NOMA Assisted Wireless Caching: Strategies and Performance Analysis

Zhiguo Ding, *Senior Member, IEEE*, Pingzhi Fan, *Fellow, IEEE*, George K. Karagiannidis, *Fellow, IEEE*, Robert Schober, *Fellow, IEEE*, and H. Vincent Poor, *Fellow, IEEE*
(*Invited Paper*)

***Abstract*— Conventional wireless caching assumes that content can be pushed to local caching infrastructure during off-peak hours in an error-free manner; however, this assumption is not applicable if local caches need to be frequently updated via wireless transmission. This paper investigates a new approach to wireless caching for situations in which the cache content has to be updated during on-peak hours. Two non-orthogonal multiple access (NOMA)-assisted caching strategies are developed, namely, the push-then-deliver strategy and the push-and-deliver strategy. In the push-then-deliver strategy, the NOMA principle is applied to push more content files to the content servers during a short time interval reserved for content pushing during on-peak hours and to provide more connectivity for content delivery, compared with the conventional orthogonal multiple access (OMA) strategy. The push-and-deliver strategy is motivated by the fact that some users' requests cannot be accommodated locally and the base station has to serve them directly. These events during the content delivery phase are exploited as opportunities for content pushing, which further facilitates the frequent update of the files cached at the content servers. It is also shown that this strategy can be straightforwardly extended to device-to-device caching, and various analytical results are developed to illustrate the superiority of the proposed caching strategies compared with OMA based schemes.**

Manuscript received September 16, 2017; revised January 15, 2018 and April 4, 2018; accepted May 3, 2018. Date of publication May 29, 2018; date of current version October 16, 2018. The work of Z. Ding was supported by the UK Engineering and Physical Sciences Research Council under grant number EP/P009719/1 and by H2020-MSCA-RISE-2015 under grant number 690750. The work of P. Fan was supported by the National Natural Science Foundation of China under grant number 61731017, and the 111 Project (No.111-2-14). The work of R. Schober was supported by the Alexander von Humboldt Professorship Program. The work of H. V. Poor was supported by the U.S. National Science Foundation under Grants CNS-1702808 and ECCS-1647198. This work was presented in part at the IEEE International Conference on Communications (ICC), Kansas City, MO, May 2018 [1]. The associate editor coordinating the review of this paper and approving it for publication was M. Tao. *(Corresponding author: Zhiguo Ding.)*

Z. Ding is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA, and also with the School of Electrical and Electronic Engineering, The University of Manchester, Manchester M13 9PL, U.K. (e-mail: zhiguo.ding@manchester.ac.uk).

P. Fan is with the Institute of Mobile Communications, Southwest Jiaotong University, Chengdu 610031, China (e-mail: pingzhifan@foxmail.com).

G. K. Karagiannidis is with the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece (e-mail: geokarag@auth.gr).

R. Schober is with the Institute for Digital Communications, Friedrich-Alexander-University Erlangen-Nurnberg, 91023 Erlangen, Germany (e-mail: robert.schober@fau.de).

H. V. Poor is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: poor@princeton.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCOMM.2018.2841929

***Index Terms*— Non-orthogonal multiple access (NOMA), wireless caching, content pushing and delivery, Poisson cluster processes (PCPs).**

I. Introduction

RECENTLY non-orthogonal multiple access (NOMA) has received significant attention as a main enabling technique for future wireless networks [2]–[4]. The key idea of NOMA is to encourage spectrum sharing among mobile nodes, which not only improves the spectral efficiency but also ensures that massive connectivity can be effectively supported. Practical concepts for implementing the NOMA principle for a single resource block, such as an orthogonal frequency division multiplexing (OFDM) subcarrier, include power domain NOMA and cognitive radio (CR) inspired NOMA [5]–[7], which provide different tradeoffs between throughput and fairness. When each user is allowed to occupy multiple subcarriers, dynamically grouping the users on different subcarriers is a challenging problem, and various multi-carrier NOMA schemes, such as sparse code multiple access (SCMA) and pattern division multiple access (PDMA) [8], [9], provide practical solutions for achieving different performance-complexity tradeoffs. Unlike single-carrier NOMA, in multi-carrier NOMA, a user's message is spread over multiple resource blocks, which requires efficient encoding schemes, such as multi-dimensional coding, to be implemented at the transmitter and low-complexity decoding schemes, such as message passing algorithms, to be used at the receivers.

NOMA has been shown to be compatible with many other advanced communication concepts. For example, several features of millimeter-wave (mmWave) communications, such as highly directional transmission, and the mismatch between the users' channel vectors and the commonly used finite resolution analog beamforming, facilitate the implementation of NOMA in mmWave networks [10], [11]. In addition, NOMA can further improve the spectral efficiency of multiple-input multiple-output (MIMO) systems. For example, MIMO-NOMA can efficiently exploit the spatial degrees of freedom of MIMO channels and, unlike single-input single-output (SISO) NOMA, is beneficial even if all users have similar channel conditions [12]–[14]. Furthermore, conventionally, when each user has a single antenna, cooperative transmission can be used to exploit spatial diversity but suffers from a reduced overall data rate, since relaying consumes extra bandwidth resources [15].

In this context, the application of NOMA can efficiently reduce the number of consumed bandwidth resource blocks, such as subcarriers and time slots, and hence improve the spectral efficiency of cooperative communications [16]–[18]. Furthermore, existing studies have also revealed a strong synergy between NOMA and CR networks, where the use of NOMA can significantly improve the connectivity for the users of the secondary network [19].

Wireless caching is another important enabling technique for future communication networks [20], [21], but little is known about the coexistence of NOMA and wireless caching. The key idea of wireless caching is to push the content in off-peak hours during the so-called *content pushing phase* close to the users before it is requested, and therefore, the users' requests can be locally served during the so-called *content delivery phase*. In fact, asking a base station (BS) to serve the users' requests directly is not preferable, not only because the maximal number of users that a BS can serve concurrently is small, but also because non-caching transmission schemes are severely constrained by the limited backhaul capacity of wireless networks. Most caching schemes can be grouped into one of two classes[1] [21], [24]. *The first class* assumes the existence of a content caching infrastructure, such as content servers, small cell BSs, etc., [25]–[27]. When caching infrastructure (e.g., content servers) is available, the objective in the content pushing phase is to push the content files to the content servers in a timely and reliable manner, before the users request these files. During the content delivery phase, an ideal situation is that all the users' requests can be locally served, without communicating with the central controller of the network, e.g., the BS. *The second class*, also known as device-to-device (D2D) caching, assumes that there is no dedicated caching infrastructure, and relies on user cooperation [28], [29]. Particularly, during the content pushing phase, all users will proactively cache some content. During the content delivery phase, a user will communicate with its BS only if none of its neighbors can help the user locally, i.e., the user cannot find its requested file in the caches of its neighbors.

A fundamental assumption made in the existing caching literature is that, in the content pushing phase, content is pushed to the content servers in an error-free manner during off-peak hours. However, performing caching only during off-peak hours is not effective if the popularity of the content is rapidly changing or the files to be cached need to be frequently updated. Typical examples for this type of content include up-to-the-minute news, sports events requiring live updates, e-commerce promotion with frequent pricing changes, newly released music videos, etc. Similarly, assuming error-free content pushing may also be questionable in many practical communication scenarios. In practice, connecting the content servers wirelessly with the BS is preferable since the cost for setting up the network is reduced and the installation of cables is avoided. Furthermore, wireless networks facilitate D2D caching, since file sharing among users in wireless networks is straightforward, whereas realizing pairwise connections in wireline networks is more difficult. However, wireless transmission is prone to noise, distortion, and attenuation, which makes error-free transmission a very strong assumption in practice. The objective of this paper is to apply the NOMA principle to wireless caching and to develop NOMA assisted caching strategies that do not require the aforementioned assumptions. The contributions of the paper are summarized as follows:

- For the case in which the content pushing and delivery phases are separated and limited bandwidth resources are periodically available for content pushing during on-peak hours, a NOMA-assisted push-then-deliver strategy is proposed. Particularly, during the content pushing phase, the BS will use the NOMA principle and push multiple files to the content servers simultaneously. A CR inspired NOMA power allocation policy is used to ensure that content files are delivered to their target content servers with the same outage probability as with conventional orthogonal multiple access (OMA) based transmission. However, by using NOMA, additional files can be pushed to the content servers simultaneously, which is important to efficiently use the limited resources reserved for content pushing and hence to improve the cache hit probability. During the content delivery phase, the use of NOMA not only improves the reliability of content delivery, but also ensures that more user requests can be served concurrently by a content server.
- The objective of the proposed push-and-deliver strategy is to provide additional bandwidth resources for content pushing. Unlike the push-then-deliver strategy, the push-and-deliver strategy seeks opportunities for content pushing during the content delivery phase. In particular, during the content delivery phase, the BS occasionally has to serve some users directly, since these users' requested files cannot be found in the local content servers. Conventionally, this is a non-ideal situation which reduces the spectral efficiency. Nevertheless, this non-ideal situation is inevitable in practice and is expected to occur frequently, as the users' requests cannot be perfectly predicted. In this paper, this non-ideal situation for content delivery is exploited as an opportunity for additional content pushing. In other words, the push-and-deliver strategy is particularly useful when the bandwidth resources reserved for content pushing are limited, but the files at the content servers need to be frequently updated. The proposed push-and-deliver strategy is also extended to D2D caching without caching infrastructure, where the caches at the D2D helpers can be refreshed

[1]We note that coded caching, in which the number of BS transmissions is reduced by exploiting the structure of the content sent during the content pushing and delivery phases, does not fall into the two considered categories [22], [23]. In particular, the principle of coded caching is to first split the files into multiple subpackets and then encode them similarly as in network coding. Some of the subpackets need to be delivered to the users directly during the content pushing phase, because coded caching requires that the users already store parts of the files prior to content delivery in order to carry out interference cancelation analogous to network coding. Therefore, *a priori* assumptions about the users' requests are crucial for coded caching, which is different from the caching strategies considered in this paper which do not need users to store parts of their requested files. In addition, for coded caching, some content delivery tasks are performed during the content pushing phase, and therefore, a clear boundary between the content pushing and delivery phases is not needed for coded caching, which is also different for the caching strategies considered in this paper.

while users are directly served by the BS. We note that the NOMA-multicasting scheme proposed in [30] can be viewed as a D2D special case of the proposed push-and-deliver strategy, if the multicasting phase in [30] is viewed as the content delivery phase. However, the impact of integrating content pushing and delivery on the cache hit probability was not investigated in [30].

- Analytical results for the cache hit probability, the transmission outage probability, and the D2D cache miss probability are derived in order to obtain a better understanding of the performance of the proposed caching strategies. Conventionally, the cache hit probability is mainly determined by the size of the caches of the content servers, instead of by transmission outages, since conventional content pushing is carried out during off-peak hours, which means that the amount of the pushed content is much larger than the size of the caches of the content servers. However, for the schemes proposed in this paper, the outage based cache hit probability is a more suitable metric for performance evaluation, as explained in the following. In particular, we assume that a short time interval is periodically reserved during on-peak hours for pushing new content to the content servers. The time interval reserved for content pushing has to be short in order to achieve high spectral efficiency and has to be shared by multiple content servers. Thus, the amount of content that can be pushed to the content servers may be much smaller than the size of the storage of the content servers. Considering the tremendous increase in storage capacity available with current technologies, this is a realistic assumption. For the considered caching scenario, the crucial issue is how to quickly push the content files to the content servers during the short time interval available for content pushing during on-peak hours. Therefore, the outage based cache hit probability is the relevant performance criterion. When caching infrastructure is available, the impact of NOMA on the content pushing phase is quantified by exploiting the joint probability density function (pdf) of the distances between the content servers and the BS, and closed-form expressions for the achieved cache hit probability are developed. The impact of NOMA on the content delivery phase is investigated by using the transmission outage probability as a performance criterion and modelling the locations of the users and the content servers as Poisson cluster processes (PCPs). Furthermore, the impact of NOMA on D2D caching is studied by modelling the effect of content pushing as a thinning Poisson point process and deriving the cache miss probability, i.e., the probability of the event that a user cannot find its requested file in the caches of its neighbours. The provided simulations verify the accuracy of the proposed analysis, and illustrate the effectiveness of the proposed NOMA based wireless caching schemes.

The remainder of the paper is organized as follows. In Section II, the considered system model, including the caching model and the spatial model, are introduced. In Section III, the NOMA-assisted push-then-deliver strategy is presented, and its impact on the content pushing and delivery phases is investigated. In Section IV, the proposed push-and-deliver strategy is developed by efficiently merging the content pushing and delivery phases, its impact on the cache hit probability is investigated, and its extension to D2D scenarios is discussed. Computer simulations are provided in Section V, and the paper is concluded in Section VI. The details of all proofs are collected in the appendix.

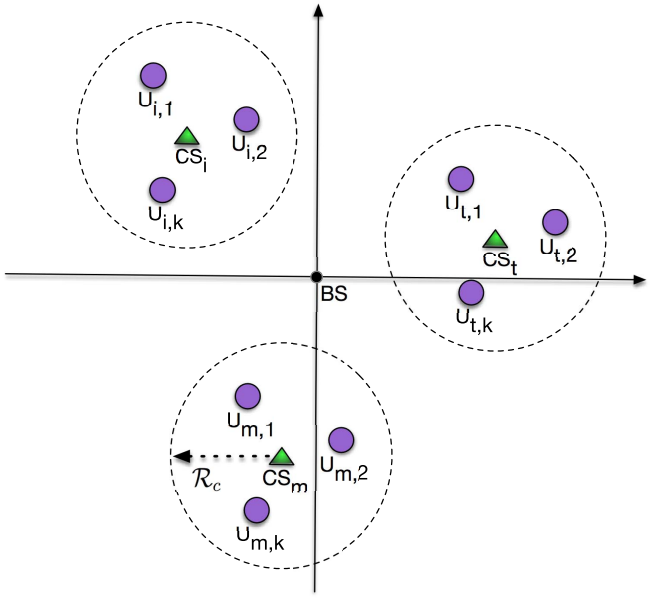


Fig. 1. An illustration of the assumed spatial model.

II. System Model

Consider a two-tier heterogeneous communication scenario, in which multiple users request cacheable content with the help of one BS and multiple content servers. The D2D scenario without caching infrastructure, e.g., content servers, will be described in Section IV.C. Assume that each user is associated with a single content server. If the file requested by a user can be found in the cache of its associated content server, this server will serve the user, which means that multiple content servers can communicate with their respective users concurrently and hence the spectral efficiency is high. However, if the file requested by a user cannot be found locally, the BS will serve the user directly, a situation that is not ideal for caching and should be avoided. The assumption that each user is associated with a single content server facilitates the use of PCP modelling, as discussed in the following subsection.

A. Spatial Clustering Model

Assume that the BS is located at the origin of a two-dimensional Euclidean plane, denoted by $\mathbb{R}^2$. As shown in Fig. 1, there are multiple content servers. The locations of the content servers and the users are modelled as PCPs. In particular, assume that the locations of the content servers are denoted by x_i and are modelled as a homogeneous Poisson point process (HPPP), denoted by Φ_c, with density λ_c, i.e., $x_i \in \Phi_c$. For notational simplicity, the location of the BS is denoted by x_0.

Each content server is the parent node of a cluster covering a disk whose radius is denoted by $\mathcal{R}_c$. Denote the content server in cluster i by CS_i. Without loss of generality, assume that there are K users associated with CS_i, denoted by $\mathrm{U}_{i,k}$. Note that users associated with the same content server are viewed as offspring nodes [31]. The offspring nodes are uniformly distributed in the disk associated with CS_i, and their locations are denoted by $y_{i,k}$. To simplify the notation, the locations of the cluster users are conditioned on the locations of their cluster heads (content severs). As such, the distance from a user to its content server is simply given by $||y_{i,k}||$, and the distance from user $\mathrm{U}_{i,k}$ to content server CS_j is denoted by $||y_{i,k} + x_i - x_j||$ [32], [33].

B. Caching Assumptions

Suppose that the files to be requested by the users are collected in a finite content library $\mathcal{F} = \{f_1, \cdots, f_F\}$. The popularity of the requested files is modelled by a Zipf distribution [34]. Particularly, the popularity of file f_l, denoted by $\mathrm{P}(f_l)$, is modelled as follows:

$$\mathrm{P}(f_l) = \frac{\frac{1}{l^\gamma}}{\sum_{p=1}^{F} \frac{1}{p^\gamma}}, \tag{1}$$

where $\gamma > 0$ denotes the shape parameter defining the content popularity skewness. We note that $\mathrm{P}(f_l)$ is the probability that a user requests file f_l. Similar to the existing wireless caching literature, [20], [21], [25]–[27], packets belonging to different files are assumed to have the same length. However, unlike the existing literature, we do not assume that the amount of information contained in the packets of different files is identical.[2] Particularly, the predetermined data rate of the packets of file f_l is denoted by R_l. We assume that packets belonging to different files have the same size but may contain different amounts of information for the following reasons. *Firstly,* the packet size is typically predefined according to practical system standards and cannot be changed. Therefore, it is reasonable to assume that all packets have the same size. *Secondly*, packets belonging to different files have different priorities and different target reception reliabilities, which requires the use of different channel coding rates for different packets. As a result, packets that have the same size do not necessarily contain the same amount of information. We note that none of the analytical results developed in this paper, except for Lemma 2, require the assumption that the packets contain different amounts of information. However, the performance for the special case of identical target data rates for all files is investigated in the simulation section.

In this paper, we assume that the popularity of the content is quasi-static, i.e., the popularity of the content is constant for a short time interval (e.g., one hour) and changes independently from one time interval to the next. However, the proposed NOMA assisted caching strategies are also applicable to scenarios in which the popularity of the content stays constant for a long time interval (e.g., one day), but changes suddenly during the on-peak hours. In these types of scenarios, conventional wireless caching mechanisms are not efficient, since content pushing is carried out during off-peak hours and hence we have to wait for a long time to update the local caches in order to reflect changes in the content popularity that happen during on-peak hours.

Finally, we assume that the BS has access to all the content files. In this paper, when content servers exist, we assume that the users have no caching capabilities. On the other hand, for the D2D assisted caching discussed in Section IV.C, it is assumed that each user has a cache.

III. Push-Then-Deliver Strategy

In this section, we consider the case in which the two caching phases, content pushing and content delivery, are separated. Unlike conventional caching which relies on the use of off-peak hours for content pushing, the proposed push-then-deliver strategy assumes that limited bandwidth resources, such as short time intervals, are periodically reserved for pushing new files to the content servers during on-peak hours. For example, every hour, a BS deployed in a large shopping mall or an airport may use a few seconds to push updated advertising and marketing videos to the content servers. The time interval reserved for content pushing during on-peak hours has to be short in order to achieve high system spectral efficiency. As will be shown, the proposed push-then-deliver strategy allows more files to be pushed within this short time interval compared to OMA.

In the following two subsections, we will demonstrate the impact of the NOMA principle on the content pushing and delivery phases, respectively.

A. Content Pushing Phase

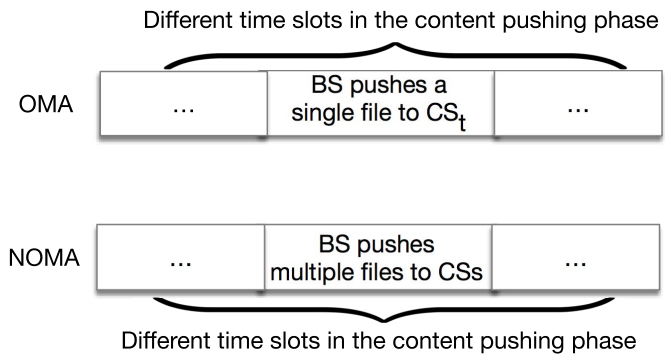
In order to have a baseline for the performance of NOMA assisted content pushing, conventional OMA based content pushing is introduced first.

