Ultra Dense Edge Caching Networks with Arbitrary User Spatial Density

Emanuele Gruppi, *Student Member, IEEE*, Kai-Kit Wong, *Fellow, IEEE*, Mohammud Z. Bocus and Woon Hau Chin, *Member, IEEE*

***Abstract*—Cache-enabled small cells can be an effective solution to deliver contents to mobile users with much lower power and latency. While the trend for getting smaller and denser cells is clear, interference will soon become unmanageable and an obstacle when the number of content requests is massive. Moreover, content request is seldom a spatially homogeneous process due to physical impediments (e.g., buidings) and social activities, which makes resource allocation for content delivery more challenging. In this paper, we consider an ultra-dense network (UDN) in which content requests are served by cache-enabled access nodes which can either be active for delivering contents to users, or inactive to reduce interference and network energy consumption. Our aim is to devise an approach that can locally adapt the caching node density and content caching probabilities to accommodate any arbitrary user density and content request for maximizing the network's successful content delivery probability (SCDP). With a non-homogeneous spatial distribution for user equipments (UEs), we find that user-load, a parameter at the access node, plays a major role in the overall optimization. Simulation results illustrate that the proposed method can obtain superior performance against the considered benchmarks, with up to 150-160% increase, and our optimized solutions effectively adapt to the spatial-dependent user density.**

***Index Terms*—Content caching, Heterogeneous network, Small cell, Stochastic geometry, Ultra dense network.**

I. Introduction

MOBILE communications networks are required to meet a variety of key performance indicators (KPIs) such as capacity, latency, energy efficiency (EE), and etc. The fifth generation (5G) mobile networks will adopt several new technologies to deliver enhanced user experience [1].

One major hurdle in future-generation mobile communication networks is the backhaul congestion. To alleviate the burden at the backhaul link, content caching has emerged as an attractive solution. Content caching also brings the benefits of shorter distance between content servers and users, greatly reducing latency and energy consumption. This approach has often been investigated in the setting of multi-tier heterogeneous networks (HetNets), see e.g., [2]–[6], which makes it easy to distinguish among different kinds of small cell base stations (SBSs) by their radii. In HetNets, a macro base station (MBS) provides coverage over a large area while remote radio heads (RRHs) or SBSs[1] are responsible for delivering high-rate data to smaller and local areas. By deploying a denser network, it is possible to have short-range low-power and low-latency transmissions delivered by edge nodes [7], [8].

Ultra dense wireless networks are becoming reality as traffic lights, lamp posts, or drones can act as access nodes serving at the network edge. Having denser cells not only shortens the communication distance of each cell but also encourages spectrum reuse to increase capacity per unit area. The network edge however becomes the main source of interference as a single content transmission has to compete against a massive number of different requests to be successful. An information-theoretic approach that fully characterizes the asymptotic limits for edge node deployment remains an open problem.

A. Literature Review

Recent years have witnessed considerable efforts on content caching in HetNets. In [9], an optimal content caching policy to maximize the diversity gain under different base station (BS) coverage models was provided. In particular, the research has been carried out to focus on the maximization of the successful content delivery probability (SCDP) as the performance metric.

In [10], a probabilistic caching model has been proposed for homogeneously distributed cache-enabled edge nodes. The authors provided a closed-form expression to optimize the SCDP over the content caching probabilities under the noise-limited case while a lower bound was used for interference-limited scenarios. Edge node cooperation was not considered and users were assigned to nodes to which they experience the best channel quality. In contrast, [11] analyzed a multicasting network of cache-enabled BSs. However, the authors considered the user-load as not dependent from the set of content caching probabilities. In this article, we provide an improved analytical derivation of user-load and its dependence on the caching probability of all the contents in the wordbook is uncovered. On the other hand, content caching with collaborative content transmission was studied as a means of improving the SCDP in [12]. Moreover, [13] considered random caching in a two-tier HetNet with SBS cooperation, providing some guidelines on the design of caching probability. Unfortunately, a major deficiency is that the load at the BSs for simultaneous content delivery was ignored. In [14], the authors provided both a lower-bound and approximation for the SCDP for a set of cooperating nodes, with the intention to investigate a

This work is supported by an EPSRC CASE studentship with Toshiba TRL.

E. Gruppi and K. K. Wong are with the Department of Electronic and Electrical Engineering, University College London, UK (e-mail: {uceeegr, kai-kit.wong}@ucl.ac.uk).

M. Bocus and W. H. Chin are with Toshiba TRL, Bristol, UK (e-mail: {zubeir.bocus, woonhau.chin}@toshiba-trel.com).

[1]In this paper, the terms 'RRH' and 'SBS' are used interchangeably.

trade-off between cooperation gain and content diversity gain. Nonetheless, the number of cooperating edge nodes was only picked as a fixed constant and its random nature was discarded. In summary, the efforts in [12]–[14] attempted to examine the SCDP by conditioning their performance metrics over an arbitrarily chosen number of cooperating edge nodes. By basing the analysis on a fixed set of cooperating content providers, the ability to properly consider the network dynamics and observe the effects of edge node density is however obscured.

A common assumption of these studies is that homogeneous point processes for modelling users spatial distribution were considered. Cooperative transmission and content caching have been widely studied under homogeneity conditions for the information sources deployment but the case for spatially dependent densities is less understood [15]–[17]. User's spatial distribution normally does not follow a homogeneous point process, and is often spatially dependent due to manmade structures such as buildings and roads, and social events. In [18], the authors considered cooperative content transmission from non-homogeneously and identically distributed sources. However, the study was limited to Thomas cluster point processes and was not able to cope with any arbitrary geometric models. This motivates our work on edge caching networks that cope with any arbitrary user spatial distributions.

UDNs using cooperative transmission and content caching are the key enabler for content-centric mobile communications but as the number of content requests increases, interference becomes a serious bottleneck. Recent research in [19]–[23] investigated the impact of UE and BS density in the form of interference for UDNs. Specifically, the authors in [19] considered the use of higher frequency bands and network densification to improve the UE rates. They also showed the possibility of saving energy and reducing interference by idling some edge nodes in a UDN. Then [20] illustrated the increase in network capacity by having a denser SBS implementation. In [21], the authors accounted for the backhaul limitation to address the user's outage probability with homogeneous small cells, providing insights over the enhanced SCDP when adjusting the access nodes' density given a fixed content caching placement strategy. Most recently, in [22], queuing theory was employed to model the movement of UEs to distinct hot-spots and evaluate the throughput and EE performance. Though spatially dependent UE density was considered, homogeneous deployment of SBS was assumed within the same hot-spot. Cooperation was also not considered, and spatially dependent downlink interference was not analyzed. In other words, each user perceives interference only from a homogeneous Poisson point process (PPP) of BSs belonging to the same cluster.

B. Contributions

Different from the previous work, our objective is to design a cache-enabled UDN that can cope with any arbitrary non-homogeneous UE density on a global scale by locally adapting the cache-enabled RRH density and the content caching strategy for enhancing the SCDP. Our model is that cache nodes can be turned on or off to adjust the RRH density for an optimal trade-off between coverage and interference. In [15], a model for the generation of non-homogeneous user distributions was considered, by means of a quantized representation of the continuous space wherein the density of users is locally constant. We propose a similar network binning approach, with the intention of being able to characterize each network's bin by means of an edge node density and local content caching policy. In the literature, a random number of caching nodes was often considered as a conditioning factor for the evaluation of a network performance metric. After the optimization is performed, the random quantity is commonly picked deterministically, with its effects empirically studied. In this paper, we show that it is important to average the SCDP over this variable to achieve better performance under multiple aspects. Our analysis is different from the literature where the number of cooperating nodes is assumed fixed which fails to account for most of the network information. Also, one of our main contributions lies in the analytical derivation of user-load for a single or random set of cooperating nodes.

The contributions of this paper are summarized as follows:

- We consider a network model that discretizes the coverage area into a finite number of bins where the RRH density and the content caching probabilities are optimized to cooperatively deliver contents to a spatially non-homogeneous content request generation.
- We derive an SCDP lower-bound which can be adopted as a metric for performance maximization. The RRH edge-node density and content caching probability are optimized via the SCDP lower bound.
- We analytically study the statistics of the user-load to account for a random set of cooperating nodes, which plays a major role in the achievable SCDP. The insight also allows us to tailor the derivation to suit different situations such as single node transmission, non-homogeneous user density, probabilistic caching model, and so on.
- We adopt a steepest ascent algorithm to jointly optimize the RRH density and content caching probabilities over the entire network's space according to local content popularity and user density, based on the SCDP lower bound and the derived gradients. Simulation results demonstrate that the proposed algorithm achieves significant SCDP performance gain compared to conventional approaches.
- Different from previous studies, our results are not conditioned on the number of cooperative edge caching nodes. Instead, by averaging out the random number of cooperative edge caching nodes from the performance metric, we obtain a solution that depends on the RRH spatial density. This allows us to optimize the decision variables valid over all the possible realizations of caching nodes.
- We show that our solution of content caching probabilities and RRH density is spatially adaptive to the user density, highlighting the need to look for locally optimized solutions to further improve the SCDP performance.

The remainder of this paper is organized as follows. Section II presents our model and the problem statement. Our choice of performance metric and its derivation are provided in Section III. Then Section IV gives details of the proposed solution to the problem. Numerical results are provided in Section V.

Finally, we conclude the paper in Section VI.

II. Preliminaries

A. Network Model

A downlink ultra-dense small-cell wireless network[2] with short-range low-power cache-enabled SBS nodes is the subject of our investigation, which comprises of equipments, commonly referred to as fog access points (F-APs),[3] cache-enabled RRHs, SBSs or simply caching nodes. The locations of RRHs are modeled as the atoms of a homogeneous PPP Φ_S with intensity function $\bar{\lambda}^{\mathrm{S}}$. Under our model, RRHs are responsible for meeting most of the requests generated by the UEs, which follow an independent non-homogeneous PPP Φ_U , described by a continuous 2D intensity function $\lambda^{\mathrm{U}}(x, y)$. The point processes are defined in a 2D space $\mathbb{R}^2$ which describes the entire network space.

Time is slotted and we focus on a single time frame during which the UEs request for contents. Those requests that experience a hit-cache with the SBS-tier within the content searching area (CSA), can directly be processed by the network's edge. A more rigorous definition of CSA will be given in Section II-B. In case a request cannot be met within the CSA, it can be either forwarded to the MBS-tier to be processed by means of backhaul resources, or fetched by closer SBSs across the network. These cases are however at the cost of extra power consumption and latency. This paper focuses on the hit-cached cases, and it is assumed the interference pattern from across the network to be tailored on the local UE density. The required contents are all considered to have unit length and be drawn from a wordbook $\mathcal{F}$ with size $|\mathcal{F}| = F$ and $\hat{p}_f \in [0, 1]$ indicates the popularity of the f-th content. The content popularity coefficients are drawn from a zipf-like distribution, with the skewness factor υ such that $\sum_{i=1}^{F} \hat{p}_i = 1$. Without loss of generality, we consider the contents to be ordered as a decreasing sequence such that $\hat{p}_1 > \hat{p}_2 > \cdots > \hat{p}_F$. Therefore, the first content corresponds to the most popular content. Content popularity usually varies much slower than the density for content requests generation, and in particular, in this paper, the content popularity is assumed static, during which our analysis and optimization are carried out.

Transmission takes place over a channel whose small-scale fading follows the Rayleigh distribution, and the path loss is modeled by a factor $r^{-\alpha}$, where r denotes the communication distance and α is the pathloss exponent. The channel also suffers from additive white Gaussian noise (AWGN).

Each edge node has a cache of size M which can store and simultaneously deliver up to M distinct contents to UEs, when a hit-cache is experienced. The probability of the f-th content being cached at each edge node is denoted as $p_f \in [0, 1]$. Our caching model is regulated by the constraint $\sum_{f=1}^{F} p_f \leq M$.

