On the Application of NOMA to Wireless Caching

Zhiguo Ding, Pingzhi Fan, George K. Karagiannidis, Robert Schober, and H. Vincent Poor

***Abstract*—This paper investigates the impact of non-orthogonal multiple access (NOMA) on wireless caching. Two NOMA caching strategies are developed, namely the *push-then-deliver strategy* and the *push-and-deliver strategy*, with the objective to improve the spectral efficiency of the two caching phases, content pushing and content delivery, compared to the conventional orthogonal multiple-access (OMA) based strategy. Both analytical and computer simulation results are provided to demonstrate the performance of the proposed caching strategies and verify the accuracy of the developed analytical results.**

I. Introduction

Recently, non-orthogonal multiple access (NOMA) has received significant attention as a key enabling technique for future wireless networks [1]–[3], and 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 [4]. In addition, NOMA was known to improve the spectral efficiency of multiple-input multiple-output (MIMO) and cooperative systems [5]–[8].

Wireless caching is another important enabling technique for future communication networks [9], [10], 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*. When caching infrastructure (e.g., content servers) is available, the aim of 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 phase of content delivery, an ideal situation is that all the users' requests can be locally served, without communicating with the central controller of the network [11]–[13].

This paper investigates the coexistence of NOMA and wireless caching, which is crucial for their joint implementation in future wireless networks, for the case when a cache infrastructure exists. In particular, we concentrate on the following two questions. The *first* question is how to realize content pushing in a timely and robust manner. Most existing works on caching assume that content can be reliably pushed to the caching infrastructure during off-peak hours. However, this assumption might not be realistic due to the dynamic nature of content popularity which implies that some of the cached content may have to be replaced by new content during peak hours. In addition, wireless transmission is prone to attenuation and various impairments. Therefore, timely and robust content pushing is critical for efficient wireless caching. The *second* question is how to cope with the non-ideal situation for wireless caching, when some users' requests have to be fetched from the base station (BS) directly. We note that for wireless caching this non-ideal situation is inevitable and is expected to occur frequently in practice, as the users' requests cannot be perfectly predicted. When this situation happens, the spectral efficiency of wireless caching is reduced, since the users' requests cannot be accommodated locally. The two NOMA-assisted caching strategies proposed in this paper address the aforementioned questions, as explained in the following.

For the case when the content pushing and delivery phases are separated, a *NOMA-assisted push-then-deliver strategy* is proposed. Particularly, during the content pushing phase, the BS will adopt the NOMA principle to push multiple files to the content servers simultaneously. A cognitive radio (CR) inspired NOMA power allocation policy is used to ensure that the most popular file is delivered to the targeted content server with the same outage probability as with orthogonal multiple access (OMA) based transmission. However, by using NOMA, additional files can be pushed to the content servers simultaneously, which significantly improves the cache hit probability. Furthermore, 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 the content server.

In addition, a *NOMA assisted push-and-deliver strategy* is proposed to efficiently combine the content pushing and delivery phases, in order to effectively cope with the situation when some users' requests have to be served by the BS directly. Although this situation is not desirable for wireless caching, it is inevitable in practice and constitutes an opportunity for the application of NOMA. Particularly, when a BS serves the user directly, i.e., it delivers a file directly to a user, the NOMA principle enables the BS to perform content delivery and content pushing simultaneously, i.e., it can push new content to the servers while serving users directly. We note that the proposed push-and-deliver strategy can be easily extended to device-to-device (D2D) caching, as shown in the journal version of this paper [14], and the NOMA-multicasting scheme proposed in [15] can be viewed as a D2D special case of the proposed strategy.

Z. Ding and H. V. Poor are with the Department of Electrical Engineering, Princeton University, Princeton, USA. Z. Ding is also with the School of Computing and Communications, Lancaster University, UK. P. Fan is with the Institute of Mobile Communications, Southwest Jiaotong University, Chengdu, China. G. K. Karagiannidis is with the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Thessaloniki, Greece. R. Schober is with the Institute for Digital Communications, Friedrich-Alexander-University Erlangen-Nurnberg (FAU), Germany.

The work of Z. Ding was supported by the UK EPSRC under grant number EP/N005597/1 and by H2020-MSCA-RISE-2015 under grant number 690750. The work of H. V. Poor was supported by U.S. National Science Foundation under Grants CNS-1702808 and ECCS-1647198. The work of R. Schober was supported in part by the Alexander von Humboldt Professorship Program of the Alexander von Humboldt Foundation.

