $\mathbf{u}_c$ for any given caching rate b_c. Let $V_c(x_c) = \sum_i g_c \, e^{-\Lambda_c^i x_c}$: it is strictly convex and decreasing and $U_c(b_c, b_{-c}, \mathbf{u}_c) = V_c(b_c/(\sum b_c + \delta))$. Hence, if all players are caching rate optimizers, the game is a variant of the Kelly mechanism [22]. The basic Kelly mechanism allocates a divisible resource among players proportionally to the players' bids, in our case the equivalent required caching rates. Here, compared to the standard formulations in literature [17], [23], [22], [16] our formulation combines three specific features which render it non standard:

- bounded compact and convex strategy set;
- $\delta > 0$ is a bidding reservation, as described in [16];
- prices may differ from player to player, i.e., the game is a generalized Kelly mechanism [22]

The game outlined above is a *generalized Kelly mechanism with reservation and bounded strategy set*.

Simultaneous Optimizers. In this case $\mathbf{u}_c = \mathbf{u}_c^*$ (see Sec. V). The structure of the game still resembles the Kelly mechanism [23]. For $M = 1$, the game corresponds to the case of caching rate optimizers. For $M \geq 2$, the fact that the game is actually a Kelly mechanism follows from [19]

Lemma 3 (Kelly form for Simultaneous Optimizers)**.** *If players are simultaneous optimizers, the game* (17) *is a generalized Kelly mechanism with reservation and bounded strategy set.*

Hence, even in the case of a simultaneous optimizer CP c, the optimal WMCR can be expressed as $U_c(b_c, b_{-c}) = U_c(b_c, b_{-c}, \mathbf{u}_c^*) = V_c(b_c/(\sum b_c + \delta))$ where $V_c(x_c)$ is convex and continuously differentiable in $x_c = \frac{b_c}{\sum b_c + \delta}$.

A. Existence and uniqueness of the Nash Equilibrium

In the general case, both CPs who are caching rate optimizers and who are simultaneous optimizers may be present. From the above discussion, the game is still a generalized Kelly mechanism with reservation and bounded strategy set.

In order to characterize the possible equilibria, we describe first the best response b_c^* of each player:

Lemma 4. *Given the opponents' strategy profile* $\mathbf{b}_{-c}$*:*
i. It holds $b_c^* = 0$ *if and only if* $\dot{U}(0, b_{-c}) > -\lambda_c$ *where*

$$\dot{U}(0, b_{-c}) = \begin{cases} -\frac{\sum_i g_c^i \Lambda_c^i u_c^i}{b_{-c} + \delta} & \textit{caching rate optimizers} \\ -\frac{g_c^1 \Lambda_c^1}{b_{-c} + \delta} & \textit{simultaneous optimizers} \end{cases}$$

ii. Let $b_c^* > 0$*, then* $b_c^* = \min\{b_c, B_c\}$*, where* $\dot{U}_c(b_c, b_{-c}) = -\lambda_c$.

The above statement follows from the fact that the objective function in (17) is convex and thus has a unique minimum in $[0, B_c]$. The expression of $\dot{U}(0, b_{-c})$ in the case of simultaneous optimizers is derived from the expression (12) reported in Lemma 1.

The zero $\mathbf{b}^* = \mathbf{0}$ and the saturated $\mathbf{b}^* = \mathbf{B}$ Nash equilibria are easily characterized in the following

Proposition 1. *i.* $\mathbf{b}^* = \mathbf{0}$ *is the unique Nash equilibrium if and only if* $g_c^1 \Lambda_c^1 < \lambda_c \delta$ *if* c *is a simultaneous optimizer and* $\sum g_c^i \Lambda_c^i < \lambda_c \delta$ *if* c *is a caching rate optimizer.*
ii. $\mathbf{b}^* = \mathbf{B}$ *is the unique Nash equilibrium if and only if it holds* $\dot{U}_c(B_c, \sum B_c + \delta) > -\lambda_c$ *for all* $c \in \mathcal{C}$.

We observe that in the original Kelly mechanism, the strategy vector $\mathbf{0}$ is never a Nash equilibrium [17], [23].

