This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCOMM.2018.2830376, IEEE Communications Letters

Coverage Performance of NOMA in Wireless Caching Networks

Zhongyuan Zhao, *Member, IEEE*, Mingfeng Xu, Weiliang Xie, Zhiguo Ding, *Senior Member, IEEE*, and George K. Karagiannidis, *Fellow, IEEE*

***Abstract*—In order to keep a balance between the transmission delay of backhaul and the spectrum efficiency of access links, the coverage performance of wireless networks with non-orthogonal multiple access (NOMA) and content caching is studied. First, an explicit expression for the coverage probability of a typical user is presented by using stochastic geometry and order statistics. This expression can provide useful insights in order to improve the coverage performance of NOMA communications. Second, a closed-form expression for the average coverage probability is derived. Finally, simulation results are provided to validate the accuracy of the analytical framework and demonstrate the performance gain, due to NOMA.**

***Index Terms*—Non-orthogonal multiple access (NOMA), content caching, coverage probability, stochastic geometry**

I. Introduction

Non-orthogonal multiple access (NOMA) is considered as a very promising technique to significantly improve the spectrum efficiency of fifth generation (5G) and beyond communication systems [1]. Since multiple users can be served simultaneously, NOMA can achieve significant performance gains, in terms of system throughput as well as connectivity. Furthermore, the outage performance of NOMA with cooperation and user pairing has been studied in [2], which show that NOMA can achieve higher transmission data rate than the conventional orthogonal multiple access (OMA) schemes. In [3] and [4], NOMA has been used in large-scale heterogeneous networks and multicast cognitive radio networks, respectively.

Although NOMA can improve the delivery efficiency of the BSs, it causes long waiting delay, since the NOMA transmissions cannot begin until all the requested messages are sent to the BSs via the backhaul. Therefore, to mitigate the increased loading of the backhaul, the joint design of NOMA and content caching has been investigated in [5] and [6]. However, due to the complicated interference effect and the non-uniform coverage, it is difficult to assess if NOMA can reduce the total transmission delay. In order to find a sophisticated balance between the transmission delay of backhaul and the spectrum efficiency of access links, we study the coverage performance of NOMA in wireless caching networks.

The main contributions of this paper can be summarized as follow: *First*, by employing stochastic geometry and order statistics, an explicit expression for the coverage probability of a typical user is derived, which can provide useful insights on how to improve the coverage performance. *Second*, a closed-form expression for the average coverage probability is presented, which shows that the transmission performance of NOMA can be guaranteed, when the data rate of backhaul is high enough. *Finally*, simulation results are provided to verify the accuracy of the analysis and to show the performance gain of NOMA in wireless caching networks.

II. System Model

Consider a downlink transmission scenario in wireless caching networks, where each BS serves multiple users. The locations of BSs are modeled as a homogeneous Poisson point process (PPP) Φ, with a given density λ. In order to characterize the fact that the users are more likely to lie close to their associated BSs, the locations of users are modeled as a Matérn cluster process Ψ. In particular, we focus on a representative cell C_i, where its coverage area can be modeled as a disc $D(B_i, r)$, where r is its radius, and the location of the serving BS B_i acts as its center, $B_i \in \Phi$. All the associated users of C_i form a point cluster ψ_i of Ψ, which are independently and uniformly distributed in $D(B_i, r)$. Moreover, the number of users belonging in ψ_i is denoted as M_i, which follows Poisson distribution with a given expectation μ, i.e., $M_i \sim \mathrm{Poi}(\mu)$. Without loss of generality, U_j is selected as a typical user, $U_j \in \psi_i$. Denoting the distance between U_j and B_i as d_j, the cumulative distribution function (CDF) and the probability density function (PDF) of d_j can be expressed as:

$$F(d_j) = \frac{d_j^2}{r^2},\ f(d_j) = \frac{2d_j}{r^2},\ d_j \in [0, r],\ U_j \in \psi_i. \quad (1)$$

In this paper, we assume that each user belonging to Ψ, requests a different content object simultaneously, while NOMA transmissions are employed to improve the spectrum efficiency. Without loss of generality, the representative cell C_i is taken as an example. All the content objects required by the users of ψ_i can be denoted as a set $\Omega_i = \{s_1, \dots, s_{M_i}\}$, where s_j is the content object required by U_j. The NOMA scheme under investigation consists of the following two phases:

Z. Zhao and M. Xu are with the Key Laboratory of Universal Wireless Communications (Ministry of Education), Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: zyzhao@bupt.edu.cn; x-u_mingf@bupt.edu.cn). W. Xie is with China Telecom Technology Innovation Center, Beijing 102209, China (e-mail: xiewl@ctbri.com.cn). Z. Ding is with the School of Electrical and Electronic Engineering, the University of Manchester Manchester, M13 9PL, UK (e-mail: zhiguo.ding@manchester.ac.uk). G. K. Karagiannidis is with the Electrical and Computer Engineering Department, Aristotle University of Thessaloniki, Thessaloniki GR-54124, Greece (email: geokarag@auth.gr).