978-1-5386-3180-5/18/$31.00 ©2018 IEEE

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. Particularly, assume that the BS is located in the origin of a two-dimensional Euclidean plane, denoted by $\mathbb{R}^2$. The locations of the content servers and the users are modelled as Poisson cluster processes (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 [16]. 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 $\mathrm{U}_{i,k}$ to CS_j is denoted by $||y_{i,k}+x_i-x_j||$ [17].

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. However, if the file requested by a user cannot be found locally, the BS will serve the user directly, a situation which is not ideal for caching and should be avoided.

More specifically, consider 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 [18]. Particularly, the popularity of file f_i, 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_i. The prefixed data rate of the packets of file f_l is denoted by R_l.

III. Push-then-Deliver Strategy

This section considers the case where the two caching phases, content pushing and content delivery, are separated, and we show that the use of NOMA can improve the reliability of both content pushing and delivery.

A. Content Pushing Phase

In order to better illustrate the performance of NOMA assisted content pushing, the conventional OMA based content pushing strategy is introduced first.

1) OMA Based Content Pushing: Without loss of generality, assume that there is only one time slot for content pushing[1]. If OMA is used, the BS broadcasts the most popular file, f_1, to the content servers. Therefore, CS_m is able to decode file f_1 with the following achievable data rate:

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

where ρ denotes the transmit signal-to-noise ratio (SNR), and $\frac{1}{L(||x_m||)}$ is the large scale path loss between CS_m and the BS located at x_0. Particularly, the following path loss model is used, $\frac{1}{L(||x_m||)}$, where $L\left(||x_m||\right) = ||x_m||^\alpha$ and α denotes the path loss exponent. We note that small scale multi-path fading is not considered for the channel gain associated with CS_m since the content servers can be deployed such that line-of-sight connections to the BS are ensured, which means that the large scale path loss is the dominant factor for signal attenuation. However, small scale fading will be considered for the channel gains of the users, since the users may not have light-of-sight connections to their respective transmitters.

2) NOMA Assisted Content Pushing: By applying the concept of NOMA, more content can be simultaneously delivered from the BS to the servers. Particularly, the BS sends the following mixture, which contains the M_s most popular files:

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

where $\bar{f}_i$ denotes the signal which represents the information contained in file f_i, α_i denotes the power allocation coefficient and $\sum_{i=1}^{M_s} \alpha_i^2 = 1$. Each content server carries 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 one, 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.

In order to compare with OMA, which pushes only one file at a time, a sophisticated power allocation policy is needed for the NOMA scheme. Without loss of generality, we assume that the content servers are ordered as follows:

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

for $1 \leq m < t$. Furthermore, we make the following quality of service (QoS) assumption, in order to facilitate the design of the power alllocation coefficients:

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

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 [19], i.e., power allocation coefficient α_1 is chosen such that f_1 can be delivered reliably to CS_t, i.e.,

$$R_{t,1}^{CP} \geq R_1. \tag{6}$$

[1]In practice, there will be multiple time slots for content pushing, and the proposed scheme can be straightforwardly extended to the multi-time-slot case by sending different files in different time slots.

This constraint results in the following choice of α_1:

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

where $\epsilon_l = 2^{R_l}-1$. As demonstrated in the performance analysis section, the use of the power allocation policy in (7) ensures that the outage probability for the NOMA based pushing strategy for the most popular file, f_1, is the same as that for OMA.

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:

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

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 the β_i are constants, which satisfy the constraint $\sum_{j=2}^{M_s}\beta_i = 1$.

3) Performance Analysis: An effective criterion to evaluate the performance of 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. Since the request probability for file l is decided by its popularity, the hit probability for a user associated with CS_m can be expressed as follows:

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

where $\text{P}_{m,i}$ denotes the outage probability of CS_m for decoding file i. Note that for the OMA case, only file 1 will be sent, and hence the corresponding OMA hit probability is simply given by

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

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

Theorem 1. *The cache hit probability achieved by the proposed NOMA assisted push-then-deliver strategy is not smaller than that of the conventional OMA based strategy, i.e.,*

$$\text{P}_m^{hit} \geq \text{P}_{m,OMA}^{hit}, \tag{11}$$

for $1 \leq m \leq t$.

