uc3m | Universidad Carlos III de Madrid

e-Archivo

This is a postprint version of the following published document:

Sciancalepore, V., et al. Enhanced content update dissemination through D2D in 5G cellular networks, in 2016 IEEE transactions on wireless communications, 15(11) *pp. 7517-7530, Nov. 2016*

DOI: https://doi.org/10.1109/TWC.2016.2604300

© 2016 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Enhanced Content Update Dissemination through D2D in 5G Cellular Networks

Vincenzo Sciancalepore, *Member, IEEE,* Vincenzo Mancuso, *Member, IEEE,*
Albert Banchs, *Senior Member, IEEE,* Shmuel Zaks and Antonio Capone, *Senior Member, IEEE*

***Abstract*—Opportunistic traffic offloading has been proposed to tackle overload problems in cellular networks. However, existing proposals only address D2D-based offloading techniques with deadline-based data propagation, and neglect content injection procedures. In contrast, we tackle the offloading issue from another perspective: the base station interference coordination problem during content injection. In particular, we focus on the dissemination of contents, and aim at the minimisation of the total transmission time spent by base stations to inject the contents into the network. We leverage the ABSF technique to keep under control intercell interference in such process. We formulate an optimisation problem, prove that it is NP-Hard and NP-Complete, and propose a near-optimal heuristic to solve it. Our algorithm substantially outperforms classical intercell interference approaches, as we evaluate through the simulation of LTE-A networks.**

***Index Terms*—eICIC, ABSF, NP-Hard, Content Dissemination, Injection, Offloading, D2D, LTE-A, Multicast**

I. Introduction

A number of web and smartphone applications have recently appeared, which cause the generation of a huge volume of traffic for mobile devices. A large fraction of the traffic generated by such applications consists in the distribution of contents such as social network updates and notifications, road traffic updates, map updates, and news feeds (e.g., *waze*, an *app* for a social network for navigation, includes all the above mentioned features).

Along with the appearance of such applications, some schemes have been recently proposed to offload the traffic generated by them in the cellular network. In particular, the device-to-device (D2D) paradigm has been proposed to assist the base station in the content distribution [1], [2], [3]: with D2D communications enabled, the base station delegates a few mobile users (*content injection*) to carry and spread contents to the other users (*content dissemination*). Although the content dissemination phase introduces delays, D2D-based content distribution is possible since it carries traffic with no strict real-time constraints, and whose content's lifetime lasts for a few minutes. Most of the currently available offloading proposals, e.g., [2], [4], focus on the characterization of content dissemination and the design of content injection strategies, but largely neglect the optimisation of radio resources in the *injection phase*, i.e., the process of injecting a content in a subset of the mobile user population, which produces bursty and periodic traffic. While this has been partially addressed, e.g., in [2], which has considered the impact of opportunistic resource utilisation in the content injection strategies, their analysis is restricted to a single cell and does not consider the interference caused by other cells, which is a key limiting factor for the deployment of dense and heterogeneous networks that are expected to appear in 5G cellular systems.

In line with the 5G view, we leverage the heterogeneity of technologies in the network to implement D2D-based offloading mechanisms, and tackle the cellular traffic offloading issue from a different and unexplored perspective: the intercell interference coordination problem. The rationale behind our approach is twofold: (i) interference is a key factor in future networks, where the single cell study case is not representative of a real network; (ii) content injection operations are impacted by network speed, which, in turn, strongly depends on intercell interference. In particular, to address the intercell interference coordination problem for 5G, in this paper we adopt the Almost Blank SubFrame (ABSF) paradigm recently defined for LTE-A [5]. This mechanism assigns resources in such a way that a subframe be *blanked* for some base stations, thus preventing their activity when the interference exceeds a threshold. A key advantage of this technique is that, by adopting a semi-distributed intercell interference coordination (ICIC) paradigm in which a central server simply announces to base stations the pattern of resources to be used, it greatly reduces the complexity of intercell interference coordination operations. While ABSF has only been proposed very recently and hence has not been thoroughly evaluated, some early studies (like our work in [6]) have shown its potential to improve cellular performance.

When scheduling the transmission of contents at base stations, our main objective is to minimise the time required for these transmissions, since (i) the faster contents are injected, the sooner they can be disseminated, and thus

Manuscript received May 11, 2015; revised December 19, 2015 and July 01, 2016; accepted August 18, 2016. Date of publication xxx xx, xxxx; date of current version xxx xx, xxxx.

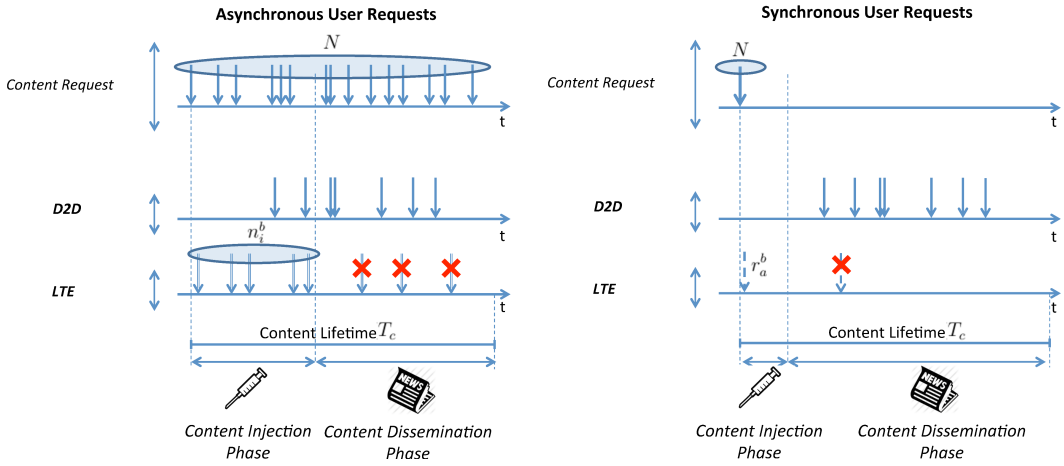
The work of this paper was performed in the context of the FP7 CROWD project (ref N^o 318115) and was followed up within the H2020 5G NORMA project (ref N^o 671584). The contributions of V. Mancuso were funded by the Ramon y Cajal grant (ref RYC-2014-01335) from the Spanish Ministry of Economy and Competitiveness.

V. Sciancalepore is with NEC Europe Ltd.

V. Mancuso and A. Banchs are with IMDEA Networks Institute and University Carlos III of Madrid, Spain.

Shmuel Zaks is with Department of Computer Science, Technion, Haifa, Israel, and was visiting IMDEA Networks while developing the work presented in this manuscript.

A. Capone is with DEIB, Politecnico di Milano, Italy.



(a) BS performs n_c^b unicast transmissions.

(b) BS performs a multicast transmission at rate r_a^b.

Fig. 1: Content Update Transmission Process for N users interested in the content c. On the left side, N users get interested in a content c at different points in time. The BS serves the first n_c^b asynchronous content requests in the Content Injection Phase using the cellular technology (LTE). Then, in the Content Dissemination Phase, users opportunistically exchange the content via D2D technologies. On the right side, we can see a particular case where BS performs a multicast transmission to a multicast group interested in the same content.

D2D-based offloading performance is optimised; and (ii) the less time required for the transmissions, the more resources are freed for other applications. We show that the problem of finding an ABSF-based scheduling algorithm that minimises the time required for content transmissions while satisfying the content deadlines is NP-Complete and NP-Hard to approximate. Thus, we design BSB, an algorithm that runs in polynomial time and achieves sub-optimal network performance, yet it outperforms the state of the art mechanisms proposed in the literature. In particular, our simulations show that BSB allows to abate the base station time devoted to content distribution by a factor 3 or larger, while boosting the ability of D2D schemes to reach the full set of content subscribers.

The contribution of our work can be summarised as follows: (i) we formulate a base station scheduling problem and we show that it is NP-Complete; (ii) we design and validate a practical algorithm for the computation of ABSF patterns; (iii) while available works on intercell interference coordination assume that a few interferes dominate the overall received interference experienced by a device, we show that, in a real network scenario, a much broader set of interferes needs to be taken into account for interference coordination; (iv) we show that channel-opportunistic D2D schemes are seriously impaired by non-ideal content injection.

The reminder of the paper is organized as follows: in Section II we present our *D2D-Assisted Content Distribution Process*, pointing out the main issues we cover in our study. Based on those features, in Section III we formulate an optimisation problem, proving its NP-Hardness and NP-Completeness, while in Section IV we introduce *BSB*, an efficient algorithm to solve the problem. Section V includes numerical results based on simulations, and Section VI provides a complete review of the state-of-the-art. Finally we conclude the paper in Section VII.

II. D2D-ASSISTED CONTENT DISTRIBUTION

In this section, we give a complete overview about the framework of our system and its building blocks. In addition, we provide a real scenario where our solution handily applies.

The overall content dissemination problem consists of two parts: injection (from the base station) and distribution (from another content subscriber, using D2D). We first consider intra-cell content injection and distribution under constrained base station resource utilisation, which yields the optimal number of content replicas to be injected to optimise the D2D-assisted distribution. Second, we control inter-cell interference so that the content injection computed in the first step is feasible. The compound effect of our approach results in maximising the data distributed to multiple groups of subscribers before content expiration.

A. Content distribution scenario

We address a LTE-A cellular scenario where N base stations are placed, each of which covers a user set $\mathcal{U}_b$, where b is the base station index. Each user subscribes a content $c \in \mathcal{C}$, with content length L_c and a deadline T_c by which the content needs to reach all subscribed users. Note that multiple users can request the same content.

An example of this scenario is the one envisaged in [7], in which users are moving in a vehicular scenario and a new available road traffic update is considered as content c. Clearly, the content needs to be delivered to the cars in sufficiently short time in order to be still useful to the users. To the aim of distributing a content to multiple users, while offloading the base station as much as possible, we exploit D2D technology communications. To control the content distribution process (see Fig. 1) we assume that a local controller is installed on each base station. The controller is in charge of deciding only the optimal number of content replicas to be injected directly by the base station. Upon users retrieve the content, they opportunistically share it (or

part of it) with other users via short-range communication technologies such as WiFi-Direct, WiFi or Bluetooth [8].