1) OMA Based Content Pushing: The content pushing phase for OMA can be divided into multiple time slots, as shown in Fig. 2(a). There are different strategies for utilizing the limited number of time slots available for content pushing during on-peak hours. In this paper, we assume that the popular content can be divided into different libraries, e.g., one library may contain popular files for sports events and another one may contain files for business or political news. Furthermore, we assume that only a single time slot is available to push files belonging to the same library, since content pushing is accomplished during on-peak hours and hence the number of time slots available for content pushing is limited.[3]

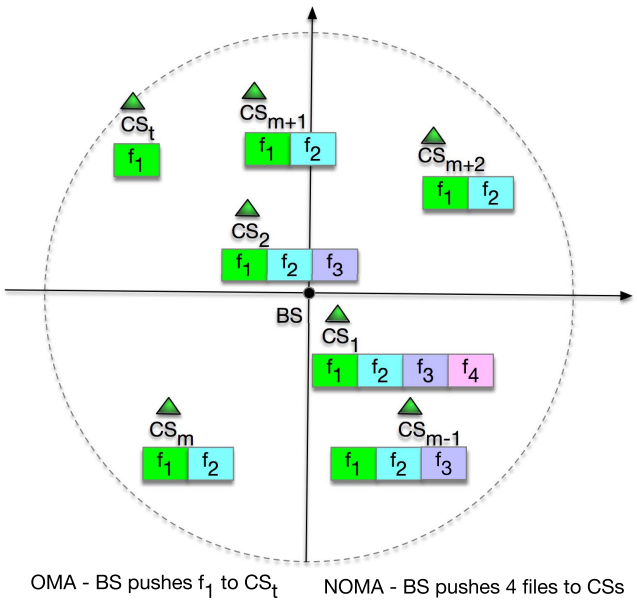
Since there is only a single time slot available to push files belonging to the same library and the use of OMA means that only a single file can be pushed in each time slot, the BS pushes the most popular file f_1 only. CS_t is able to decode

[2]In this paper, one packet refers to one symbol frame after channel coding and modulation. In other words, one packet contains not only information but also redundancy due to the use of channel coding and modulation.

[3]We note however that if multiple time slots are available for OMA based content pushing, sophisticated algorithms to optimally utilize the multiple time slots and to schedule the files for content pushing are needed. The design of such algorithms is beyond the scope of this paper.



(a) A general illustration of the content pushing phase



(b) An example for the outcome of content pushing

Fig. 2. An illustration of the impact of NOMA on content pushing. "CS" denotes a content server in the subfigures. For the example shown in subfigure (b), it is assumed that CS$_m$ is closer to the BS than CS$_t$, for $1 < m < t$. In OMA, a single file is pushed to CS$_t$, and in NOMA, the BS pushes a superimposed mixture consisting of four files, where content servers closer to the BS are likely able to decode more pushed files.

file f_1 with the following achievable data rate:

$$R_{t,OMA}^{CP} = \log\left(1 + \rho \frac{1}{L\left(||x_t - x_0||\right)}\right), \tag{2}$$

where ρ denotes the transmit signal-to-noise ratio (SNR), and $\frac{1}{L(||x_t||)}$ is the large scale path loss between CS$_t$ and the BS located at x_0. Particularly, the following path loss model is used, $\frac{1}{L(||x_t||)}$, where $L\left(||x_t||\right) = ||x_t||^{\alpha}$ and α denotes the path loss exponent. For a large scale network, the probability that $||x_t - x_0|| < 1$ is very small, and therefore, the simplified unbounded path loss model is used in this paper [35], [32], [33]. Nevertheless, the presented analytical results can be extended to other path loss models, e.g., $L\left(||x_t||\right) = ||1+x_t||^{\alpha}$ or $L\left(||x_t||\right) = \max\{1, ||x_t||^{\alpha}\}$, in a straightforward manner. We note that small scale multi-path fading is not considered for the channel gain associated with CS$_t$ since the content servers can be deployed such that line-of-sight connections to the BS are ensured, which means that large scale path loss is the dominant factor for signal attenuation. Since the large scale path loss is expected to change very slowly over time, the duration of the channel coherence time is large and hence a single file can be transmitted to the content servers within one time slot. Since only large scale path loss is assumed for the channels from the BS to the content servers, all the content servers that have shorter distances to the BS than CS$_t$ can also decode file f_1 if CS$_t$ can decode the file correctly. Therefore, the broadcast nature of wireless channels is exploited to ensure that one transmission to CS$_t$ also benefits the other content servers. We note that small scale fading will be considered for the channel gains associated with the users, since the users may not have line-of-sight connections to their transmitters.

2) NOMA Assisted Content Pushing: Similar to the OMA case, in NOMA, the content pushing phase is also divided into multiple time slots. Again, we assume that popular files are grouped into different libraries and a single time slot is available for pushing the files belonging to the same library. However, while in OMA, during a single time slot, only the most popular file can be pushed, in NOMA, the M_s most popular files belonging to the same library can be pushed, as illustrated in Fig. 2(b).[4] Particularly, the BS superimposes the M_s most popular files as follows:

$$s = \sum_{i=1}^{M_s} \alpha_i \bar{f}_i, \tag{3}$$

where $\bar{f}_i$ denotes the signal that represents the information contained in file f_i, α_i denotes the real-valued power allocation coefficient and $\sum_{i=1}^{M_s} \alpha_i^2 = 1$. The content servers carry out successive interference cancellation (SIC). The SIC decoding order is determined by the priority of the files, i.e., a more popular file, f_i, will be decoded before a less popular file, f_j, $i < j$. Suppose that the files f_j, $j < i$, have been decoded and subtracted correctly by content server CS$_m$. In this case, CS$_m$ can decode the next most popular file, f_i, with the following data rate:

$$R_{m,i}^{CP} = \log\left(1 + \frac{\rho\alpha_i^2 \frac{1}{L(||x_m - x_0||)}}{\rho \frac{1}{L(||x_m - x_0||)} \sum_{j=i+1}^{M_s} \alpha_j^2 + 1}\right). \tag{4}$$

If $R_{m,i}^{CP} \geq R_i$, then file f_i can be decoded and subtracted correctly at CS$_m$.

For a fair comparison with OMA, which pushes only one file at a time, a sophisticated power allocation policy is needed for the NOMA scheme, as discussed in the following. Without loss of generality, we assume that the content servers are ordered

[4]We note that when multiple time slots are available for pushing files belonging to the same library, the proposed NOMA assisted push-then-deliver strategy can still be applied. In particular, the popular files belonging to the same library can be further divided into different groups, with multiple files in each group. Then, the files in the same group can be pushed within one time slot by using the NOMA principle, which means that more files can be pushed compared to OMA. Sophisticated file scheduling algorithms need to be designed in order to optimally use the available time slots and improve the cache hit probability. This is an important direction for future research.

as follows:

$$\frac{1}{L\left(||x_1 - x_0||\right)} \cdots \geq \frac{1}{L\left(||x_m - x_0||\right)} \geq \cdots \geq \frac{1}{L\left(||x_t - x_0||\right)} \geq \cdots, \quad (5)$$

for $1 \leq m < t$. Furthermore, since the considered time slot is used to push f_1 to CS_t, we make the following quality of service (QoS) assumption, in order to facilitate the design of the power allocation coefficients:

QoS Target*: The most popular file, f_1, needs to reach the t-th nearest content server (CS_t).*

Both the OMA and NOMA transmission schemes need to ensure this QoS target. Therefore, the CR inspired power allocation policy can be used for NOMA [7], i.e., power allocation coefficient α_1 is chosen such that f_1 can be delivered reliably to CS_t, i.e.,

$$R_{t,1}^{CP} \geq R_1. \quad (6)$$

This constraint results in the following choice of α_1:

$$\alpha_1^2 = \min\left\{1, \frac{\epsilon_1\left(\rho\frac{1}{L(||x_t - x_0||)} + 1\right)}{\rho(1+\epsilon_1)\frac{1}{L(||x_t - x_0||)}}\right\}, \quad (7)$$

where $\epsilon_l = 2^{R_l} - 1$. The use of the power allocation policy in (7) ensures that the outage probability for pushing file, f_1, to CS_t is the same as that for OMA. The reason is that if there is an outage in OMA, α_1 becomes one, i.e., all the power is allocated to f_1. Or in other words, additional files are pushed in NOMA only if f_1 is pushed to CS_t successfully.

Since $\sum_{j=1}^{M_s} \alpha_j^2 = 1$, (7) implies that the sum of the power allocation coefficients, excluding α_1, is constrained as follows:

$$\sum_{j=2}^{M_s} \alpha_j^2 = \max\left\{0, \frac{\rho\frac{1}{L(||x_t - x_0||)} - \epsilon_1}{\rho(1+\epsilon_1)\frac{1}{L(||x_t - x_0||)}}\right\}. \quad (8)$$

The constraint in (7) is sufficient to guarantee the successful delivery of f_1 to the t-th nearest content server. How the remaining power shown in (8) is allocated to the other files, f_i, $i \neq 1$, does not affect the delivery of f_1. Therefore, in this paper, it is assumed that the portion allocated to f_i, $i \neq 1$, is fixed, i.e., $\alpha_i^2 = \beta_i P_r$, where $P_r = \max\left\{0, \frac{\rho\frac{1}{L(||x_t - x_0||)} - \epsilon_1}{\rho\ (1+\epsilon_1)\frac{1}{L(||x_t - x_0||)}}\right\}$ and β_i are constants, which satisfy the constraint $\sum_{j=2}^{M_s} \beta_i = 1$. Note that the coefficients, β_i, indicate how the remaining transmission power, after the power for f_1 has been deducted is allocated to the additional files.

3) Performance Analysis: An important criterion for evaluating content pushing is the cache hit probability which is the probability that, during the content delivery phase, a user finds its requested file in the cache of its associated content server.[5] Since the request probability for file l is determined by its popularity, the hit probability for a user associated with CS_m can be expressed as follows:

$$\mathrm{P}_m^{hit} = \sum_{i=1}^{M_s} \mathrm{P}(f_i)(1 - \mathrm{P}_{m,i}), \quad (9)$$

where $\mathrm{P}_{m,i}$ denotes the outage probability of CS_m for decoding file i. Note that for the OMA case, only file f_1 will be sent, and hence the corresponding OMA hit probability is simply given by

$$\mathrm{P}_{m,OMA}^{hit} = \mathrm{P}(f_1)(1 - \mathrm{P}_{m,1}^{OMA}), \quad (10)$$

where $\mathrm{P}_{m,1}^{OMA}$ denotes the outage probability of CS_m for decoding file f_1. The following theorem reveals the benefit of using NOMA for content pushing.

Theorem 1: The cache hit probability achieved by the proposed NOMA assisted push-then-deliver strategy is always larger than or at least equal to that of the conventional OMA based strategy, i.e.,

$$\mathrm{P}_m^{hit} \geq \mathrm{P}_{m,OMA}^{hit}, \quad (11)$$

for $1 \leq m \leq t$.

Proof: See Appendix A. □

Remark 1: Theorem 1 demonstrates that the proposed NOMA assisted caching strategy outperforms OMA based caching. The superior performance of the NOMA assisted caching strategy originates from the fact that multiple content files are pushed concurrently during the content pushing phase.

Remark 2: Only the t nearest content servers are of interest in (11), i.e., $1 \leq m \leq t$, which is due to our assumption that the BS aims to push the most popular file, f_1, to CS_t.

Remark 3: As shown in Appendix A, the key step to prove the theorem is to show $\mathrm{P}_{m,1}^{OMA} = \mathrm{P}_{m,1}$, i.e., the outage performance of NOMA for decoding f_1 at CS_m is the same as that of OMA. If f_1 is viewed as the message to the primary user in a CR NOMA system, this observation about the equivalence between the outage performances of NOMA and OMA is consistent with the results in [7] and [13].

While the use of the CR power allocation policy guarantees that CS_t can decode f_1, this also implies that the outage performance at CS_m depends on the channel conditions of CS_t. This means that for calculation of the outage probability, $\mathrm{P}_{m,i}$, the joint distribution of the ordered distances of CS_t and CS_m to the BS is needed. The following lemma provides an analytical expression for this joint distribution.

Lemma 1: Denote the distance between the BS and the i-th nearest content server by r_i. The joint pdf of r_m and r_t is given by

$$f_{r_m,r_t}(x,y) = 4y(\lambda_c\pi)^t e^{-\lambda_c\pi y^2}\frac{x^{2m-1}(y^2 - x^2)^{t-m-1}}{(t-m-1)!(m-1)!}. \quad (12)$$

Proof: See Appendix B. □

Remark 4: We note that Lemma 1 is general and can be applied to any two HPPP nodes that are ordered according to their distances to the origin.

Remark 5: It is worth pointing out that the joint pdf obtained in [36] is a special case of Lemma 1, when $m = 1$ and $t = 2$.

[5]We note that retransmission for content pushing, where after decoding failure, CS_t requests the retransmission of file f_1 by the BS, is not considered in this paper. However, investigating the impact of file retransmission on the caching performance is an important direction for future research.

Since the cache hit probability is a function of the outage probability, we provide the outage performance for content pushing in the following lemma.

Lemma 2: Assume $\epsilon_{M_s} \geq \epsilon_1$. The outage probability of CS$_n$, $1 \leq n \leq t$, for decoding f_1 is given by

$$\mathrm{P}_{n,1} = e^{-\lambda_c \pi \left(\frac{\rho}{\epsilon_1}\right)^{\frac{2}{\alpha}}} \sum_{k=0}^{n-1} \frac{(\lambda_c \pi)^k \left(\frac{\rho}{\epsilon_1}\right)^{\frac{2k}{\alpha}}}{k!}. \tag{13}$$

The outage probability of CS$_t$ for decoding f_i, $2 \leq i \leq M_s$, is given by

$$\mathrm{P}_{t,i} = e^{-\lambda_c \pi \left(\frac{\epsilon_1}{\rho} + \frac{(1+\epsilon_1)}{\rho \phi_i}\right)^{-\frac{2}{\alpha}}} \sum_{k=0}^{t-1} \frac{(\lambda_c \pi)^k \left(\frac{\epsilon_1}{\rho} + \frac{(1+\epsilon_1)}{\rho \phi_i}\right)^{-\frac{2k}{\alpha}}}{k!}, \tag{14}$$

where $\phi_i = \min\left\{\frac{\bar{\xi}_2}{\epsilon_2}, \cdots, \frac{\bar{\xi}_i}{\epsilon_i}\right\}$, $\bar{\xi}_i = \left(\beta_i - \epsilon_i \sum_{j=i+1}^{M_s} \beta_j\right)$ for $2 \leq i < M_s$, and $\bar{\xi}_{M_s} = \beta_{M_s}$.

The outage probability of CS$_m$, $1 \leq m < t$, for decoding f_i, $2 \leq i \leq M_s$, is given by

$$\begin{aligned}
&\mathrm{P}_{m,i} \\
&\approx \mathrm{P}_{t,1} + \frac{4(\lambda_c \pi)^t}{(t-m-1)!(m-1)!} \sum_{p=0}^{t-m-1} (-1)^p \binom{t-m-1}{p} \\
&\quad \times \sum_{l=1}^{N} \frac{\pi(\tau_2 - \tau_1)}{2N} f_m\left(\frac{\tau_2 - \tau_1}{2} w_l + \frac{\tau_2 + \tau_1}{2}\right) \sqrt{1 - w_l^2},
\end{aligned}$$

where $\tau_1 = \left(\frac{\rho \phi_i}{1+\epsilon_1+\epsilon_1 \phi_i}\right)^{\frac{1}{\alpha}}$, $\tau_2 = \left(\frac{\epsilon_1}{\rho}\right)^{-\frac{1}{\alpha}}$, N denotes the parameter for Chebyshev-Gauss quadrature, $w_l = \cos\left(\frac{2l-1}{2N}\pi\right)$, $g(y) = \left(\frac{(1+\epsilon_1)}{\phi_i(\rho - \epsilon_1\ y^\alpha)}\right)^{-\frac{1}{\alpha}}$, and

$$\begin{aligned}
f_m(y) = &\frac{e^{-\lambda_c \pi y^2} y^{2(t-m-1)-2p+1}}{2m+2p} \\
&\times \left(y^{2m+2p} - (g(y))^{2m+2p}\right).
\end{aligned} \tag{15}$$

Proof: See Appendix C. □

Remark 6: By using the closed-form expressions developed in Lemma 2, the outage probabilities for content pushing can be evaluated in a straightforward manner, and computationally challenging Monte Carlo simulations can be avoided. The high computational complexity of Monte Carlo simulations can be illustrated by the following example. Consider simulation of an outage probability of 10^{-4}. In this case, we need to carry out at least one million independent experiments in order to obtain an accurate estimate. Furthermore, since the distribution of the content servers follows a Poisson point process, for each experiment, we need to generate the locations of hundreds of content servers, and calculate the distances between the content servers and the BS, which further increases the complexity of Monte Carlo simulations. Therefore, using the analytical results in Lemma 2 for performance evaluation yields a significant reduction in computational complexity.

Remark 7: We note that the approximations in Lemma 2 are obtained by applying the Chebyshev-Gauss approximation. The Chebyshev-Gauss approximation parameter, N, is used to achieve a tradeoff between accuracy and complexity. Particularly, if $N \to \infty$, the approximation error goes to zero. The numerical results provided in Section V show that a choice of $N = 20$ is sufficient to obtain an accurate approximation.

Remark 8: We note that, in Lemma 2 it is assumed that the target data rates and the power allocation coefficients are chosen to ensure $\bar{\xi}_i > 0$. Otherwise, an outage will always happen for decoding file f_i, $i \geq 2$, at the content servers.

Remark 9: In Lemma 2, it is also assumed that $\epsilon_{M_s} \geq \epsilon_1$, in order to avoid a trivial case for the integral calculation; see also (74) in Appendix C. This assumption means that the target data rate for file f_{M_s} is larger than that for file f_1, which is important to the performance gain of the NOMA assisted strategy over the OMA based strategy, as explained in the following. Recall that the CR power allocation policy in (7) ensures that CS$_t$ can decode the pushed file f_1. If there is any power left after satisfying the needs of CS$_t$, the BS will use the remaining power to push additional files. If the target data rate of f_1, R_1, is very large, meeting the decoding requirement of CS$_t$ becomes challenging, i.e., it is likely that there is not much power available for pushing additional files. In this case, the use of the proposed push-then-deliver strategy will not offer much performance gain compared to OMA. In other words, a large R_1 is not ideal for applying the proposed NOMA based push-then-deliver strategy. The proposed strategy is most effective when the target data rate of f_1 is small.

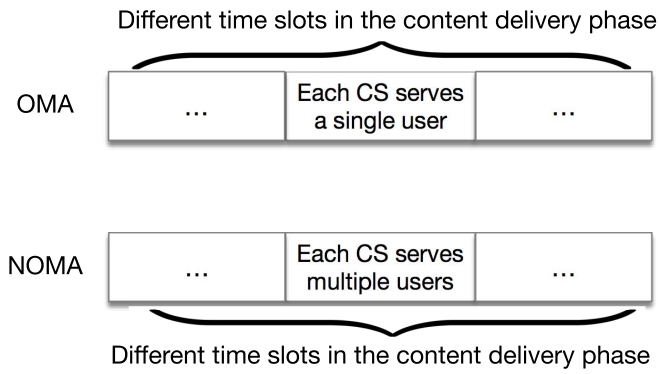
B. Content Delivery Phase

In the previous subsection, the cache hit probability for content delivery has been analyzed. However, the event that a user can find its requested file in the cache of its associated content server is not equivalent to the event that this user can receive the file correctly, due to the multi-path fading and path loss attenuation that affect its link to the content server. Hence, in this subsection, the impact of NOMA on the reliability of content delivery is investigated. Similar to the previous subsection, the conventional OMA based content delivery strategy is described first as a benchmark scheme.