The transmit power at every edge node, if active, is assumed to have unit power, i.e., $P^{\mathrm{S}} = 1$. Cooperative transmission of the same content from multiple caching nodes is adopted to exploit the cooperation gain. This approach requires little signalling between the cooperating edge nodes to work out the amount of bandwidth to perform the transmission. We allow each caching node to operate on the same frequency band of total bandwidth B [Hz]. A multicasting scheme for content dissemination is employed to perform multipoint-to-multipoint transmissions over the usable bandwidth. A frequency division multiple access (FDMA) scheme is adopted by each caching node to equally split the bandwidth into the number of distinct content requests (which is referred to as the user-load in this paper) for delivering. The user-load at an arbitrary node depends on (i) the probabilistic caching model, (ii) content popularity, (iii) UE density and (iv) RRH caching node density. This will be studied in the subsequent sections.

We assume that both the UE density $\lambda^{\mathrm{U}}(x, y)$ and content popularity $\hat{\boldsymbol{p}} = \{\hat{p}_1, \ldots, \hat{p}_F\}$ are known a-priori. The edge nodes are to be switched between an active or idle state to achieve certain RRH density, which is optimized according to the content popularity and the UE density, to balance between cooperation and diversity gain, energy consumption as well as interference. At the same time, probabilistic proactive caching is performed at the active RRHs for further improving the SCDP. To derive an achievable lower-bound of the target SCDP, we condition our performance metric on the existence of at least one edge node that have cached the required content. Further details will be provided in Section III-C.

B. UE Non-Homogeneity

To best cater for the non-homogeneous nature of UE distribution, we partition the entire space of network coverage into square-shaped bins. The center of each bin is regarded as a representative user for that bin, with its coordinates (x_n, y_n). We refer to this location as a user candidate location (UCL), a reference for all the users within the same bin in terms of user density dependent parameters such as signal-to-interference plus noise ratio (SINR) and user-load at the set of cooperating edge nodes. It is noted that although hexagonal shapes are usually adopted for tessellation to mimic circular coverage of radio signals, the use of square-shaped bins is chosen for simplicity. In addition, we define the CSA for a UCL as the squared space of side $2d$ over which a content is requested. It is assumed that the nodes within the CSA are the possible content providers for the reference UCL whereas the nodes outside, if active, cause interference. The CSA also serves to provide the boundary for cooperative transmission.[4]

For a given UCL, (x_n, y_n), the average number of UEs within its CSA, $\mathcal{D}_n$, is given by

$$U_n = \int \int_{(x,y)\in\mathcal{D}_n} \lambda^{\mathrm{U}}(x, y)\, dxdy. \tag{1}$$

Therefore, given the area of each bin, denoted by, $\mathrm{Area}(\mathcal{D}_n) =$

[2]In HetNets, MBSs are present to provide coverage with the aid of SBSs, forming a multi-tier structure. In this paper, the inclusion of MBS is omitted in our problem formulation for simplicity but some discussion will be provided in Section III-E to extend our work to HetNets with MBSs.

[3]By the term 'fog', we indicate a network architecture that adopts near-user edge devices to carry out a significant amount of storage and communication.

[4]In [15], similar network binning was used to emulate a non-homogeneous point process of UEs. Under this approach, local probabilistic caching model, caching node density, content popularity and all the statistical parameters involved are the same for all the UEs within the CSA of a reference user.

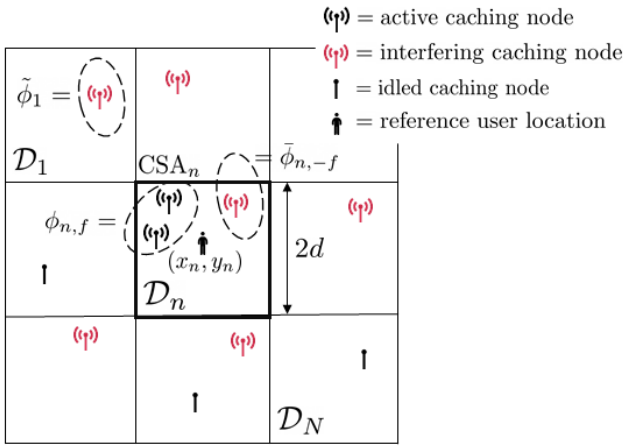

Fig. 1: The grid-based discretization of the space, where the sets $\phi_{n,f}$, $\bar{\phi}_{n,-f}$ and $\tilde{\phi}_i$ will be defined in Section II-D.

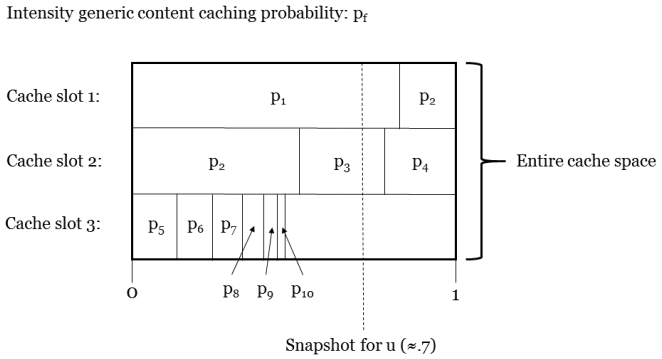

Fig. 2: Content caching without replacement with $F = 10$ and $M = 3$. In this example, the cache contents are 1 and 3.

$4d^2$, we can use

$$\lambda_n^{\mathrm{U}} = \left(\frac{U_n}{\text{Area}(\mathcal{D}_n)} \right) = \frac{U_n}{4d^2} \tag{2}$$

to approximate any arbitrary 2D continuous UE density function at the n-th network bin.

In doing so, the discretization of the entire coverage area into bins can be easily applied to compute any global performance metric of the entire network. Based on the discretized network space, the PPPs for the RRHs Φ_S and UEs Φ_U are modeled as a homogeneous PPP within each network bin.

The grid-based discretization of the coverage space with a representative n-th UCL and CSA is shown in Fig. 1. Focusing on the highlighted UCL, we see that the edge nodes belonging to other squares are not included in the content transmission but considered as the source of interference, if active. This shows that it may be desirable to move some edge nodes to an idle state to reduce interference. In this work, a CSA corresponds to a single network bin for simplicity sake.

C. Content Caching

To make the best out of the available resources, a probabilistic content caching model without replacement is adopted, and the proposed content caching placing process in [9] will be used. At each access node, the whole cache space $M \in \mathbb{N}_+$ is split into unit-lengthed chunks, and subsequently filled by the intensity of each probability p_i. Fig. 2 shows such setup when $M = 3$, where $\sum_{f=1}^{F} p_f < M$. To create a snapshot of the cache, a uniform random variable is drawn, i.e., $u \sim \mathcal{U} \in [0, 1]$ and the intersection between u and the cached probabilities determines the set of stored contents.

D. Content Delivery

When a content request is initiated from a UCL, the request is extended over its CSA. Those edge nodes inside the CSA that will experience a hit-cache are in charge of processing the request. Note that the number of cooperating nodes is a random variable whose probability mass function (pmf) follows a Poisson distribution. To keep the complexity low at the RRHs, non-coherent joint transmission will be considered from the edge nodes. This cooperation technique has been widely used in HetNets due to its low complexity and limited communication overhead [13]. We therefore consider channel state information at the receiver (CSIR) to be available for coherent detection of cooperative content delivery while the channel state information at the transmitter (CSIT) is not exploited. Since we are dealing with a random number of cooperating edge nodes, managing the estimation of the channel knowledge for a random number of cooperative transmitters would be highly complex and less practical. Our results can be interpreted as a performance lower-bound for coherent cooperative transmission. Assuming a user at the UCL (x_n, y_n) asking for content f and having unit-power transmissions at the edge nodes, the received signal power for that user can be expressed as

$$P_{n,f} = \left| \sum_{k \in \phi_{n,f}} h_{n,k} r_{n,k}^{-\frac{\alpha}{2}} \right|^2 + \sum_{\bar{k} \in \bar{\phi}_{n,-f}} \left| h_{n,\bar{k}} \right|^2 r_{n,\bar{k}}^{-\alpha} + \sum_{i \neq n} \omega_i \sum_{\tilde{k} \in \tilde{\phi}_i} \left| h_{i,\tilde{k}} \right|^2 r_{i,\tilde{k}}^{-\alpha} + W, \tag{3}$$

where $\phi_{n,f}$ denotes the set of cooperating edge nodes serving the requested content, $\bar{\phi}_{n,-f}$ denotes the set of interfering edge nodes transmitting different contents within the same CSA, and $\tilde{\phi}_n \triangleq \phi_{n,f} \cup \bar{\phi}_{n,-f}$. Thus, $\tilde{\phi}_i$ for $i \neq n$ corresponds to the set of all the interfering edge nodes outside the CSA of interest. Note that the interference originated from a given edge node outside the CSA is present only if that edge node is active and there is at least one UE within its CSA. As a consequence, we scale the interference power from the i-th CSA by the probability $\omega_i = 1 - e^{-\lambda_i^{\mathrm{U}} 4d^2}$. This allows to model the interference pattern to adapt to the spatially dependent user density. For simplicity sake, all the edge nodes that have not cached the required content within the CSA are considered as interferers. Furthermore, $h_{n,k}$ denotes the fading channel coefficient which is assumed to follow a circularly symmetric zero-mean unit-variance complex Gaussian distribution, $r_{n,k}$ is the distance between the k-th edge node and the reference user at the n-th UCL such that $r_{n,k}^{-\alpha} = \max(d_0, r_{n,k}^{-\alpha})$ where

$d_0 = 1$ is the reference distance, and W is the noise power.

Although the network is initially given as a homogeneous PPP of RRHs with intensity function $\bar{\lambda}^{\mathrm{S}}$, we aim to adapt the regional intensity function for each UCL, i.e., to have λ_n^{S} for UCL (x_n, y_n), so that the network performance can be maximized. For the PPPs $\phi_{n,f}$, $\bar{\phi}_{n,-f}$, $\widetilde{\phi}_i (i \neq n)$, we have the following intensity measures:[5]

$$\Lambda(\phi_{n,f}) = \mathbb{E}\left[\left|\phi\left(\lambda_n^{\mathrm{S}} p_{n,f} 4d^2\right)\right|\right] \equiv \mathbb{E}\left[\left|\phi\left(\mu_{n,f}\right)\right|\right], \tag{4a}$$

$$\Lambda(\bar{\phi}_{n,-f}) = \mathbb{E}\left[\left|\bar{\phi}\left(\lambda_n^{\mathrm{S}}(1-p_{n,f}) 4d^2\right)\right|\right] \equiv \mathbb{E}\left[\left|\bar{\phi}\left(\bar{\mu}_{n,f}\right)\right|\right], \tag{4b}$$

$$\Lambda(\widetilde{\phi}_i) = \mathbb{E}\left[\left|\widetilde{\phi}\left(\lambda_i^{\mathrm{S}} 4d^2\right)\right|\right] \equiv \mathbb{E}\left[\left|\widetilde{\phi}\left(\tilde{\mu}_i\right)\right|\right], \text{for } i \neq n. \tag{4c}$$

From (4c), note that the interference from across the network is not influenced by its local probabilistic content caching model but only by the cardinality of the sets $\widetilde{\phi}_i$. From (3), we can write the SINR for the reference user at UCL n as

$$\gamma_{n,f} = \frac{\left|\sum_{k\in\phi_{n,f}} h_{n,k} r_{n,k}^{-\frac{\alpha}{2}}\right|^2}{\underbrace{\sum_{\bar{k}\in\bar{\phi}_{n,-f}} \left|h_{n,\bar{k}}\right|^2 r_{n,\bar{k}}^{-\alpha} + \sum_{i\neq n} \omega_i \sum_{\tilde{k}\in\tilde{\phi}_i} \left|h_{i,\tilde{k}}\right|^2 r_{i,\tilde{k}}^{-\alpha} + W}_{\text{Interference } \mathcal{I}_{n,f}}}. \tag{5}$$

A successful content delivery is deemed to occur if a target rate for transmission ρ is achieved. That is,

$$\mathcal{E}_{n,f} \triangleq \left\{\frac{B}{\Xi_{n,f}} \log_2(1+\gamma_{n,f}) \geq \rho\right\} = \left\{\gamma_{n,f} \geq 2^{\frac{\rho\Xi_{n,f}}{B}} - 1\right\}, \tag{6}$$

in which B denotes the whole available bandwidth, and $\Xi_{n,f}$ indicates the perceived load at the set of cooperating edge nodes, respectively.