The work of Z. Zhao and M. Xu was supported in part by the National Natural Science Foundation of China (Grant No. 61501045), the State Major Science and Technology Special Projects (Grant No. 2017ZX03001014). The work of Z. Ding was supported by the UK EPSRC under grant number EP/L025272/1 and by H2020-MSCA-RISE-2015 under grant number 690750.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCOMM.2018.2830376, IEEE Communications Letters

1) Backhaul Transmission Phase: In order to support NOMA transmissions, B_i should obtain all the content objects of Ω_i via the orthogonal backhaul link use[1]. Due to the employment of content cache Π_i at B_i, the burden of the backhaul link can be mitigated since some content objects of Ω_i can be stored locally, and the random caching strategy is employed in this paper. Assuming that all the content objects are of the same size L, the transmission delay of the backhaul link is decided by the number of requested content objects that are not cached at the BS, which can be expressed as:

$$D_{\mathrm{BH}} = \frac{(M_i - N_i)L}{r_{\mathrm{BH}}} = \frac{(1-P_{\mathrm{hit}})M_i L}{r_{\mathrm{BH}}}, \tag{2}$$

where r_{BH} denotes the data rate of the backhaul link, N_i is the number of local cached content objects, M_i is the total number of requested content objects, and $P_{\mathrm{hit}} = N_i/M_i$ is the hit ratio of Π_i. As shown in (2), the transmission delay of the backhaul link is identical for each the requested content object.

2) NOMA Transmission Phase: By employing NOMA, B_i can serve all the users of ψ_i simultaneously, and its transmitted message can be expressed as $t_i = \sum_{m=1}^{M_i} \sqrt{\rho_m} s_m$, where ρ_m denotes the transmit power of s_m, $\sum_{m=1}^{M_i} \rho_m = \rho$, and ρ is the total transmit power at each BS. Then, the observation of U_j can be expressed as:

$$\begin{aligned} y_j \;\; = \;\; & d_j^{-\alpha/2} h_j \sqrt{\rho_j} s_j + \sum_{k \neq j, s_k \in \Omega_i} d_j^{-\alpha/2} h_j \sqrt{\rho_k} s_k \\ & + \sum_{B_n \in \Phi / B_i} l_n^{-\alpha/2} g_n t_n + w_j, \end{aligned} \tag{3}$$

where h_j denotes the Rayleigh fading coefficient of the link between U_j and B_i, g_n is defined similarly for the interfere BS B_n, $h_j,\ g_n \sim \mathcal{CN}(0,1)$, d_j and l_n denote the distance between U_j and its associated BS B_i and the interfere BS B_n, respectively, $n \neq i$, w_j is the additive white Gaussian noise at U_j, and α is the pathloss exponent.

To mitigate the impact of intra-cell interference, SIC is implemented at each user. Without loss of generality, we assume that the user index j determines the detection order of message s_j, which means that $s_1, \ldots, s_{j-1}$ should be removed from the observation at U_j before detecting s_j. As introduced in [1], a conventional method is to establish the detection order by using instantaneous channel state information (CSI). However, it causes a large amount of signaling, which leads to a decrease of spectrum efficiency. In [7], the pathloss, which can sketch the key feature of channel gains efficiently, is employed instead to keep a balance between the signaling overhead and coverage performance of NOMA scheme. Therefore, the pathloss of each user associated with B_i is in an ascending order, i.e., $d_1^{-\alpha} \leqslant \cdots \leqslant d_{M_i}^{-\alpha}$.

In this paper, we focus on the interference limited scenario, and the received signal to interference ratio (SIR) of s_m at U_j can be expressed as:

$$\gamma_{j,m} = \frac{\rho_m d_j^{-\alpha} |h_j|^2}{\sum_{k=m+1}^{M_i} \rho_k d_j^{-\alpha} |h_j|^2 + \sum_{B_n \in \Phi/B_i} \rho l_n^{-\alpha} |g_n|^2}. \tag{4}$$

[1]Please note that the content objects may be stored at different content center, and thus there is no chance to use NOMA in this phase.