In our case, it may be the Nash equilibrium because the MNO's usage of the cache ($\delta > 0$). In fact, the physical interpretation is provided by condition i. in Prop. 1. No CP has incentive to start caching when the marginal utility is below $\lambda_c \delta$. This is the value of the cache share reserved to the MNO operations. Conversely, at low prices a saturated Nash equilibrium $\mathbf{b}^* = \mathbf{B}$ is expected.

In the general case, the presence of a bounded strategy set requires a specific proof for the uniqueness of the Nash equilibrium, as seen in the following.

Theorem 2. *A unique Nash equilibrium exists for the game.*

Here, we describe a brief outline of the full proof, which is found in [19]. In order to prove the existence of Nash equilibria of the game, it is sufficient to observe that:

- the multistrategy set is a convex compact subset of $\mathbb{R}^C$;
- $U_c(b_c, b_{-c}, \mathbf{u}_c)$ is convex conditionally to the opponents' strategy, both for simultaneous and caching rate optimizers.

Hence, the existence of Nash equilibria is a direct consequence of the result of Rosen [24] for n–persons concave games. From Prop. 1, Nash equilibria of the type $\mathbf{b}^* = \mathbf{0}$ or $\mathbf{b}^* = \mathbf{B}$ are always unique. In the other cases, uniqueness can be derived by extending an argument [16] to the case of a bounded strategy set. However, the original argument requires cost functions to be twice continuously differentiable in b_c. For simultaneous optimizers, this just holds piecewise and a continuity argument needs to be derived. Finally, uniqueness applies also when part of the players are simultaneous optimizers and the others are caching-rate optimizers.

From the proof of Thm. 2 a simple bisection algorithm calculates the unique solution of the game. In the numerical section we shall further characterize the game by describing the pricing operated by the MNO and the convergence to the Nash equilibrium when CPs are myopic cost minimizers.

VII. Numerical Results

In this section we first validate the models' assumptions against a real world scenario. Then, we focus on the single player's actions, having fixed the remaining players' strategies. Finally we provide numerical characterization of the game introduced in the previous section [2].

[2]Both the Python scripts and the dataset can be downloaded at https://www.dropbox.com/s/mm1hja2dbp4tw0x/caching_scripts.tar.gz?dl=0

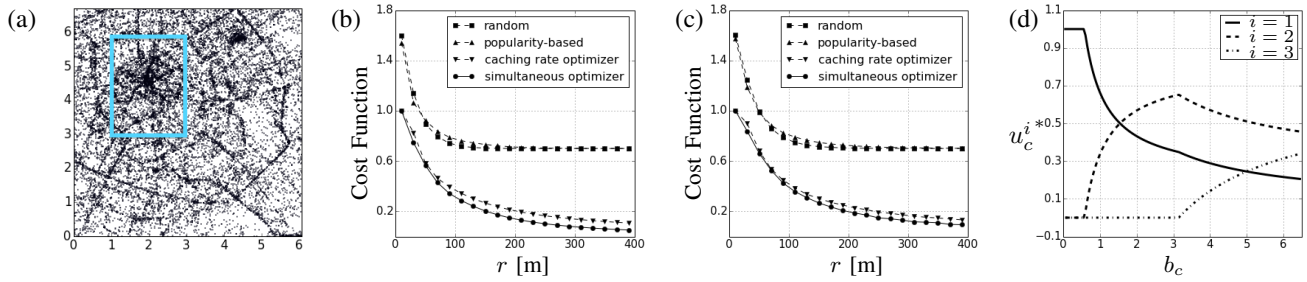

Fig. 3. (a) Milan downtown base stations deployment, detail of the area considered; (b) CP optimal cost: theoretical prediction (c) CP optimal cost: outcome of the simulation (d) CP optimal caching policy for varying b_c and fixed value of $r = 210$ m. Settings are: $M = 3$, $g_c^i = 0.589, 0.294, 0.118$, $\delta = 2$, $b_c = 70$, $b_{-c} = 300$, $N = 10000$, $N_c^i = 1000, 4000, 10000$.