It is possible that the achievable rate far exceeds the target rate. In this case, the interference caused by those edge nodes can be reduced by turning off some nodes while the target rate is still met. The benefit is twofold as both interference and power consumption can be reduced. This is one of the intuitions of this work that attempts to adapt the spatial intensity of the RRHs by selectively idling some edge nodes.

III. Performance Metric

Considering the event (6), we can express the SCDP conditioned on a large number of network parameters as (7) (see top of next page).

To better model the SCDP metric, we first focus on the desired signal power term assuming $K_{n,f} \triangleq |\phi_{n,f}|$ cooperating edge nodes. Thus, we have

$$Z \triangleq \left|\sum_{k=1}^{K_{n,f}} h_{n,k} r_{n,k}^{-\frac{\alpha}{2}}\right|^2 = |X+iY|^2 = X^2 + Y^2, \tag{8}$$

where X and Y each follow $\mathcal{N}(0, \sigma_{n,f}^2 \equiv \frac{1}{2}\sum_{k=1}^{K_{n,f}} r_{n,k}^{-\alpha})$. It follows that

$$Z \sim \sigma_{n,f}^2 \mathcal{X}_2^2 \sim 2\sigma_{n,f}^2 \exp(1), \tag{9}$$

[5]The expected values are made on stochastic point processes. Therefore, the mean is made over the random location of the atoms within the area over which the process is defined and over the number of atoms of the process.

where $\mathcal{X}_2^2$ refers to the Chi-squared distribution with 2 degrees of freedom, and $\exp(1)$ is the standard exponential distribution. As such, the probability of the event $\mathcal{E}_{n,f}$ to occur is

$$\begin{aligned}
&\Pr(\mathcal{E}_{n,f}) \\
&= \Pr\left(\bar{Z} \geq \frac{1}{\sigma_{n,f}^2} \underbrace{\frac{(\mathcal{I}_{n,f}+W)}{2}(2^{\frac{\rho\Xi_{n,f}}{B}}-1)}_{\widetilde{\rho}} \,\middle|\, \begin{matrix}\mathcal{I}_{n,f},\\ \sigma_{n,f}^2,\\ \Xi_{n,f}\end{matrix}\right) \\
&= \mathbb{E}\left[\int_{\frac{\widetilde{\rho}}{\sigma_{n,f}^2}}^{\infty} e^{-x}\, dx \,\middle|\, \widetilde{\rho}, \sigma_{n,f}^2\right] = \mathbb{E}\left[e^{-\frac{\widetilde{\rho}}{\sigma_{n,f}^2}} \,\middle|\, \widetilde{\rho}, \sigma_{n,f}^2\right],
\end{aligned} \tag{10}$$

where $\bar{Z}$ corresponds to the standard exponentially distributed random variable. The SCDP of a single UCL can be found by averaging the expression (10) over the indicated conditioning variables and combining the results for all contents $f \in \mathcal{F}$. Unfortunately, no closed-form expression can be obtained for the averaged result. For this reason, we resort to Jensen's inequality that leads to

$$\begin{aligned}
&\mathrm{SCDP}_{n,f} \\
&= \mathbb{E}_{\mathcal{I}_{n,f},\sigma_{n,f}^2,\Xi_{n,f}}\left[\exp\left(-\frac{(2^{\frac{\rho\Xi_{n,f}}{B}}-1)}{2}\frac{(\mathcal{I}_{n,f}+W)}{\sigma_{n,f}^2}\right)\right] \\
&\geq \exp\left(-\mathbb{E}_{\mathcal{I}_{n,f},\sigma_{n,f}^2,\Xi_{n,f}}\left[\frac{(2^{\frac{\rho\Xi_{n,f}}{B}}-1)}{2}\frac{(\mathcal{I}_{n,f}+W)}{\sigma_{n,f}^2}\right]\right).
\end{aligned} \tag{11}$$

Before we evaluate the lower bound of SCDP for the f-th content given by Jensen's inequality in (11), we look at several important parameters of our model.

A. The Scaled Probabilistic Content Caching Model

When a given content f is being considered, it is assumed that a hit-cache has occurred. The knowledge of the realization of this event conditions the content caching probabilities at the set of cooperating edge nodes. Given that the generic f request has been cached, and the constraint over the cache size $\sum_{f=1} p_f \leq M$, the set of content caching probabilities becomes

$$p_{f'|f} = \begin{cases} p_{f'} \times \left(\frac{\sum_{\bar{f}=1}^{F} p_{\bar{f}} - \eta - 1}{\sum_{\bar{f}=1}^{F} p_{\bar{f}} - \eta - p_f}\right) & \text{if } f' \neq f, \\ 1 & \text{if } f' = f, \end{cases} \tag{12}$$

where η denotes the number of contents with caching probability of one. The scaling factor, introduced in (12), is important to correct the caching probabilities when content f is considered cached and the whole set of content caching probabilities changes accordingly. This adopted scaled version of the set of content caching probabilities is necessary to correctly evaluate the user-load and hence the effective bandwidth used by the set of cooperating nodes. Note that we have omitted the location index n as the discussion is the same regardless of the UCL. In the sequel, unless otherwise stated, the reference UCL 0 will be considered.

$$\Pr(\mathcal{E}_{n,f}) = \Pr\left(\frac{\left|\sum_{k\in\phi_{n,f}} h_{n,k} r_{n,k}^{-\frac{\alpha}{2}}\right|^2}{\sum_{\bar{k}\in\bar{\phi}_{n,-f}} |h_{n,\bar{k}}|^2 r_{n,\bar{k}}^{-\alpha} + \sum_{i\neq n} \omega_i \sum_{\tilde{k}\in\tilde{\phi}_i} |h_{n,\tilde{k}}|^2 r_{n,\tilde{k}}^{-\alpha} + W} \geq 2^{\frac{\rho\Xi_{n,f}}{B}} - 1 \,\middle|\, \begin{array}{l} \{r_{n,k}\}_{\forall k}, \{h_{n,k}\}_{\forall k}, \\ \phi_{n,f}, \bar{\phi}_{n,-f}, \{\tilde{\phi}_i\}_{i\neq n}, \Xi_{n,f} \end{array} \right) \quad (7)$$

B. User Load with K Cooperating Caching Nodes

In our model, it is assumed that the set of cooperating nodes has access to operate over the whole available bandwidth B when delivering content f. However, the effective bandwidth usage for the single content transmission depends on the user-load experienced by the whole set of edge nodes that cooperate to perform the transmission. If we now define the user-load for a generic active edge node k as $\xi_k \in \{1, 2, \ldots, M\}$ which represents the number of distinct simultaneous contents to be delivered by that node, then the amount of bandwidth that can be used to deliver a single content for that node would be given by B/ξ_k. If there are multiple edge nodes cooperatively delivering the same content, according to our joint transmission approach, all the cooperative edge nodes will need to occupy a common portion of bandwidth to deliver the content to the UE.[6] Specifically, let us say that $\mathcal{C}$ is the set of some cooperating edge nodes of interest. Then for $k \in \mathcal{C}$, those cooperating nodes should use

$$\frac{B}{\max_{k\in\mathcal{C}} \xi_k} \equiv \frac{B}{\Xi} \text{ bandwidth} \quad (13)$$

to deliver the same requested content. Note that the subscripts n and f have been dropped here for conciseness. Clearly, Ξ is a random variable and we can work out

$$\begin{aligned} \Pr(\Xi = m) &= \Pr(\Xi \leq m) - \Pr(\Xi \leq m-1) \\ &= F_{\xi_k}(m)^K - F_{\xi_k}(m-1)^K, \end{aligned} \quad (14)$$

where $F_{\xi_k}(m)$ stands for the cumulative density function (cdf) of the user-load for a single generic edge node ξ_k. Also, it is known that for discrete random variables, we can write

$$F_{\xi_k}(M) = \sum_{m=1}^{M} \Pr(\xi_k = m). \quad (15)$$

As such, if we can obtain the pmf of ξ_k, then we will be able to derive the pmf of the user-load for a set of collaborating edge nodes Ξ. As our discussion is always based on the condition that the f-th content is cached, ξ_k is certainly at least one. To derive $\Pr(\xi_k = m)$, we first define the sets of indices for hit-cache contents as ζ, missed-cache contents as $\bar{\zeta}$ and not-cached contents as $\tilde{\zeta}$. Considering a generic $\xi_k = m$ with $m > 1$, it means that $m-1$ contents from the set $\mathcal{F} \setminus f$ contribute to the user-load. These contents belong to the set ζ and the total number of possible index combinations belonging to this set is

$$\binom{|\mathcal{F} \setminus f|}{m-1} = \binom{F-1}{m-1}. \quad (16)$$

[6]A centralized approach is considered and we assume that resource allocation is suitably done to ensure that the same portion of bandwidth is used by all the cooperating edge nodes for delivering the same content.

Given an instance of ζ, we can write that the total number of possible index combinations for $\bar{\zeta}$ stands as

$$\binom{|\mathcal{F} \setminus \{f, \zeta\}|}{M-m} = \binom{F-m}{M-m}, \quad (17)$$

with the remaining elements $\mathcal{F} \setminus \{f, \zeta, \bar{\zeta}\}$ that define the indices in $\tilde{\zeta}$. Note that when computing the pmf of ξ_k for $m = M$ (i.e., all the contents in the cache contribute to the user-load), the corresponding set $\bar{\zeta}$ would be empty. Similarly, when computing the pmf for $m = 1$ (i.e., no contents in the cache except f contribute to the user-load), we would consider an empty ζ. According to which of the three sets the content belongs to, for a generic $\tilde{f} \in \mathcal{F} \setminus f$, we have, at UCL 0, the probabilities:

$$\begin{aligned} \text{hit-cache} &\rightarrow \left(1 - e^{\lambda_0^{\mathrm{U}} \hat{p}_{\tilde{f}} 4d^2}\right) p_{\tilde{f}|f} \\ \text{missed-cache} &\rightarrow e^{\lambda_0^{\mathrm{U}} \hat{p}_{\tilde{f}} 4d^2} p_{\tilde{f}|f} \\ \text{not-cache} &\rightarrow 1 - p_{\tilde{f}|f} \end{aligned} \quad (18)$$

where $p_{\tilde{f}|f}$ is defined in (12), and $\hat{p}_f$ denotes the global content popularity for content f which is known a priori and is not location dependent. Clearly, $\sum_{f=1}^{F} \hat{p}_f = 1$. A visual representation of an example with $m = 2$ is provided in Fig. 3. As we can see, the shaded box is the sure-cached content f. Moreover, the cache line separates the cached contents from the not-cached contents while the user-load line is used to distinguish the hit-cached contents from the missed-cached contents. In this example, one content experiences a hit-cache $|\zeta| = m - 1 = 1$, one experiences a missed-cache $|\bar{\zeta}| = M - m = 1$ while the remaining contents are not cached $|\tilde{\zeta}| = F - M = 5$. The probability of occurrence of each possible index combination can be found by

$$\underbrace{\left(1 - e^{\lambda_0^{\mathrm{U}} \hat{p}_i 4d^2}\right) p_{i|f}}_{i\in\zeta} \times \underbrace{e^{\lambda_0^{\mathrm{U}} \hat{p}_j 4d^2} p_{j|f}}_{j\in\bar{\zeta}} \times \underbrace{\prod_{k\in\tilde{\zeta}} (1 - p_{k|f})}_{k\in\tilde{\zeta}}. \quad (19)$$

