Heterogeneous Networks Relying on Full-Duplex Relays and Mobility-Aware Probabilistic Caching

Le Thanh Tan, *Member, IEEE,* Rose Qingyang Hu, *Senior Member, IEEE* and Lajos Hanzo, *Fellow, IEEE*

***Abstract*—Joint optimal resource allocation and probabilistic caching design is conceived for Device-to-Device (D2D) communications in a heterogeneous wireless network (HetNet) relying on full-duplex (FD) relays. In particular, popular contents can be cached at user devices as well as at relays that are located close to users. A user may request the contents of interest from another user via D2D communications and also from a nearby relay equipped with FD radios. If the requested contents are not found in the buffers of other users/relays within the coverage range, users may opt for connecting to the macro base station (MBS) via a relay by using FD communication. Furthermore, we propose a beneficial mobility-aware coded caching philosophy for D2D communications in the HetNet considered. Especially, we model the mobility pattern of users as discrete random jumps and exploit coded caching for improving the throughput attained. Subsequently, we develop mathematical models for analyzing the throughput in the presence of edge caching, where both the system-level co-channel interference and the FD self-interference are considered. We circumvent the high complexity of stochastic optimization by developing low-complexity optimization. Finally, numerical results are presented to illustrate the theoretical findings developed in the paper and quantify the throughput gains attained.**

***Index Terms*—User mobility; coded caching; particle swarm optimization; assignment algorithm; full-duplex communications.**

I. Introduction

The flawless multimedia applications supported by the next-generation networks will demand an unprecedented capacity expansion for supporting a high quality of service (QoS) in cellular networks. To tackle these challenges, advanced technologies are required for enhancing the capacity and QoS in cellular networks [1], [2]. Furthermore, recent studies reveal that different contents require different levels of priority [3], [4]. Explicitly, the most popular contents are requested by the majority of users, whilst the remaining large portion of the contents impose rather infrequent access demands [3].

In particular, 5G New Radio (NR) is an emerging wireless standard that will become the foundation of next generation mobile networks. With its early commercial deployments planned for 2018-2020, 5G technologies promise to deliver peak data rates of 10 Gbit/s. 5G will provide new architectures, new deployment and new services such as IoT ecosystem [1], [5]. As a representative technology of 5G, cloud radio access network (C-RAN) still faces many challenges, such as long latency due to the long distance from devices to the cloud, fronthaul/backhaul bandwidth limitation, high energy consumption and etc. As such, fog/edge-computing lately becomes a promising 5G technology, which puts a substantial amount of storage, communication and computation at the edge closer to the end users [6]. To make 5G NR a commercial reality, there should be an advantageous edge-computing architecture supporting new technologies, namely caching placement, device to device (D2D) communications and full-duplex (FD) communications in conjunction with the 3GPP's "New Radio" in the mmWave spectrum.

D2D communications are gaining increased attention in heterogeneous networks (HetNets) relying on the caching of popular contents at various local devices, as studied in [7]–[13]. At the time of writing, wireless cooperative caching constitutes one of the most widely studied paradigms, where both the local user terminals and the relay nodes can cooperatively store the multimedia contents [7]–[18]. In [7], relay nodes with caching capability are introduced to deliver the stored messages cooperatively with the BS, yielding a low delay. In [8], Ji *et al.* considered the combined effect of using both coding in the delivery phase for achieving "coded multicast gain", and spatial reuse as a benefit of local short-range D2D communication. In [18], Wang *et al.* carried out the theoretical analysis of the push-based content delivery methods, where the most popular contents are pushed through broadcasting to alleviate the cellular data bottleneck. A pair of scenarios were considered, where caching was performed either at a small BS (SBS), or directly at the user terminals, which communicate using D2D communications [19]. However, there is a paucity of studies in which both the relays and the users have a caching capacity simultaneously. To cooperatively perform caching at both user- and relay-levels, one must carefully select the contents to be cached at these levels and hence to enhance the probability of cache hits as well as the overall system performance.

Furthermore, FD radio technology has been considered as a promising next-generation technology conceived for successfully improving the capacity and reducing the transmission delay [20]–[27]. In theory, FD communications is capable of doubling the throughput of the BS-relay-user link, because a FD radio can transmit and receive data simultaneously on the same frequency band [20]. However, the major issue lies in the

Manuscript received February 27, 2018; revised January 07, 2019; accepted March 02, 2019. The research of L. T. Tan and R. Q. Hu was supported in part by National Science Foundation under grants NeTS 1423348 and EARS 1547312, in part by Natural Science Foundation of China under grant 61728104, and in part by Intel Corporation. L. Hanzo would like to acknowledge the financial support of the EPSRC projects EP/Noo4558/1, EP/PO34284/1, COALESCE, of the Royal Society's Global Challenges Research Fund Grant as well as of the European Research Council's Advanced Fellow Grant QuantCom. The editor coordinating the review of this paper and approving it for publication is D. Niyato.

L. T. Tan and R. Q. Hu are with the Department of Electrical and Computer Engineering, Utah State University, Logan, Utah 84322-4120, USA. Emails: {tan.le, rose.hu}@usu.edu.

L. Hanzo is with the School of Electronics and Computer Science, University of Southampton, UK. Email: lh@ecs.soton.ac.uk.

interference footprint of FD links, which is generally larger than that of its half-duplex (HD) counterpart. In particular, having many FD links in the network may impose severe interference on the other nodes in addition to the strong self-interference at the relay's receivers, which can significantly reduce the feasibility of the aggressive spatial frequency reuse targeted by network densification. Additionally, theoretical throughput doubling of FD operations at the device level can not scale to the network level, unless further sophisticated measurements are taken [20]. Solutions targeting a reduction of the FD interference footprint have also been proposed for improving the FD throughput gain scalability [23]. However, the potential benefits brought in by the adoption of caching, for example, the overall reduction of the aggregate network interference, must be considered, if relays operate in the FD mode.

The effects of user mobility on the caching placement strategies have also received plenty of attention [28]–[30]. Wang *et al.* [28] developed a framework of mobility-aware coded caching and demonstrated that coded caching schemes achieve higher performance than their un-coded counterparts. Wang *et al.* [29] also designed caching for mobile devices coexisting with D2D communication networks. They model the user mobility pattern by taking into account the inter-contact times of different users, based on which they proposed a mobility-aware caching placement policy for maximizing the data offloading ratio. Liu *et al.* [30] designed a framework for mobility-aware coded probabilistic caching, which was applied to MEC-enabled small-cell networks. However, this work only considered caching at the SBS level, hence there was no cooperation during content caching and sharing.

By contrast, in this paper, we make a further bold step in terms of designing, analyzing and optimizing the cooperative coded caching placement at both user- and relay-levels as well as the resource allocation whilst considering the constraints of user mobility and self-interference at the relays and of the interference footprint of numerous coexisting FD links. We assume that both the macro BSs (MBSs) and the users operate in the HD mode, whereas the relays operate in the FD mode. In a nutshell, the contributions of this paper can be summarized as follows.

1) We model the HetNets supporting D2D communications and mobility-aware coded probabilistic caching, where the encoded segments are cached at both user- and relay-levels.

2) We formulate a joint optimal caching and resource allocation problem of maximizing the system's throughput under several constraints, such as limited storage capacities at both user- and relay-levels, the content popularity, the transmit powers at users and relays, and the quality of FD self-interference cancellation.

3) We analyze the throughput and propose a near-optimal caching scheme for mitigating the complexity of joint optimal caching placement and resource allocation. In particular, we first develop an algorithm based on particle swarm optimization (PSO) [32] for configuring the parameters of caching placement and resource allocation under the given content assignment sets for both the SBSs and the users. Subsequently, we propose overlapping and non-overlapping content assignment algorithms for determining the content assignment sets for both the SBSs and the users.

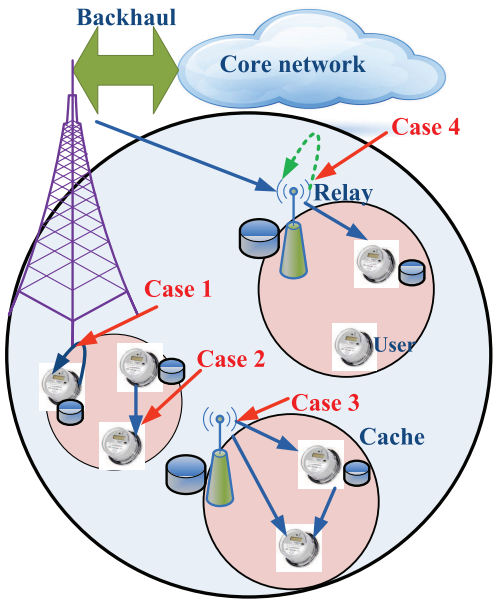

Fig. 1. Caching-aided HetNet with D2D communications.

4) We present numerical results for characterizing the performance of the proposed algorithms and the throughput gains by using the optimal parameter configurations found for both caching and user mobility.

The outline of this paper is as follows. Section II describes the system model. In Section III, coded caching based mobility-aware D2D communications and FD relays are presented. We analyze the throughput of the proposed overlapping caching content placement scheme and formulate the related optimization problem in Section IV. The reduced-complexity suboptimal scheme is formulated in Section V, while Section VI presents our performance results. The concluding remarks are drawn in Section VII.

II. System Models

In this section, we first introduce a two-tier HetNet supporting underlay D2D communications and edge caching, as illustrated in Fig. 1, where the relays and the users are spatially distributed according to two mutually independent homogeneous Poisson Point Processes (PPPs) with densities of λ_R and λ_U, respectively.

A. Network Architecture

A two-tier HetNet consisting of MBSs and low-power relays (or SBSs) associated with underlay D2D communications is considered. A relay with caching capability can transmit the cached contents to the users within the coverage radius R_R. It also exploits its FD capability to forward the requested contents to a user while simultaneously receiving the contents from the BS. All the users are assumed to have caching capabilities. A portion of the users, say α, $0 \leq \alpha \leq 1$, request the contents and act as the receivers. The remaining users

with a portion of $(1-\alpha)$ act as the content transmitters. A user can send the requested contents to the requesting user by using D2D communications, provided that the requesting user is within the user's transmission range of R_U, $R_U < R_R$. We assume furthermore that the distribution of caching-enabled users obeys the PPP with a density of $\alpha\lambda_U$.

The popularity distribution vector of the contents is denoted by $q = [q_1, \ldots, q_i, \ldots, q_N]$, where N is the total number of contents and q_i is the access probability for the i-th content. Let us assume that the contents are arranged in a descending order of popularity and their distribution follows a Zipf distribution with a parameter γ [3]. So the popularity of content i can be written as [18]

$$q_i = \frac{1}{i^\gamma} \frac{1}{\sum_{j=1}^{N} \frac{1}{j^\gamma}}. \tag{1}$$

Without loss of generality, all the N contents are assumed to have an equal size L and the caching capacity of all relays is the same, namely LM_R, where $M_R < N$. The vector $\mathbf{P}_R = \left[p_1^R, p_2^R, \ldots, p_i^R, \ldots, p_N^R\right]$ is defined as the set of caching placement probabilities at the relays, where p_i^R is the proportion of relays that cache content i, $0 \le p_i^R \le 1$ for $i = 1, 2, \ldots, i, \ldots, N$. The distribution of relays that have content i follows the PPP with a density of $\lambda_R p_i^R$. Furthermore, all the users are assumed to have the same caching capacity of LM_U. The vector $\mathbf{P}_U = \left[p_1^U, p_2^U, \ldots, p_i^U, \ldots, p_N^U\right]$ is defined as the caching placement at the users, where p_i^U is the proportion of users that have the content i, $0 \le p_i^U \le 1$ for $i = 1, 2, \ldots, N$. Similarly, the distribution of users that have content i also follows the PPP with a density $\alpha\lambda_U p_i^U$.

In this paper, we consider the overlapping-caching placement, where a common set of contents is cached at both SBSs and users. Here, the overlapping set of contents that are cached at both the SBSs and the users significantly improves the throughput performance in the mobility-aware framework. We also consider two other methods: *i)* common caching placements where both the SBSs and the users cache the same set of contents; and *ii)* separate caching placements where the SBSs and the users cache different sets of contents.

B. Mobility Model

In this paper, we adopt the contact time (or sojourn time) t and connect frequency ρ for characterizing the user mobility pattern. Although the user positions can change any time, the user moving range is relatively small during a short time. We assume that within the contact time t^U, a mobile user remains connected to the same user and within the contact time t^R, a mobile user remains connected to the same SBS. In fact, we assume $t^U < t^R$, since the two users that are connected via D2D communications may move at the same time, while the SBSs normally remain static. After the contact time t, a user can be connected to another new device at a new position and then remains connected with the new contact for the time observation t. If users move at a high speed, the contact time can be short, compared to a rather longer contact time at a low speed.