Proof: Please refer to Appendix A in [14]. ■

Remark 1: 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.

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 is impacted by the channel conditions of CS_t. This means that for the calculation of the outage probability, $\text{P}_{m,i}$, 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 probability density function (pdf) of r_m and r_t is given by*

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

Proof: Please refer to Appendix B in [14]. ■

Remark 2: It is worth pointing out that the joint pdf obtained in [20] is a special case of Lemma 1, when $m = 1$ and $t = 2$.

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*

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

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

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

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

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

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

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

Proof: Please refer to Appendix C in [14]. ■

Remark 3: In Lemma 2, it is assumed that the targeted 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 4: 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, as shown in [14], and the assumption can be justified as follows. If R_1 is large, the use of the CR power allocation policy means that most of the transmission power will be consumed by f_1, and hence, not much power will be left for pushing additional files. In other words, the case with a large R_1 is not an ideal situation for applying the proposed NOMA pushing strategy.

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. Hence, in this subsection, the impact of NOMA on the reliability of content delivery is investigated. Similar to the previous subsection, the OMA based strategy is described first as a benchmark scheme.

1) OMA Based Content Delivery: For the OMA case, during the content delivery phase, each content server randomly schedules a single user whose request is available locally in the cache of the server. We assume that each content server can find a user to serve, and all the content servers transmit simultaneously, which facilitates the used PCP modelling.

2) NOMA Assisted Content Delivery: If the NOMA principle is applied in the content delivery phase, each content server can serve two users[2]. Assume that the two users are ordered based on their distances to their associated content servers. The weak user, denoted by $\mathrm{U}_{m,1}$, is inside a ring with radii $\mathcal{R}_s$ and $\mathcal{R}_c$, $\mathcal{R}_s < \mathcal{R}_c$. The strong user, denoted by $\mathrm{U}_{m,2}$, is in 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:

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

where $\bar{f}_{j,l}$ denotes the signal which represents the information contained in file $f_{j,l}$, α_l denotes the NOMA power allocation coefficient, $n_{m,k}$ is the additive complex Gaussian noise, and $h_{j,mk}$ denotes the Rayleigh fading channel coefficient between CS_j and $\mathrm{U}_{m,k}$. 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 notations consistent, the power allocation coefficients are still denoted by α_i.

[2]The two-user case is focused since two-user NOMA based downlink transmission has been proposed for long term evolution (LTE) Advanced [21].

As a result, $\mathrm{U}_{m,1}$ will treat its partner's message as noise and decode its own message $f_{m,1}$ with the following 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 $\mathrm{I}_{inter}^{m,1} = \sum_{x_j \in \Phi_c \setminus m} \frac{|h_{j,m1}|^2}{L(||y_{m,1}+x_m-x_j||)}$. 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 strong 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 $\mathrm{SINR}_{m,1}^{1}$, i.e., $\mathrm{SINR}_{m,2}^{1} = \frac{\frac{\alpha_1^2 |h_{m,m2}|^2}{L(||y_{m,2}||)}}{\frac{\alpha_2^2 |h_{m,m2}|^2}{L(||y_{m,2}||)} + \mathrm{I}_{inter}^{m,2} + \frac{1}{\rho}}$, and the inter-cluster interference, $\mathrm{I}_{inter}^{m,2}$, is defined similarly to $\mathrm{I}_{inter}^{m,1}$. If $\log(1+\mathrm{SINR}_{m,2}^{1}) > R_1$, 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:

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

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

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

and

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

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:*

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

where $\tilde{\tau} = \min\left\{\frac{\alpha_1^2 - \epsilon_1 \alpha_2^2}{\epsilon_1}, \frac{\alpha_2^2}{\epsilon_2}\right\}$, $q(s) = \exp\left(-2\pi\lambda_c \frac{s^{\frac{2}{\alpha}}}{\alpha} B\left(\frac{2}{\alpha}, \frac{\alpha-2}{\alpha}\right)\right)$, *$B(\cdot)$ 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:

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

Proof: Please refer to Appendix D in [14]. ■

IV. Push-and-Deliver Strategy

A situation which is undesirable but inevitable for wireless caching is that a user's request cannot be accommodated by its local content server and hence the BS has to serve the user directly. Conventionally, when this situation happens, the spectrum efficiency of wireless caching is reduced. The proposed push-and-deliver strategy treats this situation as an opportunity for the application of NOMA. In particular, consider a time slot which is dedicated to user $\mathrm{U}_{m,k}$. 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 servers, 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, $1 \leq i \leq M_s$. Assume that f_0 and f_i, $1 \leq i \leq M_s$, belong to different sets of the file library, in order to avoid correlation among these files and to simplify the expression for the cache hit probability. In order to obtain tractable analytical results, it is assumed that $\mathrm{U}_{m,k}$ is randomly selected from the offsprings of CS_m.