Summing up all the contribution of the combinations will give $\Pr(\xi_k = 2)$ for the example in Fig. 3. As a result, we can obtain the generic pmf of ξ_k as

$$\begin{aligned} \Pr(\xi_k = m) = \sum_{c(m)} \prod_{i\in\zeta_c} \left(1 - e^{\lambda_0^{\mathrm{U}} \hat{p}_i 4d^2}\right) p_{i|f} \times \\ \sum_{g(c(m))} \prod_{j\in\bar{\zeta}_g} e^{\lambda_0^{\mathrm{U}} \hat{p}_j 4d^2} p_{j|f} \prod_{k\in\tilde{\zeta}_g} (1 - p_{k|f}), \end{aligned} \quad (20)$$

where $c(m)$ specifies a combination of hit-cached indices such that $\xi_k = m$, $g(c(m))$ indicates a combination of indices as a function of $c(m)$ and the sets ζ_c, $\bar{\zeta}_g$ and $\tilde{\zeta}_g$ are defined as before except they are now specific to a given combination,

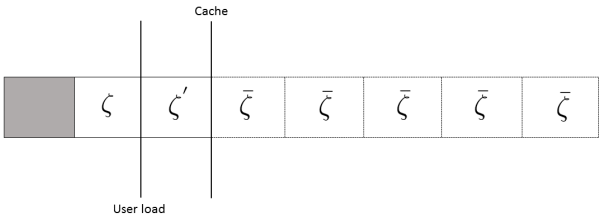

Fig. 3: A visual representation of the sets ζ, $\bar{\zeta}$, $\tilde{\zeta}$ for a wordbook and cache size of $F = 8$ and $M = 3$ and with user-load $\xi_k = 2$.

either $c(m)$ for ζ or $g(c(m))$ for $\bar{\zeta}_g$ and $\tilde{\zeta}_g$. By replacing (20) into (15), the pmf of the user-load at the set of cooperating edge nodes can be found from (14).

Note that because of the way we define $p_{f'}$ and $p_{f'|f}$ in the probabilistic caching model, the probability (20) may have a scaling issue, but this can be easily fixed by

$$\Pr(\xi_k = m) \leftarrow \frac{\Pr(\xi_k = m)}{\sum_{i=1}^{M} \Pr(\xi_k = i)}. \tag{21}$$

C. A Zero Truncated Poisson (ZTP) Distribution

As described, the number of edge nodes follows a Poisson distribution across the network. As a consequence, there exists a probability of not having any caching nodes at the network edge to perform the required transmission, and in this case the content will have to be fetched from the nearest MBS over the backhaul link. For a more complete analysis, this would mean that the latency as well as the power consumption for the MBS will need to be accounted for. Our objective in this paper is however on the benefits of using the cache-enabled edge nodes in terms of the SCDP. Thus, we focus on the lower bound of SCDP in (11) where the probability is conditioned on the fact that there is at least one edge node (i.e., $K \geq 1$) able to provide the required content, successfully or not.

Because the case $K = 0$ is not valid in our analysis, we adopt the ZTP distribution when accounting for the number of cooperating caching nodes, thus removing the case $K = 0$. Hence, given μ as the mean of the general Poisson distribution $f(k;\mu)$, the ZTP pmf, $g(k;\mu)$, can be expressed as

$$g(k;\mu) = \Pr(K = k | K > 0) = \frac{f(k;\mu)}{1 - f(0,\mu)} = \frac{\mu^k}{k!(e^{\mu} - 1)}, \tag{22}$$

where $\mu_{n,f} = p_{n,f}\lambda_n^{\mathrm{S}} 4d^2$ if the n-th CSA for content f is considered. The ZTP pmf will be useful when retrieving the unconditioned lower bound of the SCDP in (11).

D. The Objective Function

In order to come up with a global performance metric that can capture all the essential parameters of the caching network, we define $\mathcal{G}(\boldsymbol{\lambda}^{\mathrm{S}}, \boldsymbol{p})$ as the metric, based on the lower bound (11) and averaged over the random variables, given by

$$\mathcal{G}(\boldsymbol{\lambda}^{\mathrm{S}}, \boldsymbol{p}) = \sum_{n=1}^{N} \lambda_n^{\mathrm{U}} \sum_{f=1}^{F} \hat{p}_f \times$$
$$\exp\left(-\mathbb{E}_{\substack{\mathcal{I}_{n,f}, \\ \{r_{n,k}\}_{k\in\phi_{n,f}}, \\ K_{n,f}, \Xi_{n,f}}} \left[\frac{1}{2}\left(2^{\frac{\rho \Xi_{n,f}}{B}} - 1\right) \frac{\mathcal{I}_{n,f} + W}{\frac{\sum_{k=1}^{K_{n,f}} r_{n,k}^{-\alpha}}{2}} \right] \right), \tag{23}$$

where $\boldsymbol{\lambda}^{\mathrm{S}} \triangleq \{\lambda_1^{\mathrm{S}}, \lambda_2^{\mathrm{S}}, \ldots, \lambda_N^{\mathrm{S}}\}$ and $\boldsymbol{p} \triangleq \{p_{n,f}\}_{\forall n,f}$ are the network parameters to be optimized for maximizing the function $\mathcal{G}$. Note that the variables $\{\mathcal{I}_{n,f}\}$, $\{K_{n,f}\}$, $\{\Xi_{n,f}\}$, depend on the choices of $\boldsymbol{\lambda}^{\mathrm{S}}$ and $\boldsymbol{p}$. The index n in (23) indicates over which UCL the SCDP is computed.

Theorem 1. *The global performance metric, $\mathcal{G}(\boldsymbol{\lambda}^{\mathrm{S}}, \boldsymbol{p})$ in (23), permits the expression*

$$\mathcal{G}(\boldsymbol{\lambda}^{\mathrm{S}}, \boldsymbol{p}) = \sum_{n=1}^{N} \lambda_n^{\mathrm{U}} \sum_{f=1}^{F} \hat{p}_f \exp\left(-\varphi_{n,f}^{\mathrm{D}} \varphi_{n,f}^{\mathrm{I}}\right), \tag{24}$$

where $\varphi_{n,f}^{\mathrm{D}}$ and $\varphi_{n,f}^{\mathrm{I}}$, (27) and (28), respectively, are given at top of next page in which the functions $\mu_{n,f}$, $\bar{\mu}_{n,f}$ and $\tilde{\mu}_i$ have been defined earlier in (4), $\boldsymbol{p}_n \triangleq (p_{n,1}, \ldots, p_{n,F})$, $F_{\xi_{n,f}}(m)$ is given by (15) with the indices n and f re-inserted to the equation, and

$$J_n(t) \triangleq \frac{1}{4d^2}\left(\iint_{\mathcal{D}_n \setminus \mathcal{B}_0} e^{-\frac{t}{(x^2+y^2)^{\alpha/2}}}\, dxdy + \pi e^{-t} \right), \tag{25}$$

where $\mathcal{B}_0$ is the circle of unit radius centered at the n-th UCL.

Proof. See Appendix. □

E. Extension with MBS Sharing the Same Frequency Bands

To consider the presence of MBSs, it is necessary to add an independent term φ^{M} to $\varphi_{n,f}^{\mathrm{I}}$ in the argument of the exponential function in (24). The derivation of φ^{M} can be easily done following the steps in Appendix as

$$\begin{aligned}\varphi^{\mathrm{M}} &= \mathbb{E}_{r_b^{\mathrm{M}}, h_b^{\mathrm{M}}, \phi_b^{\mathrm{M}}}\left[\sum_{b=1}^{N_{\mathrm{M}}} \left|h_b^{\mathrm{M}}\right|^2 (r_b^{\mathrm{M}})^{-\alpha} \right] \\ &= \mathbb{E}_{r^{\mathrm{M}}, h^{\mathrm{M}}, N_{\mathrm{M}}}\left[N_{\mathrm{M}} \left|h^{\mathrm{M}}\right|^2 (r^{\mathrm{M}})^{-\alpha} \right] \\ &= \frac{\mu^{\mathrm{M}}}{\mathcal{D}}\left(\iint_{\mathcal{D}\setminus\mathcal{B}_0} (x^2+y^2)^{-\alpha/2}\, dxdy + \pi \right)\end{aligned} \tag{26}$$

where μ^{M}, r^{M} and h^{M} are, respectively, the average number of MBSs, the random link-distance and the channel fading coefficient. Note that the indices for location (i.e., n) is no longer needed because the same density for the MBSs is considered over the entire coverage area.

$$\varphi^{\mathrm{D}}_{n,f}(\lambda^{\mathrm{S}}_n, \boldsymbol{p}_n) = \frac{1}{e^{\mu_{n,f}}-1}\int_0^{\infty}\sum_{m=1}^{M}(2^{\frac{\rho m}{B}}-1)\left(e^{\mu_{n,f}F_{\xi_{n,f}}(m)J_n(t)} - e^{\mu_{n,f}F_{\xi_{n,f}}(m-1)J_n(t)}\right)dt, \tag{27}$$

$$\varphi^{\mathrm{I}}_{n,f}(\boldsymbol{\lambda}^{\mathrm{S}}, p_{n,f}) = \bar{\mu}_{n,f}\bar{J}(n) + \sum_{i\neq n}\omega_i\tilde{\mu}_i\widetilde{J}(i) + W, \tag{28}$$

$$\begin{aligned}\frac{\partial \mathcal{G}_n}{\partial p_{n,f}} = \lambda^{\mathrm{U}}_n \hat{p}_f \exp\left(-\varphi^{\mathrm{D}}_{n,f}\varphi^{\mathrm{I}}_{n,f}\right)\times \\ \Bigg[-\left(-\frac{\varphi^{\mathrm{I}}_{n,f}\frac{\partial\mu_{n,f}}{\partial p_{n,f}}e^{\mu_{n,f}}}{(e^{\mu_{n,f}}-1)^2}\int_0^{\infty}\sum_{m=1}^{M}(2^{\frac{\rho m}{B}}-1)\left(e^{\mu_{n,f}F_{\xi_{n,f}}(m)J_n(t)} - e^{\mu_{n,f}F_{\xi_{n,f}}(m-1)J_n(t)}\right)dt\right. \\ + \frac{\varphi^{\mathrm{I}}_{n,f}}{e^{\mu_{n,f}}-1}\int_0^{\infty}\sum_{m=1}^{M}(2^{\frac{\rho m}{B}}-1)\left[\begin{array}{c} J_n(t)\left(F_{\xi_{n,f}}(m)\frac{\partial\mu_{n,f}}{\partial p_{n,f}} + \mu_{n,f}\frac{\partial F_{\xi_{n,f}}(m)}{\partial p_{n,f}}\right)e^{\mu_{n,f}F_{\xi_{n,f}}(m)J_n(t)} \\ -J_n(t)\left(F_{\xi_{n,f}}(m-1)\frac{\partial\mu_{n,f}}{\partial p_{n,f}} + \mu_{n,f}\frac{\partial F_{\xi_{n,f}}(m-1)}{\partial p_{n,f}}\right)e^{\mu_{n,f}F_{\xi_{n,f}}(m-1)J_n(t)}\end{array}\right]dt \\ \left.+\frac{\varphi^{\mathrm{D}}_{n,f}}{4d^2}\frac{\partial\bar{\mu}_{n,f}}{\partial p_{n,f}}\left(\iint_{\mathcal{D}_n}(x^2+y^2)^{-\alpha/2}dxdy+\pi\right)\right)\Bigg] - \lambda^{\mathrm{U}}_n\sum_{\tilde{f}\neq f}\hat{p}_{\tilde{f}}\exp\left(-\varphi^{\mathrm{D}}_{n,\tilde{f}}\varphi^{\mathrm{I}}_{n,\tilde{f}}\right)\left[\frac{\varphi^{\mathrm{I}}_{n,\tilde{f}}}{(e^{\mu_{n,\tilde{f}}}-1)}\times\right. \\ \left.\int_0^{\infty}\sum_{m=1}^{M}(2^{\frac{\rho m}{B}}-1)\mu_{n,\tilde{f}}J_n(t)\left(\frac{\partial F_{\xi_{n,\tilde{f}}}(m)}{\partial p_{n,f}}e^{\mu_{n,\tilde{f}}F_{\xi_{n,\tilde{f}}}(m)J_n(t)} - \frac{\partial F_{\xi_{n,\tilde{f}}}(m-1)}{\partial p_{n,f}}e^{\mu_{n,\tilde{f}}F_{\xi_{n,\tilde{f}}}(m-1)J_n(t)}\right)dt\right]\end{aligned} \tag{29}$$