User mobility is modeled by discrete random jumps with the corresponding intensity characterized by the average sojourn time between jumps. The distribution of the sojourn time can be reasonably modeled by an exponential function as in [28]. Therefore, the probability density function (PDF) of the sojourn time in user-to-user connections and SBS-to-user connections is denoted by $p^U(t^U)$ and $p^R(t^R)$, respectively, which can be expressed as follows:

$$p^U(t^U) = \rho^U \exp\left(-\rho^U t^U\right), \tag{2}$$

$$p^R(t^R) = \rho^R \exp\left(-\rho^R t^R\right), \tag{3}$$

where ρ^U and ρ^R are the relative connection frequencies, which account for the mobility intensities.

We assume that the transmission rate of a direct D2D transmission link or SBS transmission is high enough for a typical user content of size L to be transmitted during the contact time t. In addition to the contact time, we define the total downloading time as $T = L/R_U + \Delta T$, where L/R_U is the time needed for content-downloading, if only D2D communication is used to download all the contents while ΔT is the additional time needed if other slower mechanisms are used to get the contents.

Let r indicate the remaining data of a requested content at the jumping instant. Consider a general scenario where it takes k jumps for a user to download a content. Thus $r = L - \sum_{i=1}^{j} d_i$ is the remaining data at jump j, where d_i is the amount of data received between jump $(i-1)$ and jump i. The *pdf* of r is calculated as follows:

$$p^U(r) = \frac{\exp\left(\frac{\sigma^U r}{\rho^U TL}\right)}{\int_{r=0}^{L} \exp\left(\frac{\sigma^U r}{\rho^U TL}\right) dr}, \tag{4}$$

$$p^R(r) = \frac{\exp\left(\frac{\sigma^R r}{\rho^R TL}\right)}{\int_{r=0}^{L} \exp\left(\frac{\sigma^R r}{\rho^R TL}\right) dr}, \tag{5}$$

where σ^U and σ^R are the control parameters that are used to adapt the mobility intensity.

C. Coded Caching Scheme

In a coded caching scheme, each content file is encoded into multiple segments by a rateless Fountain code and a requested content can be recovered by collecting any k_c encoded segments, which are a subset of the complete set of segments, cached in the local user storage or in the SBS storage. In other words, any packet of the encoded segments is useful for recovering the content. If a user cannot collect enough encoded segments from the local cache either via D2D communication or by SBS transmission within the tolerable downloading time T, the cellular BS will send the missing data to the user.

In contrast to the complete caching scheme, we denote the cache placement at the SBSs by using the coded caching scheme as $\mathbf{M}_R = \left[m_1^R, m_2^R, \ldots, m_N^R\right]$, $0 \le m_i^R \le 1$ for $i = 1, 2, \ldots, N$, where m_i^R refers to the percentage of content i cached in SBS. The cache storage constraint at the SBSs can be expressed as $\sum_{i \in \mathcal{F}^R} m_i^R p_i^R L \le S_R$, where $\mathcal{F}^R$ is the set of files that are cached at SBSs, i.e., $m_i^R \ne 0, p_i^R \ne 0$. In a similar way, let $\mathbf{M}_U = \left[m_1^U, m_2^U, \ldots, m_N^U\right]$ represent

the cache placement at the users having a caching capability for each content, where m_i^U is the percentage of content i cached in each user, and $0 \leq m_i^U \leq 1$ for $i = 1, 2, ..., N$. The cache storage constraint at the cache-enabled users can then be written as $\sum_{i \in \mathcal{F}^U} m_i^U p_i^U L \leq S_U$, where $\mathcal{F}^U$ is the set of files that are cached at the users (i.e. $m_i^U \neq 0, p_i^U \neq 0$).

D. Content Access Protocol

The content access protocol includes four possible cases, which are presented in the following. An example including all four cases is given in Fig. 1.

1) Local cache hit: A requesting user has the requested content from its own cache.
2) D2D cache hit: In the scenario that the user's local storage does not have the requested content, the requesting user checks the nearby users within its transmit coverage radius R_U for the content. If there is at least one user that has the requested content, the user will establish the D2D transmission session with the nearest user having the cached content.
3) Relay cache hit: When the requested content is not available at the local cache or at the nearby users, the user would search nearby relays within a radius of R_R. If there is at least one relay, whose storage has the requested content, the user will download the content from that relay.
4) Cache miss: If the requested content is not available in any of the above cases, the user will connect to the BS via a nearby relay. This case represents a cache miss. Here at least one relay within a distance of R_R performs FD communications to simultaneously receive the content from the BS and to transmit it to the requesting user for reducing the access latency and also for saving power.

Remark 1: We employ the homogeneous PPP in modeling the spatial distributions of relays and users. Although this point process model generally does not work the best for the repulsive behavior of nodes owing to the effect of node correlation, it is sufficient to model the wireless network for our study [34]. Moreover, this approach enables us to generate better insights into the mobility-aware edge caching problem and is sufficient to ensure tractability. The extension of the model to more advanced general point processes such as Ginibre point process and Poisson hard-core hole process will be considered in our future works. Relevant implementations that were published in some recent contributions [33]–[35] would be useful for these further studies.

III. Coded Caching Based Mobility-Aware D2D Communications and Full Duplex Relays in HetNets

A. Assumptions

We first introduce the major assumptions used in our models.

1) If the requesting user gets connected a new user due to user mobility but it is still in the coverage of the current SBS, it can request the remaining content from the new connected user instead of getting it from the current SBS. Moreover, if the requesting user finishing fetching the contents stored by the connected D2D user within a duration less than the contact time, the remaining time is used for receiving the remaining coded data from the SBS within the coverage; if the requesting user finishing fetching the contents stored at the SBS within the duration less than the SBS-user contact time (and the user-user contact time), the remaining time is used for receiving the remaining coded data from the MBS via the SBS by using the FD technology.
2) For simplicity, we only consider the case that the amount of coded data m_i^R for file i cached at the SBSs is always higher than that for the same file cached at the users, m_i^U. The extension to different storage scenarios can be readily performed, although it entails complex derivation. This assumption makes sense, because the storage capacity of SBSs is larger than that of users. Furthermore, caching the common files at both the SBSs and the users helps us to improve the throughput gain in our mobility-aware caching framework. This is the first step of investigating this method, hence we set aside the more complex assignment of coded content segments for our future work.
3) All the users cache the same set of contents and the same amount of coded segments for each file, but different segments. Similarly, all the SBSs cache the same set of contents and for each cached content, they cache the same number of coded segments but different segments. It may be readily generalized to scenarios, where different users and SBSs cache different contents and different amount of coded segments for each content. However, again the derivations are set aside for our future work.

B. Parameter Calculation

1) Cache hit probabilities $p_{hit,i}^U$, $p_{hit,i}^R$, and $p_{hit,i}$: Based on [31], the distribution of file i at user-level obeys the PPP with density $(1-\alpha)\lambda_U p_i^U$. The probability of the event that file i is cached by other users within the coverage range of the requested user is given by [8]:

$$p_{hit,i}^U = 1 - \exp\left(-\pi (1-\alpha) \lambda_U p_i^U R_U^2\right). \quad (6)$$

Similarly, the distribution of file i at relay-level follows the PPP with density $\lambda_R p_i^R$. The probability of the event that i is cached by the relay within the coverage range of the requested user is given as

$$p_{hit,i}^R = 1 - \exp\left(-\pi \lambda_R p_i^R R_R^2\right). \quad (7)$$

2) Content transmission success probabilities $p_{succ,i}^U$, $p_{succ,i}^R$ and $p_{succ,i}^{FD}$: We now define $p_{succ,i}^U$, $p_{succ,i}^R$ and $p_{succ,i}^{FD}$ in this section, while their detailed derivations are presented in the next section. To determine $p_{succ,i}^U$, let j_0 denote a requesting user, which receives the content from the nearest user, $\bar{j}_0$, by using D2D communications. Let us define Θ_j as the cache miss event at relay j, and $\mathbf{1}_{\Theta_j}$ as the indicator function ($\mathbf{1}_{\Theta_j} = 1$ if cache miss event occurs at j, otherwise $\mathbf{1}_{\Theta_j} = 0$).

Relay j uses FD communications to forward the requested content from the BS to its requesting user, once a cache miss occurs. The fading channel spanning from transmitter j to receiver k, h_{jk}, follows a Rayleigh fading distribution with $\mathcal{CN}(0,1)$. So in case 2, the success probability of D2D content transmission for file i is given by

$$p_{succ,i}^{U} = \Pr\left\{\text{SINR}_{j_0}^{U} > \phi\right\}, \tag{8}$$

where ϕ is the predetermined SINR threshold. $\text{SINR}_{j_0}^{U}$ is the SINR at user j_0 formulated as

$$\text{SINR}_{j_0}^{U} = \frac{P_D h_{\bar{j}_0 j_0}^2 d_{\bar{j}_0 j_0}^{-\beta_1}}{\sigma^2 + I_{j_0}}, \tag{9}$$

where $\bar{j}_0$ is the nearest user having the cached content; $d_{\bar{j}_0 j_0}$ is the distance between j_0 and $\bar{j}_0$; β_1 is the pathloss exponent of D2D transmission; and σ^2 is the thermal noise power. The total interference I_{j_0} is expressed as

$$I_{j_0} = \sum_{j \in \Phi_{j_0}^{U} \setminus \{\bar{j}_0 j_0\}} P_D h_{j j_0}^2 d_{j j_0}^{-\beta_1} + \sum_{k \in \Phi_{j_0}^{R}} P_R h_{k j_0}^2 d_{k j_0}^{-\beta_1} + \sum_{l \in \Phi_{j_0}^{R}} P_{BS} h_{BS_l j_0}^2 d_{BS_l j_0}^{-\beta_2} \mathbf{1}_{\Theta_l}. \tag{10}$$

The first term in (10) is the interference caused by user j, which is in the coverage area of user j_0, i.e. $j \in \Phi_{j_0}^{U}$ (where $\Phi_{j_0}^{U}$ is the set of these active users). The second term in (10) represents the interference caused by the transmission of relay $k \in \Phi_{j_0}^{R}$, where $\Phi_{j_0}^{R}$ is the set of active relays that can cause interference to j_0. The third term in (10) represents the interference caused by the specific BS that user j_0 is associated with, BS_l (relay l is associated with BS_l). Relay l is in the coverage area of user j_0 but performs FD communications to help other users. Note that P_{BS} is the transmit power of the BS and β_2 is the pathloss exponent between the BS and the users.

Similarly, we calculate the success probability $p_{succ,i}^{R}$ of a content transmission by relay k_0 in case 3, yielding:

$$p_{succ,i}^{R} = \Pr\left\{\text{SINR}_{j_0}^{RU} > \phi\right\}. \tag{11}$$

To elaborate, $\text{SINR}_{j_0}^{RU}$ is the SINR at user j_0 wrt relay k_0, which can be expressed as

$$\text{SINR}_{j_0}^{RU} = \frac{P_R h_{k_0 j_0}^2 d_{k_0 j_0}^{-\beta_1}}{\sigma^2 + I_{k_0 j_0}}. \tag{12}$$

The interference $I_{k_0 j_0}$ is calculated as

$$I_{k_0 j_0} = \sum_{j \in \Phi_{j_0}^{U} \setminus j_0} P_D h_{j j_0}^2 d_{j j_0}^{-\beta_1} + \sum_{k \in \Phi_{j_0}^{R} \setminus k_0} P_R h_{k j_0}^2 d_{k j_0}^{-\beta_1} + \sum_{l \in \Phi_{j_0}^{R} \setminus k_0} P_{BS} h_{BS_l j_0}^2 d_{BS_l j_0}^{-\beta_2} \mathbf{1}_{\Theta_l}. \tag{13}$$

Furthermore, the success of a full duplex transmission $p_{succ,i}^{FD}$, when a cache miss occurs (case 4), is determined as

$$p_{succ,i}^{FD} = \Pr\left\{\text{SINR}_{j_0}^{FD} > \phi \text{ and } \text{SINR}_{R}^{FD} > \phi\right\}. \tag{14}$$

Here, the transmissions rely on two hops, where the first hop is the communication from the BS to the relay and the second hop is the communication from the relay to the user. Furthermore, $\text{SINR}_{j_0}^{FD}$ is the SINR wrt nearby relay k_0 at user j_0 formulated as:

$$\text{SINR}_{j_0}^{FD} = \frac{P_R h_{k_0 j_0}^2 d_{k_0 j_0}^{-\beta_1}}{\sigma^2 + I_{k_0 j_0}^{FD}}. \tag{15}$$

The interference $I_{k_0 j_0}^{FD}$ is calculated as

$$I_{k_0 j_0}^{FD} = \sum_{j \in \Phi_{j_0}^{U} \setminus j_0} P_D h_{j j_0}^2 d_{j j_0}^{-\beta_1} + \sum_{k \in \Phi_{j_0}^{R} \setminus k_0} P_R h_{k j_0}^2 d_{k j_0}^{-\beta_1} + \sum_{l \in \Phi_{j_0}^{R}} P_{BS} h_{BS_l j_0}^2 d_{BS_l j_0}^{-\beta_2} \mathbf{1}_{\Theta_l}. \tag{16}$$

Note that the difference between $I_{k_0 j_0}$ and $I_{k_0 j_0}^{FD}$ arises from the fact that $I_{k_0 j_0}^{FD}$ includes the interference caused by transmissions between the BS and relay k_0.