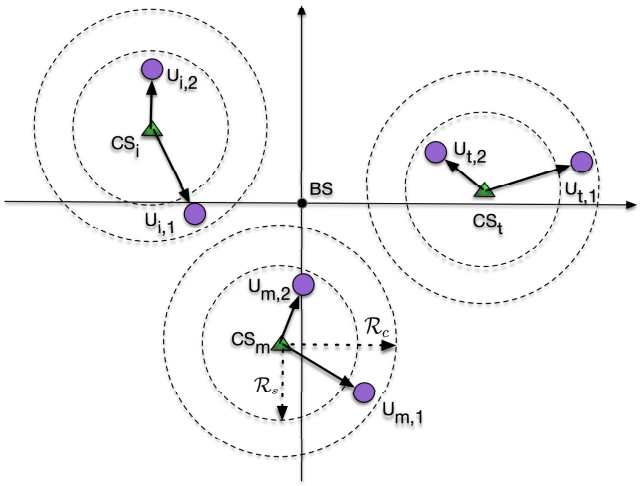
1) OMA Based Content Delivery: Similar to the content pushing phase, the content delivery phase is also divided into multiple time slots, as shown in Fig. 3(a). During each time slot, for the OMA case, each content server randomly schedules a single user whose requested file is available in its cache. We assume that each content server can find a user to serve for this OMA based content delivery, and multiple content servers help their associated users simultaneously.

2) NOMA Assisted Content Delivery: If the NOMA principle is applied in the content delivery phase, each content server can serve two users.[6] Thereby, it is assumed that each

[6]We focus on the case with two users since the content delivery phase is analogous to the conventional downlink case and two-user NOMA based downlink transmission has been proposed for long term evolution (LTE) Advanced [37]. The analytical results presented in this paper can be extended to the case with more than two users by dividing the disc covered by a content server into multiple rings. In practice, the number of users to be served simultaneously needs to reflect a practical tradeoff between system complexity and throughput.



(a) A general illustration for the content delivery phase



(b) An example for the outcome of content pushing

Fig. 3. An illustration of the impact of NOMA on content delivery. In OMA, each content server serves a single user. By using NOMA, an additional user can be served.

content server can find at least two users whose requests can be accommodated locally.[7] This assumption is applicable to high-density wireless networks, such as networks deployed in sport stadiums or airports, where the number of users is much larger than the number of content servers. We note that this assumption constitutes the worst case for the reception reliability of the users. In fact, content servers that do not have any user to serve will not cause interference to the users served by other content servers. Without loss of generality, assume that the two users are ordered based on their distances to the associated content server. As shown in Fig. 3, the far user, which is far from the content server and is denoted by $\mathrm{U}_{m,1}$, is inside a ring bounded by radii $\mathcal{R}_s$ and $\mathcal{R}_c$, $\mathcal{R}_s < \mathcal{R}_c$. The near user, which is close to the content server and is denoted by $\mathrm{U}_{m,2}$, is inside a disc with radius $\mathcal{R}_s$. Without loss of generality, denote the file requested by $\mathrm{U}_{m,k}$ by $f_{m,k}$, $f_{m,k} \in \mathcal{F}$. Each content server broadcasts a superposition signal containing two messages, and $\mathrm{U}_{m,k}$, which is associated with CS_m, receives the following signal:

$$y_{m,k} = \underbrace{\frac{h_{m,mk}}{\sqrt{L\left(||y_{m,k}||\right)}} \sum_{l=1}^{2} \alpha_l \bar{f}_{m,l}}_{\text{Signals from CS}_m} + \underbrace{\sum_{x_j \in \Phi_c \setminus m} \frac{h_{j,mk}}{\sqrt{L\left(||y_{m,k} + x_m - x_j||\right)}} \sum_{l=1}^{2} \alpha_l \bar{f}_{j,l}}_{\text{Signals from interfering clusters}} + n_{m,k}, \tag{16}$$

where $\bar{f}_{j,l}$ denotes the signal that represents the information contained in file $f_{j,l}$, α_l denotes the NOMA power allocation coefficient, $n_{m,k}$ is additive complex Gaussian noise, and $h_{j,mk}$ denotes the fading channel coefficient between CS_j and $\mathrm{U}_{m,k}$. We assume that the channels between content servers and users suffer from both large scale path loss and small scale multi-path fading. In particular, $h_{j,mk}$ is assumed to be quasi-static Rayleigh fading, which means that the channels stay constant for one block and change independently from one block to the next.[8] In order to obtain tractable analytical results, fixed power allocation is used, instead of CR power allocation, and it is assumed that all content servers use the same fixed power allocation coefficients. In order to keep the notation consistent, the power allocation coefficients are still denoted by α_i. We note that the simulation results provided in Section V show that the use of this fixed power allocation can still ensure that NOMA outperforms OMA for both users.

$\mathrm{U}_{m,1}$ will treat its partner's message as noise and decode its own message $f_{m,1}$ with the following signal-to-interference-plus-noise ratio (SINR):

$$\mathrm{SINR}_{m,1}^{1} = \frac{\frac{\alpha_1^2 |h_{m,m1}|^2}{L(||y_{m,1}||)}}{\frac{\alpha_2^2 |h_{m,m1}|^2}{L(||y_{m,1}||)} + \mathrm{I}_{inter}^{m,1} + \frac{1}{\rho}}, \tag{17}$$

where

$$\mathrm{I}_{inter}^{m,1} = \sum_{x_j \in \Phi_c \setminus m} \frac{|h_{j,m1}|^2}{L\left(||y_{m,1} + x_m - x_j||\right)}.$$

In practice, the content servers are expected to use less transmission power than the BS, but for notational simplicity, ρ is still used to denote the ratio between the transmission power of the content servers and the noise power. In Section V, for the presented computer simulation results, different transmission powers are adopted for the BS and the content servers.

The near user, $\mathrm{U}_{m,2}$, intends to first decode its partner's message with the data rate $\log(1 + \mathrm{SINR}_{m,2}^{1})$, where $\mathrm{SINR}_{m,2}^{1}$ is defined similarly to $\mathrm{SINR}_{m,1}^{1}$,

[7]The proposed content delivery strategy can be opportunistically applied whenever there is a content server with two users to serve. The only case where it is not applicable is if all content servers in the network have no user or just one user to serve. This unlikely scenario may happen if the number of files is large compared to the number of users and the file popularity distribution is not very concentrated. However, this scenario is not ideal for the application of wireless caching in general, with or without NOMA, since the cache hit probability is small for this scenario. We note that we did not make any assumption regarding which particular user makes a request to its content server. This is the reason why the locations and fading channels of the users are modelled as random.

[8]For the case in which the coherence time of the channels between the content servers and the users is too short to deliver an entire file, we note that unlike for content pushing, more time slots are available for content delivery during on-peak hours. Therefore, a content server can allocate multiple time slots to one user, where one file is split into multiple subfiles and each subfile is transmitted within one fading block.

i.e., $\mathrm{SINR}^1_{m,2} = \frac{\frac{\alpha_2^2|h_{m,m2}|^2}{L(||y_{m,2}||)}}{\frac{\alpha_2^2|h_{m,m2}|^2}{L(||y_{m,2}||)}+\mathrm{I}^{m,2}_{inter}+\frac{1}{\rho}}$, and the inter-cluster interference, $\mathrm{I}^{m,2}_{inter}$, is defined similarly to $\mathrm{I}^{m,1}_{inter}$. If $\log(1+\mathrm{SINR}^1_{m,2}) > R_1$, i.e., $\mathrm{U}_{m,2}$ can decode its partner's message successfully, $\mathrm{U}_{m,2}$ will remove $f_{m,1}$ and decode its own message with the following SINR:

$$\mathrm{SINR}^2_{m,2} = \frac{\frac{\alpha_2^2|h_{m,m2}|^2}{L(||y_{m,2}||)}}{\mathrm{I}^{m,2}_{inter}+\frac{1}{\rho}}. \quad (18)$$

The outage probabilities of the two users are defined as follows:

$$\mathrm{P}^1_{m,1} = \mathrm{P}(\log(1+\mathrm{SINR}^1_{m,1}) < R_1), \quad (19)$$

and

$$\mathrm{P}^2_{m,2} = 1 - \mathrm{P}(\log(1+\mathrm{SINR}^1_{m,2}) > R_1, \log(1+\mathrm{SINR}^2_{m,2}) > R_2). \quad (20)$$

The following lemma provides closed-form expressions for these outage probabilities.

Lemma 3: The outage probability of $U_{m,2}$ can be expressed as follows:

$$\mathrm{P}^o_{m,2} \approx 1 - \sum_{n=1}^{N} \bar{w}_n e^{-\frac{c_{n,\mathcal{R}_s}\frac{1}{\rho}}{\tilde{\tau}}} q\left(\frac{c_{n,\mathcal{R}_s}}{\tilde{\tau}}\right), \quad (21)$$

where $\tilde{\tau} = \min\left\{\frac{\alpha_1^2-\epsilon_1\alpha_2^2}{\epsilon_1}, \frac{\alpha_2^2}{\epsilon_2}\right\}$, $q(s) = \exp\left(-2\pi\lambda_c \frac{s^{\frac{2}{\alpha}}}{\alpha} B\left(\frac{2}{\alpha}, \frac{\alpha-2}{\alpha}\right)\right)$, $B(\cdot,\cdot)$ *denotes the beta function,* $\bar{w}_n = \frac{\pi}{2N}\sqrt{1-w_n^2}(w_n+1)$, w_n *is defined in Lemma 2, and* $c_{n,r} = \left(\frac{r}{2}w_n + \frac{r}{2}\right)^{\alpha}$.

The outage probability of $U_{m,1}$ can be expressed as follows:

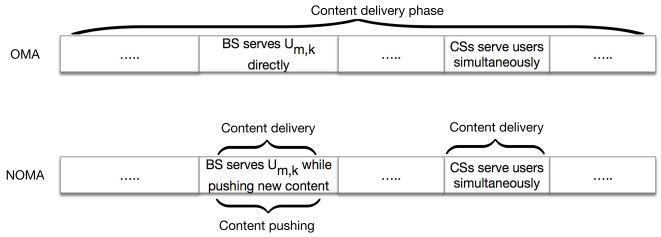
$$\mathrm{P}^o_{m,1} \approx 1 + \frac{\mathcal{R}_s^2}{\mathcal{R}_c^2-\mathcal{R}_s^2}\sum_{n=1}^{N} \bar{w}_n e^{-\frac{c_{n,\mathcal{R}_s}\frac{\epsilon_1}{\rho}}{\alpha_1^2-\epsilon_1\alpha_2^2}} q\left(e^{-\frac{c_{n,\mathcal{R}_s}\epsilon_1}{\alpha_1^2-\epsilon_1\alpha_2^2}}\right) - \frac{\mathcal{R}_c^2}{\mathcal{R}_c^2-\mathcal{R}_s^2}\sum_{n=1}^{N} \bar{w}_n e^{-\frac{c_{n,\mathcal{R}_c}\frac{\epsilon_1}{\rho}}{\alpha_1^2-\epsilon_1\alpha_2^2}} q\left(e^{-\frac{c_{n,\mathcal{R}_c}\epsilon_1}{\alpha_1^2-\epsilon_1\alpha_2^2}}\right). \quad (22)$$

Proof: See Appendix D. □

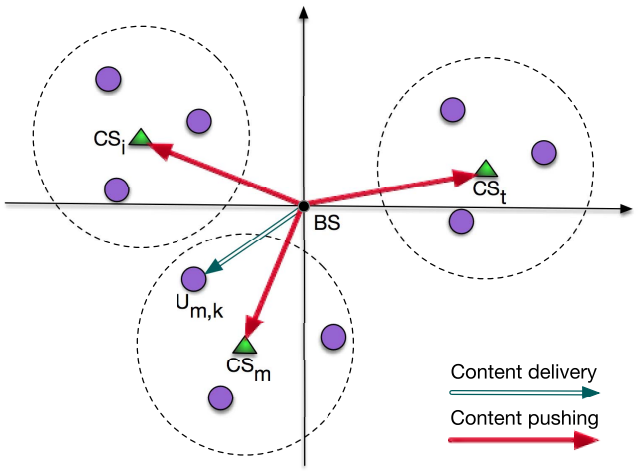
Remark 10: In the previous subsection, the CR power allocation policy is used and this type of power allocation ensures that the NOMA outage performance of the far user, $\mathrm{U}_{m,1}$, is the same as that for OMA. Since fixed power allocation coefficients are used in this subsection for content pushing, the performance of the far user is no longer guaranteed, but surprisingly, our simulation results indicate that the use of NOMA can still yield an outage performance gain for the far user, compared to OMA, as shown in Section V.

IV. Push-and-Deliver Strategy

While the proposed push-then-deliver strategy ensures that more files can be pushed to the content servers during a short content pushing phase, it still relies on the same principle as conventional caching in the sense that content pushing and content delivery are separately carried out. Therefore, if the time interval between two adjacent content pushing phases is large, the caches of the content servers can be updated only infrequently. If new content arrives during the content delivery phase, the use of both conventional caching and the proposed push-then-deliver strategy means that the BS has to wait until the next content pushing phase in order to update the caches of the content servers. In contrast, the proposed push-and-deliver strategy provides an efficient mechanism for frequently updating the files cached at the content servers by exploiting opportunities for content pushing during the content delivery phase as illustrated in Fig. 4(a). In particular, such opportunities arise when the BS has to serve a user directly during the content delivery phase, as explained in the following.



(a) General principle of push-and-deliver



(b) An illustration of push-and-deliver

Fig. 4. An illustration of the proposed push-and-deliver strategy.

In particular, consider a time slot that is dedicated to user $\mathrm{U}_{m,k}$, as shown in Fig. 4(b). During this time slot, if OMA is used, only this user can be served by the BS directly. However, the use of the NOMA principle offers the opportunity to also push new content to the content servers,[9] i.e., the BS sends a superposition signal containing the file requested by $\mathrm{U}_{m,k}$, denoted by f_0, and the M_s most popular files pushed by the BS, denoted by f_i, $1 \le i \le M_s$. Assume that f_0 and f_i, $1 \le i \le M_s$, belong to different file libraries, and the newly pushed files are useful to all the content servers, in order to avoid

[9]We note that the BS cannot apply the proposed push-and-deliver strategy in each time slot, as the need to serve a user directly is a prerequisite. Particularly, the proposed strategy is applied in time slots in which a user cannot find the requested file at its local content server. More specifically, in these time slots, the BS pushes new content to the content servers, while serving the user directly. In other words, the push-and-deliver strategy can be applied whenever a user has to be served directly by the BS.

correlation among these files and to simplify the expression for the cache hit probability. This assumption is reasonable in practice and can be justified by using the following scenario as an example. Assume that after a period following the initial content pushing phase, one user needs to be served by the BS directly, and during this period, a new library of files, e.g., a set of videos for breaking news, has arrived at the BS to be pushed to the content servers. None of the content servers has had a chance to cache these files yet. With the proposed push-and-deliver strategy, the BS can push the new files to the content servers, while serving the requesting user directly, i.e., waiting for the next content pushing phase is avoided. We note that, in the case where the M_s most popular files of one library have already been pushed, we can either push files from a new library that has not been scheduled before, or keep scheduling files from the same library, by simply ignoring the previously pushed content and transmitting files that have not been sent yet.

A. Performance Analysis

Following steps similar to those in the previous section, the data rate of $\mathrm{U}_{m,k}$ for decoding its requested file, f_0, which is directly sent by the BS, is given by

$$\mathrm{R}_{m,k} = \log\left(1 + \frac{\frac{\alpha_0^2|h_{mk}|^2}{L(||y_{m,k}+x_m||)}}{\sum_{l=1}^{M_s}\frac{\alpha_l^2|h_{mk}|^2}{L(||y_{m,k}+x_m||)} + \frac{1}{\rho}}\right), \quad (23)$$

and each content server, CS_m, can decode the additionally pushed file f_i with the following data rate:

$$\mathrm{R}_m^l = \log\left(1 + \frac{\frac{\alpha_l^2}{L(||x_m||)}}{\sum_{l=i+1}^{M_s}\frac{\alpha_l^2}{L(||x_m||)} + \frac{1}{\rho}}\right), \quad (24)$$

if R_m^j is larger than R_j, for $0 \le j \le i-1$, where R_l denotes the target data rate of f_l. Again, small scale multi-path fading is not considered in the channel model for CS_m, as we assume that the large scale path loss is dominant in this case, but small scale fading is considered for the users' channels. Note that the indices of the power allocation coefficients α_i start from 0, due to file f_0. Compared to the distance between CS_m and the BS, the corresponding distance between $\mathrm{U}_{m,k}$ and the BS has a very complicated pdf, as shown in the following subsection. Therefore, in order to obtain tractable analytical results, fixed power allocation coefficients α_i will be used, instead of the CR based ones. The outage probabilities of the user and the content servers will be studied in the following subsections, respectively.

1) Performance of the User: The main challenge in analyzing the outage performance at the user is the complicated expression for the pdf of the distance $||y_{m,k} + x_m||$. First, we define $\bar{z}_{m,k} = \frac{|h_{mk}|^2}{L(||y_{m,k}+x_m||)}$. The outage probability at the user can be expressed as follows:

$$\begin{aligned}\mathrm{P}_{m,k}^1 &= \mathrm{P}(\mathrm{R}_{m,k} < R_0) = \mathrm{P}\left(\bar{z}_{m,k} < \frac{\epsilon_0}{\rho\zeta_1}\right)\\ &= \mathcal{E}_{L(||y_{m,k}+x_m||)}\left\{1 - e^{-L(||y_{m,k}+x_m||)\frac{\epsilon_0}{\rho\zeta_0}}\right\}, \quad (25)\end{aligned}$$

where $\zeta_l = \alpha_l^2 - \epsilon_l\sum_{j=l+1}^{M_s}\alpha_j^2$ for $0 \le l < M_s$, and $\zeta_{M_s} = \alpha_{M_s}^2$. Again, it is assumed that the power allocation coefficients and the target data rates are carefully chosen to ensure that ζ_l is positive.

In order to derive the pdf of $||y_{m,k} + x_m||$, we first define $r_m = ||x_m||$ and also a function

$$g(r_m, r) = \frac{2r\arccos\frac{r_m^2+r^2-\mathcal{R}_c^2}{2r_m r}}{\pi\mathcal{R}_c^2}.$$

Conditioned on r_m, the pdf of $||y_{m,k} + x_m||$ is given by [38]

$$f_{||y_{m,k}+x_m||}(r|r_m) = g(r_m, r), \quad (26)$$

for $r_m - \mathcal{R}_c \le r \le r_m + \mathcal{R}_c$, if $r_m > \mathcal{R}_c$. Otherwise, we have

$$\begin{aligned}&f_{||y_{m,k}+x_m||}(r|r_m)\\ &= \begin{cases}2\pi r, & \text{if } r \le \mathcal{R}_c - r_m\\ 2\pi r - g(r_m, r), & \text{if } \mathcal{R}_c - r_m < r \le \sqrt{\mathcal{R}_c^2 - r_m^2}\\ g(r_m, r), & \text{if } \sqrt{\mathcal{R}_c^2 - r_m^2} < r \le \mathcal{R}_c + r_m.\end{cases} \quad (27)\end{aligned}$$

In order to avoid the trivial cases, which lead to $r = 0$, i.e., the user is located at the same place as the BS, we assume that no content server can be located inside the disc, denoted by $\mathcal{B}(x_0, \delta\mathcal{R}_c)$, i.e., a disc with the BS located at its origin and radius $\delta\mathcal{R}_c$ with $\delta > 1$, which means that $r_m \ge \delta\mathcal{R}_c$ for all $m \ge 1$. Therefore, only the expression in (26) needs to be used since r_m is strictly larger than $\mathcal{R}_c$.