In this context, the main objective of this paper is to design a strategy to deliver content to users that minimises the total resources required by base stations, as this frees resources that can be used by other applications. An additional benefit of our approach is that power consumption of base stations decreases, since it depends on the total activity time of base stations. In order to achieve the objective, we need to address the following two challenges:

Intra-BS optimisation: the controller selects the optimal number of users to which delivers content replicas through cellular legacy transmissions. This ensures that (i) the content reaches as much as possible subscribers in the cell by the deadline and, (ii) resources required from the base stations are optimised.

Inter-BS optimisation: direct injections performed by each base station using the cellular technology need to be scheduled by accounting for inter-cell interference, i.e., by coordinating base stations, thus guaranteeing a bound on the total time required by these transmissions.

Note that the focus of the above challenges excludes D2D operations. However, D2D distribution plays a fundamental role in the system under investigation, so we study its performance in the Appendix and we use the results while presenting the mechanism that we propose for intra-BS and inter-BS optimisation problems.

B. Intra-BS content distribution

The content distribution process for a particular content may be divided into two phases: (i) *content injection* and (ii) *content dissemination*, hereafter described in details. Users placed under the coverage of base station b get interested in content c randomly, according to a normal distribution with average μ. Content validity period lasts T_c seconds and users may get interested only in a valid content c. We assume that the maximum number of interested users is equal to N, corresponding to the popularity index of the content c [9]. For the sake of simplicity, we suppose the same popularity index N for every content provided in the network.[1]

In the first phase, namely *content injection*, base stations transmit unicastly contents to each interested user asking for those updates. Specifically, the BS controller properly decides n_c^b the maximum number of unicast transmissions per content c BS b can perform. Then, the phase ends when exactly n_c^b interested users, called *injected users*, have received the content directly from base station b, e.g., upon n_c^b users get interested in the content. In the second phase, namely *content dissemination*, the content is spread opportunistically into the network via D2D technologies to those users which could not download the content directly from the base station. Although the two phases may overlap, this does not affect our analysis as the total time spent to deliver content replica to interested users does not change, as already proven in [10].

The number n_c^b of injected users plays a key-role in driving the system to an efficient working point. On the one hand, the more the number of injected users, the more the time required by the base station to perform the content transmission. On the other hand, if the number of injected users n_c^b has not been designed properly, most of the users asking for the content will not be reached within the content lifetime (T_c).

Therefore, we introduce a 2-dimensional Markov chain, where each state $S_j(t)$ is the total *number of content replica* distributed in the network at time t given j users interested in the content, regardless of the specific users carrying those replica. Assuming a homogeneous mobility model where users get in touch each other following an average inter-contact rate λ and getting interested in a content according to an average rate μ, we obtain the average number of users with content at the end of the content lifetime T_c as follows

$$\mathbb{E}[S_j(T_c)] = \sum_{x=1}^{N} x \; p_x(T_c - d_{in}). \tag{1}$$

where $p_x(t)$ is the probability to stay in the state x at time t. The homogeneous mobility assumption could be easily relaxed by introducing an interesting rate μ^t as function of the time t, e.g., peak hours or night hours. While this amendment could bring additional complexity in the analysis, it does not affect the computation of Eq. (1), as the probability $p_x(t)$ will be derived from the enhanced Markov chain model. For further details we refer the reader to the Appendix.

Eq. (1) provides a function returning the average number of users with the content after the content lifetime expiration (T_c), based on the number of injected nodes (n_c^b) and the number of interested users in that content j. Based on such information, BS b decides the number of injected nodes n_c^b per content c by solving the following optimisation problem

Problem INJECTION:

$$\begin{aligned} \text{maximise} \quad & \sum_{c=1}^{|\mathbb{C}|} \log(\eta_c), \\ \text{subject to} \quad & \sum_{c=1}^{|\mathbb{C}|} \tfrac{n_c^b \; L_c}{T_c} \leq \alpha C_b; \\ & n_c^b \in \{1...N\}, \end{aligned}$$

where the content throughput is defined as $\eta_c = L_c \; \frac{\mathbb{E}[S_j(T_c)]}{T_c}$, while αC_b identifies the available resources at the base station side. In other words, base station b finds the optimal n_c^b per content to ensure that the base station capacity constraint is fulfilled. Note that the use of *log* in the above formulation raises non-linear issues, but it helps to properly account for the fairness across content throughputs according to the *proportional fair* paradigm.

The optimisation of Problem INJECTION can be easily linearised and solved by means of a commercial solver. Moreover, due to scalability issue, very large instances of the problem can be approached through a simple heuristic,

[1]Nonetheless, we can readily derive equivalent results for heterogeneous content population indexes N_c depending on the content c.

providing an affordable trade-off between accuracy and complexity. Specifically, to linearize Problem INJECTION we sample the logarithmic function into a limited number of values, as only a discrete set of n_c^b values are considered for the optimisation. We obtain a matrix $\zeta = \{\zeta_{i,n}\}$ of $[|\mathbb{C}| \times |\mathbb{N}|]$ size, where $\zeta_{c,n} = \log(\eta_c)$, with $n = n_c^b$. Therefore, assuming the same content length L and lifetime T, $\forall c \in \mathbb{C}$, we can rewrite Problem INJECTION as follows:

Problem INJECTION-LIN:

$$\begin{aligned} \text{maximise} \quad & \sum_{c=1}^{|\mathbb{C}|} \sum_{n=1}^{|\mathbb{N}|} s_{c,n}\ \zeta_{c,n}, \\ \text{subject to} \quad & \sum_{c=1}^{|\mathbb{C}|} \sum_{n=1}^{|\mathbb{N}|} s_{c,n} \cdot n \leq K; \\ & \sum_{n=1}^{|\mathbb{N}|} s_{c,n} \leq 1, \forall c \in \mathbb{C}; \\ & s_{c,n} \in \{0,1\}, \end{aligned}$$

where $K = \alpha\ C_b \frac{T}{L}$, while $s_{c,n}$ is a binary value indicating with 1 whether n nodes are initially injected with content c, or 0 otherwise. In other words, we aim at choosing the optimal set of injected nodes values n_c^b (selecting one value per content), guaranteeing that the capacity constraint of the base station is efficiently fulfilled. When the number of available contents $|\mathbb{C}|$ or the content popularity index $|\mathbb{N}|$ tend to huge numbers, solving this problem may take very long time. Given that Problem INJECTION-LIN can be easily mapped into a generalized assignment problem, as heuristic to solve the problem we can use an extended version of the Hungarian Algorithm [11] to provide a near-optimal solution in reasonable time.

It is worth noting that the content distribution process can be readily extended to other scenarios, such as synchronous content update subscriptions [12], where user interest rate μ tends to infinite. In such scenarios users covered by base station b subscribe a content update c arranging distinct content interesting groups per cell (multicast groups), as depicted in Fig. 1(b). A new content will be issued every T_c seconds to any multicast group by any base station in the network. Each user subscribes only one single content update and the multicast groups are disjoint. Given that the multicast operation requires a transmission at the least user rate of all multicast receivers in the group [13], for a given multicast rate only a part of the users in the group will be able to decode the message (i.e., those whose channel condition enables them to receive at the chosen rate). Therefore, during the *content injection* phase, upon a new content update is available, the BS controller decides the rate r_c^b at which multicast transmissions must be performed. The *content dissemination* phase starts spreading the content (or part of it) opportunistically in the group to reach those users which have not received the content during the injection phase. Also in this case, the choice of the multicast rate for the initial injection involves the following trade-off: (i) if the selected multicast rate is too low, the number of bits injected will be small and thus efficiency will be low, (ii) however, if the selected rate is too high, the initial injection will only involve few users and hence content is unlikely to spread to all subscribed users by the content lifetime T_c. Therefore, BS b needs to optimally solve Problem INJECTION, where the number of injected nodes n_c^b is computed as a function of rate r_c^b, as studied in [2].

C. Inter-BS scheduling

Following the previous explanations, during the *content injection* phase, the content reaches only n_c^b users. Moreover, such injections cause interference due to the presence of multiple base stations. To address this problem, we adopt the ABSF paradigm, which has been shown to provide improved performance in presence of inter-cell interference [6], [14], [15].

On average, if $\mathcal{C}$ is the population of active contents, a single base station b needs to perform d_b content transmissions, where $d_b = \sum_{c \in \mathcal{C}} n_c^b$. Content requests arrive asynchronously, even though contents are made available at regular intervals T_c, whose duration represents the content's lifetime. In addition, a base station serves all its users with unicast transmissions, applying a scheduler with equal rate, i.e., all users with pending transmissions are scheduled and receive the same data rate on a per-TTI basis. The achievable throughput t_u of each user u in subframe i depends on its signal-to-noise-ratio:

$$t_u(i) = B_T \log_2 \left(1 + \frac{S_u^b(i)}{N_0 + \sum_{j \neq b} I_u^j(i) x_{ij}} \right) \qquad (2)$$

where B_T is the used bandwidth, S_u^b is the useful signal received by user u from the serving base station b, N_0 is the background noise, I_u^j is the interference created by the base station j toward user u, and x_{ij} is a binary value which indicates whether the base station j is scheduled in the subframe i. We define $w_u, u \in 1, .., d_b$, as the set of positive coefficients representing the fraction of resources allocated to active user u in a subframe, such that equal rates are achieved:

$$w_p t_p = w_q t_q, \quad \forall p, q \in \mathcal{U}_b, \quad \text{s.t.:} \sum_{p=1}^{d_b} w_p = 1. \qquad (3)$$

Therefore, the coefficients w_u can be computed (in each subframe i) as follows:

$$w_u(i) = \frac{\frac{1}{t_u(i)}}{\sum_{k=1}^{d_b} \frac{\delta_{ki}}{t_k(i)}}, \qquad (4)$$

where δ_{ki} is 1 if transmission k is ongoing in subframe i, and it is 0 otherwise. With the above, the throughput of user u is $w_u(i) t_u(i)$ in subframe i.

III. Base station transmission time minimisation

Here, we formulate the inter-BS scheduling introduced before as an optimisation problem, and show that it is NP-Complete and NP-hard to approximate. Then, we provide a sufficient condition to solve the problem, which we will leverage to generate ABFS patterns (see Section IV).

A. Problem formulation

The efficiency of the content dissemination depends on the speed of the content injection process, and therefore our goal when designing the inter-BS scheduling is to minimise the time needed to inject the content, as expressed in the following optimisation problem:

Problem BS-SCHEDULE

Input:
A collection of N base stations $B = \{1, 2, \cdots, N\}$, and distinct transmission entities[a] $O = \{o_1^b, o_2^b, \cdots, o_{d_b}^b\}$ associated with base station $b \in B$. Positive constants N_0, τ, Θ, L_c, B_T. Integer $Z > 0$. For a generic entity o associated with base station b: $S_o^b(i)$, $w_o(i)$ and $I_o^j(i)$ for every $j \in B \setminus \{b\}$ and every $i = 1, 2, \cdots, Z$.