SINR_{R}^{FD} is the SINR at relay k_0 wrt the BS, which is formulated as:

$$\text{SINR}_{R}^{FD} = \frac{P_{BS} h_{BS k_0}^2 d_{BS k_0}^{-\beta_2}}{\sigma^2 + I_{BS k_0}^{FD}}. \tag{17}$$

The interference $I_{BS k_0}^{FD}$ is calculated as

$$I_{BS k_0}^{FD} = \sum_{j \in \Phi_{k_0}^{RU} \setminus j_0} P_D h_{j k_0}^2 d_{j k_0}^{-\beta_1} + h_{k_0 k_0}^2 \xi P_R + \sum_{k \in \Phi_{k_0}^{R} \setminus k_0} P_R h_{k k_0}^2 d_{k k_0}^{-\beta_1} + \sum_{l \in \Phi_{k_0}^{R}} P_{BS} h_{BS_l k_0}^2 d_{BS_l k_0}^{-\beta_2} \mathbf{1}_{\Theta_l}. \tag{18}$$

Here, the first term in (18) is the interference caused by the transmission of user j, which is in the set $\Phi_{k_0}^{RU}$ falling in the coverage of relay k_0. The second term in (18) presents the self-interference caused by power leakage from the transmission of relay k_0, and ξ characterizes the quality of self-interference cancellation [23]. The third term in (18) represents the interference caused by the transmission of relay $k \in \Phi_{k_0}^{R}$, where $\Phi_{k_0}^{R}$ is the set of active relays within the coverage of the considered relay k_0. The fourth term in (18) represents the interference between the BS, namely BS_l and the relay l. Here, relay l is within relay k_0's coverage and performs FD communications to help the other users.

3) Calculation of $p_{succ,i}^{U}$, $p_{succ,i}^{R}$, $\bar{p}_{succ,i}^{R}$, $p_{succ,i}^{FD}$ and $\bar{p}_{succ,i}^{FD}$: We study two scenarios: *a)* in the first one, the SBSs and the users are assigned to cache distinct sets of contents and it is termed as the non-overlapping scenario; and *b)* in the second one, the SBSs and the users can cache the common set of contents, which is referred to as the overlapping scenario. Let us define the following parameters. The bandwidth of the MBS, SBSs and users are W_b, W_r, W_u. We assume furthermore that $W_b = \omega_1 W_r = \omega_2 W_u$, where $\omega_1, \omega_2 \in [0,1]$, $\omega_1 > \omega_2$, i.e., the channel capacity of the backhaul is lower than that of the downlink.

For ease expression, we omit the subscript i for the calculated parameters (such as $p_{hit,i}^{U}$, $p_{hit,i}^{R}$, $p_{succ,i}^{R}$, $p_{succ,i}^{U}$ and etc), when we consider the calculation for the content i in the

following. We define that $\mathcal{F}^U$ and $\mathcal{F}^R$ are the sets of files that are cached at the users and the SBSs, respectively; $\mathcal{F}^{RU}$ is the set of files that are cached at the users and at the SBSs ($\mathcal{F}^{RU} = \mathcal{F}^U \cup \mathcal{F}^R$); $\mathcal{F}^{RU\backslash U}$ is the set of files that are cached at the SBSs but not at the users ($\mathcal{F}^{RU\backslash U} = \left(\mathcal{F}^U \cup \mathcal{F}^R\right) \backslash \mathcal{F}^U$); $\mathcal{F}^{RU\backslash R}$ is the set of files that are cached at the users but not at the SBSs ($\mathcal{F}^{RU\backslash R} = \left(\mathcal{F}^U \cup \mathcal{F}^R\right) \backslash \mathcal{F}^R$).

The calculations of coverage probabilities ($p^U_{succ,i}$, $p^R_{succ,i}$, $\bar{p}^R_{succ,i}$, $p^{FD}_{succ,i}$ and $\bar{p}^{FD}_{succ,i}$) are omitted and can be found in the technical report [40].

IV. Throughput Analysis

In this section, we analyze the throughput of our overlapping caching content placement technique. The non-overlapping caching placement can be readily obtained by omitting the cases that the considered content i belongs the common set of the SBSs and the users. Then, the throughput optimization is formulated under the constraints of caching placement and resource allocation.

A. Definition of Throughput

Let us define $\mathcal{NT}$ as the total throughput per unit area, which is calculated as the achieved rate, when the requested contents are successfully received in one unit area. Moreover, $\mathcal{T}$ denotes the total throughput per user, which is calculated as the achieved rate by the tagged requesting user, when its requested contents are successfully received. In the following, we perform throughput analysis for the throughput per user (or just called the throughput for simplicity). The throughput per unit area is thus calculated by $\mathcal{NT} = \alpha\lambda_U\mathcal{T}$.

B. Throughput Calculation

$\mathcal{T}$ can be expressed as

$$\mathcal{T} = \mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3 + \mathcal{T}_4, \tag{19}$$

where $\mathcal{T}_1$, $\mathcal{T}_2$, $\mathcal{T}_3$ and $\mathcal{T}_4$ are the throughputs achieved for four cases that are derived as follows.

1) Case 1: If the user j_0 requests the content i, $i \notin \mathcal{F}^U \cup \mathcal{F}^R$, it can be served by any SBS within its coverage area. In particular, the served SBS exploits the FD communication to simultaneously download the requested content from the MBSs and to forward it to the requesting user. So the throughput achieved is given as

$$\mathcal{T}_1 = \Pr\left\{i \notin \mathcal{F}^U \cup \mathcal{F}^R\right\} p^{FD,0}_{succ} R_{FD}, \tag{20}$$

where

$$\Pr\left\{i \notin \mathcal{F}^U \cup \mathcal{F}^R\right\} = \sum_{i \in \mathcal{F}\backslash\mathcal{F}^{RU}} q_i, \tag{21}$$

$$R_{FD} = W_b \log_2\left(1 + \frac{\text{SINR}^{FD}_R \text{SINR}^{FD}_{j_0}}{1 + \text{SINR}^{FD}_R + \text{SINR}^{FD}_{j_0}}\right), \tag{22}$$

and $p^{FD,0}_{succ}$ is calculated in [40]. Note that we use the lower bound ϕ for the SINR (such as $\text{SINR}^{FD}_{j_0}$, SINR^{FD}_R) when we calculate the rates in numerical results for simplicity. From the following, we consider the case that the requested content i belongs the set of contents stored in the SBSs and/or users.

2) Case 2: If the user j_0 requests the content i after it experiences multiple jumps, then the remaining data volume, r, is smaller than the coded data stored by the user, i.e. we have $r < m^U L < m^R L$. Recall that we assume $m^U < m^R$, implying that the capacity of the SBSs is higher than that of the users. Then, the throughput achieved is given by

$$\mathcal{T}_2 = \Pr\left\{r < m^U L\right\}\left(\mathcal{T}_{2.1} + \mathcal{T}_{2.2} + \mathcal{T}_{2.3}\right), \tag{23}$$

where $\mathcal{T}_{2.1}$, $\mathcal{T}_{2.2}$, and $\mathcal{T}_{2.3}$ correspond to the following three sub-cases.

a: Case 2.1

If $i \in \mathcal{F}^{RU\backslash R}$ ($\mathcal{F}^{RU\backslash R} = \left(\mathcal{F}^U \cup \mathcal{F}^R\right) \backslash \mathcal{F}^R$), there are two possible scenarios: *i)* a high transmission rate of R_U will be achieved if the requesting user successfully downloads the content sought from a nearby user; and *ii)* the transmission rate will be R_{FD} if the nearby users do not cache content i, hence the requesting user will be served by a SBS within the coverage area. We assume that an advanced FD SBS is used, which obtains content i from the MBSs and concurrently forwards it to the requesting user in the second case. Thus, the throughput achieved for content i is given by

$$\mathcal{T}_{2.1} = \Pr\left\{i \in \mathcal{F}^{RU\backslash R}\right\}\left[p^U_{succ} R_U + \left(1 - p^U_{succ}\right)\bar{p}^{FD}_{succ} R_{FD}\right], \tag{24}$$

where $R_U = W_u \log_2\left(1 + \text{SINR}^{j_0}_U\right)$, while R_{FD} is from Eq. (22), and $\bar{p}^{FD}_{succ}$ as well as p^U_{succ} can be determined according to [40].

b: Case 2.2

Similarly, if $i \in \mathcal{F}^{RU\backslash U}$ ($\mathcal{F}^{RU\backslash U} = \left(\mathcal{F}^U \cup \mathcal{F}^R\right) \backslash \mathcal{F}^U$), a high transmission rate R_R will be achieved (if the requesting user successfully downloads the requested content cached in the SBS); and the transmission rate will be R_{FD} (if the nearby users do not cache content i and the requesting user will be served by any of the SBSs within the coverage area). Then the throughput associated with content i is written as

$$\mathcal{T}_{2.2} = \Pr\left\{i \in \mathcal{F}^{RU\backslash U}\right\}\left[p^R_{succ} R_R + \left(1 - p^R_{succ}\right)\bar{p}^{FD}_{succ} R_{FD}\right], \tag{25}$$

where $R_R = W_r \log_2\left(1 + \text{SINR}^{j_0}_R\right)$, and R_{FD} is from Eq.(22). Then, p^R_{succ} and $\bar{p}^{FD}_{succ}$ are calculated as in [40].

c: Case 2.3

If $i \in \mathcal{F}^U \cap \mathcal{F}^R$, there are three possible scenarios: *i)* the transmission rate will be R_U, if the requesting user successfully downloads the content sought from a nearby user; *ii)* otherwise, a high transmission rate of R_R will be achieved (if the user successfully downloads the requested content cached by the SBS); and *iii)* if the user fails to download content i from any of the nearby users or from the SBS caching i, it will be served by any of the SBSs within the coverage area (which does not cache content i) and the transmission rate will be R_{FD}. Hence, the throughput provided for content i can be expressed as

$$\begin{aligned}\mathcal{T}_{2.3} = \Pr\left\{i \in \mathcal{F}^U \cap \mathcal{F}^R\right\}\big[&p^U_{succ} R_U + \left(1 - p^U_{succ}\right) p^R_{succ} R_R \\ &+ \left(1 - p^U_{succ}\right)\left(1 - p^R_{succ}\right)\bar{p}^{FD}_{succ} R_{FD}\big].\end{aligned} \tag{26}$$

3) Case 3: If the user j_0 requests the content i after it experiences multiple jumps, then the remaining data volume is larger than the coded data stored in the user's cache but smaller than that stored in the SBS's cache, i.e. we have $m^U L < r < m^R L$. The throughput achieved is given by

$$\mathcal{T}_3 = \mathcal{T}_{3.1} + \mathcal{T}_{3.2} + \mathcal{T}_{3.3}, \tag{27}$$

where $\mathcal{T}_{3.1}$, $\mathcal{T}_{3.2}$, and $\mathcal{T}_{3.3}$ correspond to the following three cases.

a: Case 3.1

We consider the scenario that $i \in \mathcal{F}^{RU\setminus R}$. Furthermore, we introduce the shorthand $t_0^U = m^U L/R_U$, $t_1^U(r) = t_0^U + (r - m^U L)/R_{FD}$, where t^U is the jumping time of user j_0, beyond which it establishes contact with a new group of users. If $t^U \leq t_0^U$, the requesting user j_0 will receive its data from a local donor user at a high transmission rate R_U. However, if $t^U > t_0^U$, user j_0 will receive the requested data from the nearby donor user at the transmission rate R_U, where the downloading duration is within the range of $\left[0, t_0^U\right]$; and then the requesting user j_0 will receive its data from the relaying SBS at the rate R_{FD} (again, the relaying SBS employs FD communications) and the download period is within the range of $\left[t_0^U, t^U\right]$. So the rate for the case of $t^U > t_0^U$ is lower than that for the case of $t^U \leq t_0^U$. The average rate for the case of $t^U > t_0^U$ can be expressed as

$$R_1(t^U, r) = \begin{cases} \frac{t_0^U R_U + \left(t^U - t_0^U\right) R_{FD}}{t^U} & t_0^U < t^U \leq t_1^U(r) \\ \frac{t_0^U R_U + \left(t_1^U(r) - t_0^U\right) R_{FD}}{t_1^U(r)} & t^U > t_1^U(r) \end{cases}. \tag{28}$$