After using the pdf of $||y_{m,k} + x_m||$, the outage probability can be expressed as follows:

$$\mathrm{P}_{m,k}^1 = 1 - \int_{\delta\mathcal{R}_c}^{\infty}\int_{z-\mathcal{R}_c}^{z+\mathcal{R}_c} e^{-\frac{\epsilon_0 r^\alpha}{\rho\zeta_0}} g(z, r)dr\bar{f}_{r_m}(z)dz, \quad (28)$$

where $\bar{f}_{r_m}(z)$ denotes the pdf of r_m. Because of the assumption that no content server can be located inside of $\mathcal{B}(x_0, \delta\mathcal{R}_c)$, the pdf of r_m is different from that in (50), but the steps of the proof for [39, Theorem 1] can still be applied to obtain the pdf of r_m. Particularly, first denote by $\mathcal{A}_r$ the ring with inner radius $\delta\mathcal{R}_c$ and outer radius r. The cumulative distribution function (CDF) of r_m can be expressed as follows:

$$\begin{aligned}\tilde{F}_{r_m}(r) &= 1 - \mathrm{P}(\#\text{ of nodes in the ring } \mathcal{A}_r < m)\\ &= 1 - \sum_{l=0}^{m-1}\frac{(\lambda_c[\pi r^2 - \pi\delta^2\mathcal{R}_c^2])^l}{l!}e^{-\lambda_c[\pi r^2 - \pi\delta^2\mathcal{R}_c^2]}.\end{aligned} \quad (29)$$

Therefore, the pdf of r_m can be calculated as follows:

$$\begin{aligned}&\bar{f}_{r_m}(r)\\ &= -2\pi\lambda_c r e^{-\lambda_c[\pi r^2 - \pi\delta^2\mathcal{R}_c^2]}\left(\sum_{l=1}^{m-1}\frac{(\lambda_c[\pi r^2 - \pi\delta^2\mathcal{R}_c^2])^{l-1}}{(l-1)!}\right.\\ &\quad\left. - \sum_{l=0}^{m-1}\frac{(\lambda[\pi r^2 - \pi\delta^2\mathcal{R}_c^2])^l}{l!}\right)\\ &= 2\pi\lambda_c^m r e^{-\lambda_c[\pi r^2 - \pi\delta^2\mathcal{R}_c^2]}\frac{[\pi r^2 - \pi\delta^2\mathcal{R}_c^2]^{m-1}}{(m-1)!}.\end{aligned} \quad (30)$$

Substituting (30) into (28), the outage probability of the user can be obtained.

2) Performance of the Content Servers: The content servers need to carry out SIC in order to decode the newly pushed files f_l. As a result, the outage probability of CS_m for decoding f_i can be expressed as follows:

$$\begin{aligned}\text{P}_m^i &= 1-\text{P}(\text{R}_m^l > R_l, \forall l \in \{0,\cdots,i\}) \\ &= \text{P}\left(L(||x_m||) > \min\left\{\frac{\rho\zeta_l}{\epsilon_l}, \forall l \in \{0,\cdots,i\}\right\}\right). \quad (31)\end{aligned}$$

By applying the assumption that $r_m \geq \delta\mathcal{R}_c$ and also the pdf in (30), the outage probability of CS_m for decoding f_i can be expressed as follows:

$$\text{P}_m^i = \sum_{l=0}^{m-1} \frac{(\lambda_c\left[\frac{\pi}{\bar{\tau}_i^2} - \pi\delta^2\mathcal{R}_c^2\right])^l}{l!} e^{-\lambda_c\left[\frac{\pi}{\bar{\tau}_i^2} - \pi\delta^2\mathcal{R}_c^2\right]}, \quad (32)$$

where $\bar{\tau}_i = \left(\frac{1}{\min\left\{\frac{\rho\zeta_l}{\epsilon_l}, \forall l \in \{0,\cdots,i\}\right\}}\right)^{\frac{1}{\alpha}}$.

Based on the outage probability P_m^i, the corresponding cache hit probability for a user associated with CS_m can be expressed as follows:

$$\text{P}_m^{hit} = \sum_{i=1}^{M_s} \text{P}(f_i)(1-\text{P}_m^i), \quad (33)$$

where f_0 has been omitted as it is the file currently requested by a user and it is assumed that f_0 and f_l, $1 \leq l \leq M_s$, belong to different libraries.

B. OMA Benchmarks

A naive OMA based benchmark is that the BS does not push new content while serving a user directly. Compared to this naive OMA scheme, the benefit of the proposed push-and-deliver strategy is obvious since new content is delivered and the cache hit probability will be improved.

A more sophisticated OMA scheme is to divide a single time slot into (M_s+1) sub-slots. During the first sub-slot, the user is served directly by the BS. From the second until the (M_s+1)-th sub-slots, the BS will individually push the files, f_i, $i \in \{1,\cdots,M_s\}$, to the content servers. Compared to this more sophisticated OMA scheme, the use of the proposed push-and-deliver strategy still offers a significant gain in terms of the cache hit probability, as will be shown in Section V.

C. Extension to D2D Caching

The aim of this subsection is to show that the concept of push-and-deliver can also be applied to D2D caching. Assume that a time slot is dedicated to a user whose request cannot be found in the caches of its neighbors, and during this time slot, the BS will send the requested file f_0 to the user directly. By applying the push-and-deliver strategy, the BS will also proactively push M_s new files, f_l, $1 \leq l \leq M_s$, to all users for future use. In other words, when the BS addresses the current demand of a user directly, the BS pushes more content files to all users, including the user that requests f_0, for future use.

In the context of D2D caching, content servers are no longer needed. Therefore, the spatial model presented in Section II needs to be revised accordingly. Particularly, it is assumed that the locations of the users are denoted by y_k and modelled as an HPPP, denoted by Φ_u, with density λ_u.

After implementing the push-and-deliver strategy, following steps similar to those in the previous subsection, for a user with distance r from the BS and Rayleigh fading channel gain h, the outage probability for decoding f_i is given by

$$\text{R}_r^i = \log\left(1+\frac{\alpha_i^2|h|^2 r^{-\alpha}}{\sum_{j=i+1}^{M_s}\alpha_j^2|h|^2 r^{-\alpha}+\frac{1}{\rho}}\right), \quad (34)$$

when $\text{R}_r^l > R_l$, for $0 \leq l \leq i-1$. If f_j, $0 \leq j \leq M_s-1$, can be decoded correctly, f_{M_s} can be decoded by this user with the following data rate:

$$\text{R}_r^{M_s} = \log\left(1+\rho\alpha_{M_s}^2|h|^2 r^{-\alpha}\right). \quad (35)$$

Consequently, for a user with distance r from the BS, the probability of successfully decoding f_i can be expressed as follows:

$$\begin{aligned}\text{P}^i(r) &= \text{P}\left(\text{R}_r^l > R_l, \forall l \in \{0,\cdots,i\}\right) \quad (36)\\ &= e^{-\bar{\tau}_i^{\alpha} r^{\alpha}}.\end{aligned}$$

Following (36), one can draw the conclusion that the locations of the users that can successfully receive f_i no longer follow the original HPPP with λ_u, but follow an inhomogeneous PPP which is thinned from the original HPPP by $\text{P}^i(r)$, i.e., the density of this new PPP is $\text{P}^i(r)\lambda_u$. By using this thinning process, the cache miss probability can be characterized as follows.

During the D2D content delivery phase, assume a newly arrived user, whose location is denoted by y_0, requests file f_i. Denote by $\mathcal{B}(y_0,d)$ a disc with radius d and origin located at y_0. This disc is the area in which the user searches for a helpful neighbor that has the requested file in its cache. For this inhomogeneous PPP, the cache hit probability for the user requesting f_i is given by

$$\begin{aligned}\text{P}_i^{hit} &= 1-\text{P}(\text{no user in } \mathcal{B}(y_0,d) \text{ caches } f_i) \\ &= 1-e^{-\Lambda_i\ (\mathcal{B}(y_0,d))}, \quad (37)\end{aligned}$$

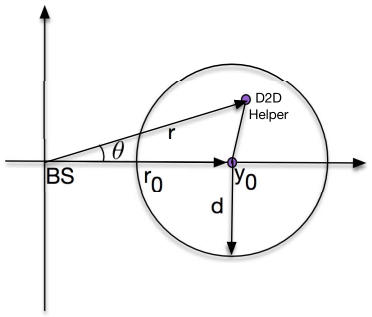
where $\Lambda_i\ (\mathcal{B}(y_0,d))$ denotes the intensity measure of the inhomogeneous PPP for the users that have f_i in their caches. In (37), the hit probability is found by determining the cache miss probability which corresponds to the event that the user cannot find its requested file in the caches of its neighbors located in the disc $\mathcal{B}(y_0,d)$. The calculation of the cache hit probability depends on the relationship between d and the distance between the observing user and the BS, denoted by r_0, as shown in the following subsections.

1) For the Case of $d < r_0$: For $d < r_0$, define $\Lambda_i\ (\mathcal{B}(y_0,d)) \triangleq \Lambda_{d\leq r_0}^i(r_0)$. The assumption $d < r_0$ means that the BS is excluded from $\mathcal{B}(y_0,d)$. Therefore, the intensity measure can be calculated as follows:

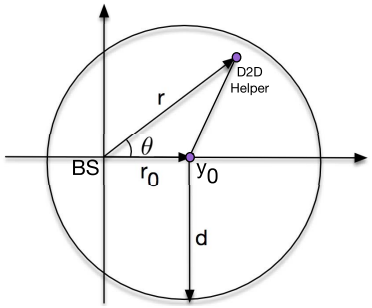
$$\Lambda_{d\leq r_0}^i(r_0) = \int\int_{r,\theta\in\mathcal{B}(y_0,d)} \text{P}^i(r)\lambda_u d\theta r dr. \quad (38)$$

As can be observed from Fig. 5, the constraint on r and θ can be expressed as follows:

$$r^2 + r_0^2 - 2r_0 r\cos\theta \leq d^2. \quad (39)$$



(a) $d < r_0$



(b) $d \geq r_0$

Fig. 5. Two possible cases between the radius of the search disc, $\mathcal{B}(y_0, d)$, and the distance between the observing user located at y_0 and the BS.

Therefore, the intensity measure can be expressed as follows:

$$\Lambda^i_{d\leq r_0}(r_0) = \int_{r_0-d}^{r_0+d} \int_{-\arccos\frac{r^2+r_0^2-d^2}{2r_0 r}}^{\arccos\frac{r^2+r_0^2-d^2}{2r_0 r}} \mathrm{P}^i\,(r)\lambda_u d\theta r dr$$
$$= 2\lambda_u \int_{r_0-d}^{r_0+d} \mathrm{P}^i(r) r \arccos\frac{r^2+r_0^2-d^2}{2r_0 r} dr. \quad (40)$$

By applying Chebyshev-Gauss quadrature, the intensity measure can be approximated as follows:

$$\Lambda^i_{d\leq r_0}(r_0) \approx 2\lambda_u d \sum_{l=1}^{N} \frac{\pi}{N} g_r\left(r_0 + d w_l\right)\sqrt{1-w_l^2}, \quad (41)$$

where $g_r(z)$ is given by

$$g_r(z) = \mathrm{P}^i(z) z \arccos\frac{z^2+r_0^2-d^2}{2r_0 z}. \quad (42)$$

2) For the Case of $d \geq r_0$: For $d \geq r_0$, define $\Lambda^i(\mathcal{B}(y_0, d)) \triangleq \Lambda_{d>r_0}$. The assumption, $d \geq r_0$, means that the BS is inside of $\mathcal{B}(y_0, d)$. Following steps similar to those in the previous case, the intensity measure can be evaluated as follows:

$$\Lambda^i_{d\leq r_0}(r_0) = \int_0^{d-r_0} \int_{-\pi}^{\pi} \mathrm{P}_i\,(r)\lambda_u d\theta r dr$$
$$+ \int_{d-r_0}^{r_0+d} \int_{-\arccos\frac{r^2+r_0^2-d^2}{2r_0 r}}^{\arccos\frac{r^2+r_0^2-d^2}{2r_0 r}} \mathrm{P}_i(r)\lambda_u d\theta r dr$$
$$\approx \frac{2\pi\lambda_u}{\alpha\bar{\tau}_i^2}\gamma\left(\frac{2}{\alpha}, \bar{\tau}_i^{\alpha}(d-r_0)^{\alpha}\right)$$
$$+ 2\lambda_u r_0 \sum_{l=1}^{N} \frac{\pi}{N} g_r\left(d + r_0 w_l\right)\sqrt{1-w_l^2}, \quad (43)$$

where $\gamma(\cdot)$ denotes the incomplete gamma function, and the approximation in the last step follows from the application of Chebyshev-Gauss quadrature.

Finally, the cache hit probability can be obtained by substituting (41) and (43) into (37).

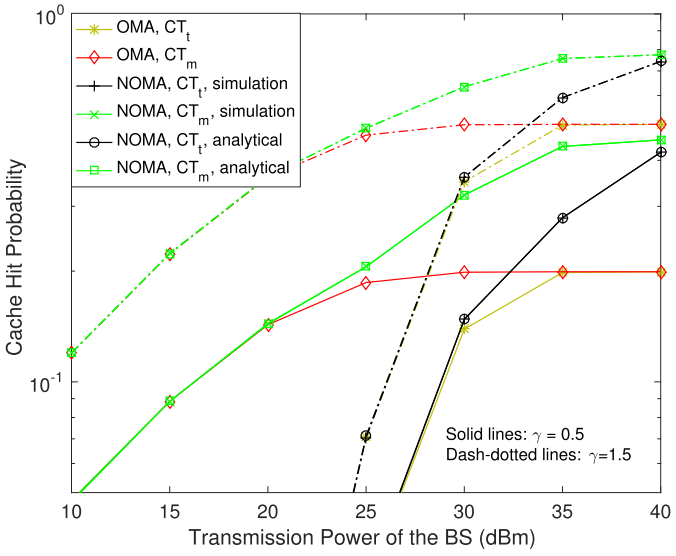
V. Numerical Studies and Discussions

In this section, the performances achieved by the proposed push-then-deliver and push-and-deliver strategies are studied by using computer simulations, where the accuracy of the developed analytical results will be also verified. The system parameters adopted for simulation and analysis are specified in the captions of the figures shown in this section.

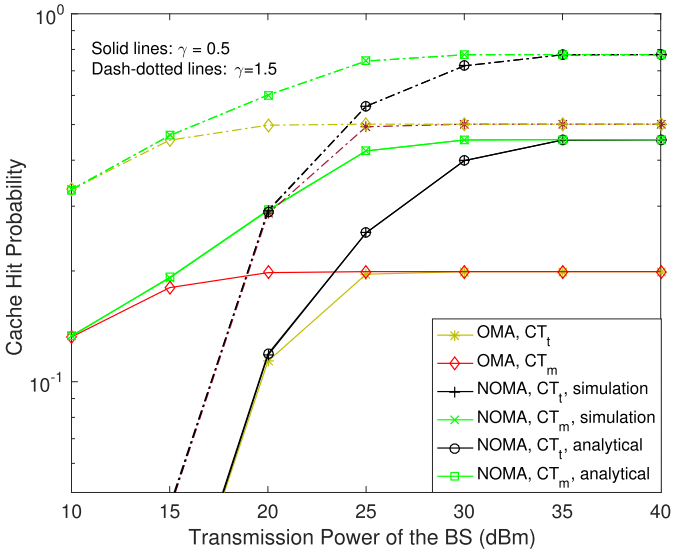
A. Performance of Push-Then-Deliver Strategy

In Figs. 6 and 7, the impact of the NOMA assisted push-then-deliver strategy on the cache hit probability is studied. The thermal noise is set as $\sigma_n^2 = -100$ dBm. For the transmission power of the BS, values between 10 dBm and 40 dBm are considered, which means that the highest BS transmission power is 10 watts. The density of the content servers is parametrized by cluster radius $\mathcal{R}_c$, i.e., $\lambda_c = \frac{0.01}{\pi\mathcal{R}_c^2}$, in order to account for the fact that the density of the content servers is affected by $\mathcal{R}_c$. By applying the NOMA principle to the content pushing phase, more content can be pushed to the content servers simultaneously, and hence, the cache hit probability is improved, compared to the OMA case, as can be observed from Fig. 6. For example, when the transmission power is 40 dBm, the shape parameter for the content popularity is $\gamma = 0.5$, and $\mathcal{R}_c = 50$ m, the use of OMA yields a hit probability of 0.2, and the use of NOMA improves this value to 0.45, which corresponds to a 100% improvement. At low SNR, NOMA and OMA yield the same performance. This is due to the use of the CR inspired power allocation policy in (7), which implies that at low SNR, all the power is allocated to file f_1, and hence, there is no difference between the OMA and NOMA schemes. Note that the curves for analysis and simulation match perfectly in Fig. 6, which demonstrates the accuracy of the developed analytical results. Furthermore, we note that the NOMA power allocation coefficients β_i are assumed to be fixed. Optimizing these power allocation coefficients dynamically according to the users' channel conditions can further enhance the performance gain of NOMA assisted caching compared to the OMA baseline scheme.

The impact of γ, the shape parameter defining the content popularity, on the hit probability is significant, as can be observed in Fig. 6. Particularly, increasing the value of γ improves the hit probability. This is because a larger value of γ implies that the first M_s files become more popular, hence ensuring the delivery of these more popular files can significantly improve the hit probability, as indicated by (9). Comparing Fig. 6(a) with Fig. 6(b), one can observe that the impact of $\mathcal{R}_c$ on the hit probability is also significant, which is due to the fact that the density of the content servers depends on $\mathcal{R}_c$. Particularly, a larger $\mathcal{R}_c$ means that the content servers are more sparsely deployed and hence it is more difficult for



(a) $\mathcal{R}_c = 100$m



(b) $\mathcal{R}_c = 50$m

Fig. 6. The cache hit probability for the push-then-deliver strategy. $N = 20$, $\alpha = 3$, $\lambda_c = \frac{0.01}{\pi \mathcal{R}_c^2}$, $t = 5$, $m = 1$, $M_s = 3$, and $R_l = 1$ bit per channel use (BPCU), for $1 \leq l \leq 3$. The power allocation coefficient for file f_1 is based on the CR power allocation policy. The power allocation coefficients for files f_2 and f_3 are $\beta_2 = \frac{3}{4}$ and $\beta_3 = \frac{1}{4}$, respectively. $|\mathcal{F}| = 10$.