$$\frac{\partial F_{\xi_{n,\tilde{f}}}(m)}{\partial p_{n,f}} = \sum_{\bar{m}=1}^{m}\frac{\partial}{\partial p_{n,f}}\left[\sum_{c(\bar{m})}\prod_{i\in\zeta_c}\left(1-e^{\lambda^{\mathrm{U}}_n\hat{p}_i 4d^2}\right)p_{n,i|\tilde{f}}\times\sum_{g(c(\bar{m}))}\prod_{j\in\bar{\zeta}_g}e^{\lambda^{\mathrm{U}}_n\hat{p}_j 4d^2}p_{n,j|\tilde{f}}\prod_{k\in\tilde{\zeta}_g}(1-p_{n,k|\tilde{f}})\right] \tag{30}$$

$$\begin{aligned}\frac{\partial\mathcal{G}}{\partial\lambda^{\mathrm{S}}_n} = \lambda^{\mathrm{U}}_n\sum_{f=1}^{F}\Big[\hat{p}_f\exp\left(-\varphi^{\mathrm{D}}_{n,f}\varphi^{\mathrm{I}}_{n,f}\right)\times \\ \Bigg[-\varphi^{\mathrm{I}}_{n,f}\left(-\frac{\frac{\partial\mu_{n,f}}{\partial\lambda^{\mathrm{S}}_n}e^{\mu_{n,f}}}{(e^{\mu_{n,f}}-1)^2}\int_0^{\infty}\sum_{m=1}^{M}(2^{\frac{\rho m}{B}}-1)\left(e^{\mu_{n,f}F_{\xi_{n,f}}(m)J_n(t)} - e^{\mu_{n,f}F_{\xi_{n,f}}(m-1)J_n(t)}\right)dt\right. \\ \left.+\frac{1}{e^{\mu_{n,f}}-1}\int_0^{\infty}\sum_{m=1}^{M}(2^{\frac{\rho m}{B}}-1)\left[\begin{array}{c} J_n(t)F_{\xi_{n,f}}(m)\frac{\partial\mu_{n,f}}{\partial\lambda^{\mathrm{S}}_n}e^{\mu_{n,f}F_{\xi_{n,f}}(m)J_n(t)} \\ -J_n(t)F_{\xi_{n,f}}(m-1)\frac{\partial\mu_{n,f}}{\partial\lambda^{\mathrm{S}}_n}e^{\mu_{n,f}F_{\xi_{n,f}}(m-1)J_n(t)}\end{array}\right]dt\right) \\ -\frac{\varphi^{\mathrm{D}}_{n,f}}{4d^2}\frac{\partial\bar{\mu}_{n,f}}{\partial\lambda^{\mathrm{S}}_n}\left(\iint_{\mathcal{D}_n}(x^2+y^2)^{-\alpha/2}dxdy+\pi\right)\Bigg]\Big] + \sum_{\tilde{n}\neq n}^{N}\lambda^{\mathrm{U}}_{\tilde{n}}\sum_{f=1}^{F}\hat{p}_f\left[\exp\left(-\varphi^{\mathrm{D}}_{\tilde{n},f}\varphi^{\mathrm{I}}_{\tilde{n},f}\right)\left(-\varphi^{\mathrm{D}}_{\tilde{n},f}\omega_n\iint_{\mathcal{D}_n}(x^2+y^2)^{-\alpha/2}dxdy\right)\right]\end{aligned} \tag{31}$$

IV. The Proposed Approach

A. The Problem and Subproblems

In this paper, our objective is to maximize the global metric (23) (and hence (24)) by adapting the RRH density and the content caching probabilities for all the UCLs. That is,

$$(\mathbb{P}_0): \quad \underset{\boldsymbol{\lambda}^{\mathrm{S}},\boldsymbol{p}}{\text{maximize}} \quad \mathcal{G}(\boldsymbol{\lambda}^{\mathrm{S}},\boldsymbol{p}) \tag{32a}$$

$$\text{subject to} \quad \sum_{f=1}^{F}p_{n,f} \leq M \tag{32b}$$

$$0 \leq p_{n,f} \leq 1 \tag{32c}$$

$$0 \leq \lambda^{\mathrm{S}}_n \leq \bar{\lambda}^{\mathrm{S}}, \tag{32d}$$

where $\bar{\lambda}^{\mathrm{S}}$ denotes the upper limit of the caching node density. Note that while the initial upper-bound on the SBS intensity function is homogeneously considered, it can potentially be adapted as a non-homogeneous upper-bound as $\bar{\lambda}^{\mathrm{S}}_n$, with no changes to be made on our method. This allows operators to better mimic the existing initial SBS distribution and eventually investigate the benefit from introducing edge nodes.

The problem $(\mathbb{P}_0)$ needs some interpretation. We observe that the linear constraints for $(\mathbb{P}_0)$ are jointly independent. The problem can be decoupled as a combination of subproblems which can be solved via an iterative algorithm. Therefore, we decompose $(\mathbb{P}_0)$ in (32) into $N+1$ subproblems, where N is the total number of distinct network bins. In particular, $(\mathbb{P}_0)$ can be solved by repeatedly finding the solutions to N

problems for the local optimum content caching probability (one for each n), and the solution for the RRH density optimization problem, in an iterative fashion. We refer to the two kinds of subproblems $(\mathbb{P}_1)$ and $(\mathbb{P}_2)$, written as

$$(\mathbb{P}_1): \quad \underset{\boldsymbol{p}_n}{\text{maximize}} \quad \mathcal{G}_n(\boldsymbol{\lambda}^{\mathrm{S}}, \boldsymbol{p}) \equiv \tag{33a}$$

$$\lambda_n^{\mathrm{U}} \sum_{f=1}^{F} \hat{p}_f \exp\left(-\varphi_{n,f}^{\mathrm{D}} \varphi_{n,f}^{\mathrm{I}}\right) \tag{33b}$$

$$\text{subject to} \quad \sum_{f=1}^{F} p_{n,f} \leq M \tag{33c}$$

$$0 \leq p_{n,f} \leq 1, \tag{33d}$$

and

$$(\mathbb{P}_2): \quad \underset{\boldsymbol{\lambda}^{\mathrm{S}}}{\text{maximize}} \quad \mathcal{G}(\boldsymbol{\lambda}^{\mathrm{S}}, \boldsymbol{p}) \text{ subject to } 0 \leq \lambda_n^{\mathrm{S}} \leq \bar{\lambda}^{\mathrm{S}}. \tag{34}$$

Although Theorem 1 gives an expression to evaluate $\mathcal{G}(\boldsymbol{\lambda}^{\mathrm{S}}, \boldsymbol{p})$, a closed-form expression is not possible, and a steepest ascent gradient based method is used when searching for the maximizers. Also, note that the constraints for both $(\mathbb{P}_1)$ and $(\mathbb{P}_2)$ are convex sets. Ideally, it would be necessary to prove that the involved objective functions are concave so that the search of a local maxima would lead to the global maxima. For both $(\mathbb{P}_1)$ and $(\mathbb{P}_2)$ we are dealing with two continuous optimization of differentiable functions over a convex set. However, the study of the concavity of $\mathcal{G}_n(\boldsymbol{\lambda}^{\mathrm{S}}, \boldsymbol{p})$ and $\mathcal{G}(\boldsymbol{\lambda}^{\mathrm{S}}, \boldsymbol{p})$ is rather arduous. We target a stationary point for the problem $(\mathbb{P}_0)$ by means of a diminishing stepsize gradient-based maximization. The obtained results will thus be compared with the numerically found optimal solutions, to validate the proposed method.

To solve $(\mathbb{P}_0)$, the two subproblems $(\mathbb{P}_1)$ and $(\mathbb{P}_2)$ allow the gradients to be determined with respect to (w.r.t.) the decision variables $\{\lambda_n^{\mathrm{S}}\}$ and $\{p_{n,f}\}$. We propose to solve the two subproblems separately and iteratively to provide the joint solution. In particular, $(\mathbb{P}_1)$ addresses the probabilistic caching problem while $(\mathbb{P}_2)$ deals with the effects of caching node density at a global scale. The pseudocode of the proposed algorithm is given as Algorithm 1.

Algorithm 1 Alternating optimization for solving $(\mathbb{P}_0)$

initialize the iteration index $t = 1$
initialize $\tau =$ some large number
initialize $\boldsymbol{p}_n^{(t)}$ from a uniform content caching probability
initialize $\boldsymbol{\lambda}^{\mathrm{S},(t)} = \bar{\lambda}^{\mathrm{S}}$
initialize $\delta_p^{\text{init}}, \delta_\lambda^{\text{init}}, \beta, \epsilon_p, \epsilon_\lambda, \kappa$
while $\tau >$ some small threshold **do**
  for $n = 1$ to N **do**
    $\delta_p \leftarrow \delta_p^{\text{init}}$
    while $\delta_p \geq \epsilon_p$ **do**
      compute $\mathbf{g}_p = \frac{\nabla \mathcal{G}_n}{\|\nabla \mathcal{G}_n\|}$
      if $\mathcal{G}_n(\boldsymbol{p}_n^{(t)} + \delta_p \mathbf{g}_p) \geq \mathcal{G}_n(\boldsymbol{p}_n^{(t)}) + \delta_p \kappa \mathbf{g}_p^T \nabla \mathcal{G}_n$ **then**
        $\boldsymbol{p}_n^{(t+1)} \leftarrow \boldsymbol{p}_n^{(t)} + \delta_p \mathbf{g}_p$
      else
        $\delta_p \leftarrow \beta \delta_p$
      end if
    end while
  end for
  $\delta_\lambda \leftarrow \delta_\lambda^{\text{init}}$
  while $\delta_\lambda \geq \epsilon_\lambda$ **do**
    compute $\mathbf{g}_\lambda = \frac{\nabla \mathcal{G}}{\|\nabla \mathcal{G}\|}$
    if $\mathcal{G}(\boldsymbol{\lambda}^{(t)} + \delta_\lambda \mathbf{g}_\lambda) \geq \mathcal{G}(\boldsymbol{\lambda}^{(t)}) + \delta_\lambda \kappa \mathbf{g}_\lambda^T \nabla \mathcal{G}$ **then**
      $\boldsymbol{\lambda}^{\mathrm{S},(t+1)} \leftarrow \boldsymbol{\lambda}^{\mathrm{S},(t)} + \delta_\lambda \mathbf{g}_\lambda$
    else
      $\delta_\lambda \leftarrow \beta \delta_\lambda$
    end if
  end while
  update $\tau = \max\{\|\boldsymbol{\lambda}^{\mathrm{S},(t+1)} - \boldsymbol{\lambda}^{\mathrm{S},(t)}\|, \|\boldsymbol{p}_n^{(t+1)} - \boldsymbol{p}_n^{(t)}\|\}$
  $t = t + 1$
end while

B. Backtracking Line Search based Optimization

As seen in Algorithm 1, the backtracking line search with Armijo-Goldstein condition [24], [25] is employed when solving $(\mathbb{P}_1)$ or $(\mathbb{P}_2)$. In the search, the objective at each iteration is to find a step size δ which allows the following Armijo-Goldstein condition at the t-th iteration to be fulfilled

$$f(\mathbf{x}^{(t)} + \delta \mathbf{g}) \geq f(\mathbf{x}^{(t)}) + \delta \times \kappa \times \underbrace{\mathbf{g}^T \nabla f(\mathbf{x}_k^{(t)})}_{\text{local slope along direction } \mathbf{g}}, \tag{35}$$

where the superscript (t) is the iteration index, $\mathbf{g}$ is a unit vector computed in the direction where a local increase occurs and $\kappa = 10^{-4}$ is the control parameter which ensures the increment to be at least a fraction κ of the Taylor approximation of f at $\mathbf{x}$. In addition, $f(\cdot)$ and $\nabla f(\cdot)$ correspond to the objective function and its gradient, respectively.