Question: Is there a scheduling of the base stations in at most Z rounds, such that

$$\Psi_o^b(Z) = \tau B_T \sum_{i=1}^{Z} x_{ib} w_o(i) \log_2 \left(1 + \frac{S_o^b(i)}{N_0 + \sum_{j \neq b} I_o^j(i) x_{ij}}\right) \geq L_c,$$

$\forall o \in \{1, .., d_b\},\ b \in \{1, .., N\}$, and

$$\sum_{b=1}^{N} T_{TOT}^b = \tau \sum_{b=1}^{N} \sum_{i=1}^{Z} x_{ib} \leq \Theta \quad ?$$

[a]Throughout all the paper, we refer with term *transmission entity* for both unicast user (u) and multicast group (a), as the same problem formulation can be easily applied to both unicast and multicast transmission types.

In Problem **BS-SCHEDULE**, each term $T_{TOT}^b = \tau \sum_{i=0}^{Z} x_{ib} = \tau Z_b$ represents the activity time of base station b (τ is the subframe duration). The term $w_o(i)$ is the generic fraction of resources reserved to a transmission entity o in subframe i. Z is the number of subframes that correspond exactly to the content lifetime interval T_c, while Θ is the upper bound for the aggregate transmission time of the system. Transmission rates are computed using Shannon capacity formula.

Although here we formulate an optimisation problem for content injection, the Markov Chain describing the dissemination process, after the injection phase is completed, reveals that our formulated problem is equivalent to the maximisation of the success probability in the content dissemination operation, with multiple contents to be distributed in parallel. Indeed, as reported in the Appendix, Eqs. (14)–(16), the probability of being in a state with more distributed pieces of content increases with the time t available for the dissemination phase. Therefore, the average number of distributed content pieces increases when the injection time is reduced.

B. Complexity of Problem *BS-SCHEDULE*

Classical wireless scheduling problems, e.g., scheduling and channel assignment, have been shown to be NP-Hard [16], [17]. However, we are the first to address the complexity of base station resource allocation with deadlines and multicast transmissions using variable rates. Specifically, we show that problem **BS-SCHEDULE** is NP-Complete when $Z \geq 3$ for bounded interferences, and for Z=2 for unbounded interferences. These NP-Completeness results apply to very special instances of the problem ($d_b = 1$ for every base station b).

Theorem 1. *Problem **BS-SCHEDULE** is NP-Complete, for any $Z \geq 3$, even when all interferences are $\in \{0, 1\}$.*

Sketch of Proof: It is clear that the problem is in NP. For the NP-Hardness we use a reduction from the problem **GCk** of graph k-coloring (see [18]). We are given an instance $I_{\textbf{GCk}} = H(V, E)$ of Problem **GCk**, and construct an instance $I_{\textbf{BS-SCHEDULE}}$ of Problem **BS-SCHEDULE**. Assume $V = \{1, 2, \cdots, n\}$. The base stations are $B = \{b_1, b_2, \cdots, b_n\}$, and the users $U = \{u_1, u_2, \cdots, u_n\}$, where for every t base station b_t is serving user u_t. In addition, $Z = k$, $N_0 = \tau = B_T = L_c = 1$, $\Theta = n$, and $S_{u_t}^{b_t}(i) = w_a(i) = 1$ for every $i = 1, 2, \cdots, Z$, $t = 1, 2, \cdots, n$. Last, for every $t = 1, 2, \cdots, n$, every $j \neq t$ and every $i = 1, 2, \cdots, Z$, $I_{u_t}^{b_j}(i) = 1$ if $(i, j) \in E$ and is 0 otherwise.

We have to show that there is a k-coloring of $I_{\textbf{GCk}}$ if and only if for $I_{\textbf{BS-SCHEDULE}}$ there is a scheduling of the base stations in at most k rounds, with $\Psi_{u_t}^{b_t}(Z) \geq 1 = L_c$, and $\sum_{t=1}^{n} T_{TOT}^{b_t} \leq n$.

Given a graph k-coloring of $I_{\textbf{BS-SCHEDULE}}$, with colors $1, 2, \cdots, k$. If a node t is colored p, then we schedule station b_t in round p, for $p = 1, 2, \cdots, k$. $\Psi_{u_t}^{b_t}(Z) = \sum_{i=1}^{Z} x_{ib} \log_2 \left(1 + \frac{1}{1 + \sum_{j \neq t} I_{u_t}^{b_j}(i) x_{ij}}\right)$ for every t. Since all base stations b_j scheduled with b_t are such that $(j, t) \notin E$, and since each base station is scheduled in exactly one round, therefore $\Psi_{u_t}^{b_t}(3) = \log_2 \left(1 + \frac{1}{1}\right) = 1$. $\sum_{t=1}^{n} T_{TOT}^{b_t} = n$ since each station is scheduled in exactly one round.

Conversely, assume that for $I_{\textbf{BS-SCHEDULE}}$ there is a general scheduling of at most k rounds, such that for each user $\Psi_{u_t}^{b_t}(k) \geq 1$ and $\sum_{t=1}^{n} T_{TOT}^{b_t} \leq n$. $\Psi_{u_t}^{b_t}(k) > 0$ implies that each user—and thus each station—is scheduled in at least one round. $\sum_{t=1}^{n} T_{TOT}^{b_t} \leq n$ implies that each station—and thus each user—is scheduled in exactly one round. Moreover, if user u_i is scheduled with user u_j, then $(i, j) \notin E$ (otherwise $\Psi_{u_i}^{b_i}(Z) < 1 = L_c$). Thereby assigning color p to nodes associated with the stations in round $p = 1, 2, \cdots, k$, results in a k-coloring of graph $I_{\textbf{GCk}}$. □

Theorem 2. *Problem **BS-SCHEDULE** is NP-Complete for $Z = 2$.*

Sketch of Proof: We use a reduction from a variation of the Partition problem. We term this Problem **MPAR**. In the Partition problem we are given integers $A = \{a_1, a_2, \cdots, a_n\}$, such that $\sum_{j=1}^{n} a_j = 2S$, and have to determine whether there exist $\{a'_1, a'_2, \cdots, a'_k\} \subseteq A$ such that $\sum_{j=1}^{k} a'_j = S$ (see [18]). In the modified version **MPAR** (that can be shown to be NP-Complete) we are given integers $A =$

$\{x_1, x_2, \cdots, x_{2n}\}$, $S > 0$, $S < x_i < 2S$ for all i, such that $\sum_{j=1}^{2n} x_j = 2(n+1)S$, and have to determine whether there exist $\{x'_1, x'_2, \cdots, x'_n\} \subseteq A$ such that $\sum_1^n x'_j = F$, where $F = (n+1)S$.

We are given an instance I of **MPAR**, and construct an instance $I_{\textbf{BS-SCHEDULE}}$ of Problem **BS-SCHEDULE** as follows. The base stations are $B = \{b_1, b_2, \cdots, b_{2n}\}$, and the users $U = \{1, 2, \cdots, 2n\}$; base station b_i is serving user i. $Z = 2$, $N_0 = F$, $\tau = B_T = L_c = 1$, $\Theta = n$, and $S_{u_t}^{b_t}(i) = 2F$, $w_a(i) = 1$ for $i = 1, 2$, $t = 1, 2, \cdots, n$. Last, for every $t = 1, 2, \cdots, n$, every $j \neq t$ and every $i = 1, 2$: $I_{u_t}^{b_j}(i) = x_j + \frac{x_i}{n-1}$.

We have to show that there is a solution to I if and only if there is a scheduling for $I_{\textbf{BS-SCHEDULE}}$ in at most 2 rounds, such that for each user $\Psi_{u_t}^{b_t}(2) \geq 1 = L_c$, and $\sum_{t=1}^n T_{TOT}^{b_t} \leq n$.

Assume there is a solution to I. Thus we assume the existence of a $\{x'_1, x'_2, \cdots, x'_n\} \subset A$ such that $\sum_1^n x'_j = F$. Schedule the base stations $b_{x'_1}, b_{x'_2}, \cdots, b_{x'_n}$ in the first round and the other n base stations in the second round. Clearly $\sum_{t=1}^n T_{TOT}^{b_t} \leq n$.

Every user t is thus scheduled in exactly one round, and thus

$$\Psi_{u_t}^{b_t}(2) = \log_2\left(1 + \frac{2F}{F + \sum\left\{x'_j + \frac{x'_i}{n-1} \middle| j=1,..,n, j\neq i\right\}}\right) = \log_2\left(1 + \frac{2F}{F+\sum_1^n x'_j}\right) = \log_2\left(1 + \frac{2F}{F+F}\right) = 1.$$

Conversely, assume a solution to $I_{\textbf{BS-SCHEDULE}}$. Since each interference is positive, and since $\sum_{t=1}^n T_{TOT}^{b_t} \leq n$, it follows that each station is scheduled in exactly one round.