A. Performance Analysis

Following similar steps as 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

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

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

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

if R_m^j is larger than R_j, for $0 \leq j \leq i-1$, where R_l denotes the targeted data rate of f_l. Again, small scale multi-path fading is not considered in the channel model of CS_m, as we assume that the large scale path loss is dominant in this case, and 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.

1) Performance of the users: The main challenge in analyzing the outage performance at the users 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 a user can be expressed as follows:

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

where $\zeta_l = \alpha_l^2 - \epsilon_l\sum_{j=l+1}^{M_s}\alpha_j^2$ for $0 \leq l < M_s$, and $\zeta_{M_s} = \alpha_{M_s}^2$. Again it is assumed that the power allocation coefficients and the targeted 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 [22]

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

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

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

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 $r_m \geq \delta\mathcal{R}_c$ for all $m \geq 1$. 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:

$$\mathrm{P}_{m,k}^1 = 1 - \int_{\delta\mathcal{R}_c}^{\infty}\int_{z-\mathcal{R}_c}^{z+\mathcal{R}_c} e^{-\frac{\epsilon_0 r^\alpha}{\rho\zeta_0}} g(z, r) dr \bar{f}_{r_m}(z) dz, \tag{27}$$

where $\bar{f}_{r_m}(z)$ denotes the pdf of r_m. Following the steps provided in [14], the cumulative distribution function (CDF) of r_m can be expressed as follows:

$$\begin{aligned}\tilde{F}_{r_m}(r) =& 1 - \mathrm{P}(\#\text{ of nodes in the ring } \mathcal{A}_r < m) \\ =& 1 - \sum_{l=0}^{m-1}\frac{(\lambda_c[\pi r^2 - \pi\delta^2\mathcal{R}_c^2])^l}{l!} e^{-\lambda_c[\pi r^2 - \pi\delta^2\mathcal{R}_c^2]},\end{aligned} \tag{28}$$

and a derivative of $\tilde{F}_{r_m}(r)$ leads to the pdf of r_m as follows:

$$\bar{f}_{r_m}(r) = 2\pi\lambda_c^m r e^{-\lambda_c[\pi r^2 - \pi\delta^2\mathcal{R}_c^2]}\frac{[\pi r^2 - \pi\delta^2\mathcal{R}_c^2]^{m-1}}{(m-1)!}. \tag{29}$$

Substituting (29) into (27), 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:

$$\begin{aligned}\mathrm{P}_m^i =& 1 - \mathrm{P}(\mathrm{R}_m^l > R_l, \forall l \in \{0, \cdots, i\}) \\ =& \mathrm{P}\left(L(||x_m||) > \min\left\{\frac{\rho\zeta_l}{\epsilon_l}, \forall l \in \{0, \cdots, i\}\right\}\right).\end{aligned} \tag{30}$$

By applying the assumption that $r_m \geq \delta\mathcal{R}_c$ and also the pdf in (29), the outage probability of CS_m for decoding f_i

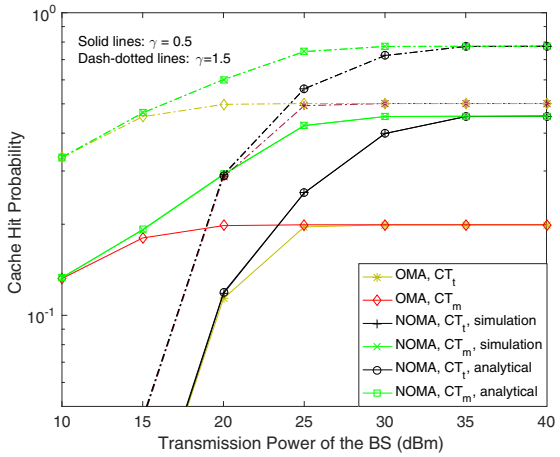

Fig. 1. The cache hit probability for the push-then-deliver strategy. $\mathcal{R}_c = 50$m, $N = 20$, $\alpha = 3$, $\lambda_c = \frac{0.01}{\pi \mathcal{R}_c^2}$, $t = 5$, $m = 1$, $M_s = 3$, and $R_l = 1$ bit per channel use (BPCU), for $1 \leq l \leq 3$. The power coefficient for file f_1 is based 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. $|\mathcal{F}| = 10$.