The initial step-sizes for the probability and density maximization problems are chosen to be $\delta_p^{\text{init}} = 0.1$ and $\delta_\lambda^{\text{init}} = \frac{\bar{\lambda}^{\mathrm{S}}}{4}$, respectively. Note that $\bar{\lambda}^{\mathrm{S}}$ denotes the maximum RRH density of the network. The search terminates if a sufficiently small step-size is reached. In our simulations, we set the stopping thresholds to be $\epsilon_p = \frac{\delta_p^{\text{init}}}{20}$ and $\epsilon_\lambda = \frac{\delta_\lambda^{\text{init}}}{20}$. Also, at the t-th iteration, if the Armijo-Goldstein condition is not met, the step-size δ will be reduced by a factor $\beta = 0.8$; otherwise, the optimizing variables will be updated by $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \delta \mathbf{g}$.

To carry out the steepest ascent algorithm, we also need the expression for the gradient $\nabla f(\mathbf{x}_k)$. For the subproblem $(\mathbb{P}_1)$, we need to know $\frac{\partial \mathcal{G}_n}{\partial p_{n,f}}$, which after some lengthy derivations gives (29), where

$$\frac{\partial \mu_{n,f}}{\partial p_{n,f}} = \lambda_n^{\mathrm{S}} 4 d^2, \tag{36}$$

$$\frac{\partial \bar{\mu}_{n,f}}{\partial p_{n,f}} = -\lambda_n^{\mathrm{S}} 4 d^2, \tag{37}$$

and $\frac{\partial F_{\xi_{n,\tilde{f}}}(m)}{\partial p_{n,f}}$ is given by (30) in which $p_{n,i|\tilde{f}}$ has been defined in (12) with the index n re-insterted in the expression.

Similarly, the gradient $\nabla f(\mathbf{x}_k)$ for the subproblem $(\mathbb{P}_2)$ over

the n-th UCL, i.e., $\frac{\partial \mathcal{G}}{\partial \lambda_n^{\mathrm{S}}}$, writes as (31), where

$$\frac{\partial \mu_{n,f}}{\partial \lambda_n^{\mathrm{S}}} = p_{n,f} 4d^2, \tag{38}$$

$$\frac{\partial \bar{\mu}_{n,f}}{\partial \lambda_n^{\mathrm{S}}} = (1 - p_{n,f}) 4d^2. \tag{39}$$

C. Performance Trade-off

There is a performance trade-off achievable by controlling the local intensity of edge nodes density. If the number of edge caching nodes storing multiple copies of the same content is increased, then the cooperation gain is increased to enhance the SCDP. However, having more cooperative edge nodes increases the experienced user-load, resulting in less bandwidth which can be exploited for the content transmission. At the same time, when RRH density is too high, it might cause too much interference outside the CSA. The optimization aims to strike a good balance by finding the appropriate edge caching node density in order to maximize the overall SCDP.

V. Simulation Results

In this section, we provide simulation results to evaluate the performance of the proposed algorithm that jointly optimizes the spatial cache node density and the content caching probability. Table I provides the values of the network parameters used in the simulations, if not stated otherwise.

Variable	Value	Description
$\lambda^{\mathrm{S}}(x,y)$	≈ 0.0893 [unit/m^2]	Initial caching node density
λ^{S}	≈ 0.0893 [unit/m^2]	Upper caching node density limit
#UE	300	Total number of users
-	140×140 [m^2]	Total network space
d	10 [m]	Half side length of each CSA
W	-174 [dBm]	Thermal noise power
F	8	Wordbook size
M	3	Cache size
υ	0.7	Skewness factor content popularity
α	3	Path-loss coefficient
B	100 [MHz]	Bandwidth
ρ	$\{2, \ldots, 30\}$ [Mbps]	Target bit-rate
P^{S}	1	Transmitting power at the edge node

TABLE I: The network parameters.

The following baselines are also considered and compared with the proposed algorithm:

1) Most popular content (MPC) caching policy with $\boldsymbol{\lambda}^{\mathrm{S}} \approx 0.0893$. This can be used together with any RRH density.
2) Uniform content (UC) caching policy with $\boldsymbol{\lambda}^{\mathrm{S}} \approx 0.0893$. This can also be used with any RRH density.
3) Content caching optimization in [11] with fixed caching node density of $\boldsymbol{\lambda}^{\mathrm{S}} \approx 0.0893$.

In Fig. 4, an example of the employed arbitrary user density over a 1D projection of the 2D network space is shown, with $\min_{n \in \mathcal{N}} \lambda_n^{\mathrm{U}} \approx 0.0037$ and $\max_{n \in \mathcal{N}} \lambda_n^{\mathrm{U}} \approx 0.0350$. User density is considered zero outside the 140×140 [m^2] network space. When zero UE density is experienced at some location, the corresponding ω_i coefficient in (28) would be zero and

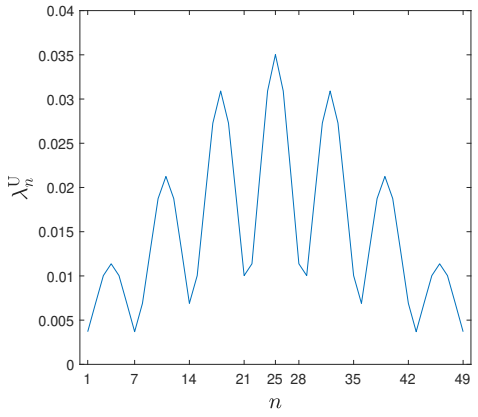

Fig. 4: A user intensity function λ_n^{U} for the generating PPP.

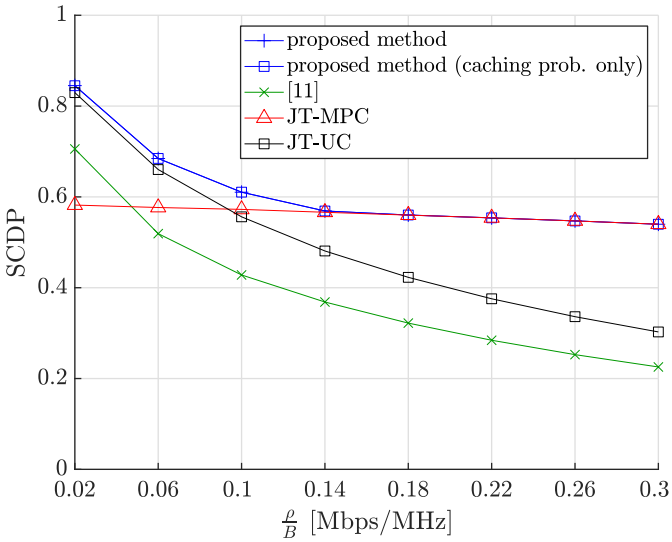

Fig. 5: The SCDP results.

no contribution to the interference is given. Note that the bin model was used only for our optimization to compute the edge node density and content caching probabilities but the SCDP results in the figures were obtained using Monte-Carlo simulations without the bin model restriction. Similarly, the following results will consider the case of no cooperating nodes, i.e., $K = 0$, avoided during the optimization of $(\mathbb{P}_0)$, as a zero contribution to the reported SCDP.

A. SCDP vs User Target Bit Rate

Fig. 5 provides the SCDP results for the proposed algorithm and baselines against the network work-load (i.e., spectral efficiency usage) ρ/B for the 7×7 network. Results show that the proposed method achieves the best SCDP compared to other benchmarks although the SCDP of the proposed method gradually decreases and converges to that of JT-MPC for high spectral efficiency usage. The proposed method's superior performance is particularly obvious when the spectral efficiency usage is on the low side, which corresponds to the case with higher network densification. Also, it is expected that at high spectrum efficiency usage, JT-MPC tends to be optimal so it makes sense to see that the proposed method converges to

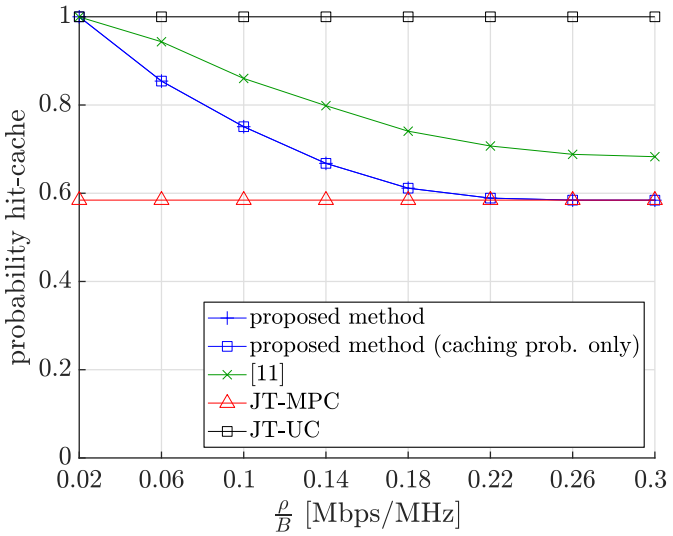

Fig. 6: The probability hit-cache results.

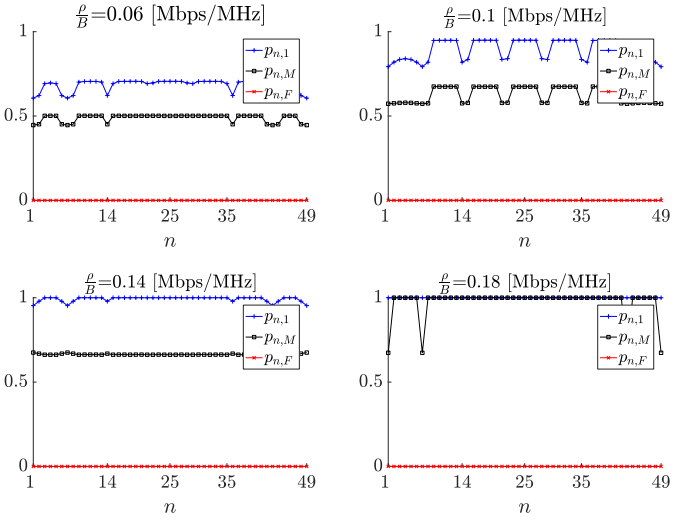

Fig. 7: The probabilistic content caching policy by the proposed model for various ρ/B. The x-axis shows the index of the UCL while the y-axis corresponds to the content index.