The transmission delay of s_m from B_i to U_j can be written as:

$$D_{j,m} = \frac{L}{r_{j,m}} = \frac{L}{b \log(1+\gamma_{j,m})}, \tag{5}$$

where b denotes the bandwidth of the access link. Please note that the rates of both the source coding and the channel coding can be adjusted dynamically due to the instantaneous channel conditions, such as the scalable video coding (SVC) and the adaptive modulation and coding (AMC) mechanisms. The achievable data rate given by (5) can capture the quality of channel conditions, and thus $D_{j,m}$ can be employed to characterize the transmission delay.

III. Coverage Performance of NOMA Transmissions

In this section, the coverage performance of NOMA is studied. In particular, the coverage probability of a typical user U_j is defined as the probability that the total transmission delay is less than a given threshold, which means that the content requirement of U_j can be responded by the serving BS B_i successfully. Due to the implementation of SIC, it is necessary to ensure that $s_1, \ldots, s_{j-1}$ can be detected successfully at U_j, which means that their received SIRs should be higher than the given thresholds. Therefore, the coverage probability of U_j can be written as:

$$P_j = \Pr\left\{D_{\mathrm{BH}} + D_{j,1} \leqslant T_1, \ldots, D_{\mathrm{BH}} + D_{j,j} \leqslant T_j\right\}, \tag{6}$$

where T_m denotes a given delay threshold of s_m.

A. Coverage Performance of A Typical User U_j

1) Coverage Probability of U_j: Recalling (2), D_{BH} can be treated as a constant when the data rate r_{BH} and the hit ratio P_{hit} are fixed. Therefore, the coverage probability of U_j is mainly decided by the delay of NOMA transmissions. Substituting (4) and (5) into (6), P_j can be rewritten as:

$$P_j = \Pr\left\{\gamma_{j,1} \geqslant 2^{\frac{L}{b(T_1 - D_{\mathrm{BH}})}} - 1, \ldots, \gamma_{j,j} \geqslant 2^{\frac{L}{b(T_j - D_{\mathrm{BH}})}} - 1\right\}. \tag{7}$$

Next, a tractable expression of P_j conditioned on d_j can be obtained by using the stochastic geometry-based system model.

Theorem *1:* When the transmit power satisfies the constraint $q_m = \rho_m - \mu_m(\sum_{k=m+1}^{M_i} \rho_k) > 0$, $m = 1, \ldots, j$, then the average coverage probability of U_j can be expressed as:

$$P_j = {}_1F_1\left[M_i - j + 1; M_i + 1; -2\pi\lambda r^2 A(\alpha) \theta_j^{2/\alpha}\right], \tag{8}$$

where $\mu_m = 2^{\frac{L}{b(T_m - D_{\mathrm{BH}})}} - 1$, $\theta_j = \max\{\eta_1, \ldots, \eta_j\}$, η_m is defined as $\eta_m = \mu_m \rho / q_m$ to simplify the expression of P_j, $A(\alpha) = \frac{\pi}{\alpha}\csc(\frac{2\pi}{\alpha})$, and ${}_1F_1(a;b;z)$ is the confluent hypergeometric function.

Proof: Substituting (4) into (7), the coverage probability of U_j can be derived as:

$$P_j = \Pr\left\{|h_j|^2 \geqslant \theta_j d_j^{\alpha} \left(\sum_{B_n \in \Phi/B_i} l_n^{-\alpha} |g_n|^2\right)\right\}. \tag{9}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCOMM.2018.2830376, IEEE Communications Letters

Note that $|h_j|^2$ and $|g_n|^2$ follow independent identically distributed exponential distributions, and thus (9) can be expressed as:

$$P_j = \mathbb{E}_{d_j}\Big\{\underbrace{\mathbb{E}_{\Phi}\Big\{\prod_{B_n\in\Phi/B_i}\frac{1}{1+\theta_j d_j^{\alpha} l_n^{-\alpha}}\Big\}}_{\mathcal{L}}\Big\}. \tag{10}$$