The throughput achieved for content i would be written as

$$\begin{aligned} \mathcal{T}_{3.1} = {} & \Pr\left\{i \in \mathcal{F}^{RU\setminus R}\right\} p_{succ}^U \\ & \times \Big[\Pr\left\{m^U L < r < m^R L\right\} \Pr\left\{t^U < t_0^U\right\} R_U \\ & \quad + \int_{r=m^U L}^{m^R L} \int_{t^U=t_0^U}^{\infty} R_1(t^U, r) p^U(t^U) p^U(r) dr dt^U \Big] \\ & + \Pr\left\{i \in \mathcal{F}^{RU\setminus R}\right\} \left(1 - p_{succ}^U\right) \\ & \quad \times \Pr\left\{m^U L < r < m^R L\right\} p_{succ}^{FD,0} R_{FD}. \end{aligned} \tag{29}$$

b: Case 3.2

Let us now consider the case of $i \in \mathcal{F}^{RU\setminus U}$. It is readily observed that this scenario is similar to Case 2.2 and the throughput achieved for content i is written as

$$\begin{aligned} \mathcal{T}_{3.2} = {} & \Pr\left\{i \in \mathcal{F}^{RU\setminus U}\right\} \Pr\left\{m^U L < r < m^R L\right\} \\ & \times \left[p_{succ}^R R_R + \left(1 - p_{succ}^R\right) \bar{p}_{succ}^{FD} R_{FD}\right]. \end{aligned} \tag{30}$$

c: Case 3.3

We continue by considering the case of $i \in \mathcal{F}^U \cap \mathcal{F}^R$. In this case, there are two sub-cases:

- Case 3.3.1: the requesting user j_0 successfully connects to the caching user with probability p_{succ}^U;
- Case 3.3.2: the user j_0 fails to connect to any caching user with a probability of $(1 - p_{succ}^U)$.

In Case 3.3.1, if $t^U \leq t_0^U$, the requesting user j_0 will receive data at a high transmission rate of R_U. This case is termed as *Case 3.3.1.1)* corresponding to throughput $\mathcal{T}_{3.3.1.1}$. However, if $t^U > t_0^U$, the user j_0 will receive data at a high transmission rate of R_U within the time range of $\left[0, t_0^U\right]$; and then from t_0^U to t^U, it will receive the remaining content at the rate of R_R or R_{FD}, depending on the availability of the caching SBSs within the coverage area. Hence there are two sub-cases for $t^U > t_0^U$: *Case 3.3.1.2)* the requesting user j_0 successfully connects to the caching SBS with a probability of p_{succ}^R corresponding to throughput $\mathcal{T}_{3.3.1.2}$; *Case 3.3.1.3)* the user j_0 fails to connect to any caching SBS with probability $(1 - p_{succ}^R)$ corresponding to throughput $\mathcal{T}_{3.3.1.3}$. Thus, the throughput achieved can be expressed as

$$\mathcal{T}_{3.3} = \Pr\left\{i \in \mathcal{F}^U \cap \mathcal{F}^R\right\} \left(\mathcal{T}_{3.3.1.1} + \mathcal{T}_{3.3.1.2} + \mathcal{T}_{3.3.1.3} + \mathcal{T}_{3.3.2}\right), \tag{31}$$

where $\mathcal{T}_{3.3.2}$ is the throughput of Case 3.3.2, while $\mathcal{T}_{3.3.1.1}$, $\mathcal{T}_{3.3.1.2}$, $\mathcal{T}_{3.3.1.2}$, and $\mathcal{T}_{3.3.2}$ correspond to the following three cases.

In Case 3.3.1.1, $\mathcal{T}_{3.3.1.1}$ is expressed as

$$\mathcal{T}_{3.3.1.1} = \Pr\left\{m^U L < r < m^R L\right\} p_{succ}^U \Pr\left\{t^U < t_0^U\right\} R_U. \tag{32}$$

In Case 3.3.1.2, the remaining content volume must be $(r - m^U L) \leq (m^R L - m^U L) < m^R L$, hence the requesting user will receive all the remaining segments from the caching SBS by $t_2^U(r)$, where $t_2^U(r) = t_0^U + (r - m^U L)/R_R$ is the maximum downloading time (i.e. $t_0^U \leq t^U \leq t_2^U(r)$). Hence, the average rate can be expressed as

$$R_2(t^U, r) = \frac{t_0^U R_U + \left(t^U - t_0^U\right) R_R}{t^U} \quad t_0^U \leq t^U \leq t_2^U(r). \tag{33}$$

Then, $\mathcal{T}_{3.3.1.2}$ can be calculated as

$$\begin{aligned} \mathcal{T}_{3.3.1.2} = {} & p_{succ}^U p_{succ}^R \\ & \times \int_{r=m^U L}^{m^R L} \int_{t^U=t_0^U}^{t_2^U(r)} R_2(t^U, r) p^U(t^U) p^U(r) dr dt^U. \end{aligned} \tag{34}$$

Here, we assume that the probability that the contact time t^U of the users is higher than the contact time t^R between the user and the SBS is very low. Hence, we ignore the case when the user will jump to another caching SBS's coverage within the duration of the user receiving the contents sought from the current caching SBS. The exact derivation covering this scenario is omitted here owing to the space limit, but motivated readers can find the details in Appendix A of [40]. We will validate the approximate derivation by using our numerical simulations.

In Case 3.3.1.3, the requesting user receives all the remaining segments from the MBS by the instant $t_1^U(r)$, where $t_1^U(r) = t_0^U + (r - m^U L)/R_{FD}$ is the maximum time (i.e. $t_0^U \leq t^U \leq t_1^U(r)$). Thus, the average rate $R_1(t^U, r)$ can be derived as in Eq. (28). Then, $\mathcal{T}_{3.3.1.3}$ can be calculated as

$$\begin{aligned} \mathcal{T}_{3.3.1.3} = {} & p_{succ}^U \left(1 - p_{succ}^R\right) \bar{p}_{succ}^R \\ & \times \int_{r=m^U L}^{m^R L} \int_{t^U=t_0^U}^{\infty} R_1(t^U, r) p^U(t^U) p^U(r) dr dt^U. \end{aligned} \tag{35}$$

In Case 3.3.2, the requesting user cannot access any nearby caching user. Hence, it can only get the contents from the

caching SBS within the coverage area, provided that a caching SBS is available. Otherwise, it can get the content from the MBS via other relaying SBSs within the coverage, which do not cache the requested content. So $\mathcal{T}_{3.3.2}$ is derived as follows:

$$\mathcal{T}_{3.3.2} = \Pr\left\{m^U L < r < m^R L\right\}\left(1 - p_{succ}^U\right) \times \left[p_{succ}^R R_R + \left(1 - p_{succ}^R\right)\bar{p}_{succ}^{FD} R_{FD}\right]. \quad (36)$$

4) Case 4: If the user j_0 requests the content i, where the remaining data volume is in the range of $r > m^R L$, the throughput achieved is given by

$$\mathcal{T}_4 = \mathcal{T}_{4.1} + \mathcal{T}_{4.2} + \mathcal{T}_{4.3}, \quad (37)$$

where $\mathcal{T}_{4.1}$, $\mathcal{T}_{4.2}$ and $\mathcal{T}_{4.3}$ correspond to the following three cases.

a: Case 4.1

We now assume that $i \in \mathcal{F}^{RU\setminus R}$. If the requesting user j_0 succeeds in accessing a nearby caching user and if the contact time $t^U < t_0^U$, then it will receive the content at the high rate of R_U; otherwise, it first receives the content at the high rate of R_U in the delay range of $\left[0, t_0^U\right]$ and at the lower rate of R_{FD} in the remaining time (the average rate is $R_1(t^U, r)$ from Eq. (28)). If the requesting user j_0 fails to access a nearby caching user, it will get the content from the MBS via the relaying SBS. Therefore, the throughput achieved can be written as

$$\mathcal{T}_{4.1} = \Pr\left\{i \in \mathcal{F}^{RU\setminus R}\right\} p_{succ}^U \times \left[\Pr\left\{r > m^R L\right\}\Pr\left\{t^U < t_0^U\right\} R_U + \int_{r=m^R L}^{L}\int_{t^U=t_0^U}^{\infty} R_1(t^U, r) p^U(t^U) p^U(r) dr dt^U\right] + \Pr\left\{i \in \mathcal{F}^{RU\setminus R}\right\}\Pr\left\{r > m^R L\right\}\left(1 - p_{succ}^U\right) p_{succ}^{FD,0} R_{FD}, \quad (38)$$

where $R_1(t^U, r)$ is from Eq. (28), while $p_{succ}^{FD,0}$ is from [40].

b: Case 4.2

Let us now assume that $i \in \mathcal{F}^{RU\setminus U}$. We define the shorthand of $t_0^R = m^R L / R_R$, $t_3^R(r) = t_0^R + (r - m^R L)/R_{FD}$, and t^R as the jumping time of user j_0 after which it contacts a new SBS. If the requesting user j_0 succeeds in accessing a nearby caching SBS and if $t^R \le t_0^R$, the user j_0 will receive its data at the high transmission rate of R_R. However, if $t^R > t_0^R$, the user j_0 will receive its data at the high transmission rate of R_R within the delay range of $\left[0, t_0^R\right]$ and at the lower rate of R_{FD} beyond the delay of t_0^R. So the average rate can be expressed as

$$R_3(t^R, r) = \begin{cases} \frac{t_0^R R_R + \left(t^R - t_0^R\right) R_{FD}}{t^R} & t_0^R < t^R \le t_3^R(r) \\ \frac{t_0^R R_R + \left(t_3^R(r) - t_0^R\right) R_{FD}}{t_3^R(r)} & t^R > t_3^R(r) \end{cases}, \quad (39)$$

If the requesting user j_0 fails to access any nearby caching SBS, it will get the content from the MBS via a nearby relaying SBS, which does not cache the content requested. Therefore, the throughput achieved can be written as

$$\mathcal{T}_{4.2} = \Pr\left\{i \in \mathcal{F}^{RU\setminus U}\right\} p_{succ}^R \times \left[\Pr\left\{r > m^R L\right\}\Pr\left\{t^R < t_0^R\right\} R_R + \int_{r=m^R L}^{L}\int_{t^R=t_0^R}^{\infty} R_3(t^R, r) p^R(t^R) p^R(r) dr dt^R\right] + \Pr\left\{i \in \mathcal{F}^{RU\setminus U}\right\}\Pr\left\{r > m^R L\right\}\left(1 - p_{succ}^R\right)\bar{p}_{succ}^{FD} R_{FD}. \quad (40)$$

c: Case 4.3

Let us now consider the case of $i \in \mathcal{F}^U \cap \mathcal{F}^R$. In this case, there are four sub-cases:

- Case 4.3.1: the requesting user j_0 successfully connects to the caching user with a probability of p_{succ}^U and to the caching SBS with a probability of p_{succ}^R;
- Case 4.3.2: the user j_0 successfully connects to the caching user with a probability of p_{succ}^U, but fails to access the caching SBS with a probability of $(1 - p_{succ}^R)$;
- Case 4.3.3: the user j_0 fails to connect to the caching user with a probability of $(1 - p_{succ}^U)$, but successfully connects to the caching SBS with a probability of p_{succ}^R;
- Case 4.3.4: the user j_0 fails to connect to any caching user and to any caching SBS with a probability of $(1 - p_{succ}^U)$ and $(1 - p_{succ}^R)$, respectively.

So the throughput achieved can be formulated as

$$\mathcal{T}_{4.3} = \Pr\left\{\mathcal{F}^U \cap \mathcal{F}^R\right\}\left(\mathcal{T}_{4.3.1} + \mathcal{T}_{4.3.2} + \mathcal{T}_{4.3.3} + \mathcal{T}_{4.3.4}\right), \quad (41)$$

where $\mathcal{T}_{4.3.1}$, $\mathcal{T}_{4.3.2}$, $\mathcal{T}_{4.3.3}$ and $\mathcal{T}_{4.3.4}$ correspond to the following three cases.