JT-MPC. The hit-cache probability results in Fig. 6 show the different approaches being taken by the various methods. As we can see, for JT-MPC, its hit-cache probability and SCDP are similar, which suggests that in this scheme, whenever there is a hit-cache, it will likely be successfully delivered. On the contrary, for JT-UC, it has a hit-cache probability of one, but not all the contents will be successfully delivered. Therefore, we can also observe that for the proposed method, it is able to increase the hit-cache probability while ensuring that almost all contents are delivered successfully and this is the reason why the proposed method is able to enhance the SCDP.

The resulting sets of content caching probabilities from our proposed method are shown in Fig. 7 for some representative cases of work-load ρ/B and content indices $f = [1, M, F]$. When user density is high, cooperation gain outweighs diversity gain by storing multiple copies of the same content. A low user density operates to exploit the diversity gain by storing more distinct contents, as reported in Fig. 7. Also, a higher rate requirement ρ will amplify the benefit of cooperation gain and prefer a more biased caching strategy based on content popularity while a smaller ρ will favour a more uniform caching strategy to benefit from content diversity. It can be observed in Fig. 8, the edge node density tends to follow the user density for content delivery. This is particularly clear at higher ρ. The reason is that at higher ρ the caching strategy tends to exploit more the cooperation gain, and decrease the edge node density in areas with low UE density, as shown in Fig. 8, to reduce the local number of cooperating nodes.

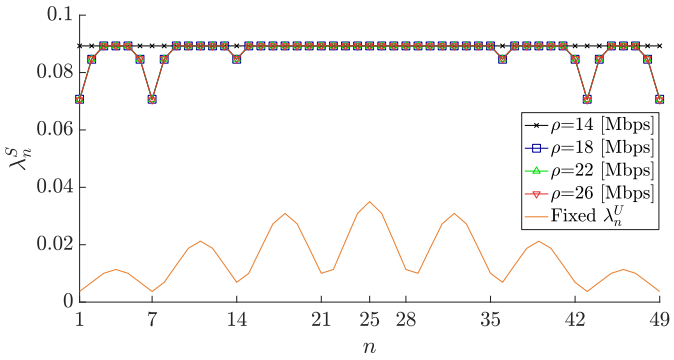

Fig. 8: The RRH density $\{\lambda_n^{\mathrm{S}}\}$ by the proposed method.

B. Network Energy Consumption

Knowing that idling caching nodes can help reducing interference, it is anticipated that the proposed algorithm can not only improve SCDP but also achieve energy saving. It is worth noticing that our objective function (i.e., a lower bound for the target global SCDP) does not explicitly take into account any measure of energy consumption. Caching nodes density and probabilistic content caching model also have a strong effect on the employed bandwidth for content transmission, and the amount of consumed energy by a network is dependent on the bandwidth over which the power is spread. To compare all the considered approaches, we provide the relevant results normalised by that achieved by the proposed method.

Fig. 9 shows the total network energy consumption of all the methods normalized by that of the proposed method. As we can see, all the benchmarks except JT-UC spend more total power than the proposed method, while it is also important to note that the proposed method has the best SCDP out of all the methods. In addition, although JT-UC spends the least overall power consumption, it has a much worse SCDP than the proposed method, as has been demonstrated before. From the results in this figure, we can compare the energy consumption performance with and without optimizing the RRH density. Recall from the results in Fig. 5 that optimizing the RRH density does not seem to provide any additional benefit for SCDP. We have now identified that the benefit of optimizing the RRH density comes in terms of energy consumption.

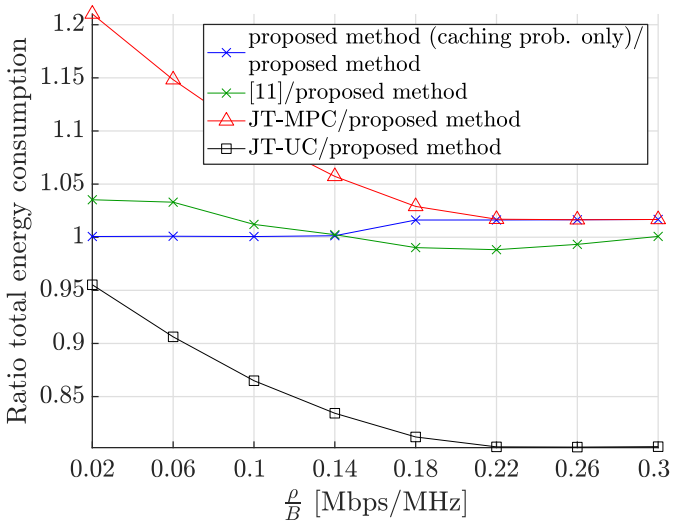

Fig. 9: Average total network power consumption for various work-load values ρ/B.

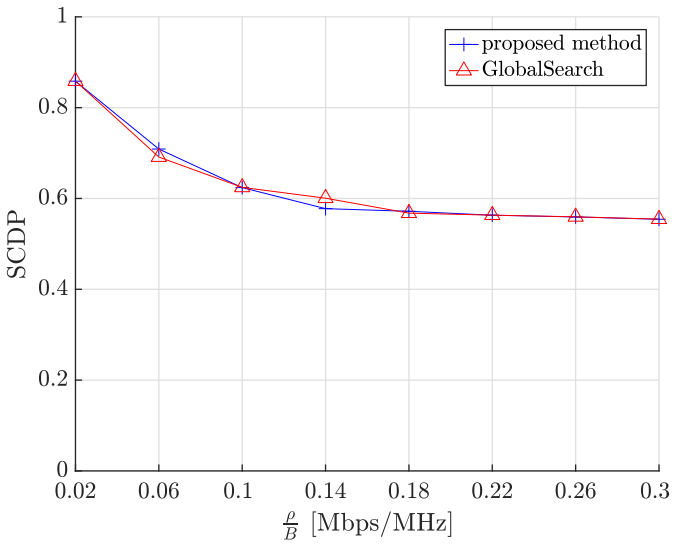

Fig. 10: Comparison of SCDP results for different methods in a 3×3 network.

C. Optimality for the Proposed Method

The proposed method finds a stationary point for maximizing the lower-bound of the SCDP. To understand the optimality of the proposed method, Fig. 10 provides the results for the SCDP obtained by the proposed method and that obtained by the function from the optimization toolbox of MATLAB. Due to the high computational complexity of GlobalSearch, we are restricted to consider only a simple 3×3 edge caching network. It can be noticed that at $\frac{\rho}{B} = [.06, .14]$ [Mbps/MHz], the results for the proposed method and GlobalSearch depart only very slightly. For the other cases both methods appear to have achieved the same SCDP performance. Based on these results, it is believed that the proposed algorithm is effective to obtain the near-optimal solution.

VI. Conclusions and Future Work

In this paper, we investigated the joint optimization of the RRH caching node density and the content caching probability for an ultra-dense content caching network where the user density is non-homogeneous. We considered a simple cooperation strategy for delivering the contents from the active RRHs and the only interference control mechanism is to idle RRHs. The optimization has been performed to maximize the lower bound of the SCDP using the steepest ascent algorithm. Simulation results have illustrated significant performance improvement in terms of the SCDP can be obtained by the proposed algorithm over conventional approaches, and revealed that the optimized RRH density and content caching probabilities can adapt very well to the non-homogeneous user spatial density. An analytic derivation of the user-load has been derived and shown to be affected by a set of key network parameters. There are future directions that deserve further effort. The optimization of a multi-objective function would shed more light on the balance of different performance metrics such as SCDP, latency or power consumption for optimizing network operations. When the costs of fetching a generic content are correctly introduced, the inclusion of MBSs would also provide more insights on the optimal network choices. Spectrum sharing between MBS and SBS is also another important direction. The effects of some time-varying parameters such as content popularity would also be of some interest, with the intention to target optimal network choices in a more dynamic way.

Appendix

The proposed Jensen's lower bound is conditioned on a set of random variables as shown in (23). Due to independence between the interference power and the desired signal power, we can work out the expected value in (23) by separately deriving the following two independent terms

$$\varphi^{\mathrm{I}}_{n,f}(\boldsymbol{\lambda}^{\mathrm{S}}, p_{n,f}) = \mathbb{E}_{\mathcal{I}_{n,f}}[\mathcal{I}_{n,f} + W], \tag{44a}$$

$$\varphi^{\mathrm{D}}_{n,f}(\lambda^{\mathrm{S}}_{n}, \boldsymbol{p}_n) = \mathbb{E}_{\substack{\{r_{n,k}\}_{k\in\phi_{n,f}},\\ K_{n,f},\Xi_{n,f}}} \left[\frac{2^{\frac{\rho\Xi_{n,f}}{B}} - 1}{\sum_{k=1}^{K_{n,f}} r_{n,k}^{-\alpha}} \right]. \tag{44b}$$

For what concerns (44a), the derivation of (43) is reported to articulate the expected value on the variables we need to average out.

For simplicity we will now refer to $\bar{n}$ to indicate the random cardinality of the set of interfering caching nodes that have not cached the f-th content within the n-th CSA whose mean is $\mathbb{E}\left[\bar{\phi}_{n,-f}\right] = \bar{\mu}_{n,f} = \lambda^{\mathrm{S}}_{n}(1 - p_{n,f})4d^2$, see (4b). Similarly, $\widetilde{n}_i$ stands for the set of interferers for the i-th CSA such that $i \neq n$ and $\mathbb{E}[\widetilde{\phi}_i] = \tilde{\mu}_i = \lambda^{\mathrm{S}}_{i} 4d^2$ as per (4c).

As a result, we obtain from (43)

$$\mathbb{E}_{\mathcal{I}_{n,f}}\left[\mathcal{I}_{n,f} + W\right] = \mathbb{E}_{\bar{n},h_n,r_n}\left[\bar{n}\,|h_n|^2\, r_n^{-\alpha}\right] + \sum_{i\neq n} \omega_i\, \mathbb{E}_{\widetilde{n}_i,h_i,r_i}\left[\widetilde{n}_i\,|h_i|^2\, r_i^{-\alpha}\right] + W, \tag{45}$$

where the subscripts $\bar{k}$ and $\tilde{k}$ have been dropped for conciseness as long as the independence of the terms of both sums allows to consider each term of the sums independently.

$$
\begin{aligned}
&\mathbb{E}_{\substack{\{r_{n,k}\}_{k\in\phi_{n,f}},\\ K_{n,f},\Xi_{n,f}}}\left[\frac{2^{\frac{\rho\Xi_{n,f}}{B}}-1}{\sum_{k=1}^{K_{n,f}} r_{n,k}^{-\alpha}}\right]\\
&=\mathbb{E}_{\{r_{n,k}\}_{k\in\phi_{n,f}},K_{n,f}}\left[\sum_{m=1}^{M}\frac{2^{\frac{\rho m}{B}}-1}{\sum_{k=1}^{K_{n,f}} r_{n,k}^{-\alpha}}\Pr(\Xi_{n,f}=m)\right]\\
&=\mathbb{E}_{\{r_{n,k}\}_{k\in\phi_{n,f}},K_{n,f}}\left[\sum_{m=1}^{M}\frac{2^{\frac{\rho m}{B}}-1}{\sum_{k=1}^{K_{n,f}} r_{n,k}^{-\alpha}}\left(F_{\xi_{n,f}}(m)^{K_{n,f}}-F_{\xi_{n,f}}(m-1)^{K_{n,f}}\right)\right]\\
&=\mathbb{E}_{\{r_{n,k}\}_{k\in\phi_{n,f}},K_{n,f}}\left[\int_0^\infty e^{-t\sum_{k=1}^{K_{n,f}} r_{n,k}^{-\alpha}}\,dt\sum_{m=1}^{M}(2^{\frac{\rho m}{B}}-1)\left(F_{\xi_{n,f}}(m)^{K_{n,f}}-F_{\xi_{n,f}}(m-1)^{K_{n,f}}\right)\right] \qquad (40)\\
&=\mathbb{E}_{\{r_{n,k}\}_{k\in\phi_{n,f}},K_{n,f}}\left[\int_0^\infty \prod_{k=1}^{K_{n,f}} e^{-tr_{n,k}^{-\alpha}}\,dt\sum_{m=1}^{M}(2^{\frac{\rho m}{B}}-1)\left(F_{\xi_{n,f}}(m)^{K_{n,f}}-F_{\xi_{n,f}}(m-1)^{K_{n,f}}\right)\right]\\
&\overset{a}{=}\mathbb{E}_{K_{n,f}}\left[\sum_{m=1}^{M}(2^{\frac{\rho m}{B}}-1)\times\int_0^\infty J_n(t)^{K_{n,f}}\left(F_{\xi_{n,f}}(m)^{K_{n,f}}-F_{\xi_{n,f}}(m-1)^{K_{n,f}}\right)\,dt\right]\\
&\overset{b}{=}\sum_{m=1}^{M}(2^{\frac{\rho m}{B}}-1)\int_0^\infty\Theta(t,m)\,dt,
\end{aligned}
$$