Assume the base stations at the first round are $b_1, b_2, \cdots, b_k$, and in the second round are $b_{k+1}, \cdots, b_{2n}$. If $k \neq n$ then one of these rounds has more than n base stations. Assume, with no loss of generality, that $k > n$. This means that $\sum\left\{x_j + \frac{x_i}{n-1} \middle| j = 1, 2, \cdots, k, j \neq i\right\} > nS + \frac{nx_i}{n-1} > F$, for every $i = 1, 2, \cdots, k$, thus $\Psi_{u_i}^{b_i}(2) < 1$, a contradiction. Therefore $k = n$. The interference of each of the users in the first (second) round is $\log_2\left(1 + \frac{2F}{F+\sum_{i=1}^n x_i}\right)$ ($\log_2\left(1 + \frac{2F}{F+\sum_{i=n+1}^{2n} x_i}\right)$). So, $\sum_{i=1}^n x_i = \sum_{i=n+1}^{2n} x_i = F$, and all interferences are 1. □

When considering the minimisation version of the problem (to determine a scheduling with smallest number of rounds), we use [19], which shows that for all $\epsilon > 0$, approximating the chromatic number of a given graph $G = (V, E), |V| = n$ within $n^{1-\epsilon}$, is NP-hard. Since coloring G with n colors is trivial, this means that this result is rather strong. Using it we show that Problem **BS-SCHEDULE** is rather difficult to approximate, as follows:

Theorem 3. *For all $\epsilon > 0$, approximating within $n^{1-\epsilon}$ the minimal number of rounds required to solve Problem **BS-SCHEDULE** with n base stations is NP-hard.*

Sketch of Proof: Following the same reduction from **GCk**, as done in the proof of Theorem 1, it is clear that the instance of **BS-SCHEDULE** can be scheduled in k rounds if and only if the given graph can be colored with k colors. Therefore the existence of an algorithm with approximation ratio $n^{(1-\epsilon)}$ for **BS-SCHEDULE** will imply the existence of an algorithm with the same approximation ratio for **GCk**. □

C. Sufficient condition for Problem *BS-SCHEDULE*

Since, as we have shown above, Problem **BS-SCHEDULE** is NP-complete and NP-hard to approximate, in the following we provide a sufficient condition that guarantees that the entire content is delivered before its lifetime, i.e., in Z subframes. Specifically, we can derive the following inequality from Eq. (4), which holds for any subframe i:

$$w_u(i)t_u(i) = \frac{1}{\sum_{k=1}^{d_b} \frac{\delta_{ki}}{t_k(i)}} \geq \frac{t_{\min}(i)}{d_b}; \tag{5}$$

where $t_{\min}(i) = \min_i\{t_u(i)\}$. If we now sum over the subframes in which the user is served within the time horizon Z, we obtain a bound for the volume of traffic V_u received by a user:

$$V_u = \tau \sum_{i=1}^{Z} x_{ib} w_u(i) t_u(i) \geq \tau \sum_{i=1}^{Z} x_{ib} \frac{t_{\min}(i)}{d_b} \geq \tau \frac{Z_b}{d_b} t^*_{\min}, \tag{6}$$

where $t^*_{\min} = \min_i\{t_{\min}(i)\} = \min_{u,i}\{t_u(i)\}$ is the minimum instantaneous rate allotted to any user, and $Z_b = \sum_{i=1}^{Z} x_{ib}$ is the number of subframes in which base station b is active. Since it is sufficient to guarantee that $V_u \geq L_c$ to guarantee that user u received the content on time, we obtain the following sufficient condition for the doability of the scheduling:

$$t^*_{\min} \geq \frac{L_c d_b}{Z_b \tau}. \tag{7}$$

In conclusion, inverting the Shannon formula from the minimum value for $t^*_{\min}$ given in Eq. (7), we deduce that it is sufficient to schedule a base station when all its scheduled transmission entities have at least the following SINR:

$$\text{SINR} \geq 2^{\frac{d_b L_c}{\tau Z_b B_T}} - 1 \doteq \texttt{TH}. \tag{8}$$

Note that the above equation defines an SINR threshold TH that depends, in addition to some constants, on the number of subframes Z_b in which base station b is allowed to transmit. Next, we derive a lower bound on Z_b for which the inter-BS scheduling guarantees that d_b content injections are doable within the deadline.

D. Lower bound for Z_b

The throughput of a base station b can be bounded by the following expression:

$$\frac{d_b L_c}{\tau \sum_{o=1}^{d_b} \sum_{i=1}^{Z} w_o(i) x_{ib}} = \frac{d_b L_c}{\tau Z_b} \leq R_{\text{MAX}}, \tag{9}$$

where R_{MAX} is the maximum transmission rate permitted in the network (e.g., $R_{\text{MAX}} = 93.24$ Mbps in an FDD LTE-A network using 20 MHz bandwidth). Therefore, there is

a lower bound for Z_b below which the content injection of d_b contents cannot be guaranteed:

$$Z_b \geq \frac{d_b L_c}{\tau R_{\text{MAX}}}, \quad \forall b \in B. \tag{10}$$

Since we aim to minimise the total transmission time, which is given by $\Theta = \tau \sum_{b \in B} Z_b$, it is reasonable to assume that an ICIC algorithm that approximates the solution of Problem **BS-SCHEDULE** will be able to complete the injection of d_b contents at base station b in a number of subframes that is close to the bound given above, i.e., $Z_b = \frac{d_b L_c}{\tau R_{\text{MAX}}}$. With this approximation, we can express the threshold TH in (8) as a function that does not depend on Z_b.

The above provides a sufficient condition to guarantee that d_b contents are delivered within their lifetime; in particular, we have found a threshold TH for the SINR of users to be scheduled. In Section IV, we leverage this result for the design of our ICIC algorithm.

E. Maximum number of contents

Before describing our heuristic for Problem **BS-SCHEDULE** in Section IV, we compute the maximum number of contents that base stations can handle. This result will be useful in Section V to evaluate eICIC schemes. To achieve our goal, we assume that all the base stations have, at least on average, the same number of contents to inject in interval T_c.

If all base stations have the same number of contents to inject, we can derive an upper bound for Z_b. The total number of subframes used by all base stations cannot exceed $\sum_{b \in B} Z_b = N Z_b$. If Z is the total number of subframes in which the content is valid, we have that $N Z_b \leq Z$ and thus, we can derive an upper bound as $Z_b \leq \frac{Z}{N}, \quad \forall b \in B$, which, jointly with (10), yields the following range for Z_b:

$$\frac{d_b L_c}{\tau R_{\text{MAX}}} \leq Z_b \leq \frac{Z}{N}, \quad \forall b \in B. \tag{11}$$

From the analysis above, we can then compute the maximum number of injectable contents that can be handled by a base station while guaranteeing that all contents are served within the deadline $T_c = \tau Z$. In particular, from (11), it is clear that the Z_b range is not empty under the following condition, which gives an upper bound for d_b:

$$d_b \leq d_b^* = \frac{\tau Z R_{\text{MAX}}}{L_c N}. \tag{12}$$

IV. Base Station Blanking Algorithm

In this section, we propose BSB (Base Stations Blanking), an algorithm to approximate the optimal solution of Problem **BS-SCHEDULE** formulated in Section III. BSB relies on the sufficient condition given by Eq. (8). Following such condition, BSB aims to find an optimal ABSF pattern, i.e., an allocation of base stations to LTE-A subframes, in which the interference is limited, so to guarantees a minimum SINR (and hence a minimum rate) to any mobile device that might receive a content from the base station on any portion of radio resources selected for that user. Note that our algorithm is meant to allocate ABSF patterns, and does not impose any *user scheduling* policy. However, since we aim to a minimum guaranteed rate for any content injection, any simple policy like Round Robin over the entire available spectrum can be used at the base station.

A schematic view of BSB is reported here. BSB runs in a LTE-A network, and requires the presence of a central controller, namely the Base Stations Coordinator (BSC), which could be run on the Mobility Management Entity) [20]. Our algorithm requires cooperation between the BSC and base stations, which can be implemented over the standard $X2$ interface [5]. The main role of BSC is to collect SINR statistics from the base stations, run BSB, and announce ABSF patterns to the base stations, as detailed in what follows:

BSB Algorithm

The BSC collects user statistics, puts all active base stations in a candidate set, and checks whether the resulting SINR for each user is above the SINR threshold TH.

If at least one user does not reach the SINR threshold:

- compute the most interfering base station b^*
- remove b^* from the candidate set,
- check the SINR of all users of the remaining base stations.

Repeat the check and remove base stations from the candidate set until all remaining users meet the SINR constraint. The resulting set of base stations is scheduled in the first subframe and inserted in a *priority*-1 list. In general, at each subframe, scheduled base stations are added to the *priority*-k list, where k is the current number of subframes enabled for a base station to transmit. All other base stations go to a *priority*-0 list.

For each successive subframe, populate the candidate set with the *priority*-0 list and repeat the operation described for the first subframe until the SINR constraint is met.

Then, for $k = 1, 2, \ldots$, in increasing order:

- add to the candidate set all base stations in the *priority*-k list,
- within *priority*-k list, remove base stations causing SINR below TH.

The algorithm stops when the priority list is empty.

The BSC issues the resulting ABSF pattern to each base station via the $X2$ interface.

In the above description, the interference caused by a base station is computed as the aggregate sum of interferences caused towards all users of all other bees station in the candidate set. The threshold TH is computed based on d_b and the lowest possible value for Z_b, given by (10). The scheduling pattern computed with BSB can range between

1 and N subframes. However, since the standard specifies that ABSF patterns should be issued every 40 subframes, the BSB pattern is repeated in order to cover a multiple of 40 subframes. The obtained sequence of scheduling patterns represents the ABSF pattern according to [5].

For each subframe, the algorithm starts by selecting the full set of base stations that have not been scheduled in previously allotted subframes. The rationale behind this choice is twofold: (i) the aggregate interference caused by a base station grows with the size of the candidate set, and thus the importance of the interference generated by a base station is more properly quantified by the full candidate set; (ii) existing ICIC algorithms suggest to mitigate interference by preventing the transmission of a few base stations, beginning with the most interfering one [6], [21], [22]. BSB complexity is dominated by the number of base stations, as stated in Theorem 4.

Theorem 4. *The complexity of BSB is* $O(U \cdot N^3)$*, where* $U = \max_{b \in B}\{U_b\}$*, and* $N = |B|$.

Sketch of Proof: The BSB algorithm runs in at most N rounds, corresponding to N allocated subframes: in the worst case, exactly one base station is allocated in exactly one subframe. At subframe $q = 1, 2, ..., N$, there are at most q priority lists. In the worst case, the *priority*-0 list contains $N - q + 1$ base stations and each other priority list contains 1 base station. Evaluating the SINR for all users of base stations in *priority*-0 requires checking all reconfigurations with $N - q + 1$, $N - q$, ..., 1 base stations in the candidate set. Checking the possibility to add to the resulting scheduled set any base station in the other priority lists is at most involving $N - q + 2$ base stations for considering *priority*-1 list, $N - q + 3$ for *priority*-2 and so on until N base stations for the last priority list. Overall, the cost per subframe is $O(U \cdot N^2)$. Therefore, in the worst case, in which N subframes are needed, the complexity is $O(U \cdot N^3)$. □

The study of complexity of our base station scheduling solution reveals that we achieve not only a polynomial algorithm, but also that our scheduler has a very low complexity, which depends on the third power of the number of base stations to coordinate and is a linear function of the number of users in the most populated cell.

V. Performance Evaluation

Here we study the impact of BSB on the performance of D2D-assisted content distribution. We benchmark the performance achieved with BSB against the one achieved under different frequency reuse schemes (in particular frequency reuse 1, 3, and 5), and against a state-of-the-art dynamic resource allocation scheme proposed for ICIC in LTE-like networks [21]. We refer to the latter as ECE. Differently from BSB, ECE assigns resource blocks rather than subframes, thus implementing a scheme for *soft fractional frequency reuse* [23].