In Case 4.3.1, if $t^U \le t_0^U$, the requesting user j_0 will receive its data at the high transmission rate of R_U. However, if $t^U > t_0^U$, the user j_0 will receive its data at the high transmission rate R_U within the delay range of $\left[0, t_0^U\right]$; while from t_0^U to t^U, it will receive the remaining content at the rate R_R or R_{FD}, depending on the availability of caching SBSs within the coverage area. There are two cases: *Case 4.3.1.1)* $m^R L < r \le m^R L + m^U L$ (the corresponding throughput $\mathcal{T}_{4.3.1.1}$) and *Case 4.3.1.2)* $r > m^R L + m^U L$ (the corresponding throughput $\mathcal{T}_{4.3.1.2}$). Therefore, $\mathcal{T}_{4.3.1} = \mathcal{T}_{4.3.1.1} + \mathcal{T}_{4.3.1.2}$.

In Case 4.3.1.1, if $t^U > t_0^U$, the user j_0 will receive the remaining content $(r - m^U L)$ at the high transmission rate R_R. The transmission duration at the rate R_R spans from t_0^U to t^U if $t_0^U < t^U < t_0^U + t_0^R$, where $t_0^R = m^R L / R_R$. Otherwise (i.e. $t^U > t_0^U + t_0^R$), the remaining content is received in the delay range of $\left[t_0^U, t_0^U + t_0^R\right]$. Note that there is no jump for the SBS-user contact before the jump for the user-user contact. Since we assume that the probability of having $t^R < t^U$ is very small, we can ignore this case. Therefore, the average rate for the case of $t^U > t_0^U$ can be expressed as

$$R_4(t^U, r) = \begin{cases} \frac{t_0^U R_U + \left(t^U - t_0^U\right) R_R}{t^U} & t_0^U < t^U < t_0^U + t_0^R \\ \frac{t_0^U R_U + t_0^R R_R}{t_0^U + t_0^R} & t^U > t_0^U + t_0^R \end{cases}. \quad (42)$$

Then, $\mathcal{T}_{4.3.1.1}$ can be calculated as

$$\mathcal{T}_{4.3.1.1} = p^U_{succ} p^R_{succ} \times \Big[\Pr\{m^U L < r < (m^R + m^U)L\}\Pr\{t^U < t^U_0\} R_U + \int_{r=m^U L}^{(m^R+m^U)L} \int_{t^U=t^U_0}^{\infty} R_4(t^U, r) p^U(t^U) p^U(r) dr dt^U \Big]. \quad (43)$$

In Case 4.3.1.2, if $t^U > t^U_0$, let us define $t^R_4(r) = t^U_0 + t^R_0 + \left(r - m^U L - m^R L\right)/R_{FD}$, which is the total downloading time by using the user-user transmission, the SBS-user transmission and the MBS-SBS-user transmission. In this case, if $t^U > t^U_0 + t^R_0$, the requesting user will receive the contents at the rate of R_R for the duration of t^R_0 and then at the rate of R_{FD} in the remaining time. Again, we note that there is no jump for the SBS-user contact before the jump for the user-user contact. Hence, the average rate for the case of $t^U > t^U_0$ can be derived as

$$R_5(t^U, r) = \begin{cases} \frac{t^U_0 R_U + \left(t^U - t^U_0\right) R_R}{t^U} & t^U_0 < t^U < t^U_0 + t^R_0 \\ \frac{t^U_0 R_U + t^R_0 R_R + \left(t^U - t^U_0 - t^R_0\right) R_{FD}}{t^U} & t^U_0 + t^R_0 < t^U \le t^R_4(r) \\ \frac{t^U_0 R_U + t^R_0 R_R + \left(t^R_4(r) - t^U_0 - t^R_0\right) R_{FD}}{t^R_4(r)} & t^U > t^R_4(r) \end{cases}.$$

Then, $\mathcal{T}_{4.3.1.2}$ can be written as

$$\mathcal{T}_{4.3.1.2} = p^U_{succ} p^R_{succ} \times \Big[\Pr\{r > (m^R + m^U)L\}\Pr\{t^U < t^U_0\} R_U + \int_{r=(m^U+m^R)L}^{L} \int_{t^U=t^U_0}^{\infty} R_5(t^U, r) p^U(t^U) p^U(r) dr dt^U \Big]. \quad (44)$$

In Case 4.3.2, if $t^U \le t^U_0$, the requesting user j_0 will receive data at the high transmission rate R_U. However, if $t^U > t^U_0$, the user j_0 will receive data at rate R_U within the delay range of $\left[0, t^U_0\right]$; and from t^U_0 to t^U, it will exploit FD communications to receive the remaining content at the rate R_{FD} via the relaying SBS, which does not cache content i. Therefore, $\mathcal{T}_{4.3.2}$ can be calculated as

$$\mathcal{T}_{4.3.2} = p^U_{succ}\left(1 - p^R_{succ}\right)\Big[\Pr\{r > m^R L\}\Pr\{t^U < t^U_0\} R_U + \bar{p}^{FD}_{succ} \int_{r=m^U L}^{L} \int_{t^U=t^U_0}^{\infty} R_1(t^U, r) p^U(t^U) p^U(r) dr dt^U \Big], \quad (45)$$

where $R_1(t^U, r)$ is from Eq. (28).

In Case 4.3.3, if $t^R \le t^R_0$, the requesting user j_0 will receive its data at the transmission rate R_R from its caching SBS. However, if $t^R > t^R_0$, the user j_0 will receive its data at rate R_R within the delay range of $\left[0, t^R_0\right]$; and from t^R_0 to t^R, it will receive the remaining content at the rate R_{FD} via the relaying SBS by using FD communications. Then, $\mathcal{T}_{4.3.3}$ can be calculated as

$$\mathcal{T}_{4.3.3} = p^R_{succ}\left(1 - p^U_{succ}\right)\Big[\Pr\{r > m^R L\}\Pr\{t^R < t^R_0\} R_R + \int_{r=m^U L}^{L} \int_{t^R=t^R_0}^{\infty} R_3(t^R, r) p^R(t^R) p^R(r) dr dt^R \Big], \quad (46)$$

where $R_3(t^R, r)$ is from Eq. (39).

In Case 4.3.4, the user j_0 cannot access any nearby caching user or SBS. Hence, it can rely on FD communications to get the content from the MBS via the relaying SBSs within the coverage, which do not cache the requested content. So $\mathcal{T}_{4.3.4}$ is expressed as follows:

$$\mathcal{T}_{4.3.4} = \Pr\left\{r > m^R L\right\}\left(1 - p^U_{succ}\right)\left(1 - p^R_{succ}\right)\bar{p}^{FD}_{succ} R_{FD}. \quad (47)$$

In summary, the flow chart in Fig. 2 elaborates the relationship of all the cases 1, 2, 3 and 4, while the flow charts in Fig. 3 present the relationship of subcases in cases 2, 3 and 4. Note that the leaf nodes represent all the cases.

Remark 2: Due to space limitation and the complexity of the problem under study, we do not consider the performance analysis of other important metrics such as delay and outage in this paper. Throughput analysis can gain insights into the investigated FD relays aided mobility-aware probabilistic caching problem, while still keeping the problem sufficiently tractable. Analysis performed in the above content can be extended to consider the delay and outage performance in the future work. Relevant results published in some recent works such as those in [14], [16], [17], [36], [37] would be useful for these further studies.

C. Problem Formulation

In this treatise, throughput is our key performance metric. We are interested in finding the optimal caching placement and resource allocation achieving the maximum throughput. Specifically, the throughput $\mathcal{T}$ is a function of the ratio α, the caching placements at both the users $(\mathbf{P}_U, \mathbf{M}_U)$ and at the relays $(\mathbf{P}_R, \mathbf{M}_R)$ as well as of the sets of cached contents at the users and the SBSs $\left(\mathcal{F}^U, \mathcal{F}^R\right)$, the transmit powers (P_D, P_R) of the users and of the relays. Then, the throughput maximization problem can be formulated as follows:

[P1]:

$$\begin{aligned} \max_{\alpha, \mathcal{F}^U, \mathcal{F}^R, \mathbf{P}_U, \mathbf{M}_U, \mathbf{P}_R, \mathbf{M}_R, P_D, P_R} \quad & \mathcal{N}\mathcal{T} \\ \text{s.t. } & \textstyle\sum_{i \in \mathcal{F}^R} m^R_i p^R_i L \le S_R;\ \sum_{i \in \mathcal{F}^U} m^U_i p^U_i L \le S_U \\ & \alpha, m^R_i, p^R_i, m^U_i, p^U_i \in [0, 1];\ P_D \le \bar{P}_D, P_R \le \bar{P}_R, \end{aligned} \quad (48)$$

where $\bar{P}_D$ and $\bar{P}_R$ are the maximum transmit powers of users and of relays, respectively. Here, the first and second constraints represent the limited cached capacities of the SBSs and of the users, respectively. The third constraint is for the parameters of caching placement and α, whilst the last constraint describes the upper bounds of the transmit powers of the users and of the SBSs. We will present our solutions of problem **P1** in the next section.

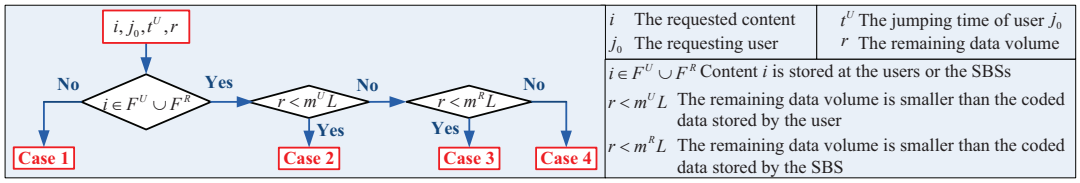

Fig. 2. The relationship of all the cases 1, 2, 3 and 4.

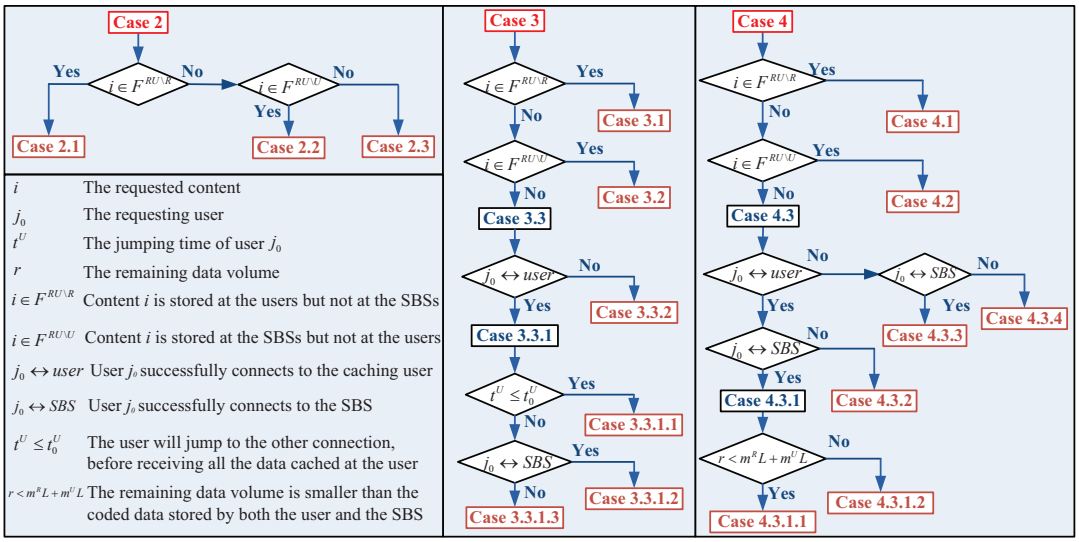

Fig. 3. The relationship of subcases in cases 2, 3 and 4.

V. Throughput Maximization

Problem **P1** of (48) is non-convex in general. Due to the nonlinear and combinatorial structure of the content assignment problem formulated, it would be impossible to obtain its optimal closed-form solution. However, we can employ the brute-force search (i.e. exhaustive search) to determine the best content assignment that results in the maximum throughput. More particularly, we can enumerate all possible content assignment solutions and then determine the best solution by comparing their throughput achieved. This solution method requires an analytical throughput model in Section IV for calculating the throughput for any particular content assignment solution. However, this method imposes an exponentially increasing complexity as a function of the number of the users, the SBSs as well as the contents. Furthermore, we would determine the optimal caching placement for both the users and the SBSs as well as the optimal transmit powers of the users and the SBSs to maximize the throughput achieved. These configuration parameters are also difficult to obtain. Therefore, we develop low-complexity suboptimal algorithms for the content assignment, caching placement and power allocation in the following sections.

A. Caching Placement and Power Allocation Solutions

In this section, we determine the caching placement and power allocation at both the users and SBSs to maximize the throughput attained, where we assume that the cached content sets at the users and the SBSs (i.e. $\mathcal{F}^U$ and $\mathcal{F}^R$, respectively) are given. Optimization of these content sets is considered in the next section. More particularly, we will solve problem **P1**, given the predetermined content sets $\mathcal{F}^U$ and $\mathcal{F}^R$. However, the problem is still challenging to solve, since the throughput in the objective function given in (19) is a complex non-linear function of the optimization variables.