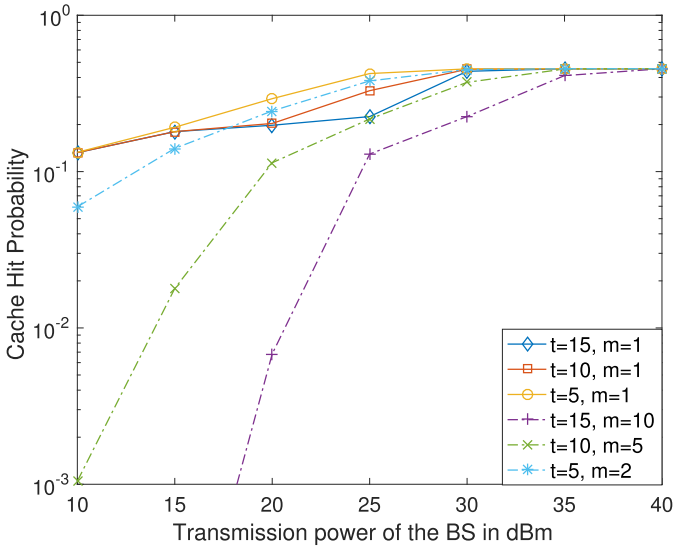
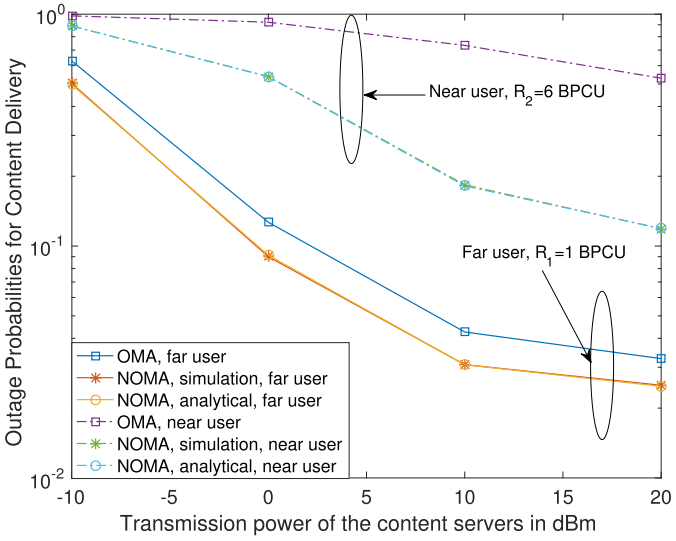


Fig. 7. The impact of the choices of m and t on the cache hit probabilities for the push-then-deliver strategy. $N = 20$, $\alpha = 3$, $\mathcal{R}_c = 50$m, $\lambda_c = \frac{0.01}{\pi \mathcal{R}_c^2}$, $M_s = 3$, $t = 5$, $m = 1$, and $R_l = 1$ BPCU, for $1 \leq l \leq 3$. The power allocation coefficient for file f_1 is based on the CR power allocation policy. The power allocation coefficients for files f_2 and f_3 are $\beta_2 = \frac{3}{4}$ and $\beta_3 = \frac{1}{4}$, respectively. $\gamma = 0.5$ and $|\mathcal{F}| = 3$. Analytical results are used to generate the figure.

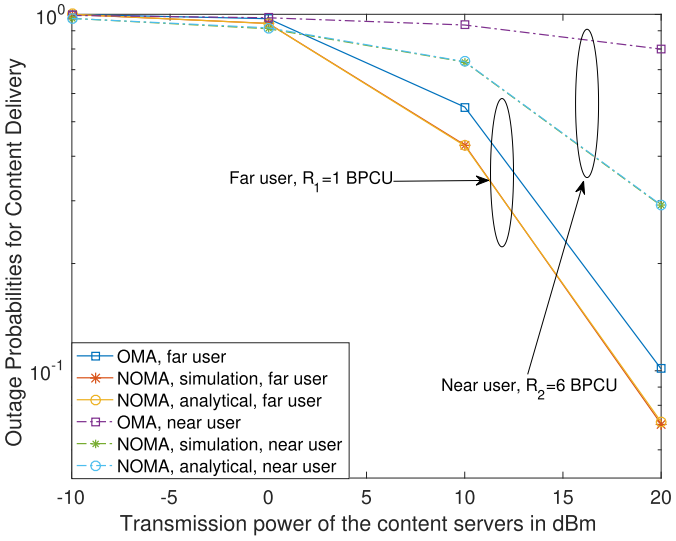
the BS to push content to the content servers, and the cache hit probability decreases.

Recall that during the time slot considered for content pushing in Section III-A, the BS pushes additional files to CS_m while ensuring that f_1 is pushed to CS_t. In Fig. 7, the impact of different choices of m and t on the cache hit probability is studied. As can be observed from the figure, increasing t will decrease the hit probability. This is again due to the use of the CR power allocation policy. In particular, a larger t means that more transmission power is needed to deliver f_1 to CS_t, and hence, less power is available for other files. An interesting observation in Fig. 7 is that the shape of the hit probability curves is not smooth. This is due to the fact that the hit probability is the summation of popularity probabilities $P(f_l)$ and these popularity probabilities are prefixed and not continuous, as shown in (1). On the other hand, for a fixed t, increasing m reduces the cache hit probability, since increasing m means that CS_m is further away from the BS and hence its reception reliability deteriorates.

In Fig. 8, the impact of using NOMA for content delivery is studied, where the rate pair $\{R_1, R_2\}$ is set to $\{1, 6\}$ BPCU to account for the fact that the near user can achieve a higher data rate. For the transmission power of a content server, values between -10 dBm and 20 dBm are considered, which reflects the fact that the content servers transmit at lower power than the BS. As can be observed from the figure, the proposed push-then-deliver strategy can improve the reliability of content delivery, particularly for the user with strong channel conditions. For example, when the path loss exponent is set to $\alpha = 3$ and the transmission power of the content servers is 20 dBm, the use of NOMA ensures that the outage probability for the far user is improved from 3.2×10^{-2} to 2.4×10^{-2}, which is a relatively small performance gain. However, the performance gap between the OMA and NOMA schemes for the near user is much larger, e.g., for the same case as considered before, the outage probability is improved from 5×10^{-1} to 1.1×10^{-2}. This observation is consistent with the existing published results on NOMA which show that NOMA is more beneficial to the near user than to the far user [7]. Note that the outage probability for content delivery has an error floor, i.e., increasing the transmission power of the content severs cannot reduce the outage probability to zero. This is because multiple content servers transmit simultaneously, and hence, content delivery becomes interference limited at high SNR. We note that the impact of the path loss exponent on the reliability of content delivery is significant, as can be observed by comparing Figs. 8(a) and 8(b). This is due to the fact that



(a) $\alpha = 3$



(b) $\alpha = 4$

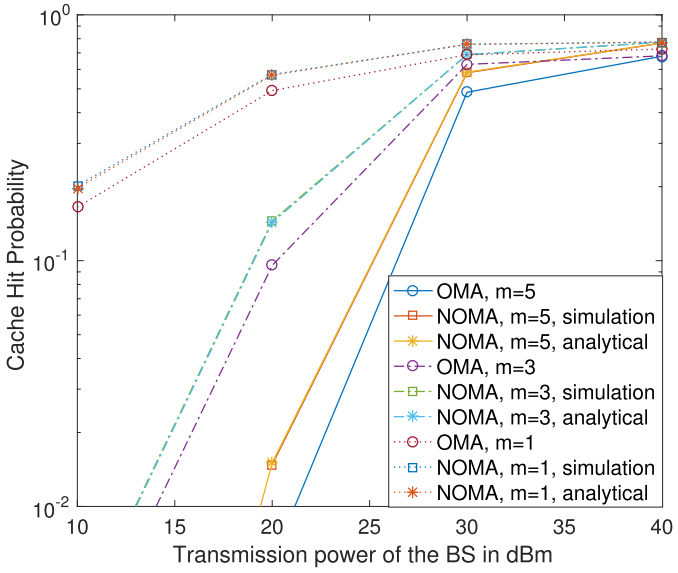
Fig. 8. The outage probabilities for content delivery for the push-then-deliver strategy. $N = 20$, $\alpha = 4$, $\mathcal{R}_c = 100$m. $\lambda_c = \frac{0.01}{\pi \mathcal{R}_c^2}$, $R_1 = 1$ BPCU, and $R_2 = 6$ BPCU. The power allocation coefficients are $\alpha_1^2 = \frac{3}{4}$ and $\alpha_2^2 = \frac{1}{4}$.

a smaller value of α results in a lower path loss, which leads to an improved reception reliability.

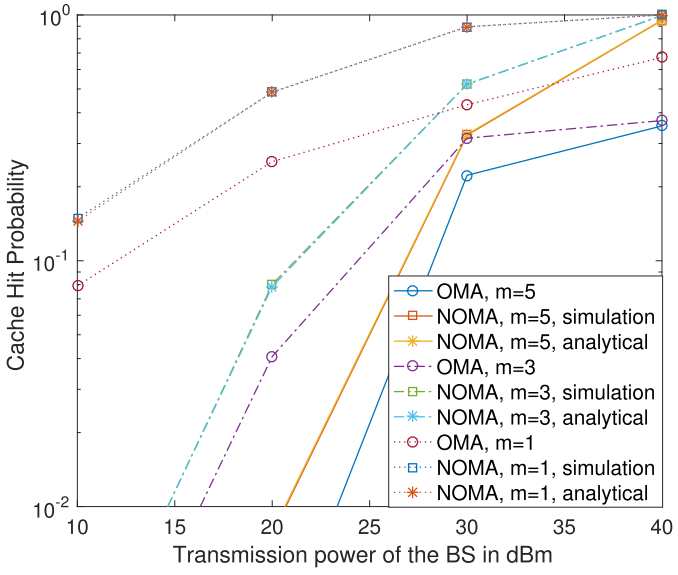
B. Performance of Push-and-Deliver Strategy

In Fig. 9, the impact of the proposed push-and-deliver strategy on the cache hit probability is studied. We employ $\delta = 1.1$ to avoid the trivial case where the user is colocated with the BS, as discussed in Section IV-A. As can be observed, the use of the proposed strategy can effectively improve the cache hit probability compared to the OMA case, which is consistent with the conclusions drawn in the previous subsection. In both sub-figures of Fig. 9, the analytical results perfectly match the simulation results, which verifies the accuracy of the developed analysis.

In Fig. 9, the impact of different choices for the popularity parameters on the cache hit probability is studied. In particular, the following two cases are considered:



(a) Case 1



(b) Case 2

Fig. 9. The cache hit probability for the proposed push-and-deliver strategy. $\mathcal{R}_c = 50$m, $N = 20$, $\gamma = 1.5$, $\alpha = 3$, $\lambda_c = \frac{0.01}{\pi \mathcal{R}_c^2}$, $M_s = 3$, $\delta = 1.1$. $R_0 = \frac{1}{8}$ BPCU, $R_1 = \frac{3}{4}$ BPCU, $R_2 = \frac{7}{8}$ BPCU, and $R_3 = \frac{11}{4}$ BPCU. The power allocation coefficients are $\alpha_0^2 = \frac{4}{8}$, $\alpha_1^2 = \frac{3}{8}$, $\alpha_2^2 = \frac{2}{8}$, and $\alpha_3^2 = \frac{1}{8}$.

- Case 1: $\mathcal{F}_1 = \{f_1, \cdots, f_{10}\}$, and the power allocation coefficient for f_l is α_l;
- Case 2: $\mathcal{F}_2 = \{f_1, \cdots, f_3\}$, and the power allocation coefficient for f_l is α_{4-l}.

The two cases correspond to two different options for mapping files with different popularities to different power levels (or equivalently SIC decoding orders), where in the first case, more popular files are assigned more power, and in the second case, less power is assigned to more popular files.

In Case 1, the performance gap between NOMA and OMA is not significant, as can be observed from Fig. 9(b). For example, when the transmit power is 40 dBm and $m = 5$, the use of OMA results in a hit probability of 0.7 and the

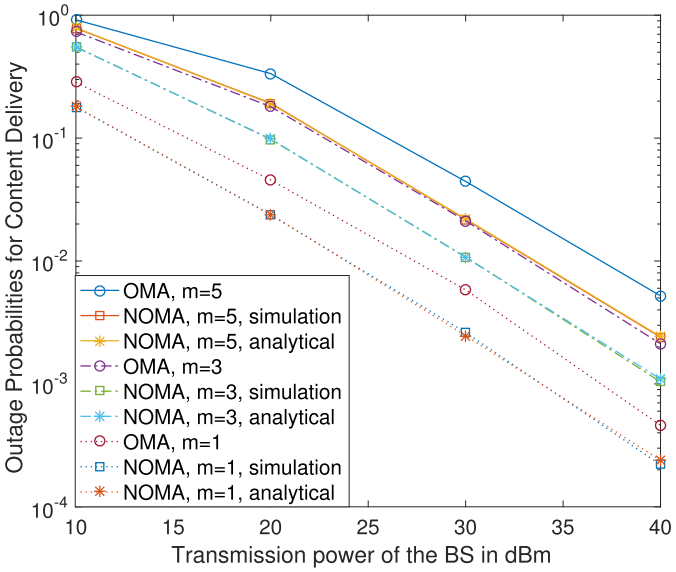
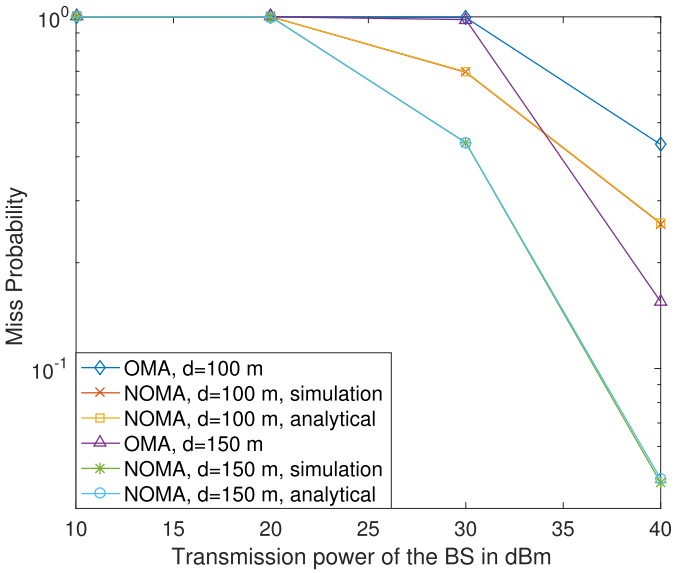
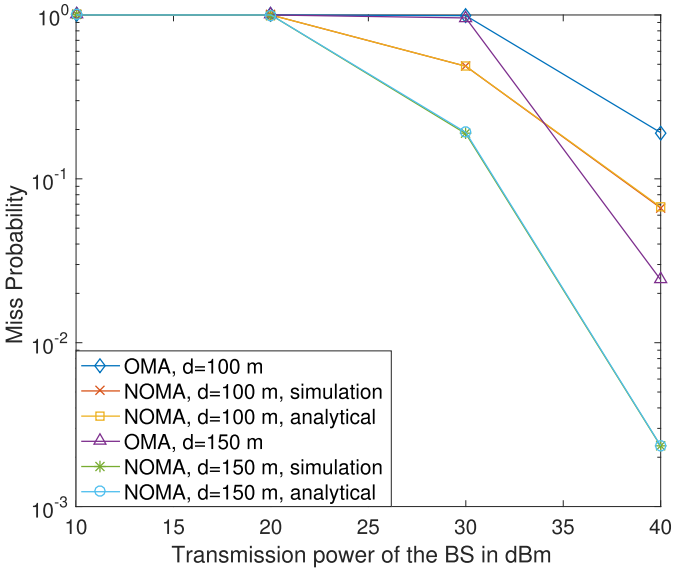


Fig. 10. The impact of the push-and-deliver strategy on content delivery. $\mathcal{R}_c = 50$m, $\gamma = 1.5$, $\alpha = 3$, $\lambda_c = \frac{0.01}{\pi \mathcal{R}_c^2}$, $M_s = 3$, $N = 20$, $\delta = 1.1$. $R_0 = \frac{1}{8}$ BPCU, $R_1 = \frac{3}{4}$ BPCU, $R_2 = \frac{7}{8}$ BPCU, and $R_3 = \frac{11}{4}$ BPCU. The power allocation coefficients are $\alpha_0^2 = \frac{4}{8}$, $\alpha_1^2 = \frac{3}{8}$, $\alpha_2^2 = \frac{2}{8}$, and $\alpha_3^2 = \frac{1}{8}$. Case 2 is considered.



(a) $\lambda_u = 5 \times 10^{-5}$



(b) $\lambda_u = 1 \times 10^{-4}$

Fig. 11. The impact of the proposed push-and-deliver strategy on the cache miss probability in D2D caching scenarios. $N = 20$, $\alpha = 3$, $y_0 = (500$ m, 500 m), $R_0 = 0.5$ BPCU, $R_1 = 4$ BPCU and $M_s = 1$. The power allocation coefficients are $\alpha_0^2 = \frac{3}{4}$ and $\alpha_1^2 = \frac{1}{4}$.

use of NOMA yields a hit probability of 0.8, where the gap is only 0.1. However, for a different set of popularity parameters, i.e., Case 2, the performance gap between OMA and NOMA is significantly increased. For example, for a transmit power of 40 dBm and $m = 5$, the performance gap between OMA and NOMA is enlarged to 0.5. The reason behind this phenomenon is as follows. Recall that the use of NOMA can significantly improve the reception reliability of the files that are decoded at the later stages of the SIC procedure, but the improvement for the files that are decoded during the first few stages of SIC is not significant. In Case 1, the first few files will get larger weights in the sum of the cache hit probability, i.e., file f_l, for a small l, has more impact on the overall performance. As a result, the gap between OMA and NOMA in Case 1 is small, since the reception reliability for decoding these files in the case of NOMA is not so different from that for OMA. On the other hand, Case 2 means that the most popular file, f_1, will be decoded last. As discussed before, the capabilities of OMA and NOMA to decode f_1 are quite different, which is the reason for the larger performance gap in Case 2.

Recall that the key idea of the push-and-deliver strategy is to perform content pushing when asking the BS to serve the users directly. Fig. 9 clearly demonstrates that this strategy can efficiently push new content to the content servers, but it does not demonstrate the impact of this strategy on content delivery, which is studied in Fig. 10. Particularly, as can be observed from the figure, the use of the proposed push-and-deliver strategy does not degrade the reception reliability of content delivery. In fact, the use of NOMA can even improve the outage probability for content delivery.

In Fig. 11, the concept of the proposed push-and-deliver strategy is extended to D2D caching scenarios. Without loss of generality, the newly arrived user is located at $y_0 = (500\text{m}, 500\text{m})$. As expected, the use of the proposed strategy can significantly reduce the miss probability, compared to the case of OMA. For example, for the case where the user density is $\lambda_u = 5 \times 10^{-5}$, a transmit power of 40 dBm, and $d = 150$ m, the use of NOMA yields a miss probability of 6×10^{-2}, whereas the miss probability for OMA is 1.6×10^{-1}, which is much worse. As can be observed from the figure, increasing the value of d can reduce the miss probability, since the area for searching for a D2D helper is increased. Another important observation is that by increasing the density of the users, the miss probability can be further reduced, since increasing the density means that more users are located in the same area

and hence it is more likely to find a D2D helper. We note that, in Fig. 11, computer simulation and analytical results match perfectly, which demonstrates again the accuracy of the developed analysis.