As concerns the system parameters adopted in our performance evaluation, we use FDD LTE-A frame specifications,

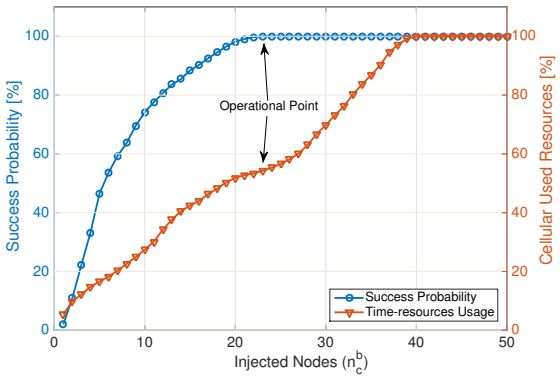


Fig. 2: Probability of delivering successfully a content for different values of injection nodes in a scenario with 5 BSs and 750 users. Cellular resources used for the injection phase are also show (averaged over the content lifetime $T_c = 100$ s).

with 20 MHz bandwidth distributed over 100 frequency chunks, resulting in 100 resource blocks per time slot, i.e., 200 resource blocks per LTE-A subframe [24]. Transmission power is fixed to 40 W, antenna gain and path loss are computed according to [25], and the spectral noise density is $3.98 \cdot 10^{-21}$ W/Hz for all nodes [26]. Modulations and coding schemes are selected according to the SINR thresholds reported in [24], while the ratio between received power (or interfering signal) and noise, for each user in the network, is computed as for Rayleigh fading, with average computed from transmission power and path loss.

D2D communications occur *outband* (i.e., on a channel not interfering with any of the base stations), and mobile devices exchange data when their distance is 30 m or less. A new content update is available synchronously for any content c, every $T_c = 100$ s. Each mobile device is interested in at most one content (whose size is 8 Mbits). Users get interested in a content at different points in time, according to a truncated normal distribution function having μ as mean value for the interesting rate. For the sake of completeness, we have also conducted simulations to evaluate the case with an infinitive μ corresponding to a content subscription case where base stations inject beforehand the contents through multicast transmissions. Background traffic is also generated in some of our experiments, consisting in uniformly random file requests, with file size 8 Mbits. Background requests are processes as new contents for single users.

As concerns the mobility of users, we use a Random Waypoint mobility model over a regular grid [27]. Mobile users are initially assigned uniformly over the area, then they choose uniformly random distributed destinations (waypoints P_u), and speeds (V_n) uniformly distributed in range $[1, 2]$ m/s, independently of past and present speed values. Then, the mobile user travels toward the newly chosen destination at constant speed V_n. Upon arrival to destination P_u, the mobile user randomly chooses another destination and speed. Note that, at the considered low speed, the resulting contact time is long (several seconds). Therefore, we assume that complete file transfers are possible during the contact time. This results in a particular contact rate λ.

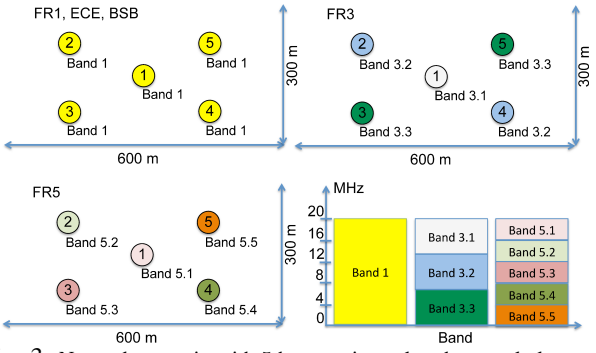


Fig. 3: Network scenario with 5 base stations placed at regularly spaced positions, and 750 users (not shown in the figure) randomly dropped into an area of 600 m × 300 m. For each tested scheme, the figure reports the BS baseband bandwidth.

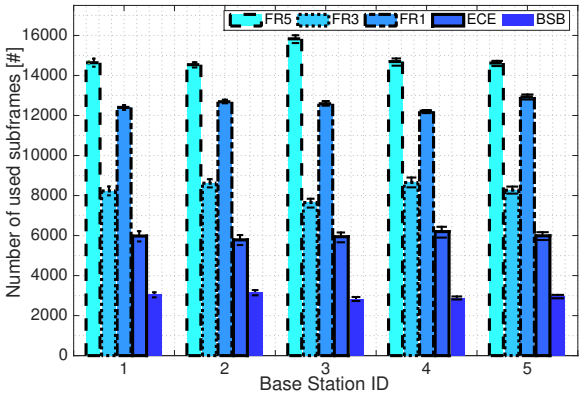


Fig. 4: Content transmission time with 5 base stations, 750 users, interesting rate $\mu = 200$, and no background traffic.

All experiments refers to a dense LTE-A deployment, with 5 overlapping cells, and several hundreds of mobile users. Each experiment includes 50 new content updates for each content, with period 100 s (i.e., the experiment simulates 5 000 s), and is repeated 20 times. Average and 95% confidence intervals are reported in the figures. When using BSB, a specific ABSF pattern is issued every 40 subframes, which perfectly complies with 3GPP standard specifications [5].

A. Injection Phase: empirical validation

The injection phase plays a key-role in driving the content dissemination process to extremely efficient conditions and we have defined Problem INJECTION to compute the optimal number of injections per content. Here, we explore the importance of such optimisation by evaluating the performance achieved by evaluating the impact of the number of injected replicas. A wrong decision on the number of injected contents brings the system to a faulty performance efficiency. Therefore, we show how that decision impacts on the system performance in terms of probability of successfully content delivering as well as the portion of offloaded base station time-resources. Fig. 2 shows the probability to receive the entire content at the end of the content lifetime (T_c) trying out different values of injections (accounting for both LTE-A and D2D transmissions), when 150 users get interested in a content. We applied on top of injection number decision our algorithm BSB to efficiently schedule the BS time-resource to intended users requiring the content. Intuitively, the more injected nodes, the more the probability that a D2D content exchange occurs, the more users will get the entire content at the end of the content lifetime. In addition, we show the portion of time-resources saved by LTE-A base stations during the content dissemination process. Whenever more than 37 injected contents are required, the system results in a critical time-resources shortage. Also, the graph highlights the operational point of our algorithm derived from Problem INJECTION, as explained in Section II-B: with 23 content transmissions the system successfully delivers the content to all interested users while significantly limiting the time-resources usage (up to 54%) per base station. Indeed, our approach uses the minimum amount of resources needed while achieving the highest achievable success probability, thus establishing an excellent trade off between performance and resource utilisation.

B. Base station transmission time and delivery success probability

We simulate the network depicted in Fig. 3, with 5 base stations and 750 mobile devices. Therefore, in the described results, scheme FR1 represents the case with no ICIC, while FR5 guarantees no interference. Our objective is to analyse in details how scheduling procedures affect the base stations offloading throughout the whole content distribution process.

For the first set of results, we evaluate the effective amount of time-resources saved by each base station while applying compared scheduling approaches. Fig. 4 shows the per-base station average transmission time in terms of transmission slots lasting 1 ms as per LTE specifications, expressed in terms of used subframes, when 200 users get interested and require a content each second. No background traffic was injected during the experiment. For the case of ECE, in which resource blocks are allotted rather than subframes, we count the total number of used resource blocks, and normalise that number with respect to the number of resource blocks per subframe. BSB clearly outperforms ECE and FR3 by a factor ~ 3, and up to ~ 5 for the case of FR5. Note that, for a fair comparison to BSB and ECE, frequency reuse schemes simulated in the experiment allocate only $1/n$, $n \in \{1, 3, 5\}$ of the available bandwidth to each base station. With the data reported in the figure, it is clear that BSB improves the results of FRn, $n \in \{3, 5\}$, by a factor $\sim n$. Therefore, we could extrapolate that modifying FR3 and FR5 schemes using n times the bandwidth used by BSB would achieve similar results as BSB. Indeed, we have validated such an intuitive result by running an experiment in which all base stations always use the entire 20 MHz bandwidth. Results show negligible performance differences (below 1%) between the schemes. However, we remark that BSB would require $1/n$ of the frequencies needed by frequency reuse schemes.

Fig. 5 shows a cumulative distribution function for the successfully delivered portion of each content, under the

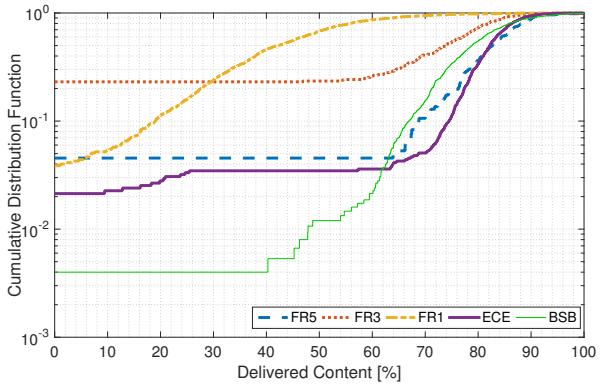


Fig. 5: Success probability of content distribution with 5 base stations, 750 users, interesting rate $\mu = 200$, and no background traffic.

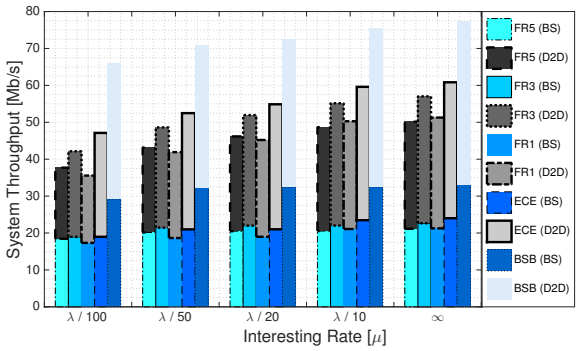


Fig. 6: System throughput for different interesting rate μ when 5 base stations and 750 users are placed, a meeting rate $\lambda = 2000$ pair/contacts/second is considered, and different scheduling procedures are applied. The last case $\mu = \infty$ provides system performance for content subscription scenario with multicast transmission.

tested schemes. BSB exhibits an impressive behaviour compared with the other solutions. Only in 0.3% of the cases BSB fails to start the content delivery, whereas in almost the 99% of the cases BSB delivers at least 50% of the content. All the other considered schemes show a much higher probability to fail to start the delivery (2% to 22% of cases). In general, FR1 and FR3 perform much worst than the others, as static frequency reuse mechanisms are not able to dynamically follow the network changes resulting in a very high probability to deliver only a few chunks of the content, whereas FR5 and ECE and BSB manage to reduce interference sensibly and so guarantee high delivery rates, although, as shown in Fig. 4, BSB operates the injection much faster.