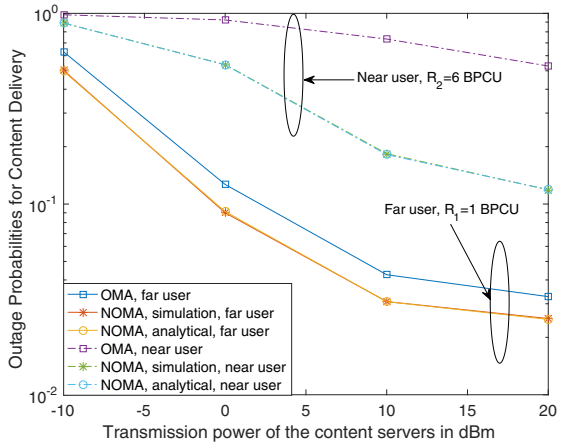

Fig. 2. The outage probabilities for content delivery for the push-then-deliver strategy. $\alpha = 3$, $N = 20$, $\alpha = 4$, $\mathcal{R}_c = 100$m. $\lambda_c = \frac{0.01}{\pi \mathcal{R}_c^2}$, $R_1 = 1$ BPCU, and $R_2 = 6$ BPCU. The power allocation coefficients are $\alpha_1^2 = \frac{3}{4}$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

can be expressed as follows:

$$\mathrm{P}_m^i = \sum_{l=0}^{m-1} \frac{(\lambda_c \left[\frac{\pi}{\bar{\tau}_i^2} - \pi\delta^2 \mathcal{R}_c^2\right])^l}{l!} e^{-\lambda_c \left[\frac{\pi}{\bar{\tau}_i^2} - \pi\delta^2 \mathcal{R}_c^2\right]}, \quad (31)$$

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:

$$\mathrm{P}_m^{hit} = \sum_{i=1}^{M_s} \mathrm{P}(f_i)(1 - \mathrm{P}_m^i), \quad (32)$$

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

B. OMA Benchmarks

As a naive OMA based benchmark, the BS may 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 can still offer a significant gain in terms of the hit probability, as will be shown in Section V.

V. Numerical Studies and Discussions

In this section, the performances achieved by the proposed two strategies are studied by using computer simulations.

In Fig. 1, 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. 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. 1. For example, when the transmission power is 40 dBm and $\gamma = 0.5$, 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. The impact of γ on the hit probability is significant, as can be observed in Fig. 1. Particularly, increasing the value of γ improves the hit probability. This is because a larger value of γ means 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).

In Fig. 2, the impact of using NOMA for content delivery is studied. 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 good channel conditions. For example, for transmission power of the content servers of 20 dBm, the use of NOMA ensures that the outage probability for the near user is improved from 5×10^{-1} to 1.1×10^{-1}, which is a significan performance gain. 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.

In Figs. 3 and 4, the impact of the proposed push-and-deliver strategy on the cache hit probability is studied. 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 the figures, the impact of different choices for the popularity parameters on the cache hit probability is also studied. In particular, the following two cases are considered:

- 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

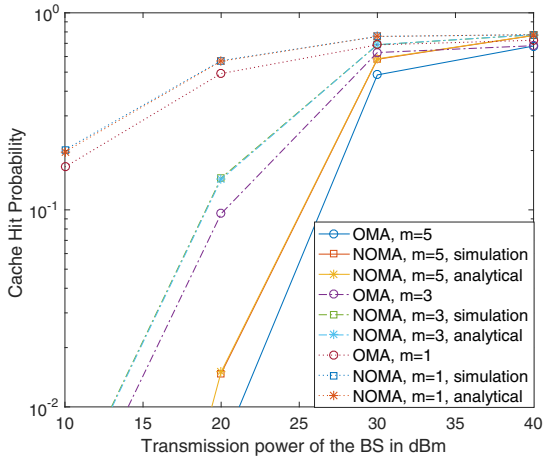

Fig. 3. The cache hit probability for the proposed push-and-deliver strategy for Case 1. $\mathcal{R}_c = 50$m, $N = 20$, $\gamma = 1.5$, $\alpha = 3$, $\lambda_c = \frac{0.01}{\pi \mathcal{R}_c^2}$, $M_s = 3$, $\delta = 1.1$. $R_0 = \frac{1}{8}$ BPCU, $R_1 = \frac{3}{4}$ BPCU, $R_2 = \frac{7}{8}$ BPCU,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