Since Φ is stationary, the location of U_j can be chosen as the origin. Based on the probability generating functional (PGFL) of PPP, $\mathcal{L}$ in (10) can be written as:

$$\begin{aligned}\mathcal{L} &= \exp\left[-2\pi\lambda\int_0^{\infty}\left(1-\frac{1}{1+\theta_j d_j^{\alpha} l_n^{-\alpha}}\right) l_n \mathrm{d}l_n\right] \\ &= \exp\left[-2\pi A(\alpha)\lambda\theta_j^{2/\alpha} d_j^2\right].\end{aligned} \tag{11}$$

$d_1,\ldots,d_{M_i}$ are a group of order statistics and follow identical distributions. Note that d_j is the (M_i+1-j)-th order statistic, and its PDF can be expressed as [10]:

$$f_{d_j}(x) = M_i\binom{M_i-1}{M_i-j}\left(\frac{x^2}{r^2}\right)^{M_i-j}\left(1-\frac{x^2}{r^2}\right)^{j-1}\frac{2x}{r^2}. \tag{12}$$

Substituting (12) into (10), P_j can be written as (8), and Theorem 1 has been proved. ■

2) Further Discussion of Theorem 1: As shown in (8), P_j is determined by θ_j. The first-order derivative of P_j with respect to θ_j can be derived as:

$$\begin{aligned}\nabla_{\theta_j} P_j = &\frac{-4\pi\lambda r^2 A(\alpha)\theta_j^{2/\alpha-1}(M_i-j+1)}{\alpha(M_i+1)}\times \\ &{}_1F_1\left[M_i-j+2; M_i+2; -2\pi\lambda r^2 A(\alpha)\theta_j^{2/\alpha}\right] \leqslant 0.\end{aligned} \tag{13}$$

Based on (13), we can obtain the following remark, which is related to the monotonicity of P_j.

Remark *2:* The coverage probability P_j is a decreasing function with respect to θ_j.

Moreover, Theorem 1 shows that the P_j is related to the detection order of U_j. Furthermore, the coverage performance of C_i is mainly decided by the coverage probability of U_{M_i}. The reasons can be explained as follows: (i) As pointed out in [8], to ensure the performance gain of NOMA, the coverage probability of U_{M_i} should be guaranteed. (ii) Based on Remark 2 and (8), when the value of θ_{M_i} is minimized, $\theta_1,\ldots,\theta_{M_i-1}$ also decrease since they are defined based on a subset of $\{\eta_1,\ldots,\eta_{M_i}\}$. Therefore, optimizing θ_{M_i} can improve the coverage performance of all users of C_i, and thus fairness can be ensured. Then, the following corollary about the coverage probability of U_{M_i} can be provided.

Corollary *3:* In order to maximize the coverage probability of U_{M_i}, the following constraint should be satisfied:

$$\eta_1 = \cdots = \eta_{M_i} = \theta_{\min}, \tag{14}$$

where η_m follows the notation given by Theorem 1. When (14) can be satisfied, the power allocated to U_j can be expressed as

$$\rho_j = \frac{\prod_{l=j}^{M_i} a_l}{\sum_{m=1}^{M_i}\left(\prod_{n=m}^{M_i} a_n\right)}\rho,\ j=1,\ldots,M_i, \tag{15}$$

where $a_n = \mu_n(1+\frac{1}{\mu_{n+1}}), n=1,\ldots,M_i-1$ and $a_{M_i}=\mu_{M_i}$. $\theta_{\min}$ can be expressed as $\theta_{\min} = \sum_{m=1}^{M_i}\left(\prod_{n=m}^{M_i} a_n\right)$.

Proof: Based on Remark 2, it is equivalent to prove that θ_{M_i} can be minimized when (14) is satisfied, and a proof by contradiction is provided. In particular, we assume that $\mathbf{p}^* = [\rho_1^*,\ldots,\rho_{M_i}^*]$ is the optimal power allocation strategy to minimize θ_{M_i}, $\sum_{j=1}^{M_i}\rho_j^* = \rho$, which should satisfy the following three constraints:

C1) To grantee the reliability of NOMA transmissions, $\mathbf{p}^*$ should follow the constraint $q_m = \rho_m - \mu_m(\sum_{k=m+1}^{M_i}\rho_k) > 0$ given by Theorem 1, $m=1,\ldots,M_i$.

C2) (14) should not be satisfied when $\mathbf{p}^*$ is employed.

C3) $\mathbf{p}^*$ can minimize the value of θ_{M_i}.