C. Throughput of base stations and of D2D exchanges

Fig. 6 shows the aggregate system throughput, expressed in terms of bits delivered per second via injection (BS) and dissemination (D2D), for a 5 base stations scenario where 750 users are placed. Each of proposed scheduling approaches is studied for a particular set of interesting rates μ (expressed in terms of interested users per second), as function of meeting rate $\lambda = 2000$ pair/contact/seconds[2]. Interestingly, we show the amount of system throughput due to the base station transmissions (both for content injection and for other kinds of traffic) while, on top of the graph, the throughput due to the dissemination phase. We want to point out two main aspects. On the one hand, the faster users get interested in the content, the lower the base station load, the more free time-resources are assigned to other kinds of traffic, the higher the D2D communication throughput. The rationale behind is pretty intuitive. When users express their interest for a content at the beginning of the period T_C, base stations can promptly inject them the content, leaving more time to the users to spread the content. In this way, much more contacts occur in the network, much more data is exchanged through D2D communication (as also confirmed in the Appendix). The extreme case is modelled when $\mu = \infty$, i.e., when all users get interested at the beginning of each period T_C. On the other hand, BSB shows an incremental gain w.r.t. the other presented approaches. For the first set of interesting rates μ, FR5 and FR1 are unable to complete the injection phase, as several transmissions are required (e.g., 97 injections for $\mu = \lambda/100$ and 65 injections for $\mu = \lambda/50$) leaving no room for other traffic. When the required injections decrease to 26 for $\mu = \lambda/10$, all scheduling schemes exhibit the same base station throughput except BSB due to the ability of scheduling other traffic. This confirms that an optimal offloading base stations procedure requires a very fast injection phase, which must be properly designed through a convenient scheduling scheme.

[2]Please note that if not differently stated, for the sake of simplicity, we assume the same meeting rate $\lambda_c = \lambda$ as well as the same interesting rate $\mu_c = \mu, \forall c \in \mathbb{C}$

D. The impact of realistic mobility models and content sizes

While the homogeneity assumption helps in modelling a closed-form solution (as expressed in the Appendix) and properly designing a powerful solution, we show here that applying heterogeneous node distributions will not negatively impact on the system performance which is in line with the generic results provided in [10] when a heterogeneous distribution is applied. Therefore, we have introduced a heterogeneous mobility model generating user contacts as follows. For any given user pair (x, y), we specify a pairwise inter-contact time $t_{x,y}$ exponentially distributed with rate $\beta_{x,y}$. Contact rates, $\beta_{x,y}$, are drawn from a Pareto distribution with mean λ (determining the average frequency of the user contacts) and standard deviation σ (indicating the heterogeneity level), as suggested in [28]. With a low heterogeneity level σ, users get in touch by following the homogeneous mobility model, such as the random waypoint model.

Fig. 7 shows the system throughput considering different levels of user contact heterogeneity σ. We have expressed σ as function of the user contact rate. The lowest heterogeneity level, e.g., $\sigma = \lambda/1000$, envisages that all possible pairs of nodes tend to have the same probability to get in touch, i.e., $\beta_{x,y} \simeq \lambda$. Notably, BSB outperforms the other approaches also under heterogeneous mobility conditions. However, although not shown in the figure, we have observed that the number of injection n_c^b computed by solving Problem INJECTION may not be optimal under non homogeneous mobility hypotheses.

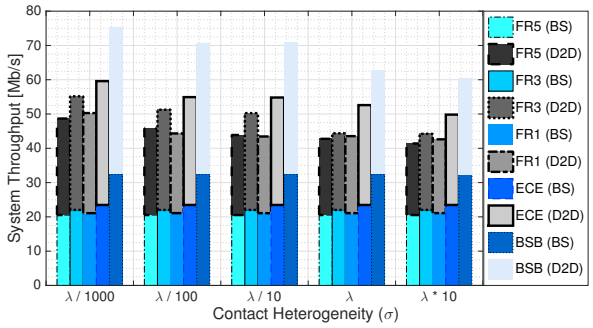


Fig. 7: System throughput for different heterogeneity levels σ when 5 base stations and 750 users are placed, a meeting rate $\lambda = 2000$ pair/contacts/second and an interesting rate $\mu = 200$ are considered, and different scheduling procedures are applied.

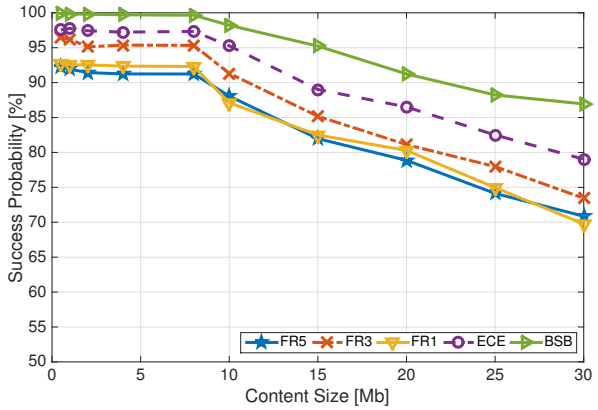


Fig. 8: Success probability of content distribution with 5 base stations, 750 users, considering different content sizes.

Nevertheless, we have observed that the system keeps a high success probability (not lower than 97%) even under high heterogeneous mobility.

Additionally, in Fig. 8, we have depicted the success probability of content delivery by considering different content sizes and fixing a content lifetime $T_c = 100$s. Intuitively, when the content size increases, the system run out of resources, which yields to unsuccessful content delivery. However, BSB shows much better performance with respect to other solutions, even when bigger contents are generated and transmitted. Note that the number of content updates successfully delivered per second, i.e., the content throughput counting only entirely delivered content updates, is computed by multiplying the content arrival rate and the success probability.

E. Impact of background traffic

To show the efficacy of BSB in more generic traffic scenarios, in addition to periodic content issues, we next simulate background file requests uniformly distributed over time at different request rates. Note that (12) expresses the maximum number of contents that can be distributed with guaranteed maximum transmission time. That expression can be also interpreted as the maximum cell load that can be handled by a base station while guaranteeing that contents will be delivered within the deadline (with each content unit used for d_b^* corresponding to an offered load $L_c/(\tau Z)$). Therefore, we expect that BSB is able to handle a background traffic equivalent to, at most,

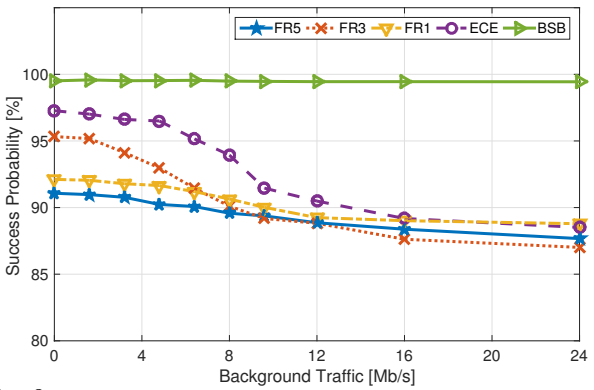


Fig. 9: Success probability of content distribution with 5 base stations, 750 users, and background traffic.

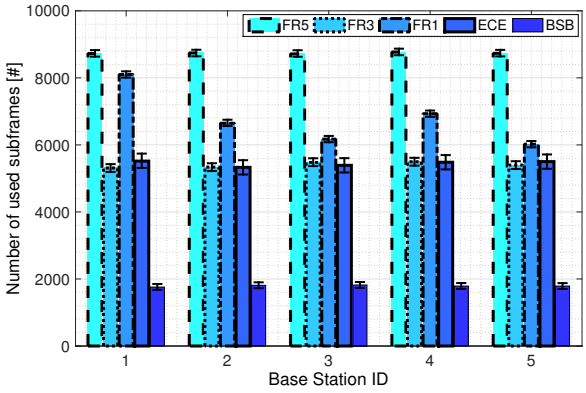


Fig. 10: Content transmission time with 5 base stations, 750 users, and no background traffic, for a content subscribe scenario with interesting rate $\mu = \infty$.

$(d_b^* - d_b) \cdot L_c/(\tau Z)$ bps. With 8-Mbit background files, $d_b = 20$, $L_c = 8$ Mbits for any content c, $\tau Z = 100$ s, and 5 base stations, the maximum background traffic is 2.125 requests per second.

In Fig. 9, we show the impact of background traffic on the probability to complete the content distribution, for various background loads. Similarly to the case in which no background traffic is injected, BSB outperforms other schemes. Interestingly, BSB is more robust to background traffic than other schemes, as shown by the fact that content delivery probability under BSB is barely affected by the background traffic. The performance of BSB starts degrading only when the offered background exceeds 3 file requests per second, which is well above 2.125 requests per second, i.e., the maximum value that guarantees the doability of content transmission within the deadline, according to (12). In contrast, frequency reuse schemes and ECE are seriously impaired by the background traffic as soon as the offered load reaches as low as 1 background file request per second.

F. Content subscription with multicast transmission

We finally assess the effect of our solution in a particular content subscription scenario in which users initially subscribe new content updates (e.g., $\mu = \infty$) and get refresh replicas every time the content is issued (every T_C seconds). This implies that base stations can easily inject the content into the network through a single multicast

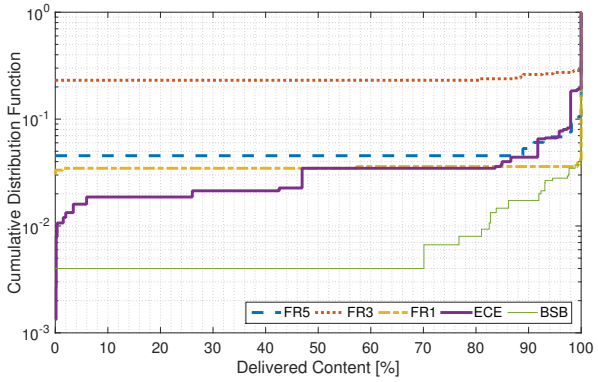


Fig. 11: Success probability of content distribution with 5 base stations, 750 users, and no background traffic, for a content subscribe scenario with interesting rate $\mu = \infty$.

transmission. The transmission rate r_c^b is properly chosen to cover as many content subscribed users as possible (see Section II-B).