Given this observation, we have devised Alg. 1 for finding the solution of this optimization problem based on the particle swarm optimization (PSO) [32]. Let us define the following vectors $\mathbf{m}^U = \{m_i^U\}_{i\in\mathcal{F}^U}$; $\mathbf{m}^R = \{m_i^R\}_{i\in\mathcal{F}^R}$; $\mathbf{p}^U = \{p_i^U\}_{i\in\mathcal{F}^U}$; $\mathbf{p}^R = \{p_i^R\}_{i\in\mathcal{F}^R}$; $\mathbf{p}^P = \{P_D/\bar{P}_D, P_R/\bar{P}_R\}$. The variable vector is defined as $\mathbf{X} = \{\mathbf{m}^U, \mathbf{m}^R, \mathbf{p}^U, \mathbf{p}^R, \mathbf{p}^P, \alpha\}$, where the length of vector $\mathbf{X}$ is $K = 2\left(\left|\mathcal{F}^U\right| + \left|\mathcal{F}^R\right|\right) + 3$. So we aim for finding the solutions of $\mathbf{X}$, where $X_i \in [0,1]$, for maximizing the throughput $\mathcal{T}(\mathbf{X})$. Note that the solutions $\mathbf{X}^*$ can be translated to the optimal solutions of $\{\mathbf{P}_U^*, \mathbf{M}_U^*, \mathbf{P}_R^*, \mathbf{M}_R^*, P_D^*, P_R^*, \alpha^*\}$.

For brevity, we describe the PSO for our problem as follows. PSO is a swarm intelligence based technique inspired by the collective behavior of social swarms of bees or birds. Here, each single solution defined as a "particle" may be viewed as a "bird" in the search space. A swarm of these particles moves through the search space at a specified velocity in order to find an optimal position. The position of particle j is $\mathbf{X}_j = [X_{j1}, X_{j2}, \ldots, X_{jK}]$ and its velocity is $\mathbf{V}_j = [V_{j1}, V_{j2}, \ldots, V_{jK}]$, where j denotes particle j and K represents the number of unknown variables given above. Firstly, a group of random particles (solutions) is used for

Algorithm 1 OPTIMIZATION OF $\{m_i^U\}, \{m_i^R\}, \{p_i^U\}, \{p_i^R\}, P_D, P_R$

for each particle $j = 1, 2, \ldots, D$ **do**
  Initialize the position $\mathbf{X}_j$ with a random number $\in [0, 1]$, the velocity V_{ji} with a random number $\in [-v_{\max}, v_{\max}]$, and the particle's best known position $\mathbf{p}_j^{best} = \mathbf{X}_j$.
  if $\mathcal{T}\left(\mathbf{p}_j^{best}\right) > \mathcal{T}\left(\mathbf{g}_j^{best}\right)$ **then**
    Update the swarm's best known position $\mathbf{g}_j^{best} = \mathbf{p}_j^{best}$.
  end if
end for
repeat
  for each particle $j = 1, 2, \ldots, D$ **do**
    for $d = 1, 2, \ldots, N$ **do**
      Pick random numbers, ϵ_1 and $\epsilon_2 \in [0, 1]$. Update the velocity using Eq. (49).
    end for
    Update the particle's position $\mathbf{X}_j = \mathbf{X}_j + \mathbf{V}_j$.
    if $\mathcal{T}\left(\mathbf{p}_j^{best}\right) < \mathcal{T}\left(\mathbf{X}_j\right)$ **then**
      Update the particle's best known position $\mathbf{p}_j^{best} = \mathbf{X}_j$.
      if $\mathcal{T}\left(\mathbf{g}_j^{best}\right) < \mathcal{T}\left(\mathbf{p}_j^{best}\right)$ **then**
        Update the swarm's best known position $\mathbf{g}_j^{best} = \mathbf{p}_j^{best}$.
      end if
    end if
    $\mathbf{g}^{best} = \underset{\mathbf{g}_j^{best}}{\operatorname{argmax}} \mathcal{T}\left(\mathbf{g}_j^{best}\right)$.
  end for
until Convergence
$\mathbf{X}^* = \mathbf{g}^{best}$.

initializing the PSO and then the optimal solution is sought by updating the consecutive generations as follows. At every iteration, there are two "best" values: *1)* the particle's best known position, $\mathbf{p}^{best}$ is the position vector of the best solution (fitness) of this particle achieved so far; *2)* the swarm's best known position, $\mathbf{g}^{best}$ is the position vector of the global "best" solution, which is tracked by the particle swarm optimizer. Then, each particle is updated by using these two "best" values, i.e. the position and velocity of the particles as follows:

$$V_{ji} = \psi V_{ji} + c_1\epsilon_1\left(p_{ji}^{best} - X_{ji}\right) + c_2\epsilon_2\left(g_{ji}^{best} - X_{ji}\right), \quad (49)$$

$$X_{ji} = X_{ji} + V_{ji}, \quad (50)$$

where V_{ji} is the velocity of particle j and X_{ji} is the current solution (or position) of particle j; c_1, c_2 are positive constants (usually, $c_1 = c_2 = 2$); and ϵ_1, ϵ_2 are a pair of independent random variables with uniform distribution between 0 and 1, which are generated at every update for each individual dimension; ψ is the inertia weight, which shows the effect of the previous velocity vector on the new vector. We summarize this procedure in Alg. 1.

B. Optimization of Cached Content Sets

Recall from the previous section that the content sets of the users and of the SBSs are assumed to be fixed to optimize the transmit power parameters and the caching probability as well as the volume of each content at the users and SBSs. We propose low-complexity greedy algorithms for finding the solution for this problem (i.e. finding the optimal assignment sets of contents for the users and SBSs), which are described in Algs. 2 and 3 [40]. There are two main parts of the algorithms: *1)* the non-overlapping content algorithm; and *2)* the overlapping content algorithm. The detailed presentation of these algorithms is omitted due to the journal's strict space limits, which can be found in the technical report [40].

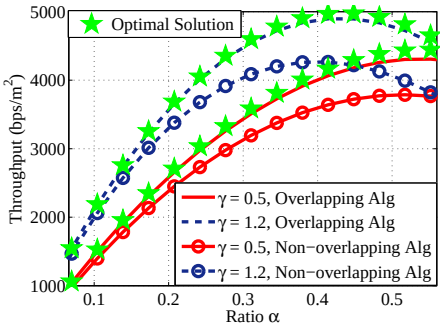

Fig. 4. Throughput per unit area vs ratio α for $S_U = 1200MB$, $S_R = 2400MB$, $\rho_R T = 0.1$, $\rho_U T = 10$, $\gamma = \{0.5, 1.2\}$, $\lambda_U = 10^{-2}$ (per m^2), $\lambda_R = 10^{-3}$ (per m^2), $\omega_1 = 0.75$ and $\omega_2 = 0.5$.

TABLE I
CPU TIME VS NUMBER OF CONTENTS N

N	10	20	30	40	50
CPU time (s)	18.1468	32.2392	37.0052	43.3206	51.1056

C. Complexity Analysis

We analyze the complexity of Alg. 2 and Alg. 3 in this subsection. Let us proceed by analyzing the steps taken in each iteration in Alg. 2. To determine the best assignment for the first content, we have to search over N contents for SBS and user sets, which involves $2N$ cases. Similarly, to assign the second content, we need to perform searching over $N-1$ contents for SBS and user sets (one content is already assigned in the first iteration). Hence, the second assignment involves $2(N-1)$ cases. Similar analysis can be applied for other assignments in later iterations. In summary, the total number of cases involved in assigning all contents for SBS and user sets is $2(N + \ldots + 2 + 1) = 2N(N+1)/2$, which is $O(N^2)$. This is the complexity of non-overlapping algorithm. To calculate the complexity of overlapping algorithm, we then perform h iterations for the Alg. 3, where h is the number of contents in the SBS and user sets, $h \leq N$. In the worst case, the complexity of Alg. 3 is $O(N)$. Considering the complexity of both algorithms, the complexity of overlapping algorithm is $O(N^2 + N) = O(N^2)$, which is much lower than that of the optimal brute-force search algorithm ($O(2^{2N})$). For Alg. 1, the convergence time is much improved, when the modification of the classical PSO is applied. One of the best way is to use the split-up in the cognitive behavior, which helps the particle remember both its previously visited best position and the worst position [32], [38]. By doing so, the computation time is reduced significantly, because the new scheme can effectively identify the valid region in the entire search space. Finally, the numerical results confirm that our proposed scheme has a low computational complexity as shown in Table I.

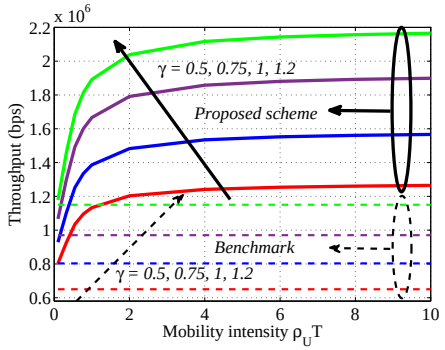

Fig. 5. Throughput vs user mobility intensity (user-user contact rate) and γ for $S_U = 1200MB$, $S_R = 2400MB$, $\rho_R T = 0.1$, $\omega_1 = 0.75$; $\omega_2 = 0.5$, $\lambda_U = 10^{-3}$ (per m^2), and $\lambda_R = 10^{-5}$ (per m^2).

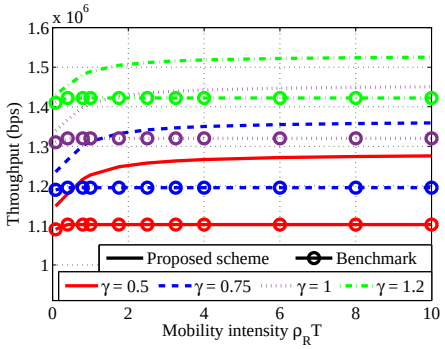

Fig. 6. Throughput vs user mobility intensity (user-SBS contact rate) and γ for $S_U = 1200MB$, $S_R = 2400MB$, $\rho_U T = 10$, $\omega_1 = 0.6$; $\omega_2 = 0.2$, $\lambda_U = 10^{-3}$ (per m^2), and $\lambda_R = 10^{-5}$ (per m^2).

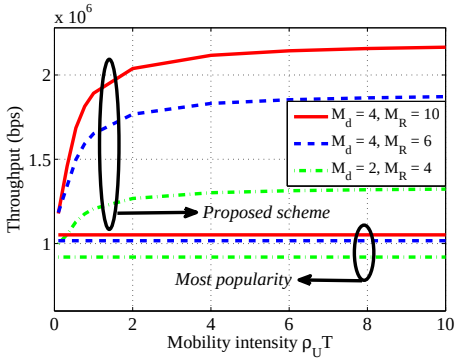

Fig. 7. Throughput vs user mobility intensity (user-user contact rate) and caching capacities of users and SBSs for $\rho_R T = 0.1$, $\gamma = 1.2$, $\omega_1 = 0.75$, $\omega_2 = 0.5$, $\lambda_U = 10^{-3}$ (per m^2), and $\lambda_R = 10^{-5}$ (per m^2).

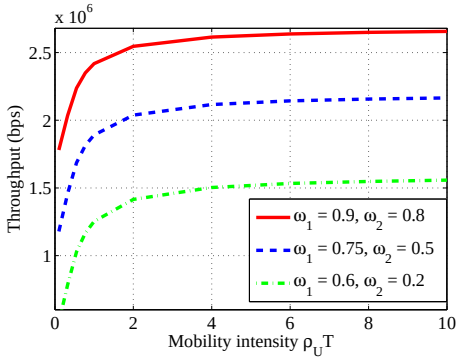

Fig. 8. Throughput vs user mobility intensity (user-user contact rate) and backhaul capacities for $S_U = 1200MB$, $S_R = 2400MB$, $\rho_R T = 0.1$, γ = 1.2, $\lambda_U = 10^{-3}$ (per m^2), and $\lambda_R = 10^{-5}$ (per m^2).

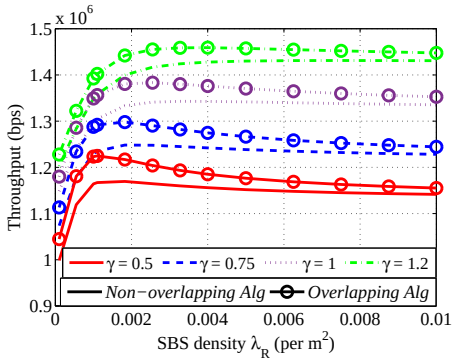

Fig. 9. Throughput vs SBS density for $S_U = 900MB$, $S_R = 1800MB$, $\rho_R T = 0.1$, $\rho_U T = 5$, $\gamma = \{0.5, 0.75, 1, 1.2\}$, $\lambda_U = 10^{-2}$ (per m^2), $\omega_1 = 0.75$ and $\omega_2 = 0.5$.