Due to *C2*, we assume that $\eta_j(\mathbf{p}^*)$ is strictly larger than $\eta_1(\mathbf{p}^*),\ldots,\eta_{j-1}(\mathbf{p}^*),\eta_{j+1}(\mathbf{p}^*),\ldots,\eta_{M_i}(\mathbf{p}^*)$, and thus $\theta_{M_i}(\mathbf{p}^*) = \eta_j(\mathbf{p}^*)$.

By adjusting power allocation to U_j and U_{j+1}, another power allocation strategy can be obtained, which can be denoted as $\mathbf{p}^\dagger = [\rho_1^*,\ldots,\rho_j^*+\Delta\rho,\rho_{j+1}^*-\Delta\rho,\ldots,\rho_{M_i}^*]$, and the following constraint of $\Delta\rho$ should be satisfied:

$$\Delta\rho < \min\left\{(\mu_{j+1}+1)\rho_{j+1}^* - \frac{\mu_{j+1}}{\mu_j}\rho_j^*, \rho_{j+1}^* - \mu_{j+1}\sum_{k=j+2}^{M_i}\rho_k^*\right\}. \tag{16}$$

Due to (16), $\mathbf{p}^\dagger$ follows the constraint given by *C1*, and the relationship between $\eta_k(\mathbf{p}^*)$ and $\eta_k(\mathbf{p}^\dagger)$ can be expressed as:

$$\begin{aligned}\eta_j(\mathbf{p}^\dagger),\ \eta_{j+1}(\mathbf{p}^\dagger) < \eta_j(\mathbf{p}^*),\ \text{and}\ \eta_m(\mathbf{p}^\dagger) = \eta_m(\mathbf{p}^*), \\ m\neq j, j+1.\end{aligned} \tag{17}$$

Note that (17) shows that $\theta_{M_i}(\mathbf{p}^*) > \theta_{M_i}(\mathbf{p}^\dagger)$, which contradicts with the assumption that $\theta_{M_i}(\mathbf{p}^*)$ is the minimum value of θ_{M_i}. Moreover, if there exist multiple $\eta_j(\mathbf{p}^*)$-s that can achieve $\theta_{M_i}(\mathbf{p}^*)$, a similar $\mathbf{p}^\dagger$ can be obtained by following (16). Therefore, (14) must be satisfied to maximize P_{M_i}.

The power allocation strategy given by (15) can be obtained directly based on (14) and $\sum_{j=1}^{M_i}\rho_j^* = \rho$, and a closed-form expression for $\theta_{\min}$ can be derived based on Lemma 1 and Appendix B in [9]. Then the proof is completed. ■

B. Coverage Performance of A Cell C_i

The average coverage probability of C_i *is* defined as $\bar{P} = \mathbb{E}_{\psi_i}\{\frac{1}{M_i}\sum_{U_j\in\psi_i} P_j\}$, where P_j follows the definition given by (6). Based on Corollary 3, a closed-form expression of $\bar{P}$ can be obtained, according to the following theorem.

Theorem *4:* When Corollary 3 is satisfied, the average coverage probability of C_i can be expressed as:

$$\bar{P} = \frac{1-\exp\left[-2\pi A(\alpha)\lambda r^2\theta_{\min}^{2/\alpha}\right]}{2\pi A(\alpha)\lambda r^2\theta_{\min}^{2/\alpha}}, \tag{18}$$

where $\theta_{\min}$ is defined in (14).

Proof: When $\theta_j = \theta_{\min}$, $\bar{P}$ can be derived as follows based on (11) and (12):

$$\bar{P} = \mathbb{E}_{M_i}\Big\{\int_0^r \underbrace{\left(\sum_{j=1}^{M_i}\frac{1}{M_i}f_{d_j}(x)\right)}_{\mathcal{J}} e^{-2\pi A(\alpha)\lambda\theta_{\min}^{2/\alpha}x^2}\mathrm{d}x\Big\}. \tag{19}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCOMM.2018.2830376, IEEE Communications Letters

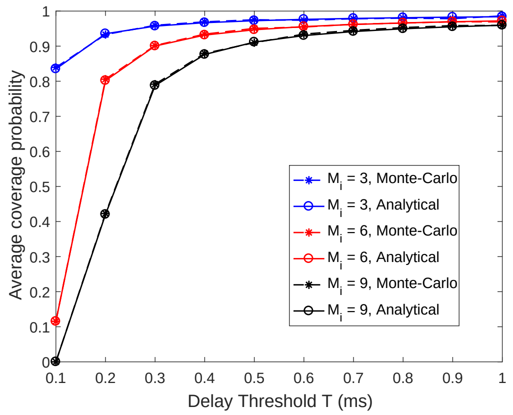

Fig. 1. The average coverage probability of NOMA.