Similarly to Fig. 4, Fig. 10 shows the per-base station average transmission time, expressed in terms of used subframes, when the simultaneous update of 5 contents is periodically distributed in the network. Even in this case, BSB outperforms all other schemes and uses a number of subframes very close to the lower edge of the interval predicted in (11). Moreover, BSB outperforms FR3 and ECE by a factor ~ 3, and more than ~ 4 for the cases of FR5 and FR1. In addition, Fig. 11 reports the cumulative distribution function of the portion of delivered contents, under the tested schemes. For this performance metric, we count the number of contents that were correctly and entirely delivered to the subscribers, and normalise to the number of subscribers. BSB emerges as the scheme that guarantees the highest content delivery probabilities, resulting in 97.24% of delivered contents, on average. Noticeably, FR1, FR3, FR5 and ECE perform *much worst* than BSB. This result points out that both static frequency planning schemes and classic resource allocation schemes are not able to cope with the interference generated in dense environments. Moreover, FR3 achieves by far the worst results. Therefore, comparing FR1 (all base stations use the same wide bandwidth) and FR3 (at most two base stations share the same bandwidth, which is $1/3$ of the one used under FR1), we argue that the interference generated by few neighbors in a dense scenario is much less important than the available bandwidth. As a consequence, spectral efficiency over wide frequency bands is key to boost network performances, while bandwidth fragmentation due to frequency planning is undesirable.

VI. Related work

Our proposal can be classified as semi-distributed [23], since it relies on a central entity that coordinates scheduling resources (ABSF patterns), while each base station remains responsible for scheduling its users. In this section, we comment on other semi-centralized ICIC schemes that have been proposed in the literature.

The authors of [21], [29] design a heuristic to allocate resource blocks when adjacent cells interfere with each other. Their approach allows the reuse of resource blocks in cell centers, while users at the cell edge, which suffer higher interference, cannot be allocated specific resource blocks, as figured out by the proposed heuristic. However, differently from our proposal, that work only considers avoiding the interference of the two most interfering base stations. As a results, we have shown in Section V that their approach is not suitable for dense networks.

Similarly, the proposal in [30] assigns resource blocks via a central entity while [31] solves the problem in a distributed manner. However, they allocate resources not only to base stations but also to users, based on backlog and channel conditions. Therefore, unlike our proposal, it results in intractable complexity.

The author of [22] uses graph theory to model network interference. That work proposes a graph coloring technique to cope with interference coordination, based on two interference graphs: one *outer* graph using global per-user interference information, and an *inner* graph using local information, available at the base station, and global constraints derived from the global graph. To reduce complexity, [22] uses genetic algorithms to seek a suboptimal resource block allocation. However, differently from BSB, that approach does not allow to use a generic user scheduler, since users are allocated according to the *inner* graph coloring problem.

In our previous work on ICIC [6], we have investigated on the optimisation of ABSF pattern allocations in a fully saturated network. However, that work does not account for content deadlines, and therefore the choice of the SINR threshold to be used in a real network was not investigated. Moreover, the resource allocation protocol proposed in [6] is far from being throughput maximal, since it is designed for achieving fairness among base stations, and so it does not guarantee the delivery of contents within a given deadline.

None of the above works tackle the impact of interference in dense scenarios, in presence of offloading traffic strategies.

VII. Conclusions

We have analysed the content dissemination process in cellular networks by shedding the light on the potential role of D2D communications and of the base station interference coordination problem. Specifically, we are the fist to analyse the *injection phase*, that is a key component of the dissemination process, yet it has been so far neglected. We have cast such a content injection problem into an optimisation problem aiming at finding the optimal number of transmissions to maximise the content replica delivery. Notably, we have proven that the injection phase critically affects the opportunistic D2D content exchange. Based on this insight, we have formulated a minimisation problem on the time required to inject contents, given the characteristics of the content dissemination and the inter-cell interference experienced by users. We have proven that the problem is NP-Complete and NP-Hard to approximate, so that scalability problems can arise in very dense cellular scenarios.

Hence, we have proposed BSB, an eICIC algorithm for LTE-A networks that efficiently approximates the solution.

Our results show that BSB substantially outperforms classical intercell interference approaches and achieves performance figures better than what achievable with (soft fractional) frequency reuse schemes. Moreover, BSB boosts the D2D opportunistic communication performance by making the injection phase quasi-ideal, i.e., by minimising the time needed to inject content replicas in the network.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their inspiring and insightful feedbacks.

REFERENCES

[1] J. Whitbeck, M. Amorim, Y. Lopez, J. Leguay, and V. Conan, "Relieving the wireless infrastructure: When opportunistic networks meet guaranteed delays," in *Proceedings of IEEE WoWMoM*, 2011.

[2] H. Cai, I. Koprulu, and N. Shroff, "Exploiting double opportunities for latency-constrained content propagation in wireless networks," *Networking, IEEE/ACM Transactions on*, pp. 1025–1037, Apr 2016.

[3] G. Zyba, G. M. Voelker, S. Ioannidis, and C. Diot, "Dissemination in opportunistic mobile ad-hoc networks: the power of the crowd," in *Proceedings of IEEE INFOCOM*, 2011.

[4] A. Asadi and V. Mancuso, "On the Compound Impact of Opportunistic Scheduling and D2D Communications in Cellular Networks," in *Proceedings of ACM MSWiM'13*, Barcelona, Spain, Nov. 2013.

[5] Third Generation Partnership Project (3GPP), "Evolved Universal Terrestrial Radio Access Network (E-UTRAN); X2 application protocol (X2AP), 3rd ed." 3GPP, technical Specification Group Radio Access Network.

[6] V. Sciancalepore, V. Mancuso, and A. Banchs, "BASICS: Scheduling Base Stations to Mitigate Interferences in Cellular Networks," in *Proceedings of 14th IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM 2013)*, 2013.

[7] J. LeBrun, C.-N. Chuah, D. Ghosal, and M. Zhang, "Knowledge-based opportunistic forwarding in vehicular wireless ad hoc networks," in *Vehicular Technology Conference (VTC) 2005-Spring*, vol. 4, May 2005, pp. 2289–2293.

[8] A. Asadi, Q. Wang, and V. Mancuso, "A survey on device-to-device communication in cellular networks," *Communications Surveys Tutorials, IEEE*, vol. 16, no. 4, pp. 1801–1819, 2014.

[9] K. Cho, M. Lee, K. Park, T. Kwon, Y. Choi, and S. Pack, "Wave: Popularity-based and collaborative in-network caching for content-oriented networks," in *Computer Communications Workshops (INFOCOM WKSHPS 2012)*, March 2012, pp. 316–321.

[10] V. Sciancalepore, D. Giustiniano, A. Banchs, and A. Hossmann-Picu, "Offloading cellular traffic through opportunistic communications: Analysis and optimization," *IEEE Journal on Selected Areas in Communications*, vol. 34, pp. 122–137, Jan 2016.

[11] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, pp. 83–97.

[12] V. Sciancalepore, V. Mancuso, A. Banchs, S. Zaks, and A. Capone, "Interference coordination strategies for content update dissemination in LTE-A," in *IEEE INFOCOM*, 2014.

[13] Third Generation Partnership Project (3GPP), "Multimedia Broadcast/Multicast Service (MBMS)," 3GPP TS 36.246 v 11.1.0, Dec. 2012.

[14] V. Sciancalepore, I. Filippini, V. Mancuso, A. Capone, and A. Banchs, "A semi-distributed mechanism for inter-cell interference coordination exploiting the absf paradigm," in *IEEE SECON*, 2015.

[15] M. Cierny, H. Wang, R. Wichman, Z. Ding, and C. Wijting, "On number of almost blank subframes in heterogeneous cellular networks," *Wireless Communications, IEEE Transactions on*, vol. 12, no. 10, pp. 5061–5073, October 2013.

[16] O. Goussevskaia, R. Wattenhofer, M. M. Halldrsson, and E. Welzl, "Capacity of arbitrary wireless networks," in *INFOCOM'09*, 2009.

[17] A. Capone, G. Carello, I. Filippini, S. Gualandi, and F. Malucelli, "Routing, scheduling and channel assignment in wireless mesh networks: Optimization models and algorithms," *Ad Hoc Networks*, vol. 8, no. 6, pp. 545–563, Aug. 2010.

[18] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*. New York, NY, USA: W. H. Freeman & Co., 1979.

[19] D. Zuckerman, "Linear degree extractors and the inapproximability of max clique and chromatic number," *Theory of Computing*, vol. 3, no. 6, pp. 103–128, 2007.

[20] Third Generation Partnership Project (3GPP), "Mobility Management Entity (MME) and Serving GPRS Support Node (SGSN) related interfaces," 3GPP TS 29.272 v 11.5.0, Dec. 2012.

[21] M. Rahman and H. Yanikomeroglu, "Enhancing cell-edge performance: a downlink dynamic interference avoidance scheme with inter-cell coordination," *IEEE Transactions on Wireless Communications*, vol. 9, no. 4, pp. 1414–1425, 2010.

[22] M. Necker, "A Novel Algorithm for Distributed Dynamic Interference Coordination in Cellular Networks," in *Proceedings of KiVS*, 2011.

[23] A. Hamza, S. Khalifa, H. Hamza, and K. Elsayed, "A Survey on Inter-Cell Interference Coordination Techniques in OFDMA-Based Cellular Networks," *IEEE Communications Surveys Tutorials*, vol. PP, no. 99, pp. 1–29, 2013.

[24] Third Generation Partnership Project (3GPP), "Evolved Universal Terrestrial Radio Access (E-UTRA), Physical Channels and Modulation," 3GPP TS 36.211 v 10.6.0, Dec. 2012.

[25] P. Kyösti, J. Meinilä, L. Hentilä, X. Zhao, T. Jämsä, C. Schneider, M. Narandzić, M. Milojević, A. Hong, J. Ylitalo, V.-M. Holappa, M. Alatossava, R. Bultitude, Y. de Jong, and T. Rautiainen, "WINNER II Channel Models," EC FP6, Tech. Rep., Sep. 2007. [Online]. Available: http://www.ist-winner.org/deliverables.html

[26] Third Generation Partnership Project (3GPP), "Physical Layer Measurements," 3GPP TS 36.214 v 11.1.0, Dec. 2012.