VI. Conclusions and Future Work

Unlike conventional wireless caching strategies which rely on the use of off-peak hours for content pushing, in this paper, the NOMA principle has been applied to wireless caching to enable the frequent update of the local caches via wireless transmission during on-peak hours. Two NOMA assisted caching strategies have been developed, namely the push-then-deliver strategy and the push-and-deliver strategy. The push-then-deliver strategy is applicable to the case when the content pushing phase and the content delivery phase are separated, and utilizes the NOMA principle independently in both phases. The developed analytical results demonstrate that the proposed NOMA assisted caching scheme can efficiently improve the cache hit probability and reduce the delivery outage probability. The push-and-deliver strategy is motivated by the fact that, in practice, it is inevitable that some user requests cannot be accommodated locally and the BS has to serve these users directly. The key idea of the push-and-deliver strategy is to merge the two phases, i.e., the BS pushes content to the content servers while simultaneously serving users directly. Furthermore, in addition to the caching scenario with caching infrastructure, e.g., content servers, we have considered D2D caching, where the use of NOMA has also been shown to yield superior performance compared to OMA. We note that the push-then-deliver and the push-and-deliver caching strategies are complementary and can be combined in practice. For example, during the content pushing phase, the push-then-deliver strategy can be applied to push as many files as possible. Then, during the content delivery phase, both strategies can be used depending on whether a user's request can be accommodated by its local content server. If a user cannot find the requested file locally, the push-and-deliver strategy can be applied. Otherwise, a content server can group multiple users whose requested files are stored locally and then apply the proposed push-then-deliver strategy.

The results in this paper open several new directions for future research. First, many system parameters, such as the number of popular files to be superimposed, the mapping between the file popularity and the SIC decoding order, as well as the NOMA power allocation coefficients, have been assumed to be fixed in this paper. Dynamically optimizing these parameters and also applying advanced content popularity models accurately predicting the users' requests can further improve the performance of NOMA assisted caching. Second, increasing the density of the users or the search area in D2D caching can increase the cache hit probability, but might also cause stronger interference during the content delivery phase, when the D2D helpers deliver the requested files to their neighbors simultaneously. Therefore, for the content delivery phase, it is important to design low-complexity algorithms for efficient scheduling of the users' requests in order to limit co-channel interference. In this context, coordinated multi-point transmission (CoMP) and cloud radio access networks (C-RANs) are interesting options for suppressing co-channel interference [25], [40]. Third, in this paper, content pushing is carried out without exploiting the structure of the content files. As shown in [22], the spectral efficiency of wireless caching can be further improved by applying coded caching, since only parts of the content files need to be cached and one multicast transmission during the content delivery phase can benefit multiple users simultaneously. Thus, enhanced caching and delivery schemes combining the benefits of the proposed NOMA based strategies and coded caching are of interest. Fourth, in this paper, the cache hit probability has been used as the performance criterion, whereas latency is another important metric [41]–[43]. The proposed caching strategies can potentially decrease the latency for content delivery. On the one hand, the proposed push-then-deliver strategy can effectively reduce the waiting time of the users for being served, since multiple users can be simultaneously served by one content server. On the other hand, with the proposed push-and-deliver strategy, the files cached at the content servers can be updated more frequently, which indirectly helps in reducing the latency of content delivery. Hence, a formal analysis of the impact of the proposed caching strategies on the latency of content delivery is needed, where various effects have to be considered, including the number of retransmissions of the content, the scheduling delay for those users which are served by the BS directly, etc.

Appendix A
Proof of Theorem 1

Recall that the NOMA cache hit probability is $\mathrm{P}_m^{hit} = \sum_{i=1}^{M_s} \mathrm{P}(f_i)(1 - \mathrm{P}_{m,i})$ and the OMA hit probability is $\mathrm{P}_{m,OMA}^{hit} = \mathrm{P}(f_1)(1 - \mathrm{P}_{m,1}^{OMA})$. Since the file popularity probabilities are positive and identical for the NOMA and OMA cases, proving $\mathrm{P}_{m,1}^{OMA} = \mathrm{P}_{m,1}$ for all CS_m, $1 \leq m \leq t$, is sufficient to prove the theorem.

Recall that each content server will carry out SIC, i.e., files j, $1 \leq j < i$, are decoded before file i is decoded. Therefore, the outage probability of CS_m for decoding file i can be expressed as follows:

$$\mathrm{P}_{m,i} = 1 - \mathrm{P}\left(f_j \text{ is decoded}, \forall j \leq i\right). \tag{44}$$

For notational simplicity, first define $z_m \triangleq \frac{1}{L(||x_m - x_0||)}$, and note that these channel gains are ordered as follows: $z_1 \geq \cdots \geq z_t$. Therefore, the outage probability can be expressed as follows:

$$\mathrm{P}_{m,i} = 1 - \mathrm{P}\left(z_m > \frac{\epsilon_l}{\rho \xi_l}, \forall l \leq i\right), \tag{45}$$

where $\xi_l = \alpha_l^2 - \epsilon_l \sum_{j=l+1}^{M_s} \alpha_j^2$.

As discussed previously, showing $\mathrm{P}_{m,1}^{OMA} = \mathrm{P}_{m,1}$ is sufficient to prove the theorem. First, we focus on the performance of CS_t. According to the definition of the CR NOMA power allocation policy, the outage probability of CS_t for decoding the most popular file, f_1, is given by

$$\begin{aligned} \mathrm{P}_{t,1} &= \mathrm{P}\left(z_t < \frac{\epsilon_1}{\rho \xi_1}\right) \\ &\overset{(a)}{=} \mathrm{P}\left(\alpha_1 = 1\right) = \mathrm{P}\left(z_t < \frac{\epsilon_1}{\rho}\right) = \mathrm{P}_{t,1}^{OMA}, \end{aligned} \tag{46}$$

where ϵ_1 and ξ_1 are used since they are valid for file f_1. Step (a) follows from the fact that an outage occurs at CS_t only if all the power is assigned to file f_1, i.e., $\alpha_1 = 1$. Therefore, regarding the capability of CS_t to decode f_1, adopting NOMA does not bring any difference, compared to OMA.

Second, the outage probability of CS_m, $1 \leq m < t$, for decoding f_1, is given by

$$\begin{aligned}\text{P}_{m,1} &= \text{P}\left(z_m < \frac{\epsilon_1}{\rho\xi_1}\right) \\ &= \text{P}\left(\alpha_1 = 1, z_m < \frac{\epsilon_1}{\rho\xi_1}\right) + \text{P}\left(\alpha_1 < 1, z_m < \frac{\epsilon_1}{\rho\xi_1}\right).\end{aligned}$$

Since the channel conditions of CS_m are better than those of CS_t, the condition that CS_t can decode f_1 correctly, i.e., $\alpha_1 < 1$, is sufficient to guarantee successful detection of f_1 at CS_m. Therefore, the outage probability can be simplified as follows:

$$\text{P}_{m,1} = \text{P}\left(\alpha_1 = 1, z_m < \frac{\epsilon_1}{\rho\xi_1}\right). \tag{47}$$

Note that the use of the CR power allocation policy in (7) complicates the expression for the outage probability, since the power allocation coefficients depend on the channel conditions of CS_t. In order to better understand the outage events, we express the event $\{z_m < \frac{\epsilon_1}{\rho\xi_1}\}$ as follows:

$$\begin{aligned}&\left\{z_m < \frac{\epsilon_1}{\rho\xi_1}\right\} \\ &= \left\{z_m < \frac{\epsilon_1}{\rho\left(1 - P_r - \epsilon_1 P_r\right)}\right\} \\ &= \left\{z_m < \frac{\epsilon_1}{\rho\left(1 - (1+\epsilon_1)\max\left\{0, \frac{\rho z_t - \epsilon_1}{\rho(1+\epsilon_1)z_t}\right\}\right)}\right\} \\ &= \left\{z_m < \frac{\epsilon_1}{\rho\left(1 - \max\left\{0, \frac{\rho z_t - \epsilon_1}{\rho z_t}\right\}\right)}\right\}.\end{aligned} \tag{48}$$

By combining (47) and (48), surprisingly probability $\text{P}_{m,1}$ can be simplified as follows:

$$\text{P}_{m,1} = \text{P}\left(z_t < \frac{\epsilon_1}{\rho}, z_m < \frac{\epsilon_1}{\rho}\right), \tag{49}$$

since $\max\left\{0, \frac{\rho z_t - \epsilon_1}{\rho\ (1+\epsilon_1)z_t}\right\} = 0$ for the case $z_t < \frac{\epsilon_1}{\rho\xi_1}$. On the other hand, it is straightforward to show that the outage probability for OMA is given by

$$\begin{aligned}\text{P}_{m,1}^{OMA} &= \text{P}\left(z_t > \frac{\epsilon_1}{\rho}, z_m < \frac{\epsilon_1}{\rho}\right) + \text{P}\left(z_t < \frac{\epsilon_1}{\rho}, z_m < \frac{\epsilon_1}{\rho}\right) \\ &= \text{P}_{m,1}.\end{aligned}$$

Therefore, the NOMA outage performance of CS_m, $1 \leq m \leq t$, for decoding f_1 is the same as that of OMA, but the use of NOMA can ensure that more content is delivered to the content servers, which proves the theorem.

Appendix B
Proof of Lemma 1

Since the content servers follow an HPPP, the pdf for the m-th shortest distance is given by [39]

$$f_{r_m}(x) = \frac{2\lambda_c^m \pi^m x^{2m-1}}{(m-1)!} e^{-\lambda_c \pi x^2}. \tag{50}$$

The conditional CDF for the t-th shortest distance, given $r_m = x$, can be expressed as follows:

$$\begin{aligned}F_{r_t|r_m}(y) &\triangleq \text{P}(r_t \leq y | r_m = x) \\ &= 1 - \text{P}(r_t > y | r_m = x).\end{aligned} \tag{51}$$

The event, $(r_t > y | r_m = x)$, corresponds to the case in which the t-th nearest content server is not located inside the ring between a larger circle with radius y and a smaller one with radius x. Or equivalently, the event, $(r_t > y | r_m = x)$, means that at most $(t - m - 1)$ content servers are inside the ring between the two circles. Therefore, the conditional CDF, $F_{r_t|r_m}(y)$, can be explicitly written as follows:

$$F_{r_t|r_m}(y) = 1 - \sum_{n=0}^{t-m-1} \text{P}(\#(\mathbf{int}(\mathcal{B}(x_0, x), \mathcal{B}(x_0, y)) = n), \tag{52}$$

where $\#(\mathcal{A})$ denotes the number of points falling into the area $\mathcal{A}$, $\mathcal{B}(x_0, x)$ denotes a disc with its origin located at x_0 and radius x, and $\mathbf{int}(\mathcal{B}(x_0, x), \mathcal{B}(x_0, y))$ denotes the ring between the boundaries of $\mathcal{B}(x_0, x)$ and $\mathcal{B}(x_0, y)$.

By applying the HPPP assumption, the conditional CDF can be found as follows:

$$F_{r_t|r_m}(y) = 1 - \sum_{n=0}^{t-m-1} (\lambda_c \pi)^n (y^2 - x^2)^n \frac{e^{-\lambda_c \pi (y^2 - x^2)}}{n!}. \tag{53}$$

In order to find the joint pdf between r_m and r_t, the conditional pdf is needed first. However, the derivative of the CDF $F_{r_t|r_m}(y)$ shown in the above equation has the following complicated form:

$$\begin{aligned}f_{r_t|r_m}(y) = &\sum_{n=1}^{t-m-1} \frac{2y(\lambda_c \pi)^n (y^2 - x^2)^{n-1}}{n!} e^{-\lambda_c \pi (y^2 - x^2)} \\ &\times \left[\lambda_c \pi (y^2 - x^2) - n\right] + 2\lambda_c \pi y e^{-\lambda_c \pi (y^2 - x^2)}.\end{aligned} \tag{54}$$

This complicated form makes the calculation of the outage probability very difficult. Instead, the steps provided in [39] can be used to obtain a much simpler form, as shown in the following. First, define $S_n = \frac{(\lambda_c \pi (y^2 - x^2))^n}{n!}$, and hence the conditional CDF obtained in (53) can be re-written as follows:

$$F_{r_t|r_m}(y) = 1 - \sum_{n=0}^{t-m-1} S_n e^{-\lambda_c \pi (y^2 - x^2)}. \tag{55}$$

After taking the derivative of the CDF and exploiting the structure of S_n, the conditional pdf can be obtained

as follows:

$$f_{r_t|r_m}(y) = 2y\lambda_c\pi e^{-\lambda_c\pi(y^2-x^2)}\left(\sum_{n=0}^{t-m-1} S_n - \sum_{n=1}^{t-m-1} S_{n-1}\right) = 2y(\lambda_c\pi)^{t-m}e^{-\lambda_c\pi(y^2-x^2)}\frac{(y^2-x^2)^{t-m-1}}{(t-m-1)!}, \quad (56)$$

which is much simpler than the expression in (54).

By applying Bayes' formula, the joint pdf between r_m and r_t can be obtained as follows:

$$f_{r_m,r_t}(x,y) = f_{r_m|r_t}(x)f_{r_t}(y) = 4y(\lambda_c\pi)^t e^{-\lambda_c\pi y^2}\frac{x^{2m-1}(y^2-x^2)^{t-m-1}}{(t-m-1)!(m-1)!}. \quad (57)$$

Note that, for the special case of $m = t-1$, the two parameters, x and y, are decoupled to yield the following simplified form for the joint pdf:

$$f_{r_m,r_t}(x,y) = \frac{4(\lambda_c\pi)^{m+1}yx^{2m-1}}{(m-1)!}e^{-\lambda_c\pi y^2}. \quad (58)$$

This completes the proof of the lemma.

Appendix C
Proof of Lemma 2

Following the steps provided in the proof of Theorem 1, the outage probability for CS_t to decode f_1 is given by

$$\text{P}_{t,1} = \text{P}\left(z_t < \frac{\epsilon_1}{\rho}\right).$$

After applying the marginal pdf of the t-th shortest distance shown in (50), $\text{P}_{t,1}$ can be calculated as follows:

$$\text{P}_{t,1} = \frac{2\lambda_c^t\pi^t}{(t-1)!}\int_{\frac{\epsilon_1^{\frac{1}{\alpha}}}{\rho^{\frac{1}{\alpha}}}}^{\infty} y^{2t-1}e^{-\lambda_c\pi y^2}dy = e^{-\lambda_c\pi\left(\frac{\rho}{\epsilon_1}\right)^{\frac{2}{\alpha}}}\sum_{k=0}^{t-1}\frac{(\lambda_c\pi)^k\left(\frac{\rho}{\epsilon_1}\right)^{\frac{2k}{\alpha}}}{k!}. \quad (59)$$

According to (49), the outage probability for CS_m to decode f_1 is given by

$$\text{P}_{m,1} = \text{P}\left(z_t < \frac{\epsilon_1}{\rho}, z_m < \frac{\epsilon_1}{\rho}\right).$$

By using the fact that $r_m \leq r_n$ and again applying the marginal distribution of r_m, the outage probability can be straightforwardly obtained as follows:

$$\text{P}_{m,1} = \text{P}\left(z_m < \frac{\epsilon_1}{\rho}\right) = e^{-\lambda_c\pi\left(\frac{\rho}{\epsilon_1}\right)^{\frac{2}{\alpha}}}\sum_{k=0}^{m-1}\frac{(\lambda_c\pi)^k\left(\frac{\rho}{\epsilon_1}\right)^{\frac{2k}{\alpha}}}{k!}. \quad (60)$$

Hence, the first part of the lemma is proved.

The outage probability for file i, $i > 1$, is more complicated than the case of f_1. The impact of the channel condition of CS_t on the outage performance of CS_m can be made explicit by expressing the individual event $\left\{z_m < \frac{\epsilon_i}{\rho\xi_i}\right\}$, $i > 1$, as follows:

$$\left\{z_m < \frac{\epsilon_i}{\rho\xi_i}\right\} = \left\{z_m < \frac{\epsilon_i}{\rho\left(\alpha_i^2 - \epsilon_i\sum_{j=i+1}^{M_s}\alpha_j^2\right)}\right\} = \left\{z_m < \frac{\epsilon_i}{\rho\bar{\xi}_i\max\left\{0, \frac{\rho z_t-\epsilon_1}{\rho(1+\epsilon_1)z_t}\right\}}\right\}, \quad (61)$$

where the last step follows from the fact that $P_r = \max\left\{0, \frac{\rho z_t-\epsilon_1}{\rho\ (1+\epsilon_1)z_t}\right\}$. Recall that $\bar{\xi}_i = \left(\beta_i - \epsilon_i\sum_{j=i+1}^{M_s}\beta_j\right)$ is a constant and not a function of the channel conditions of CS_t. Therefore, the outage probability of CS_t for decoding f_i, $i > 1$, is given by

$$\text{P}_{t,i} = \text{P}\left(\alpha_1 = 1, z_t < \max\left\{\frac{\epsilon_1}{\rho\xi_1}, \cdots, \frac{\epsilon_i}{\rho\xi_i}\right\}\right) + \text{P}\left(\alpha_1 < 1, z_t < \max\left\{\frac{\epsilon_1}{\rho\xi_1}, \cdots, \frac{\epsilon_i}{\rho\xi_i}\right\}\right). \quad (62)$$

Note that $\alpha_1 = 1$ corresponds to the event that all the power is allocated to f_1. Therefore, $z_t < \max\left\{\frac{\epsilon_1}{\rho\xi_1}, \cdots, \frac{\epsilon_i}{\rho\xi_i}\right\}$ is always true if $\alpha_1 = 1$, and therefore, the outage probability can be simplified as follows:

$$\text{P}_{t,i} = \text{P}\left(\alpha_1 = 1\right) + \text{P}\left(\alpha_1 < 1, z_t < \max\left\{\frac{\epsilon_2}{\rho\xi_2}, \cdots, \frac{\epsilon_i}{\rho\xi_i}\right\}\right). \quad (63)$$

Note that when $\alpha < 1$, the expression for the event $\{z_t < \frac{\epsilon_i}{\rho\xi_i}\}$ in (61) can be simplified as follows:

$$\left\{z_t < \frac{\epsilon_i}{\rho\xi_i}\right\} = \left\{z_t < \frac{\epsilon_i}{\rho\bar{\xi}_i\frac{\rho z_t-\epsilon_1}{\rho(1+\epsilon_1)z_t}}\right\}. \quad (64)$$

Therefore, the outage probability can be rewritten as follows:

$$\text{P}_{t,i} = \text{P}\left(\alpha_1 = 1\right) + \text{P}\left(\alpha_1 < 1, z_t < \max\left\{\frac{\epsilon_j}{\rho\bar{\xi}_j\frac{\rho z_t-\epsilon_1}{\rho\ (1+\epsilon_1)z_t}}, 2 \leq j \leq i\right\}\right) = \text{P}\left(z_t < \frac{\epsilon_1}{\rho}\right) + \text{P}\left(z_t > \frac{\epsilon_1}{\rho}, z_t < \frac{\epsilon_1}{\rho} + \frac{(1+\epsilon_1)}{\rho\phi_i}\right). \quad (65)$$

By applying the marginal pdf for the t-th shortest distance, the outage probability for CS_t to decode f_i can be obtained as follows:

$$\text{P}_{t,i} = e^{-\lambda_c\pi\left(\frac{\epsilon_1}{\rho}+\frac{(1+\epsilon_1)}{\rho\phi_i}\right)^{-\frac{2}{\alpha}}}\sum_{k=0}^{t-1}\frac{(\lambda_c\pi)^k\left(\frac{\epsilon_1}{\rho}+\frac{(1+\epsilon_1)}{\rho\phi_i}\right)^{-\frac{2k}{\alpha}}}{k!}. \quad (66)$$

Hence, the second part of the lemma is proved.