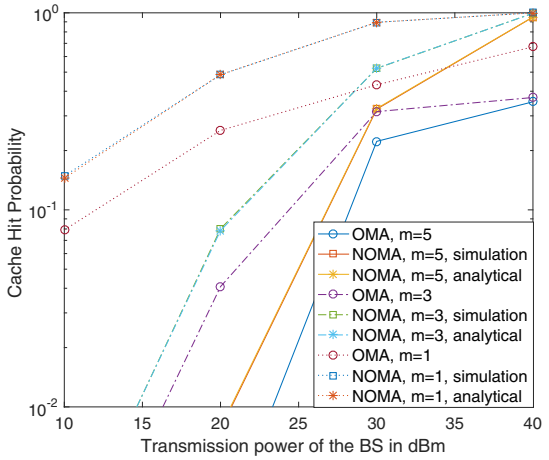

Fig. 4. The cache hit probability for the proposed push-and-deliver strategy for Case 2.

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. 3. However, for a different set of popularity parameters, i.e., Case 2, the performance gap between OMA and NOMA is significantly increased. The reason behind this phenomenon is as follows. Recall that the use of NOMA can significantly improve the reception reliability of the files which are decoded at the later stages of the SIC procedure, whereas the improvement for the files which 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 for decoding f_1 are quite different, which is the reason for the larger performance gap in Case 2.

VI. Conclusions

In this paper, the application of the NOMA principle to wireless caching has been studied. Two NOMA assisted caching strategies have been developed. The presented numerical and analytical results demonstrate that the proposed NOMA assisted caching schemes can efficiently improve the cache hit probability and reduce the delivery outage probability, compared to the conventional OMA based caching strategies.

References

[1] 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, 2017.

[2] "5G radio access: Requirements, concepts and technologies," NTT DOCOMO, Inc., Tokyo, Japan, 5G Whitepaper, Jul. 2014.

[3] "5G innovation opportunities- a discussion paper," techUK, London, 5G Whitepaper, Aug. 2015.

[4] Z. Ding, P. Fan, and H. V. Poor, "Random beamforming in millimeter-wave NOMA networks," *IEEE Access*, (to appear in 2017).

[5] 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.

[6] 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.

[7] 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, 2017.

[8] 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. PP, no. 99, pp. 1–1, 2017.

[9] 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.

[10] 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.

[11] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans. Wireless Commu.*, vol. 15, no. 9, pp. 6118–6131, Sept. 2016.

[12] 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, Sept. 2016.

[13] 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.

[14] Z. Ding, P. Fan, G. K. Karagiannidis, R. Schober, and H. V. Poor, "NOMA assisted wireless caching: Strategies and performance analysis," *IEEE Trans. Commun.*, (submitted) Available on-line at arXiv:1709.06951.

[15] Z. Zhao, M. Xu, Y. Li, and M. Peng, "A non-orthogonal multiple access (NOMA)-based multicast scheme in wireless content caching networks," *IEEE J. Sel. Areas Commun.*, vol. PP, no. 99, pp. 1–1, 2017.

[16] M. Haenggi, *Stochastic Geometry for Wireless Networks*. Cambridge University Press, Cambridge, UK, 2012.

[17] 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.

[18] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inform. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

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

[20] F. Baccelli and A. Giovanidis, "A stochastic geometry framework for analyzing pairwise-cooperative cellular networks," *IEEE Trans. Wireless Commu.*, vol. 14, no. 2, pp. 794–808, Feb. 2015.

[21] 3rd Generation Partnership Project (3GPP), "Study on downlink multiuser superposition transmission for LTE," Mar. 2015.

[22] 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. on Commun.*, Paris, France, May 2017, pp. 1–7.