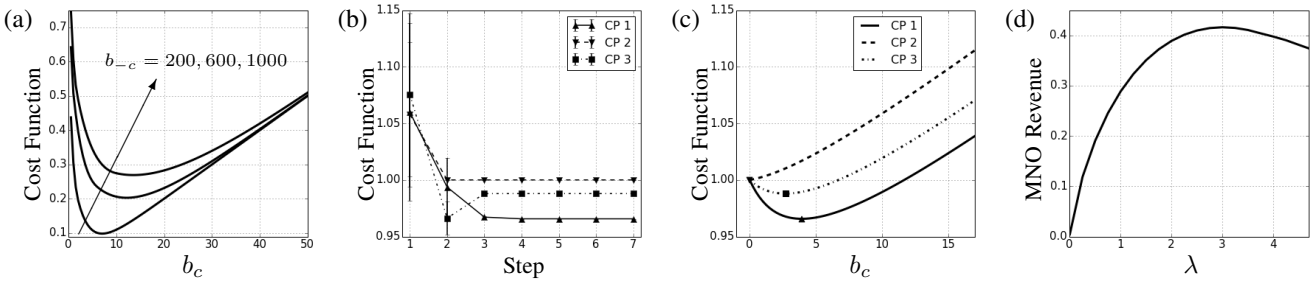

Fig. 4. (a) Cost function for simultaneous optimizer CP as b_c varies, all parameters are the same as in Fig 3(d); (b) Dynamic of the cost function for a 3-simultaneous optimizers game. Settings are: $g_1 = [0.18, 0.27, 0.55]$, $g_2 = [0.3, 0.6, 0.1]$, $g_3 = [0.6, 0.1, 0.3]$, $\Lambda_1^1 = \Lambda_1^2 = \Lambda_1^3 = 0.3$, $\Lambda_2^1 = \Lambda_2^2 = \Lambda_2^3 = 0.1$, $\Lambda_3^1 = \Lambda_3^2 = \Lambda_3^3 = 0.2$ (c) detail of the corresponding restpoint; (d) revenue of the MNO for increasing uniform price. Parameters are: $g_1 = [0.3, 0.2, 0.5]$, $g_2 = [0.3, 0.5, 0.2]$, $g_3 = [0.29, 0.36, 0.35]$, $N = 70$, $N_1^i = 600$, $N_2^i = 700$, $N_3^i = 500$, $\delta = 2$, $r = 73$m.

Point Process. Our model assumes that SCs are distributed according to a spatial Poisson process of given intensity Λ. Hence, we have tested the performance of the optimal caching policy when the SCs spatial deployment does not adhere to the Poisson point distribution assumption. In order to do so, the theoretical results are compared with the outcome of a simulation performed over a real dataset. The dataset (source http://opencellid.org/) is the sample distribution of the cell towers deployed in downtown Milan over a 2×3 Kms area, as depicted in Fig. 3(a). It includes the location of 4717 cell towers corresponding to $\Lambda = 786.2$ base stations per square Km. The distribution of base stations in a very densely populated urban area has been used as a reasonable approximation for a SC deployment.

The sample spatial density Λ has been used in the model in order to evaluate, under the same spatial density of SCs, the theoretical CP's cost function for increasing values of the covering radius $0 \leq r \leq 400$m in the following cases (see Fig. 3(b)): a) the CP performs a uniformly *random* caching policy $u_c^i = 1/3$, $i = 1, 2, 3$ for constant caching rate b_c b) the CP performs a *popularity-based* caching policy, i.e., $u_c^i := g_c^i / \sum g_c^i$, for constant b_c c) the CP is a *caching rate optimizer* adopting a popularity based caching policy d) the CP is a *simultaneous optimizer*.

The results in Fig. 3(c) refer to a simulation encompassing the same strategies under the sample point distribution of Fig. 3(a). The simulation has been performed by repeatedly selecting a random UE position in the playground, and measuring the sampling frequency of missed cache events upon requesting contents from SCs within the UE's radio range.