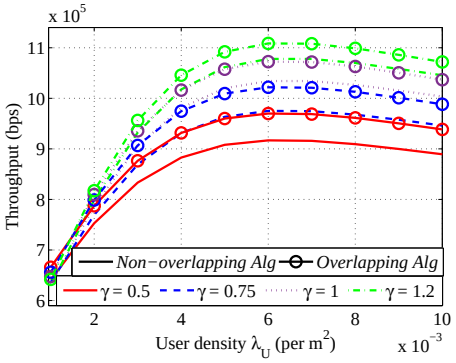

Fig. 10. Throughput vs user density for $S_U = 900MB$, $S_R = 1800MB$, $\rho_R T = 0.25$, $\rho_U T = 0.25$, $\gamma = \{0.5, 0.75, 1, 1.2\}$, $\lambda_R = 10^{-5}$ (per m^2), $\omega_1 = 0.75$ and $\omega_2 = 0.5$.

VI. Numerical Results

This section presents our numerical results for illustrating the throughput of the proposed resource allocation and caching placement. The key parameters of the proposed scheme are chosen as follows [39], unless stated otherwise: the portion of users requesting content is $\alpha = 0.5$; λ_U is in the range of $\left[10^{-4}, 10^{-3}\right]$ (per m^2); λ_R is in the range of $\left[10^{-6}, 10^{-5}\right]$ (per m^2); $\lambda_{BS} = 10^{-7}$ (per m^2); $L = 300MB$; $S_U = 1200MB$; $S_R = 2400MB$; N = 30; γ = 1.2; R_U = 15 m; R_R = 150 m; R_{BS} = 500 m; $\bar{P}_D$ = 0.15 mW; $\bar{P}_R$ = 1.5 mW; $\bar{P}_{BS}$ = 1 W; σ^2 = -140 dBm; ϕ = 0dB; $\xi = 0.01$; $W_u = 10MHz$; $\omega_1 = 0.75$; $\omega_2 = 0.5$.

Firstly, the impact of the ratio α on the throughput per unit area is illustrated in Fig. 4. The parameters are set as $S_U = 1200MB$, $S_R = 2400MB$, $\rho_R T = 0.1$, $\rho_U T = 10$, $\gamma = \{0.5, 1.2\}$, $\lambda_U = 10^{-2}$, $\lambda_R = 10^{-3}$, $\omega_1 = 0.75$ and $\omega_2 = 0.5$. We consider a pair of algorithms. The first one is our proposed scheme, where we use both Algs. 2 and 3 [40] for the

content assignments (we call it *Overlapping Alg* for brevity). The second one is termed as the *Non-overlapping Alg*, which does not use Alg. 3 for content assignments. The throughput gain of the *Overlapping Alg* over the *Non-overlapping Alg* is recorded in Fig. 4 for different values of α and γ. It is clearly observed that all the curves recorded for different γ values first increase to the maximum throughput and then decrease upon increasing the ratio α. This trend can be explained as follows. Here, we consider a dense network with $\lambda_U = 10^{-2}$ per m^2. When the ratio α is small, the number of requesting users is small but the number of transmitting users is large. Thus, not all the transmitting users need to serve requesting users. As the demand is less than the response, the total throughput per unit area is small. When α increases, the number of requesting users increases, while the number of transmitting users decreases. This helps to reduce the number of redundant transmitting users, hence the total throughput per unit area increases. When α reaches the optimal value α^* (for example, $\alpha^* = 0.44$ for $\gamma = 1.2$ in the *Overlapping Alg*), the content demand and the content response meet the balance and the throughput is maximized. If α continues to increase from this optimal α^*, the number of requesting users increases, while the number of active transmitting users decreases. A certain percentage of requesting users must download the requested contents from the nearby SBSs or the BS at the low rate. As a result, the total throughput per unit area decreases. Furthermore, if the number of requesting users is higher, there are more FD links helping the users in the case of cache miss events. Thus, the system-level co-channel interference and FD self-interference are substantially increased, which also contributes to the throughput degradation. We also demonstrate the efficacy of the proposed algorithms by comparing their throughput performance with those obtained by the optimal brute-force search algorithms. This figure confirms that our proposed algorithm achieves very close to the optimal solutions, while its complexity is much lower than that of the optimal brute-force search. This can be observed in Table I and in Section V-C.

In Fig. 4, the optimal solution α^* is in the range of $[0.35, 0.55]$. So from now on, we fix α to 0.5 and optimize the throughput per unit area with respect to the remaining parameters for simplicity. With a constant value of α, we can optimize the throughput per user $\mathcal{T}$ instead of the throughput per unit area $\mathcal{NT}$ since we have $\mathcal{T} = \mathcal{NT}/(\alpha\lambda_U)$.

Fig. 5 shows the throughput $\mathcal{T}$ versus the user-mobility intensity $\rho_U T$ for $S_U = 1200MB$, $S_R = 2400MB$, $\rho_R T = 0.1$, $\omega_1 = 0.75$; $\omega_2 = 0.5$, $\lambda_U = 10^{-3}$ (per m^2), $\lambda_R = 10^{-5}$ (per m^2), when varying γ. Here, we make a comparison with the most popular caching scheme [7], where users and SBSs prefer to cache the specific group of contents having the highest requesting probability. In particular, every user/relay caches the group of most popular files with probability one, and the remaining files with probability zero. In Fig. 5 we observe that the throughput increases when the user mobility becomes more intense (i.e. when $\rho_U T$ is higher). Furthermore, the throughput is more sensitive with respect to the content popularity skew γ. We can also see in Fig. 5 that our proposed scheme outperforms the most popular caching scheme [7]. Since the most popular contents are cached by SBSs and users in the popular caching policy, and both the SBSs and the users operate in an independent manner, the throughput gain is small. By contrast, our overlapping content assignment offers an excellent throughput gain, since both the SBSs and the users have more chance for cooperation and sharing. Furthermore, the proposed scheme fully exploits the content diversity and the mobility-induced diversity to achieve a throughput gain.

Similarly, the impact of user mobility on our caching strategy at different content popularity skew γ is demonstrated in Fig. 6. Here, we study the influence of the user-SBS contact rate $\rho_R T$, when we set the network parameters to $S_U = 1200MB$, $S_R = 2400MB$, $\rho_U T = 10$, $\omega_1 = 0.6$; $\omega_2 = 0.2$, $\lambda_U = 10^{-3}$, and $\lambda_R = 10^{-5}$ (per m^2). We have two observations from the throughput versus user mobility intensity in Figs. 5 and 6: *1)* when the users move at a low velocity, the throughput of the most popular caching strategy [7] is close to that of our proposed scheme; **2)** at higher content popularity skew γ, the gap between the most popular caching strategy and our proposed scheme becomes lower, but not equal to zero. This is because when the user mobility is low, there are less opportunities for the requesting user to jump to new caching SBSs and to contact new caching users. To glean the minimum number of coded segments required for successful decoding, every user and SBS would cache more segments (m_i^U and m_i^R) for the most popular contents i. However, our proposed PSO-aided scheme (Alg. 1) and our overlapping content assignment (Alg. 3 in [40]) mitigate this undesired impact. In particular, the overlapping content assignment algorithm coordinates the overlapping sets of contents both for the users and for the SBSs (i.e. $\mathcal{F}^U \cap \mathcal{F}^R$). Additionally, the PSO controls $\left(m_i^U, p^U, i\right)$ for the users and $\left(m_i^R, p_i^R\right)$ for the SBSs according to the set of contents obtained from the overlapping content assignment algorithm. Furthermore, the detrimental impact is negligible, when the user-mobility is high (even for the case, when the popularity content is skewed).

Fig. 7 shows the throughput $\mathcal{T}$ versus user mobility intensity parameterized by the caching capacities of the users and the SBSs. The parameter setting is as follows: $\rho_R T = 0.1$, $\gamma = 1.2$, $\omega_1 = 0.75$, $\omega_2 = 0.5$, $\lambda_U = 10^{-3}$, and $\lambda_R = 10^{-5}$ (per m^2). For ease presentation, we define the caching capacities (M_d, M_R) of the users and the SBSs as $M_d = S_U/L$, $M_R = S_R/L$. It is seen in Fig. 7 that when the cache capacities of users and SBSs increase, the throughput also increases, because the users and SBSs can more readily cooperate and share coded data. The proposed scheme contributes to the improved performance, because we harmonize the design of content-caching of both the users and of the SBSs. Furthermore, the power control techniques of the relays and the users help to reduce the interference footprints caused by the FD links and the self-interference caused by the power leakage at the FD relays.

The impact of (ω_1, ω_2) on the throughput is illustrated in Fig. 8. We consider the scenario having the same parameters, except for $S_U = 1200MB$, and $S_R = 2400MB$. The parameters (ω_1, ω_2) are set to $(\omega_1, \omega_2) = \{(0.6, 0.2), (0.75, 0.5), (0.9, 0.8)\}$, which correspond to low, medium and high backhaul as well as SBS-user link capa-

bilities. In Fig. 8, we compare the low and high backhaul as well as SBS-user link capabilities, where the former is a quarter of the latter. Somewhat unexpectedly, the throughput is not sorely degraded, because we employ optimized power allocation for FD communications for the case of cache miss, which mitigates the limited backhaul problem of the network. Furthermore, we beneficially exploit the user mobility, distributed storage and caching cooperation both amongst the users and the SBSs to achieve a high throughput gain.

The impact of SBS density on our proposed scheme is illustrated in Fig. 9. The scenario setting is as $S_U = 900MB$, $S_R = 1800MB$, $\rho_R T = 0.1$, $\rho_U T = 5$, $\gamma = \{0.5, 0.75, 1, 1.2\}$, $\lambda_U = 10^{-2}$, $\omega_1 = 0.75$ and $\omega_2 = 0.5$. Again, we consider a pair of algorithms, i.e. *Overlapping Alg* and *Non-overlapping Alg*. The throughput gain of the *Overlapping Alg* over the *Non-overlapping Alg* is recorded in Fig. 9 for different values of λ_R and γ. We can see that all the curves recorded for different γ values first increase to the maximum throughput and then decrease upon increasing the SBS density λ_R. We can explain this trend as follows. Here, we consider a dense network, where $\lambda_U = 10^{-2}$, and a low caching storage of $S_U = 900MB$, $S_R = 1800MB$. Due to the low caching storage both at the SBSs and the users, the requesting users must be served by several SBSs for the high content demand of dense networks. Otherwise, they would connect to the macro BSs to obtain the contents requested. So if the number of SBSs increases, the cache hit probability increases and hence a throughput gain is attained. However, when the number of SBSs reaches a specific threshold value, the throughput gain is not significantly improved. Furthermore, if the SBS density is higher, there are more FD links to help the users in the case of cache miss events. Thus, the system-level co-channel interference and FD self-interference are substantially increased. As a result, the throughput is degraded.

Similarly, in Fig. 10, we study the influence of user density on our proposed scheme. We set the same network parameters, except for $S_U = 900MB$, $S_R = 1800MB$, $\rho_R T = 0.25$, $\rho_U T = 0.25$, $\gamma = \{0.5, 0.75, 1, 1.2\}$, $\lambda_R = 10^{-5}$, $\omega_1 = 0.75$ and $\omega_2 = 0.5$. We also observe that the *Non-overlapping Algs* achieve noticeably lower throughput than that attained by their overlapping counterparts (i.e. *Overlapping Algs*). Again the curves recorded for different γ values first increase to the maximum throughput and then decrease upon increasing the user density λ_U. This is because if the number of users increases, then the requesting user has more opportunities to be helped by the others. Thus, the throughput is increased significantly. However, there is also a specific maximum number of the users for which the throughput gain becomes insignificant. Moreover, beyond a certain number of users, the requesting users must solicit assistance from the MBSs especially for a low cache storage. Hence both the interference and the FD self-interference are increased, since multiple links become active, which further decreases the throughput attained.

VII. Conclusions

We developed a framework of joint optimal resource allocation and mobility-aware probabilistic caching placement for HetNets relying on FD relays. In order to achieve a high throughput, we exploited the user mobility for caching at both SBSs and local users using the proposed the coded caching philosophy. We further developed a pair of content assignment algorithms for throughput maximization: *1)* the non-overlapping content assignment algorithm and *2)* the overlapping content assignment algorithm. Stochastic geometry based model was developed to analyze the throughput achieved by the overlapping content assignment algorithm. Due to the complexity of the analytical expressions, it remains an open and challenging task to obtain the optimal closed-form solutions. Instead, we developed low-complexity greedy algorithms to solve the optimization problem. The numerical results provide insights as to how and why significant performance gains can be achieved by the optimal configuration for the proposed scheme.