The outage probability for CS_m to decode f_i, $i > 1$, is the most difficult to obtain among the probabilities

shown in the lemma. This probability can be first expressed as follows:

$$\mathrm{P}_{m,i} = \mathrm{P}\left(\alpha_1 = 1, z_m < \max\left\{\frac{\epsilon_1}{\rho\xi_1}, \cdots, \frac{\epsilon_i}{\rho\xi_i}\right\}\right) + \mathrm{P}\left(\alpha_1 < 1, z_m < \max\left\{\frac{\epsilon_1}{\rho\xi_1}, \cdots, \frac{\epsilon_1}{\rho\xi_i}\right\}\right). \quad (67)$$

Note that $\alpha_1 = 1$ results in the situation in which no power is allocated to f_j, $j > 1$, which means that the event $z_m < \max\left\{\frac{\epsilon_1}{\rho\xi_1}, \cdots, \frac{\epsilon_i}{\rho\xi_i}\right\}$ always happens, if $\alpha_1 = 1$. In addition, by using the fact that $r_m \leq r_t$, the outage probability can be simplified as follows:

$$\mathrm{P}_{m,i} = \mathrm{P}\left(z_t < \frac{\epsilon_1}{\rho}\right) + \underbrace{\mathrm{P}\left(z_t > \frac{\epsilon_1}{\rho}, z_m < \max\left\{\frac{\epsilon_2}{\rho\xi_2}, \cdots, \frac{\epsilon_i}{\rho\xi_i}\right\}\right)}_{Q_1}. \quad (68)$$

Note that $z_t > \frac{\epsilon_1}{\rho}$ guarantees $z_m > \frac{\epsilon_1}{\rho\xi_1}$, as $z_t \leq z_m$ and $z_t > \frac{\epsilon_1}{\rho\xi_1}$ is equivalent to $z_t > \frac{\epsilon_1}{\rho}$. However, $z_t > \frac{\epsilon_1}{\rho}$ does not guarantee $z_m > \frac{\epsilon_j}{\rho\xi_j}$, $j > 1$. Recall that conditioned on $z_t > \frac{\epsilon_1}{\rho}$, the term $\frac{\epsilon_j}{\rho\xi_j}$, $j > 1$, can be simplified as follows:

$$\frac{\epsilon_i}{\rho\xi_i} = \frac{\epsilon_i}{\bar{\xi}_i \frac{\rho z_t - \epsilon_1}{(1+\epsilon_1)z_t}}. \quad (69)$$

Therefore, the term Q_1 can be calculated as follows:

$$Q_1 = \mathrm{P}\left(z_t > \frac{\epsilon_1}{\rho}, z_m < \max\left\{\frac{\epsilon_2}{\rho\xi_2}, \cdots, \frac{\epsilon_i}{\rho\xi_i}\right\}\right) = \mathrm{P}\left(z_t > \frac{\epsilon_1}{\rho}, z_m < \frac{(1+\epsilon_1)}{\phi_i\left(\rho - \frac{\epsilon_1}{z_t}\right)}\right). \quad (70)$$

After applying the path loss model, z_t (z_m) can be replaced by the distance between the BS and CS_t (CS_m), and the outage probability can be expressed as follows:

$$Q_1 = \mathrm{P}\left(y < \left(\frac{\epsilon_1}{\rho}\right)^{-\frac{1}{\alpha}}, x > \left(\frac{(1+\epsilon_1)}{\phi_i\left(\rho - \epsilon_1 y^\alpha\right)}\right)^{-\frac{1}{\alpha}}\right), \quad (71)$$

where x denotes the distance between the BS and CS_t and y denotes the distance between the BS and CS_m. However, there is an extra constraint on y as follows:

$$\left(\frac{\epsilon_1}{\rho}\right)^{-\frac{1}{\alpha}} > \left(\frac{(1+\epsilon_1)}{\phi_i\left(\rho - \epsilon_1 y^\alpha\right)}\right)^{-\frac{1}{\alpha}}, \quad (72)$$

which leads to the following constraint on y:

$$y^\alpha > \frac{\rho}{\epsilon_1}\left[1 - \frac{1+\epsilon_1}{\epsilon_1\phi_i}\right]. \quad (73)$$

To better understand this constraint, the term $\frac{1+\epsilon_1}{\epsilon_1\phi_i}$ is rewritten as follows:

$$\frac{1+\epsilon_1}{\epsilon_1\phi_i} = \frac{1+\epsilon_1}{\epsilon_1 \min\left\{\frac{\bar{\xi}_2}{\epsilon_2}, \cdots, \frac{\bar{\xi}_{M_s}}{\epsilon_{M_s}}\right\}} \geq \frac{1+\epsilon_1}{\epsilon_1 \frac{\bar{\xi}_{M_s}}{\epsilon_{M_s}}} = \frac{\epsilon_{M_s} 2^{R_1}}{\epsilon_1 \bar{\xi}_{M_s}}, \quad (74)$$

where $\bar{\xi}_{M_s} \leq 1$ and $2^{R_1} \geq 1$ hold. The only uncertainty for the comparison between the term $\frac{1+\epsilon_1}{\epsilon_1\phi_i}$ and 1 is caused by the relationship between ϵ_1 and ϵ_{M_s}. In the lemma, it is assumed that $\epsilon_1 \leq \epsilon_{M_s}$. As a result, the constraint in (73) is always satisfied since $\frac{1+\epsilon_1}{\epsilon_1\phi_i} \geq 1$. However, the probability in (71) also implies the following constraint:

$$y > \left(\frac{(1+\epsilon_1)}{\phi_i\left(\rho - \epsilon_1 y^\alpha\right)}\right)^{-\frac{1}{\alpha}}. \quad (75)$$

This leads to the following constraint on y:

$$y > \left(\frac{\rho\phi_i}{1+\epsilon_1+\epsilon_1\phi_i}\right)^{\frac{1}{\alpha}} \triangleq \tau_1. \quad (76)$$

After understanding the ranges of x and y, we can now apply the joint pdf to calculate the outage probability, which yields the following:

$$Q_1 = \int_{\tau_1}^{\tau_2} \int_{\left(\frac{(1+\epsilon_1)}{\phi_i(\rho - \epsilon_1 y^\alpha)}\right)^{-\frac{1}{\alpha}}}^{y} f_{r_m, r_t}(x, y) dx dy, \quad (77)$$

where τ_2 is defined in the lemma. To facilitate the calculation of this integral, the joint pdf is rewritten as follows:

$$f_{r_m, r_t}(x, y) = \frac{4(\lambda_c\pi)^t}{(t-m-1)!(m-1)!} e^{-\lambda_c\pi y^2} \sum_{p=0}^{t-m-1} (-1)^p \times \binom{t-m-1}{p} y^{2(t-m-1)-2p+1} x^{2m+2p-1}. \quad (78)$$

Now, we can apply the joint pdf which yields the following:

$$Q_1 = \frac{4(\lambda_c\pi)^t}{(t-m-1)!(m-1)!} \sum_{p=0}^{t-m-1} (-1)^p \binom{t-m-1}{p} \times \int_{\tau_1}^{\tau_2} f_m(y) dy, \quad (79)$$

where $f_m(\cdot)$ is defined in the lemma. One can apply Chebyshev-Gauss quadrature to obtain the following approximation for Q_1:

$$Q_1 \approx \frac{4(\lambda_c\pi)^t}{(t-m-1)!(m-1)!} \sum_{p=0}^{t-m-1} (-1)^p \binom{t-m-1}{p} \times \sum_{l=1}^{N} \frac{\pi(\tau_2 - \tau_1)}{2N} f_m\left(\frac{\tau_2 - \tau_1}{2} w_l + \frac{\tau_2 + \tau_1}{2}\right) \sqrt{1 - w_l^2}. \quad (80)$$

Substituting (80) and (60) into (68), the third part of the lemma is proved.

APPENDIX D
PROOF OF LEMMA 3

Because the two users associated with the same content server are located in different regions inside the disc with radius $\mathcal{R}_c$, the density functions for their channel gains are different, and therefore, the two users' outage probabilities will be calculated separately in the following subsections.

A. The Outage Performance at $U_{m,2}$

First define the composite channel gain as $z_{m,k} \triangleq \frac{|h_{m,mk}|^2}{L(||y_{m,k}||)}$, for $k \in \{1,2\}$. Recall that, for a user that is uniformly distributed in a disc with radius r, the CDF of its composite channel gain, which includes the effects of small scale Rayleigh fading and path loss, can be expressed as follows [6]:

$$F_r(z) \approx \sum_{n=1}^{N} \bar{w}_n \left(1 - e^{-c_{n,r} z}\right), \tag{81}$$

and the corresponding pdf of the channel gain is $f_r(z) \approx \sum_{n=1}^{N} \bar{w}_n c_{n,r} e^{-c_{n,r} z}$. Recall that $\mathrm{U}_{m,2}$ is uniformly distributed in a disc with radius $\mathcal{R}_s$, and therefore, the CDF and pdf of the channel gain of $\mathrm{U}_{m,2}$ are simply given by $F_{\mathcal{R}_s}(z)$ and $f_{\mathcal{R}_s}(z)$ by replacing r with $\mathcal{R}_s$. The reason for using the approximated form in (81) is that both the approximated CDF and pdf are in the form of exponential functions. In the following, we will show that these exponential functions will signficiantly simplify the application of the probability generating functional (PGFL).

With the definition of $z_{m,k} \triangleq \frac{|h_{m,mk}|^2}{L(||y_{m,k}||)}$, the SINR of $\mathrm{U}_{m,2}$ for decoding the first message, $f_{m,1}$, is given by

$$\mathrm{SINR}_{m,2}^{1} = \frac{\alpha_1^2 z_{m,2}}{\alpha_2^2 z_{m,2} + \mathrm{I}_{inter}^{m,2} + \frac{1}{\rho}}. \tag{82}$$

Similarly, the SINR of $\mathrm{U}_{m,2}$ for decoding its own message, $f_{m,2}$, can be rewritten as follows:

$$\mathrm{SINR}_{m,2}^{2} = \frac{\alpha_2^2 z_{m,2}}{\mathrm{I}_{inter}^{m,2} + \frac{1}{\rho}}. \tag{83}$$

Therefore, the outage probability of $\mathrm{U}_{m,2}$ for decoding its own message can be expressed as follows:

$$\begin{aligned}
&\mathrm{P}_{m,2}^{o} \\
&= 1 - \mathrm{P}\left(\log(1+\mathrm{SINR}_{m,2}^{l}) > R_l, l \in \{1,2\}\right) \\
&= \mathcal{E}_{\mathrm{I}_{inter}^{m,2}}\left\{\mathrm{P}\left(z_{m,2} < \max\left\{\frac{\epsilon_1 \mathrm{I}_{inter}^{m,2} + \frac{\epsilon_1}{\rho}}{\alpha_1^2 - \epsilon_1 \alpha_2^2}, \frac{\epsilon_2 \mathrm{I}_{inter}^{m,2} + \frac{\epsilon_2}{\rho}}{\alpha_2^2}\right\}\right)\right\},
\end{aligned}$$

where $\mathcal{E}_x\{\cdot\}$ denotes the expectation with respect to x. In order to facilitate the application of the PGFL, the outage probability is first rewritten as follows:

$$\begin{aligned}
&\mathrm{P}_{m,2}^{o} \\
&= \mathcal{E}_{\mathrm{I}_{inter}^{m,2}}\left\{\mathrm{P}\left(z_{m,2} < \max\left\{\frac{\mathrm{I}_{inter}^{m,2} + \frac{1}{\rho}}{\frac{\alpha_1^2 - \epsilon_1 \alpha_2^2}{\epsilon_1}}, \frac{\mathrm{I}_{inter}^{m,2} + \frac{1}{\rho}}{\frac{\alpha_2^2}{\epsilon_2}}\right\}\right)\right\} \\
&= \mathcal{E}_{\mathrm{I}_{inter}^{m,2}}\left\{\mathrm{P}\left(z_{m,2} < \frac{\mathrm{I}_{inter}^{m,2} + \frac{1}{\rho}}{\min\left\{\frac{\alpha_1^2 - \epsilon_1 \alpha_2^2}{\epsilon_1}, \frac{\alpha_2^2}{\epsilon_2}\right\}}\right)\right\}.
\end{aligned} \tag{84}$$

After using the approximated expression for the pdf of $z_{m,2}$, the outage probability can be approximated as follows:

$$\begin{aligned}
\mathrm{P}_{m,2}^{o} &\approx \mathcal{E}_{\mathrm{I}_{inter}^{m,2}}\left\{\sum_{n=1}^{N} \bar{w}_n \left(1 - e^{-c_{n,\mathcal{R}_s} \frac{\mathrm{I}_{inter}^{m,2} + \frac{1}{\rho}}{\min\left\{\frac{\alpha_1^2 - \epsilon_1 \alpha_2^2}{\epsilon_1}, \frac{\alpha_2^2}{\epsilon_2}\right\}}}\right)\right\} \\
&\approx 1 - \sum_{n=1}^{N} \bar{w}_n e^{-\frac{c_{n,\mathcal{R}_s} \frac{1}{\rho}}{\min\left\{\frac{\alpha_1^2 - \epsilon_1 \alpha_2^2}{\epsilon_1}, \frac{\alpha_2^2}{\epsilon_2}\right\}}} \\
&\quad \times \mathcal{E}_{\mathrm{I}_{inter}^{m,2}}\left\{e^{-\frac{c_{n,\mathcal{R}_s} \mathrm{I}_{inter}^{m,2}}{\min\left\{\frac{\alpha_1^2 - \epsilon_1 \alpha_2^2}{\epsilon_1}, \frac{\alpha_2^2}{\epsilon_2}\right\}}}\right\}.
\end{aligned} \tag{85}$$

Denote the Laplace transform of $\mathrm{I}_{inter}^{m,2}$ by $\mathcal{L}_{\mathrm{I}_{inter}^{m,2}}(s)$. Then, the outage probability can be rewritten as follows:

$$\begin{aligned}
\mathrm{P}_{m,2}^{o} &\approx 1 - \sum_{n=1}^{N} \bar{w}_n e^{-\frac{c_{n,\mathcal{R}_s} \frac{1}{\rho}}{\min\left\{\frac{\alpha_1^2 - \epsilon_1 \alpha_2^2}{\epsilon_1}, \frac{\alpha_2^2}{\epsilon_2}\right\}}} \\
&\quad \times \mathcal{L}_{\mathrm{I}_{inter}^{m,2}}\left(\frac{c_{n,\mathcal{R}_s}}{\min\left\{\frac{\alpha_1^2 - \epsilon_1 \alpha_2^2}{\epsilon_1}, \times \frac{\alpha_2^2}{\epsilon_2}\right\}}\right).
\end{aligned} \tag{86}$$

Therefore, the outage probability can be calculated if the Laplace transform of $\mathrm{I}_{inter}^{m,2}$ is known. Particularly, the Laplace transform of $\mathrm{I}_{inter}^{m,2}$, $\mathcal{L}_{\mathrm{I}_{inter}^{m,2}}(s)$, can be first expressed as follows:

$$\begin{aligned}
&\mathcal{L}_{\mathrm{I}_{inter}^{m,2}}(s) \\
&= \mathcal{E}\left\{\prod_{x_j \in \Phi_c \setminus x_m} \exp\left(-s \frac{|h_{j,m2}|^2}{L\left(||y_{m,2} + x_m - x_j||\right)}\right)\right\}.
\end{aligned}$$

By using the assumption that $h_{j,m2}$ is Rayleigh distributed, the small scale fading gain can be averaged out in the expression, and the Laplace transform can be expressed as follows:

$$\mathcal{L}_{\mathrm{I}_{inter}^{m,2}}(s) = \mathcal{E}\left\{\prod_{x_j \in \Phi_c \setminus x_m} \frac{1}{\frac{s}{L(||y_{m,2} + x_m - x_j||)} + 1}\right\}. \tag{87}$$

By applying Campbell's theorem and the PFGL [31], [32], [44], the Laplace transform can be simplified as follows:

$$\begin{aligned}
&\mathcal{L}_{\mathrm{I}_{inter}^{m,2}}(s) \\
&= \exp\left(-\lambda_c \int_{\mathbb{R}^2} \left(1 - \mathcal{E}_{y_{m,2}}\left\{\frac{1}{\frac{s}{L(||y_{m,2} + x_m - x||)} + 1}\right\}\right) dx\right),
\end{aligned} \tag{88}$$

which contains a two-dimensional integral with respect to an HPPP point x. Denote the pdf of $y_{m,2}$, $y_{m,2} \in \mathcal{B}(x_m, \mathcal{R}_s)$, by $f_{y_{m,2}}(y)$, where we recall that $\mathcal{B}(x_m, \mathcal{R}_s)$ denotes the disc with radius $\mathcal{R}_s$ and its origin located at x_m. Therefore, the Laplace transform can be expressed as follows:

$$\begin{aligned}
\mathcal{L}_{\mathrm{I}_{inter}^{m,2}}(s) &= \exp\left(-\lambda_c \int_{\mathcal{B}(x_m, \mathcal{R}_s)} f_{y_{m,2}}(y) \right. \\
&\quad \left. \times \int_{\mathbb{R}^2} \left(1 - \frac{1}{\frac{s}{L(||y + x_m - x||)} + 1}\right) dx dy\right).
\end{aligned} \tag{89}$$

Following steps similar to those in [31]–[33] and [44], the substitution of $y + x_m - x \rightarrow x'$ can be used to simplify the expression of the Laplace transform as follows:

$$\begin{aligned}\mathcal{L}_{\mathrm{I}_{inter}^{m,2}}(s) &= \exp\left(-\lambda_c \int_{\mathcal{B}(x_m,\mathcal{R}_s)} f_{y_{m,2}}(y) \right. \\ &\quad \left. \times \int_{\mathbb{R}^2} \left(1 - \frac{1}{\frac{s}{L(||x'||)}+1}\right) dx'dy\right) \\ &= \exp\left(-\lambda_c \int_{\mathcal{B}(x_m,\mathcal{R}_s)} f_{y_{m,2}}(y) 2\pi \right. \\ &\quad \left. \times \int_0^{\infty} \left(1 - \frac{1}{\frac{s}{L(r)}+1}\right) r dr dy\right). \end{aligned} \tag{90}$$

After applying the beta function [45], the Laplace transform can be obtained as follows:

$$\begin{aligned}\mathcal{L}_{\mathrm{I}_{inter}^{m,2}}(s) &= \exp\left(-\lambda_c \int_{\mathcal{B}(x_m,\mathcal{R}_s)} f_{y_{m,2}}(y) 2\pi \frac{s^{\frac{2}{\alpha}}}{\alpha} \right. \\ &\quad \left. \times \mathrm{B}\left(\frac{2}{\alpha}, \frac{\alpha-2}{\alpha}\right) dy\right) \\ &= \exp\left(-2\pi\lambda_c \frac{s^{\frac{2}{\alpha}}}{\alpha} \mathrm{B}\left(\frac{2}{\alpha}, \frac{\alpha-2}{\alpha}\right)\right), \end{aligned} \tag{91}$$

where the last equality follows from the fact that the integral with respect to y is not a function of x. Substituting (91) into (86), the first part of the lemma is proved.