[27] J. Yoon, M. Liu, and B. Noble, "Random waypoint considered harmful," in *Proceedings of IEEE INFOCOM 2003*, pp. 1312–1321.

[28] T. Karagiannis, J.-Y. Le Boudec, and V. M., "Power law and exponential decay of intercontact times between mobile devices," *Mobile Computing, IEEE Transactions on*, vol. 9, no. 10, pp. 1377–1390, Oct 2010.

[29] M. Rahman and H. Yanikomeroglu, "Multicell Downlink OFDM Subchannel Allocations Using Dynamic Inter-cell Coordination," in *Proceedings of IEEE Global Telecommunications Conference (GLOBECOM 07)*, 2007, pp. 5220–5225.

[30] G. Li and H. Liu, "Downlink Radio Resource Allocation for Multi-Cell OFDMA System," *IEEE Transactions on Wireless Communications*, vol. 5, no. 12, pp. 3451–3459, 2006.

[31] J. Zheng, Y. Cai, Y. Liu, Y. Xu, B. Duan, and X. Shen, "Optimal power allocation and user scheduling in multicell networks: Base station cooperation using a game-theoretic approach," *Wireless Communications, IEEE Transactions on*, vol. 13, no. 12, pp. 6928–6942, Dec 2014.

APPENDIX: STOCHASTIC ANALYSIS OF THE PROCESS

The content dissemination process is described with the 2-dimensional Markov chain, as depicted in Fig 12. We define $S_j(t)$ as the total *number of content replica* distributed in the network at time t given j users interested in the content, regardless of the specific users carrying those replica. A homogeneous mobility model is assumed, users get in touch each other following an average inter-contact rate λ and get interested in a content according to an average rate μ. As soon as the homogeneous assumption is relaxed, the interesting average rate is replaced with a μ^t to account for different timeframe t as well as the arrival rate λ with $\beta_{x,y}$ to account for different pairs' behaviours, as shown in Section V-D. Therefore, transition rates depend on the j amount of users interested in the content as well as on the number of users which have already obtained the content. Finally, the number of users which have received the content directly from the base station is represented by value n_c^b (number of injected nodes). Varying the number of injected nodes n_c^b the Markov chain is slightly affected, considering as first column only those S_j states whose

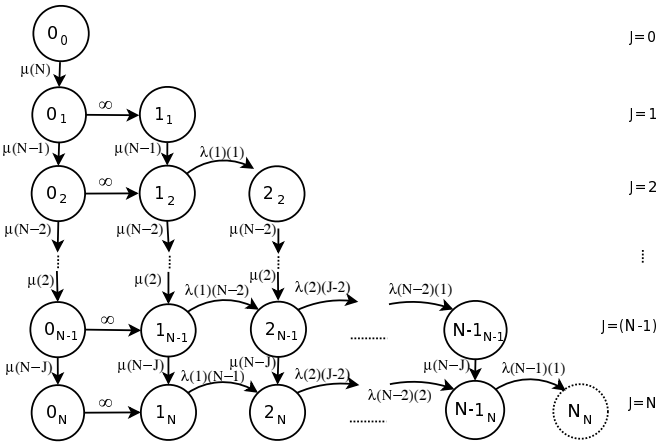


Fig. 12: Markov chain explaining the content dissemination phase performed by base station b for content c. Users get interested in the content with an average rate equal to μ, while getting in contact with a λ intercontact rate. The number of initial injected node is $n_c^b = 1$.

the number of users with the content is equal or greater than the number of injected nodes. Note that, since we assume that a contact between users allows to transfer entire content updates, state transitions in the chain do not absorb multiple contacts/messages being exchanged in different time instants.

Therefore, in order to solve the 2-dimensional Markov chain of Fig. 12, we write the forward Kolmogorov equations as follows (for all indexes $0 < i \leq j$, with $j \in \{1, \ldots, N\}$):

$$\begin{aligned} \dot{p}_{i_j}(t) = \quad & \lambda\,(i-1)(j-(i-1))\; p_{i-1_j}(t) \\ & +\mu\,(N-(j-1))\; p_{i_{j-1}}(t) \\ & -(\mu(N-j)+\lambda i(J-i))\; p_{i_j}(t). \end{aligned} \tag{13}$$

Please note that we enumerate with vector $\mathbb{S} = \{S_j(t)\}$, all the states of the Markov chain, starting counting by rows from the first state $\{S_1 = 1\}$, while we use vector $\vec{K}$ to represent the set of the unique indexes associated to every state $S(t)$, regardless the amount of users interested in the content. Please note that vectors $\mathbb{S}$ and $\vec{K}$ are time-independent. Indeed, $|\vec{K}| < |\mathbb{S}| = \frac{N(N+1)}{2}$.

Let $\vec{P}_J(t) = [p_{1_1}(t), p_{1_2}(t), p_{2_2}(t), ..., p_{S_j}(t), ..., p_{N_N}(t)]$ the set of probabilities at time t to be in each of the states $S_j \in \mathbb{S}$, while $\vec{P}(t) = [p_1(t), p_2(t), ..., p_x(t), ..., p_N(t)]$ the set of probabilities to have $S(t)$ users with the content at time t, where $p_x(t) = \sum_{j=1}^{x} p_{x_j}(t)$. To solve the set of Kolmogorov equations we can rewrite (13) as follows:

$$\vec{P}(t) = e^{-(\mathrm{L}\lambda+\mathrm{M}\mu)\mathrm{t}}\; \vec{C} \tag{14}$$

where C is a null vector of $(1 \times N)$ size, with only one non-zero value equal to 1 corresponding to the starting state index, while L and M are square matrices with $(N \times N)$ size. We define the structure of those matrices as follows:

$$L = \begin{vmatrix} 0 & & & \\ & l_J & \cdots & \\ & \vdots & & \vdots \\ & & \cdots & l_N \end{vmatrix},$$

where $l_J = \{l_{iz}\}$, and $l_{iz} = \begin{cases} (i)(J-i) & \text{if } i = z, \\ -(i)(J-i) & \text{if } i = z+1, \\ 0 & \text{otherwise;} \end{cases}$

$$M = \begin{vmatrix} m_1 & & & & \\ & m_J & \cdots & & \\ & \vdots & & \vdots & \\ & & \cdots & m_{N-1} & \\ & & & & 0 \end{vmatrix},$$

where $m_J = \{m_{iz}\}$, and $m_{iz} = \begin{cases} (N-J) & \text{if } i = z, \\ -(N-J) & \text{if } i = z+J, \\ 0 & \text{otherwise.} \end{cases}$

Note that for matrix indices we use the same order as reported in vector $\vec{K}$. This is important in order to have a general scheme to create those matrices. In matrix L we can identify $N-1$ square blocks l_J with $[J \times J]$ size. Considering Fig. 12 as a reference Markov chain, each of those blocks provides the transition rates of any single row of the Markov chain due to user meetings (except the first row). The longer the row, the larger the block, the more transition rate values. In matrix M we can identify $N-1$ non-singular blocks with $[J \times 2J]$ size, which take into account the transition rates due to new request from an interested user. Indeed, we obtain the average number of users with content at the end of the content lifetime T_c as follows

$$\mathbb{E}[S_j(T_c)] = \sum_{i=1}^{N} i\; p_i(T_c - d_{in}). \tag{15}$$

Neglecting the content transmission time with respect to the time between two users get interested in the content, the time elapsed after injecting n_b^i content replica is, on average, $d_{in} = \sum_{u=0}^{n_c^b-1} \frac{1}{\mu(N-u)}$. Indeed, using (14) in (15), we obtain

$$\mathbb{E}[S_j(T_c)] = \vec{V}\; e^{-(\mathrm{L}\lambda+\mathrm{M}\mu)(T_c-d_{in})}\; \vec{C} \tag{16}$$

where $\vec{V}$ is similar to $\vec{K}$, except that it includes only states with at least n_c^b users holding the content replica.

Vincenzo Sciancalepore (S'11-M'15) received his M.Sc. degree in Telecommunications Engineering and Telematics Engineering in 2011 and 2012, respectively, whereas in 2015, he received a double Ph.D. degree. From 2011 to 2015 he was Research Assistant at IMDEA Networks, focusing on inter-cell coordinated scheduling for LTE-Advanced networks and device-to-device communication. Currently, he is a Research Scientist at NEC Laboratories Europe in Heidelberg, focusing his activity on network virtualization and network slicing challenges.

Vincenzo Mancuso (S'02-M'06) is Research Associate Professor at IMDEA Networks, Madrid, Spain. He built his research experience by working with University of Palermo (Italy), from which he received a Ph.D. in Telecommunications in 2005, Rice University (Houston, TX, USA), and INRIA Sophia Antipolis (France). He leads the Opportunistic Architectures Lab at IMDEA Networks. His research activity focuses on analysis, design, and experimental evaluation of protocols and architectures for wireless networks.

Albert Banchs (M'04–SM'12) received the M.Sc. and Ph.D. degrees from the Polytechnic University of Catalonia (UPC-BarcelonaTech) in 1997 and 2002, respectively. He is currently Full Professor with the University Carlos III of Madrid (UC3M), and has a double affiliation as Deputy Director of the IMDEA Networks institute. Before joining UC3M, he was at ICSI Berkeley in 1997, at Telefonica I+D in 1998, and at NEC Europe Ltd. from 1998 to 2003. Prof. Banchs is Editor of IEEE Transactions on Wireless Communications and IEEE/ACM Transactions on Networking. His research interests include the performance evaluation and algorithm design in wireless and wired networks.

Shmuel Zaks holds BSc (cum laude), 1971, and MSc, 1972, in Mathematics from the Technion, Israel, and PhD, 1979, in Computer Science from the University of Illinois at Urbana, Champaign. He is a Professor with the Department of Computer Science, Technion, and is the incumbent of the Joan Callner-Miller Chair in Computer Science. He is the author of over 180 journal and conference publications. His research areas include distributed computing, graph and combinatorial algorithms, discrete mathematics, and theory of networking with an emphasis on optical networks.

Antonio Capone is Full Professor at Politecnico di Milano (Technical University of Milan), where he is the director of the ANTLab. His expertise is on networking and his main research activities include radio resource management in wireless networks, traffic management in software defined networks, network planning and optimization. On these topics he has published more than 200 peer-reviewed. He serves in the TPC of major conferences in networking, he is editor of IEEE Trans. on Mobile Computing, Computer Networks, and Computer Communications, and he was editor of ACM/IEEE Trans. on Networking from 2010 to 2014.