References

[1] R. Q. Hu, Y. Qian, "An energy efficient and spectrum efficient wireless heterogeneous network framework for 5G systems," *IEEE Commun. Mag.*, vol. 52 no.5, pp. 94-101, May 2014.

[2] L. Wei, R. Q. Hu, Y. Qian, G. Wu, "Enable device-to-device communications underlaying cellular networks: challenges and research aspects," *IEEE Commun. Mag.*, vol. 52, no. 6, pp. 90-96, June 2014.

[3] M. Cha, H. Kwak, P. Rodriguez, Y. Y. Ahn and S. Moon, "Analyzing the video popularity characteristics of large-scale user generated content systems," in *IEEE/ACM Trans. Netw.*, vol. 17, no. 5, pp. 1357-1370, Oct. 2009.

[4] P. Rodriguez, C. Spanner and E. W. Biersack, "Analysis of web caching architectures: hierarchical and distributed caching," in *IEEE/ACM Trans. Netw.*, vol. 9, no. 4, pp. 404-418, Aug. 2001.

[5] L. D. Xu, W. He and S. Li, "Internet of things in industries: A survey," in *Trans. Ind. Informat.*, vol. 10, no. 4, pp. 2233-2243, Nov. 2014.

[6] The Network, "Cisco delivers vision of fog computing to accelerate value from billions of connected devices," [Online]. Available: http://newsroom.cisco.com/press-release-content?articleId=1334100

[7] N. Golrezaei, A. F. Molisch, A. G. Dimakis and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," in *IEEE Commun. Mag.*, vol. 51, no. 4, pp. 142-149, April 2013.

[8] M. Ji, G. Caire and A. F. Molisch, "Fundamental limits of caching in wireless D2D networks," *IEEE Trans. Inf. Theory*, vol. 62, no. 2, pp. 849-869, Feb. 2016.

[9] M. Ji, G. Caire and A. F. Molisch, "Wireless device-to-device caching networks: Basic principles and system performance," in *IEEE J. Sel. Areas Commun.*, vol. 34, no. 1, pp. 176-189, Jan. 2016.

[10] Q. Li, W. Shi, X. Ge and Z. Niu, "Cooperative edge caching in software-defined hyper-cellular networks," in *IEEE J. Sel. Areas Commun.*, vol. 35, no. 11, pp. 2596-2605, Nov. 2017.

[11] M. M. Amiri and D. Gunduz, "Cache-aided content delivery over erasure broadcast channels," in *IEEE Trans. Commun.*, vol. 66, no. 1, pp. 370-381, Jan. 2018.

[12] M. M. Amiri, Q. Yang and D. Gunduz, "Decentralized caching and coded delivery with distinct cache capacities," in *IEEE Trans. Commun.*, vol. 65, no. 11, pp. 4657-4669, Nov. 2017.

[13] L. Zhang, M. Xiao, G. Wu and S. Li, "Efficient scheduling and power allocation for d2d-assisted wireless caching networks," in *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2438-2452, June 2016.

[14] J. Li, H. Chen, Y. Chen, Z. Lin, B. Vucetic and L. Hanzo, "Pricing and resource allocation via game theory for a small-cell video caching system," in *IEEE J. Sel. Areas Commun*, vol. 34, no. 8, pp. 2115-2129, Aug. 2016.

[15] J. Hu, L. L. Yang and L. Hanzo, "Energy-efficient cross-layer design of wireless mesh networks for content sharing in online social networks," in *IEEE Trans. Veh. Tech.*, vol. 66, no. 9, pp. 8495-8509, Sept. 2017.

[16] J. Li, Y. Chen, Z. Lin, W. Chen, B. Vucetic and L. Hanzo, "Distributed caching for data dissemination in the downlink of heterogeneous networks," in *IEEE Trans. Commun.*, vol. 63, no. 10, pp. 3553-3568, Oct. 2015.

[17] Y. Chen, M. Ding, J. Li, Z. Lin, G. Mao and L. Hanzo, "Probabilistic small-cell caching: Performance analysis and optimization," in *IEEE Trans. Veh. Tech.*, vol. 66, no. 5, pp. 4341-4354, May 2017.

[18] K. Wang, Z. Chen, and H. Liu, "Push-based wireless converged networks for massive multimedia content delivery," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2894–2905, May 2014.

[19] M. Gregori, J. Gómez-Vilardebo, J. Matamoros and D. Gunduz, "Wireless content caching for small cell and D2D networks," in *IEEE J. Sel. Areas Commun*, vol. 34, no. 5, pp. 1222-1234, May 2016.

[20] A. Sabharwal, P. Schniter, D. Guo, D. W. Bliss, S. Rangarajan, and R. Wichman, "In-band full-duplex wireless: Challenges and opportunities," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 9, pp. 1637-1652, Sep. 2014.

[21] Z. Zhang, K. Long, A. V. Vasilakos, and L. Hanzo, "Full-duplex wireless communications: challenges, solutions, and future research directions," *Proc. IEEE*, vol. 104, no. 7, pp. 1369-1409, Jul. 2016.

[22] U. Siddique, H. Tabassum and E. Hossain, "Downlink spectrum allocation for in-band and out-band wireless backhauling of full-duplex small cells," in *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3538-3554, Aug. 2017.

[23] L. T. Tan and L. B. Le, "Design and optimal configuration of full-duplex MAC protocol for cognitive radio networks considering self-interference," in *IEEE Access*, vol. 3, pp. 2715-2729, 2015.

[24] P. Semasinghe, E. Hossain and S. Maghsudi, "Cheat-proof distributed power control in full-duplex small cell networks: A repeated game with imperfect public monitoring," in *IEEE Trans. Commun.*, to appear, Dec. 2017.

[25] A. Shojaeifard, K. K. Wong, M. D. Renzo, G. Zheng, K. A. Hamdi and J. Tang, "Massive mimo-enabled full-duplex cellular networks," in *IEEE Trans. Commun.*, vol. 65, no. 11, pp. 4734-4750, Nov. 2017.

[26] H. Tabassum, A. H. Sakr and E. Hossain, "Analysis of massive mimo-enabled downlink wireless backhauling for full-duplex small cells," in *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2354-2369, June 2016.

[27] A. H. Sakr and E. Hossain, "On user association in multi-tier full-duplex cellular networks," in *IEEE Trans. Commun.*, vol. 65, no. 9, pp. 4080-4095, Sept. 2017.

[28] R. Wang, X. Peng, J. Zhang, and K. B. Letaief, "Mobility-aware caching for content-centric wireless networks: modeling and methodology," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 77–83, 2016.

[29] R. Wang, J. Zhang, S. H. Song, and K. B. Letaief, "Mobility-aware caching in D2D networks," in *IEEE Trans. Commun.*, vol. 16, no. 8, pp. 5001-5015, May 2017.

[30] X. Liu, J. Zhang, X. Zhang and W. Wang, "Mobility-aware coded probabilistic caching scheme for MEC-enabled smallcell networks," in *IEEE Access*, vol. 5, pp. 17824-17833, 2017.

[31] M. Haenggi, and R. K. Ganti, "Interference in large wireless networks," Foundations and Trends in Networking, vol. 3, no. 2, pp. 127-248, 2009.

[32] J. Kennedy, "Particle swarm optimization," *Encyclopedia of machine learning*, Springer US, pp. 760-766, 2011.

[33] H. B. Kong, P. Wang, D. Niyato and Y. Cheng, "Modeling and analysis of wireless sensor networks with/without energy harvesting using Ginibre point processes," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 3700-3713, March 2017.

[34] I. Flint, H. B. Kong, N. Privault, P. Wang and D. Niyato, "Analysis of heterogeneous wireless networks using Poisson hard-core hole process," in *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, pp. 7152-7167, Nov. 2017.

[35] H. B. Kong, I. Flint, P. Wang, D. Niyato and N. Privault, "Exact performance analysis of ambient RF energy harvesting wireless sensor networks with Ginibre point process," in *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3769-3784, Dec. 2016.

[36] M. Ji, G. Caire, and A. F. Molisch, "The throughput-outage tradeoff of wireless one-hop caching networks," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6833-6859, 2015.

[37] B. Banerjee and C. Tellambura, "Study of mobility in cache-enabled wireless heterogeneous networks," *Proc. IEEE WCNC2017*.

[38] A. I. Selvakumar and K. Thanushkodi, "A new particle swarm optimization solution to nonconvex economic dispatch problems," in *IEEE Trans. Power Syst.*, vol. 22, no. 1, pp. 42-51, Feb. 2007

[39] N. Lee, X. Lin, J. G. Andrews and R. W. Heath, "Power control for D2D underlaid cellular networks: Modeling, algorithms, and analysis," in *IEEE J. Sel. Areas Commun.*, vol. 33, no. 1, pp. 1-13, Jan. 2015.

[40] "Heterogeneous networks relying on full-duplex relays and mobility-aware probabilistic caching," technical report. Online: https://www.dropbox.com/s/ua0dh6m4zgjtdec/JSACTechreport.pdf?dl=0

Le Thanh Tan (S'11–M'15) received his B.Eng. and M.Eng. degrees from Ho Chi Minh University of Technology, Vietnam, in 2002 and 2004, respectively, and PhD degree from Institut National de la Recherche Scientifique, Canada in 2015. He is currently with Department of Electrical & Computer Engineering, Utah State University. His research interests include artificial intelligence, machine learning, Internet of Things, vehicular networks, 5G wireless communications, software defined networking, information centric networking and edge/fog/cloud computing. He has served on TPCs of different international conferences including IEEE ICC, CROWNCOM, VTC, PIMRC and etc.

Rose Qingyang Hu (S'95-M'98-SM'06) received the B.S. degree from the University of Science and Technology of China, the M.S. degree from New York University, and the Ph.D. degree from the University of Kansas. She is a currently a Professor with the Electrical and Computer Engineering Department and an Associate Dean for Research with the College of Engineering, Utah State University. Besides a decade academia experience, she has more than 10 years of R&D experience with Nortel, Blackberry, and Intel as a Technical Manager, a Senior Wireless System Architect, and a Senior Research Scientist, actively participating in industrial 3G/4G technology development, standardization, system level simulation, and performance evaluation. Her current research interests include next-generation wireless communications, wireless system design and optimization, Internet of Things, cloud computing/fog computing, wireless system modeling, and performance analysis. She has published extensively in top IEEE journals and conferences and holds numerous patents in her research areas. She is an IEEE Communications Society Distinguished Lecturer Class 2015-2018. She was a recipient of Best Paper Awards from the IEEE GLOBECOM 2012, the IEEE ICC 2015, the IEEE VTC Spring 2016, and the IEEE ICC 2016. She served as the TPC Co-Chair for the IEEE ICC 2018. She is currently serving on the editorial boards for the IEEE Transactions on Wireless Communications, the IEEE Transactions on Vehicular Technology, the IEEE Communications Magazine and the IEEE Wireless Communications.

Lajos Hanzo (http://www-mobile.ecs.soton.ac.uk) FREng, F'04, FIET, Fellow of EURASIP, received his 5-year degree in electronics in 1976 and his doctorate in 1983 from the Technical University of Budapest. In 2009 he was awarded an honorary doctorate by the Technical University of Budapest and in 2015 by the University of Edinburgh. In 2016 he was admitted to the Hungarian Academy of Science. During his 40-year career in telecommunications he has held various research and academic posts in Hungary, Germany and the UK. Since 1986 he has been with the School of Electronics and Computer Science, University of Southampton, UK, where he holds the chair in telecommunications. He has successfully supervised 119 PhD students, co-authored 18 John Wiley/IEEE Press books on mobile radio communications totalling in excess of 10 000 pages, published 1800+ research contributions at IEEE Xplore, acted both as TPC and General Chair of IEEE conferences, presented keynote lectures and has been awarded a number of distinctions. Currently he is directing a 60-strong academic research team, working on a range of research projects in the field of wireless multimedia communications sponsored by industry, the Engineering and Physical Sciences Research Council (EPSRC) UK, the European Research Council's Advanced Fellow Grant and the Royal Society's Wolfson Research Merit Award. He is an enthusiastic supporter of industrial and academic liaison and he offers a range of industrial courses. He is also a Governor of the IEEE ComSoc and VTS. He is a former Editor-in-Chief of the IEEE Press and a former Chaired Professor also at Tsinghua University, Beijing. For further information on research in progress and associated publications please refer to http://www-mobile.ecs.soton.ac.uk