B. The Outage Performance at $U_{m,1}$

Recall that $\mathrm{U}_{m,1}$ is located inside a ring with $\mathcal{R}_s$ as the inner radius and $\mathcal{R}_c$ as the outer radius. Therefore, the CDF of this user's channel gain needs to be calculated differently compared to that of $\mathrm{U}_{m,2}$. First, by using the assumptions that the user is uniformly distributed inside the ring and the fading gain is Rayleigh distributed, the CDF of $z_{m,1}$ can be expressed as follows [46]:

$$\begin{aligned}F_{z_{m,1}}(z) &= \frac{2}{\mathcal{R}_c^2 - \mathcal{R}_s^2} \int_{\mathcal{R}_s}^{\mathcal{R}_c} \left(1 - e^{-r^{\alpha}z}\right) r dr \\ &= \frac{1}{\mathcal{R}_c^2 - \mathcal{R}_s^2} \left[\mathcal{R}_c^2 \frac{2}{\mathcal{R}_c^2} \int_0^{\mathcal{R}_c} \left(1 - e^{-r^{\alpha}z}\right) r dr \right. \\ &\quad \left. - \mathcal{R}_s^2 \frac{2}{\mathcal{R}_s^2} \int_0^{\mathcal{R}_s} \left(1 - e^{-r^{\alpha}z}\right) r dr\right]. \end{aligned} \tag{92}$$

Comparing this with [6, eq. (3)], one can find that the approximated form shown in (81) can be applied to each term in the above expression, and hence the CDF can be approximated as follows:

$$F_{z_{m,1}}(z) = \frac{1}{\mathcal{R}_c^2 - \mathcal{R}_s^2} \left[\mathcal{R}_c^2 F_{\mathcal{R}_c}(z) - \mathcal{R}_s^2 F_{\mathcal{R}_s}(z)\right]. \tag{93}$$

Following steps similar to those in the previous subsection, the outage probability of $\mathrm{U}_{m,1}$ for decoding $f_{m,1}$ can be obtained as follows:

$$\mathrm{P}_{m,1}^{o} = \mathcal{E}_{\mathrm{I}_{inter}^{m,1}} \left\{ \mathrm{P}\left(z_{m,1} < \frac{\epsilon_1 \mathrm{I}_{inter}^{m,1} + \frac{\epsilon_1}{\rho}}{\alpha_1^2 - \epsilon_1 \alpha_2^2}\right)\right\}. \tag{94}$$

After using the approximated expression for the pdf of $z_{m,1}$, the outage probability can be approximated as follows:

$$\begin{aligned}&\mathrm{P}_{m,1}^{o} \\ &\approx \frac{\mathcal{R}_c^2}{\mathcal{R}_c^2 - \mathcal{R}_s^2} \mathcal{E}_{\mathrm{I}_{inter}^{m,1}} \left\{ \sum_{n=1}^{N} \bar{w}_n \left(1 - e^{-c_{n,\mathcal{R}_c} \frac{\epsilon_1 \mathrm{I}_{inter}^{m,1} + \frac{\epsilon_1}{\rho}}{\alpha_1^2 - \epsilon_1 \alpha_2^2}}\right)\right\} \\ &\quad - \frac{\mathcal{R}_s^2}{\mathcal{R}_c^2 - \mathcal{R}_s^2} \mathcal{E}_{\mathrm{I}_{inter}^{m,1}} \left\{ \sum_{n=1}^{N} \bar{w}_n \left(1 - e^{-c_{n,\mathcal{R}_s} \frac{\epsilon_1 \mathrm{I}_{inter}^{m,1} + \frac{\epsilon_1}{\rho}}{\alpha_1^2 - \epsilon_1 \alpha_2^2}}\right)\right\} \\ &\approx 1 + \frac{\mathcal{R}_s^2}{\mathcal{R}_c^2 - \mathcal{R}_s^2} \sum_{n=1}^{N} \bar{w}_n e^{-\frac{c_{n,\mathcal{R}_s} \frac{\epsilon_1}{\rho}}{\alpha_1^2 - \epsilon_1 \alpha_2^2}} \mathcal{E}_{\mathrm{I}_{inter}^{m,1}} \left\{ e^{-\frac{c_{n,\mathcal{R}_s} \epsilon_1 \mathrm{I}_{inter}^{m,1}}{\alpha_1^2 - \epsilon_1 \alpha_2^2}}\right\} \\ &\quad - \frac{\mathcal{R}_c^2}{\mathcal{R}_c^2 - \mathcal{R}_s^2} \sum_{n=1}^{N} \bar{w}_n e^{-\frac{c_{n,\mathcal{R}_c} \frac{\epsilon_1}{\rho}}{\alpha_1^2 - \epsilon_1 \alpha_2^2}} \mathcal{E}_{\mathrm{I}_{inter}^{m,1}} \left\{ e^{-\frac{c_{n,\mathcal{R}_c} \epsilon_1 \mathrm{I}_{inter}^{m,1}}{\alpha_1^2 - \epsilon_1 \alpha_2^2}}\right\}. \end{aligned} \tag{95}$$

It is straightforward to show that the Laplace transform of $\mathrm{I}_{inter}^{m,1}$ is the same as that of $\mathrm{I}_{inter}^{m,2}$. Therefore, substituting (91) with (95), the second part of the lemma is proved.

REFERENCES

[1] Z. Ding, P. Fan, G. Karagiannidis, R. Schober, and H. V. Poor, "On the application of NOMA to wireless caching," in *Proc. IEEE Int. Conf. Commun.*, Kansas City, MO, USA, May 2018.

[2] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. Bhargava, "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181–2195, Oct. 2017.

[3] "5G radio access: Requirements, concepts and technologies," NTT DOCOMO, Inc., Tokyo, Japan, 5G White Paper, Jul. 2014.

[4] "5G innovation opportunities—A discussion paper," techUK, London, U.K., 5G White Paper, Aug. 2015.

[5] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE Veh. Technol. Conf.*, Dresden, Germany, Jun. 2013, pp. 1–5.

[6] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014.

[7] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.

[8] H. Nikopour and H. Baligh, "Sparse code multiple access," in *Proc. IEEE Int. Symp. Pers. Indoor Mobile Radio Commun.*, London, U.K., Sep. 2013, pp. 332–336.

[9] X. Dai *et al.*, "Successive interference cancelation amenable multiple access (SAMA) for future wireless communications," in *Proc. IEEE Int. Conf. Commun. Syst.*, Coimbatore, India, Nov. 2014, pp. 222–226.

[10] Z. Ding, P. Fan, and H. V. Poor, "Random beamforming in millimeter-wave NOMA networks," *IEEE Access*, vol. 5, pp. 7667–7681, 2017.

[11] Z. Ding, L. Dai, R. Schober, and H. V. Poor, "NOMA meets finite resolution analog beamforming in massive MIMO and millimeter-wave networks," *IEEE Commun. Lett.*, vol. 21, no. 8, pp. 1879–1882, Aug. 2017.

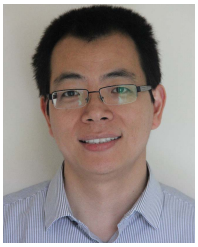
[12] J. Choi, "Minimum power multicast beamforming with superposition coding for multiresolution broadcast and application to NOMA systems," *IEEE Trans. Commun.*, vol. 63, no. 3, pp. 791–800, Mar. 2015.

[13] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 537–552, Jan. 2016.

[14] X. Chen, Z. Zhang, C. Zhong, and D. W. K. Ng, "Exploiting multiple-antenna techniques for non-orthogonal multiple access," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2207–2220, Oct. 2017.

[15] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *IEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 3062–3080, Dec. 2004.

[16] Z. Ding, M. Peng, and H. V. Poor, "Cooperative non-orthogonal multiple access in 5G systems," *IEEE Commun. Lett.*, vol. 19, no. 8, pp. 1462–1465, Aug. 2015.
[17] B. Zheng, X. Wang, M. Wen, and F.-J. Chen, "NOMA-based multi-pair two-way relay networks with rate splitting and group decoding," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2328–2341, Oct. 2017.
[18] M. Xu, F. Ji, M. Wen, and W. Duan, "Novel receiver design for the cooperative relaying system with non-orthogonal multiple access," *IEEE Commun. Lett.*, vol. 20, no. 8, pp. 1679–1682, Aug. 2016.
[19] Y. Liu, Z. Ding, M. Elkashlan, and J. Yuan, "Nonorthogonal multiple access in large-scale underlay cognitive radio networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 10152–10157, Dec. 2016.
[20] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82–89, Aug. 2014.
[21] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Commun. Mag.*, vol. 51, no. 4, pp. 142–149, Apr. 2013.
[22] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.
[23] X. Xu and M. Tao, "Modeling, analysis, and optimization of coded caching in small-cell networks," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3415–3428, Aug. 2017.
[24] Z. Chen and M. Kountouris, "D2D caching vs. small cell caching: Where to cache content in a wireless network?" in *Proc. Int. Workshop Signal Process. Adv. Wireless Commun.*, Edinburgh, U.K., Jul. 2016, pp. 1–6.
[25] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6118–6131, Sep. 2016.
[26] N. Zhao, X. Liu, F. R. Yu, M. Li, and V. C. M. Leung, "Communications, caching, and computing oriented small cell networks with interference alignment," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 29–35, Sep. 2016.
[27] F. Cheng, Y. Yu, Z. Zhao, N. Zhao, Y. Chen, and H. Lin, "Power allocation for cache-aided small-cell networks with limited backhaul," *IEEE Access*, vol. 5, pp. 1272–1283, 2017.
[28] 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.
[29] R. Wang, J. Zhang, S. H. Song, and K. B. Letaief, "Mobility-aware caching in D2D networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 5001–5015, Aug. 2017.
[30] Z. Zhao, M. Xu, Y. Li, and M. Peng, "A non-orthogonal multiple access-based multicast scheme in wireless content caching networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2723–2735, Dec. 2017.
[31] M. Haenggi, *Stochastic Geometry for Wireless Networks*. Cambridge, U.K.: Cambridge Univ. Press, 2012.
[32] K. Gulati, B. L. Evans, J. G. Andrews, and K. R. Tinsley, "Statistics of co-channel interference in a field of Poisson and Poisson-Poisson clustered interferers," *IEEE Trans. Signal Process.*, vol. 58, no. 12, pp. 6207–6222, Dec. 2010.
[33] Y. J. Chun, M. O. Hasna, and A. Ghrayeb, "Modeling heterogeneous cellular networks interference using Poisson cluster processes," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 10, pp. 2182–2195, Oct. 2015.
[34] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.
[35] J. Venkataraman, M. Haenggi, and O. Collins, "Shot noise models for outage and throughput analyses in wireless ad hoc networks," in *Proc. IEEE Military Commun. Conf.*, Washington, DC, USA, Oct. 2006, pp. 1–7.
[36] F. Baccelli and A. Giovanidis, "A stochastic geometry framework for analyzing pairwise-cooperative cellular networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 2, pp. 794–808, Feb. 2015.
[37] *Study on Downlink Multiuser Superposition Transmission for LTE*, document, 3rd Generation Partnership Project, Mar. 2015.
[38] J. Tang, G. Chen, J. P. Coon, and D. E. Simmons, "Distance distributions for matern cluster processes with application to network performance analysis," in *Proc. IEEE Int. Conf. Commun.*, Paris, France, May 2017, pp. 1–7.
[39] M. Haenggi, "On distances in uniformly random networks," *IEEE Trans. Inf. Theory*, vol. 51, no. 10, pp. 3584–3586, Oct. 2005.
[40] A. Liu and V. K. N. Lau, "Cache-enabled opportunistic cooperative MIMO for video streaming in wireless systems," *IEEE Trans. Signal Process.*, vol. 62, no. 2, pp. 390–402, Jan. 2014.
[41] J. Liu, B. Bai, J. Zhang, and K. B. Letaief, "Cache placement in Fog-RANs: From centralized to distributed algorithms," *IEEE Trans. Wireless Commu.*, vol. 16, no. 11, pp. 7039–7051, Nov. 2017.
[42] Y. Cao, M. Tao, F. Xu, and K. Liu, "Fundamental storage-latency tradeoff in cache-aided MIMO interference networks," *IEEE Trans. Wireless Commu.*, vol. 16, no. 8, pp. 5061–5076, Aug. 2017.
[43] A. Sengupta, R. Tandon, and O. Simeone, "Fog-aided wireless networks for content delivery: Fundamental latency tradeoffs," *IEEE Trans. Inf. Theory*, vol. 63, no. 10, pp. 6650–6678, Oct. 2017.
[44] Z. Qin, Y. Liu, G. Y. Li, and J. A. McCann, "Modelling and analysis of low-power wide-area networks," in *Proc. IEEE Int. Conf. Commun.*, Paris, France, May 2017, pp. 1–7.
[45] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products*, 6th ed. New York, NY, USA: Academic, 2000.
[46] H. Wang, S. Ma, T.-S. Ng, and H. V. Poor, "A general analytical approach for opportunistic cooperative systems with spatially random relays," *IEEE Trans. Wireless Commun.*, vol. 10, no. 12, pp. 4122–4129, Dec. 2011.



Zhiguo Ding (S'03–M'05–SM'15) received the B.Eng. degree in electrical engineering from the Beijing University of Posts and Telecommunications in 2000 and the Ph.D. degree in electrical engineering from the Imperial College London in 2005. From 2005 to 2018, he was with Queen's University Belfast, Imperial College, Newcastle University, and Lancaster University. From 2012 to 2018, he was an Academic Visitor at Princeton University. Since 2018, he has been with The University of Manchester as a Professor in communications.

His research interests are 5G networks, game theory, cooperative and energy harvesting networks, and statistical signal processing. He was an Editor of the IEEE WIRELESS COMMUNICATION LETTERS and the IEEE COMMUNICATION LETTERS from 2013 to 2016. He is serving as an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and the *Journal of Wireless Communications and Mobile Computing*. He received the Best Paper Award from the IET ICWMC in 2009 and the IEEE WCSP-2014, the EU Marie Curie Fellowship 2012–2014, the Top IEEE TVT Editor 2017, and the IEEE Heinrich Hertz Award 2018.

Pingzhi Fan (M'93–SM'99–F'15) received the M.Sc. degree in computer science from the Southwest Jiaotong University, China, in 1987, and the Ph.D. degree in electronic engineering from Hull University, U.K., in 1994. He is currently a Professor and the Director of the Institute of Mobile Communications, Southwest Jiaotong University, China. He has been a Visiting Professor with Leeds University, U.K., since 1997 and a Guest Professor with Shanghai Jiaotong University since 1999. He was a recipient of the U.K. ORS Award in 1992 and the Outstanding Young Scientist Award (NSFC) in 1998, the Chief Scientist of a National 973 Research Project (MoST) 2012–2016.

He has over 280 research papers published in various international journals and eight books (including edited). He has invented 22 granted patents. His research interests include vehicular communications, wireless networks for big data, signal design, and coding. He is a fellow of the IET, CIE, and CIC. He is an IEEE VTS Distinguished Lecturer (2015–2019). He also served as a Board Member of IEEE Region 10, IET (IEE) Council, and the IET Asia–Pacific Region. He served as the general chair or the TPC chair for a number of international conferences. He is the guest editor or an editorial member of several international journals. He is the Founding Chair of the IEEE VTS BJ Chapter, the IEEE ComSoc CD Chapter, and the IEEE Chengdu Section.

George K. Karagiannidis (M'96–SM'03–F'14) was born in Pithagorion, Greece. He received the University Diploma (5 years) and the Ph.D. degrees in electrical and computer engineering from the University of Patras in 1987 and 1999, respectively. From 2000 to 2004, he was a Senior Researcher with the Institute for Space Applications and Remote Sensing, National Observatory of Athens, Greece. In 2004, he joined the Aristotle University of Thessaloniki, Greece, as a Faculty Member, where he is currently a Professor with the Electrical and Computer Engineering Department and the Director of the Digital Telecommunications Systems and Networks Laboratory. He is also an Honorary Professor with South West Jiaotong University, Chengdu, China.

He has authored or co-authored over 450 technical papers published in scientific journals and presented at international conferences. He has also authored the Greek edition of a book on *Telecommunications Systems* and co-authored the book *Advanced Optical Wireless Communications Systems* (Cambridge Publications, 2012). His research interests are in the broad area of digital communications systems and signal processing, with emphasis on wireless communications, optical wireless communications, wireless power transfer and applications, molecular and nanoscale communications, stochastic processes in biology, and wireless security.

Dr. Karagiannidis has been involved as the general chair, the technical program chair, and a member of the technical program committees in several IEEE and non-IEEE conferences. He was an Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS and the *EURASIP Journal of Wireless Communications and Networks*, a Senior Editor of the IEEE COMMUNICATIONS LETTERS, and a Guest Editor of the IEEE SELECTED AREAS IN COMMUNICATIONS several times. From 2012 to 2015, he was the Editor-in Chief of the IEEE COMMUNICATIONS LETTERS. He is one of the highly cited authors across all areas of electrical engineering and recognized as the 2015, 2016, and 2017 Web of Science Highly Cited Researcher.

Robert Schober (M'01–SM'08–F'10) received the Diplom (Univ.) and the Ph.D. degrees in electrical engineering from the Friedrich-Alexander University of Erlangen-Nuremberg (FAU), Germany, in 1997 and 2000, respectively. From 2002 to 2011, he was a Professor and the Canada Research Chair with The University of British Columbia (UBC), Vancouver, Canada. Since 2012, he has been an Alexander von Humboldt Professor and the Chair for digital communication with FAU. His research interests fall into the broad areas of communication theory, wireless communications, and statistical signal processing.

Dr. Schober is a fellow of the Canadian Academy of Engineering and the Engineering Institute of Canada. He received several awards for his work, including the 2002 Heinz Maier-Leibnitz Award of the German Science Foundation, the 2004 Innovations Award of the Vodafone Foundation for Research in Mobile Communications, the 2006 UBC Killam Research Prize, the 2007 Wilhelm Friedrich Bessel Research Award of the Alexander von Humboldt Foundation, the 2008 Charles McDowell Award for Excellence in Research from UBC, the 2011 Alexander von Humboldt Professorship, the 2012 NSERC E.W.R. Steacie Fellowship, and the 2017 Wireless Communications Recognition Award by the IEEE Wireless Communications Technical Committee. He is currently the Chair of the Steering Committee of the IEEE TRANSACTIONS ON MOLECULAR, BIOLOGICAL AND MULTISCALE COMMUNICATION, a member of the Editorial Board of the PROCEEDINGS OF THE IEEE, and a Member at Large of the Board of Governors of ComSoc and the ComSoc Director of Journals. From 2012 to 2015, he served as the Editor-in-Chief for the IEEE TRANSACTIONS ON COMMUNICATIONS. He was listed as a 2017 Highly Cited Researcher by the Web of Science and a Distinguished Lecturer of the IEEE Communications Society (ComSoc).

H. Vincent Poor (M'77–SM'82–F'87) received the Ph.D. degree in EECS from Princeton University in 1977. From 1977 to 1990, he was on the faculty of the University of Illinois at Urbana–Champaign. Since 1990, he has been on the faculty at Princeton, where he is currently the Michael Henry Strater University Professor of Electrical Engineering. From 2006 to 2016, he served as the Dean of Princeton's School of Engineering and Applied Science. He has also held visiting appointments at several other universities, including most recently at Berkeley and Cambridge. His research interests are in the areas of information theory and signal processing, and their applications in wireless networks, energy systems and related fields. Among his publications in these areas is the recent book *Information Theoretic Security and Privacy of Information Systems* (Cambridge University Press, 2017).

Dr. Poor is a member of the National Academy of Engineering and the National Academy of Sciences and a foreign member of the Chinese Academy of Sciences, the Royal Society, and other national and international academies. He received the Marconi Award and Armstrong Award from the IEEE Communications Society in 2007 and 2009, respectively. Recent recognition of his work includes the 2017 IEEE Alexander Graham Bell Medal, Honorary Professorships at Peking University and Tsinghua University, both conferred in 2017, and a D.Sc. *honoris causa* from Syracuse University in 2017.