By comparing the results in Fig. 3(b) and Fig. 3(c), we observe that the Poisson distribution – as expected due to the non-uniform spatial density of the sample real-world deployment – tends to slightly underestimate the cost incurred by CPs. However, the theoretical and the simulated results are very close and the relative performance of the caching policies match the prediction of the theoretical model. This result confirms that the proposed model performs well even in real world scenarios: under a non-Poisson point process for the SC spatial distribution a rational optimizing player would choose the proposed optimal strategy over other possible strategies.

Cost function. In the next experiment we describe the optimal caching policy (Fig. 3(d)) and the cost function (Fig. 4(a)) in the case $M = 3$. In particular, Fig. 3(d) reports on the characteristic waterfilling structure of the optimal caching as the parameter b_c increases. As predicted by the model, the water-filling solution has a threshold structure. The value of b_c determines the content classes that are active: for large b_c all content classes are cached, whereas for small values only a few do. In Fig. 4(a) we have reported the typical convex shape of the cost function for increasing values of b_{-c}. It is worth noting how the actions of opponents, reflected in the value of b_{-c}, affect the shape of c's cost function.

Convergence to the Nash equilibrium. In Fig. 4(b) and Fig. 4(c) we have simulated a game of 3 CPs who are simultaneous optimizers. They are *myopic* players: each one of them, chosen

at random, optimizes his own cost function based on the opponents' profile. Numerical simulations show that, after a small number of iterations, the game stabilizes quickly at the same restpoint irrespective of initial strategies. As depicted in Fig. 4(c) the restpoint is indeed a minimum for each CP's cost function, i.e., it is the Nash equilibrium of the game. From this behavior, the game appears to have the finite improvement property [25], even though we cannot identify analytically a potential for the game. If true, the system would indeed converge to the Nash equilibrium when each player optimizes independently his own cost function against the opponents.

Finally, Fig 4(d) depict the daily revenue of the MNO at the Nash equilibrium $\mathbf{b}^*$ as a function of the caching price λ, uniform for all CPs. Because the MNO's total revenue $\sum_c \lambda \cdot b_c^*$ depends on the Nash equilibrium, she could optimize her revenue by leveraging the CPs' cost structure. We observe numerically that the total revenue appears to have a unique maximum at a certain maximizer price λ^*. This suggests the existence of a unique Stackelberg equilibrium for the proposed scheme, which permits to compute the global restpoint of the system when both CPs and MNO behave strategically.

VIII. Conclusions

Mobile edge caching will enhance the delivery of contents such as, e.g., videos, music files, online games, by reducing latency and round-trip-time. It will empower premium connectivity services while avoiding backhaul congestion. We model the competition of CPs for the caching service made available by a MNO. Our model captures several features, including popularity of contents, spatial distribution of small cells, competition for cache memory and price effects. CPs optimize the allocation of contents in order to reduce customers' missed cache rate. Best response of single CPs are of waterfilling type according to demand rates and class priority. The validity of the caching policy optimization has been tested on real-world traces. Finally, competition for the shared caching memory is formulated as a convex n–persons game: CPs trade off the expected missed cache rate for the memory price. It is *a non-smooth* Kelly mechanism with reservation and bounded strategy set – for which new existence and uniqueness of Nash equilibrium are provided. When CPs are myopic optimizers, convergence to the Nash equilibrium is showed numerically.

Furthermore, the game appears to have a unique Stackelberg equilibrium, a relevant feature for the MNO in order to maximize her revenue. Online learning of the optimal price over time will be part of future works.

References

[1] "Cisco visual networking index: Global mobile data traffic forecast update (20152020) white paper)," CISCO, White Paper, February 3 2016.

[2] J. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. Soong, and J. Zhang, "What will 5G be?" *IEEE JSAC*, vol. 32, no. 6, pp. 1065–1082, June 2014.

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

[4] C. Ge, N. Wang, S. Skillman, G. Foster, and Y. Cao, "QoE-Driven DASH video caching and adaptation at 5G mobile edge," in *Proc. of ACM Information-Centric Networking*, Kyoto, Japan, September 26-28 2016, pp. 237–242.

[5] MEC ETSI Industry Specification Group, "ETSI DGS/MEC-IEG004: Mobile-Edge Computing (MEC) – Service Scenarios," Available online:http://www.etsi.org/technologies-clusters/technologies/mobile-edge-computing.