$$
\Theta(t,m)=\sum_{K_{n,f}=1}^{\infty}J_n(t)^{K_{n,f}}\left(F_{\xi_{n,f}}(m)^{K_{n,f}}-F_{\xi_{n,f}}(m-1)^{K_{n,f}}\right)\times\frac{\mu_{n,f}^{K_{n,f}}}{(e^{\mu_{n,f}}-1)K_{n,f}!}, \qquad (41)
$$

$$
\begin{aligned}
\Theta(t,m)&=\frac{1}{e^{\mu_{n,f}}-1}\sum_{K_{n,f}=0}^{\infty}\left[\frac{\left(\mu_{n,f}F_{\xi_{n,f}}(m)J_n(t)\right)^{K_{n,f}}}{K_{n,f}!}-\frac{\left(\mu_{n,f}F_{\xi_{n,f}}(m-1)J_n(t)\right)^{K_{n,f}}}{K_{n,f}!}\right] \qquad (42)\\
&=\frac{1}{e^{\mu_{n,f}}-1}\left(e^{\mu_{n,f}F_{\xi_{n,f}}(m)J_n(t)}-e^{\mu_{n,f}F_{\xi_{n,f}}(m-1)J_n(t)}\right)
\end{aligned}
$$

$$
\mathbb{E}_{\mathcal{I}_{n,f}}\left[\mathcal{I}_{n,f}+W\right]=\mathbb{E}_{\{h_{n,k}\}_{\forall k},\{r_{n,k}\}_{\forall k},\bar{\phi}_{n,-f},\widetilde{\phi}_i}\left[\sum_{\bar{k}\in\bar{\phi}_{n,-f}}\left|h_{n,\bar{k}}\right|^2 r_{n,\bar{k}}^{-\alpha}+\sum_{i\neq n}\omega_i\sum_{\tilde{k}\in\widetilde{\phi}_i}\left|h_{i,\tilde{k}}\right|^2 r_{i,\tilde{k}}^{-\alpha}\right]+W, \qquad (43)
$$

It is known that the distribution of the squared absolute value of a circular symmetric Gaussian random variable writes as an exponential distribution $\exp(1)$. Also, it is easy to see that the expected value for the standard exponential random variable $|h|^2$ results to be $\int_0^\infty x\exp(-x)\,dx=1$. Thus, we can uncondition (45) w.r.t. the link distances and cardinalities of the two sets as

$$
\mathbb{E}_{\mathcal{I}_{n,f}}[\mathcal{I}_{n,f}+W]=\bar{\mu}_{n,f}\bar{J}(n)+\sum_{i\neq n}\omega_i\tilde{\mu}_i\widetilde{J}(i)+W, \qquad (46)
$$

where

$$
\begin{cases}
\bar{J}(n)=\dfrac{1}{4d^2}\left(\iint_{\mathcal{D}_n\setminus\mathcal{B}_0}(x^2+y^2)^{-\alpha/2}\,dxdy+\pi\right),\\
\widetilde{J}(i)=\dfrac{1}{4d^2}\iint_{\mathcal{D}_i}(x^2+y^2)^{-\alpha/2}\,dxdy.
\end{cases} \qquad (47)
$$

where $\mathcal{B}_0$ denotes the circle of unit radius centered at the UCL under investigation.

As such, we have averaged out all the random variables previously highlighted in (43) and therefore retrieved the expected value in (44a).

From Section III-B, we define the user-load of a set of jointly cooperating caching nodes $K_{n,f}$ at the n-th UCL for content f as

$$
\Xi_{n,f}=\max\{\xi_{n,f,1},\xi_{n,f,2},\ldots,\xi_{n,f,K_{n,f}}\}. \qquad (48)
$$

We can write the pmf of $\Xi_{n,f}$ in terms of the user-load perceived by the single caching node $\xi_{n,f}$ as

$$
\Pr(\Xi_{n,f}=m)=F_{\xi_{n,f}}(m)^{K_{n,f}}-F_{\xi_{n,f}}(m-1)^{K_{n,f}},
$$

where we have the number of cooperating nodes $K_{n,f}$ whose mean is $\mathbb{E}[\phi_{n,f}]=\mu_{n,f}=\lambda_n^{\mathrm{S}}p_{n,f}4d^2$ as per (4a). Therefore, we can work out (44b) as (40) where $\Theta(t,m)$ is defined in (41).

Regarding (40), (a) follows from averaging over the link distance and (b) follows from averaging over the ZTP distribution of the set of the cooperating nodes. It is easy to see

that for $K_{n,f}=0$, we have

$$J_n(t)^0(F_{\xi_{n,f}}(m)^0 - F_{\xi_{n,f}}(m-1)^0)\frac{\mu_{n,f}^0}{(e^{\mu_{n,f}}-1)0!}=0.$$

Therefore, we can express the sum over $K_{n,f}$ to simply start from 0. This allows us to write the unconditioning part over $K_{n,f}$ as a series for exponential functions. Thus, we get (42).

By substituting (42) into (40), $\varphi_{n,f}^{\mathrm{D}}(\lambda_n^{\mathrm{S}},\boldsymbol{p}_n)$ is finally defined as

$$\begin{aligned}\varphi_{n,f}^{\mathrm{D}}(\lambda_n^{\mathrm{S}},\boldsymbol{p}_n) &= \mathbb{E}_{\{r_{n,k}\}_{k\in\phi_{n,f}},K_{n,f},\Xi_{n,f}}\left[\frac{2^{\frac{\rho\Xi_{n,f}}{B}}-1}{\sum_{k=1}^{K_{n,f}} r_{n,k}^{-\alpha}}\right]\\ &= \sum_{m=1}^{M}\left(e^{\frac{\rho m}{B}}-1\right)\int_0^{\infty}\Theta(t,m)\;dt.\end{aligned}$$

REFERENCES

[1] A. Gupta and R. K. Jha, "A survey of 5g network: Architecture and emerging technologies," *IEEE Access*, vol. 3, pp. 1206–1232, July 2015.

[2] M. Peng, D. Liang, Y. Wei, J. Li, and H.-H. Chen, "Self-configuration and self-optimization in lte-advanced heterogeneous networks," *IEEE Commun. Mag.*, vol. 51, no. 5, pp. 36–45, May 2013.

[3] A. Khandekar, N. Bhushan, J. Tingfang, and V. Vanghi, "Lte-advanced: Heterogeneous networks," in *Proc. EW'10*, pp. 978–982, 2010.

[4] M. Peng, C. Wang, J. Li, H. Xiang, and V. K. N. Lau, "Recent advances in underlay heterogeneous networks: Interference control, resource allocation, and self-organization," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 700–729, Mar. 2015.

[5] H.-S. Jo, Y. J. Sang, P. Xia, and J. G. Andrews, "Heterogeneous cellular networks with flexible cell association: A comprehensive downlink sinr analysis," *IEEE Trans. Wireless Commun.*, vol. 11, no. 10, pp. 3484–3495, Oct. 2012.

[6] M. Peng, Y. Li, J. Jiang, J. Li, and C. Wang, "Heterogeneous cloud radio access networks: A new perspective for enhancing spectral and energy efficiencies," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 126–135, Dec. 2014.

[7] M. Kamel, W. Hamouda, and A. Youssef, "Ultra-dense networks: A survey," *IEEE Commun. Surveys Tuts*, vol. 18, no. 4, pp. 2522–2545, Oct.-Dec. 2016.

[8] M. Peng, Y. Li, Z. Zhao, and C. Wang, "System architecture and key technologies for 5g heterogeneous cloud radio access networks," *IEEE Net.*, vol. 29, no. 2, pp. 6–14, Mar. 2015.

[9] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," in *Proc. IEE ICC'15*, pp. 1–6, Jun. 2015.

[10] S. H. Chae and W. Choi, "Caching placement in stochastic wireless caching helper networks: Channel selection diversity via caching," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 6626–6637, Oct. 2016.

[11] Y. Cui, D. Jiang, and Y. Wu, "Analysis and optimization of caching and multicasting in large-scale cache-enabled wireless networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 5101–5112, Jul. 2016.

[12] Z. Chen, J. Lee, T. Q. Quek, and M. Kountouris, "Cooperative caching and transmission design in cluster-centric small cell networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 3401–3415, May 2017.

[13] W. Wen, Y. Cui, F.-C. Zheng, S. Jin, and Y. Jiang, "Random caching based cooperative transmission in heterogeneous wireless networks," *IEEE Trans. Commun.*, vol. 66, no. 7, pp. 2809–2825, Jul. 2018.

[14] S. H. Chae, T. Q. Quek, and W. Choi, "Content placement for wireless cooperative caching helpers: A tradeoff between cooperative gain and content diversity gain," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6795–6807, 2017.

[15] M. Newton and J. Thompson, "Classification and generation of non-uniform user distributions for cellular multi-hop networks," in *Proc. IEEE ICC'06*, pp. 4549–4553, Jun. 2006.

[16] S. Baroudi and Y. R. Shayan, "Analytical evaluation of outage probability based on signal to interference ratio for gaussian-distributed users," in *Proc. IEEE ISCC'15*, pp. 841–844, Jul. 2015.

[17] ——, "Percentage of gaussianly distributed users with adequate quality of service in a circular cell," in *Proc. IEEE CCECE'14*, pp. 1–4, May 2014.

[18] R. Amer, H. ElSawy, J. Kibiłda, M. M. Butt, and N. Marchetti, "Cooperative transmission and probabilistic caching for clustered D2D networks," *arXiv preprint arXiv:1811.11099*, 2018.

[19] D. López-Pérez, M. Ding, H. Claussen, and A. H. Jafari, "Towards 1 gbps/ue in cellular systems: Understanding ultra-dense small cell deployments," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2078–2101, Jun. 2015.

[20] S. Samarakoon, M. Bennis, W. Saad, M. Debbah, and M. Latva-Aho, "Ultra dense small cell networks: Turning density into energy efficiency," *IEEE J. Select. Areas Commun.*, vol. 34, no. 5, pp. 1267–1280, Apr. 2016.

[21] E. Baştuğ, M. Bennis, M. Kountouris, and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs," *EURASIP J. Wireless Commun. Netw.*, no. 1, p. 41, Feb. 2015.

[22] J. Ye, X. Ge, G. Mao, and Y. Zhong, "5g ultra-dense networks with non-uniform distributed users," *IEEE Trans. Veh. Technol.*, vol. 67, no. 3, pp. 2660–2670, Mar. 2018.

[23] A. Gotsis, S. Stefanatos, and A. Alexiou, "Ultradense networks: The new wireless frontier for enabling 5g access," *IEEE Veh. Technol. Mag.*, vol. 11, no. 2, pp. 71–78, Jun. 2016.

[24] L. Armijo, "Minimization of functions having lipschitz continuous first partial derivatives," *Pacific J. Math.*, vol. 16, no. 1, pp. 1–3, 1966.

[25] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.