[6] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of distributed caching in D2D wireless networks," in *Proc. of IEEE ITW*, Sevilla, Spain, Sept. 9-13 2013, pp. 1–5.

[7] F. Pantisano, M. Bennis, W. Saad, and M. Debbah, "Cache-aware user association in backhaul-constrained small cell networks," in *Proc. of IEEE WiOPT*, Hammamet, Tunisia, May 12-16 2014, pp. 37–42.

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

[9] H. J. Kang and C. G. Kang, "Mobile device-to-device (D2D) content delivery networking: A design and optimization framework," *Journal of Communications and Networks*, vol. 16, no. 5, pp. 568–577, Oct 2014.

[10] A. Sengupta, S. Amuru, R. Tandon, R. Buehrer, and T. Clancy, "Learning distributed caching strategies in small cell networks," in *Proc. of IEEE ISWCS*, Barcelona, Spain, Aug. 26-29 2014, pp. 917–921.

[11] J. Hachem, N. Karamchandani, and S. Diggavi, "Multi-level coded caching," in *Proc. of IEEE INFOCOM*, Hong-Kong, RPC, April 26th - June 1st 2015, pp. 756 – 764.

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

[13] M. Dehghan, A. Seetharam, B. Jiang, T. He, T. Salonidis, J. Kurose, D. Towsley, and R. K. Sitaraman, "On the complexity of optimal routing and content caching in heterogeneous networks," in *Proc. of IEEE INFOCOM*, Hong-Kong, RPC, April 26 - May 1 2015, pp. 936–944.

[14] J. Li, W. Chen, M. Xiao, F. Shu, and X. Liu, "Efficient video pricing and caching in heterogeneous networks," *IEEE Transactions on Vehicular Technology*, vol. PP, no. 99, pp. 1–1, 2015.

[15] S. Li, J. Xu, M. van der Schaar, and W. Li, "Popularity-driven content caching," in *Proc. of IEEE INFOCOM*, San Francisco, CA, 10-15 April 2016.

[16] R. Maheswaran and T. Başar, "Efficient signal proportional allocation (ESPA) mechanisms: Decentralized social welfare maximization for divisible resources," *IEEE J.Sel. A. Commun.*, vol. 24, no. 5, pp. 1000–1009, Sep. 2006.

[17] A. Reiffers-Masson, Y. Hayel, and E. Altman, "Game theory approach for modeling competition over visibility on social networks," in *Proc. of IEEE COMSNETS*, Bangalore, India, Jan 7-10 2014, pp. 1–6.

[18] E. Altman, M. K. Hanawal, and R. Sundaresan, "Generalising diagonal strict concavity property for uniqueness of nash equilibrium," *Indian Journal of Pure and Applied Mathematics*, vol. 47, no. 2, pp. 213–228, 2016. [Online]. Available: http://dx.doi.org/10.1007/s13226-016-0185-4

[19] F. De Pellegrini, A. Massaro, L. Goratti, and R. El-Azouzi, "Competitive Caching of Contents in 5G Edge Cloud Networks (Extended Version)," *CoRR*, vol. abs/612.01593, December 2016. [Online]. Available: https://128.84.21.199/abs/1612.01593v1

[20] T. Başar and G. Olsder, *Dynamic Noncooperative Game Theory*, 2nd ed. Society for Industrial and Applied Mathematics, 1998.

[21] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[22] Y. Yang, R. T. B. Ma, and J. C. S. Lui, "Price differentiation and control in the Kelly mechanism," *Elsevier Perf. Evaluation*, vol. 70, no. 10, pp. 792–805, October 2013.

[23] R. Johari, "Efficiency loss in market mechanisms for resource allocation," Ph.D. dissertation, Dept. of Electrical Eng. and Comp. Science, M.I.T., Cambridge, MA, 2004.

[24] J. B. Rosen, "Existence and uniqueness of equilibrium points for concave N-person games," *Econometrica*, vol. 33, no. 3, July 1965.

[25] D. Monderer and L. S. Shapley, "Potential games," *Games and Economic Behavior*, vol. 14, no. 1, pp. 124 – 143, 1996.