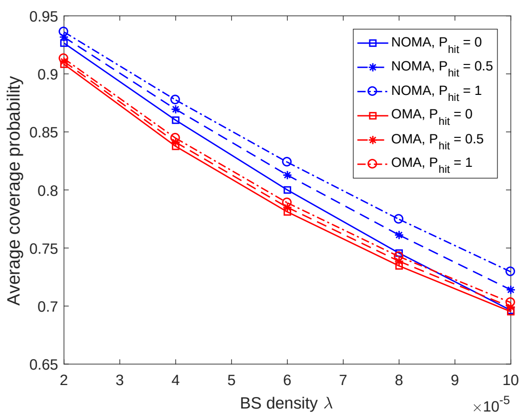

Fig. 2. Comparison between NOMA and OMA schemes.

Moreover, $\mathcal{J}$ in (19) can be expressed as:

$$\mathcal{J} = \sum_{j=1}^{M_i} \binom{M_i - 1}{M_i - j} \left(\frac{x^2}{r^2}\right)^{M_i - j} \left(1 - \frac{x^2}{r^2}\right)^{j-1} \frac{2x}{r^2} = \frac{2x}{r^2}. \tag{20}$$

Substituting (20) into (19), $\bar{P}$ can be written as (18), and Theorem 4 has been proved. ■

As shown in (18), $\bar{P}$ is a decreasing function with respect to $\theta_{\min}$, which is related to the number of users participating NOMA transmissions and the threshold of total transmission delay of each content object. Moreover, $1 - \exp[-2\pi A(\alpha)\lambda r^2 \theta_{\min}^{2/\alpha}] \sim 2\pi A(\alpha)\lambda r^2 \theta_{\min}^{2/\alpha}$ as $\theta_{\min}$ approaches 0, and thus the average coverage probability of C_i approaches 1.

IV. Simulation Results and Discussions

The simulation results are provided in this section. The coverage radius of each BS is set as $r = 15$ m, the bandwidth is $B = 10$ MHz, the path loss exponent is $\alpha = 3$. The size of each content object is set as $L = 1$ kB, and the data rate of backhaul is $r_{\rm BH} = 100$ Mbps. The number of local cached content objects is 2, and the BS density is $\lambda = 5 \times 10^{-5}$.

In Fig. 1, the average coverage probability given by Theorem 4 is plotted. As shown in the figure, the average coverage probability increases as T increases. Moreover, the analytical results match the Monte Carlo results perfectly, which verifies the accuracy of Theorem 4. In Fig. 2, a comparison between NOMA and OMA schemes is provided. The spectrum efficiency of OMA scheme is set as $\tau = 1/M_i$ to ensure the fairness of the comparison, and other settings are the same as NOMA scheme. Since there exists a tradeoff of the transmission delay between the access and backhaul links in the NOMA scheme, the implementation of caching can reduce the transmission delay of backhaul links, and ensures that NOMA can achieve the performance gain. As shown in Fig. 2, the coverage performance of NOMA is always better than that of OMA when $P_{\rm hit} = 1$, and the performance gain of NOMA is reduced as $P_{\rm hit}$ decreases.

V. Conclusions

In this paper, the coverage performance of NOMA in wireless caching networks has been studied. First, the coverage probability of a typical user has been analyzed by employing stochastic geometry and order statistics, which can provide useful insights on how to improve the coverage performance of NOMA. Second, a closed-form expression of the average coverage probability has been provided. Finally, simulation results are provided to verify the accuracy of analytical results and show the performance gain of NOMA in wireless caching networks. In our future research, sophisticated user scheduling and pairing mechanisms should be designed to improve the performance gains and the flexibility of NOMA transmissions.

References

[1] L. Dai, B. Wang, Y. Yuan, S. Han, C.-L. I, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74-81, Sep. 2015.

[2] Y. Liu, M. Derakhshani, and S. Lambotharan, "Outage analysis and power allocation in uplink non-orthogonal multiple access systems, " to appear in *IEEE Commun. Lett.*, [Online]. Available: http://ieeexplore.ieee.org/document/7842433/.

[3] Y. Liu, Z. Qin, M. Elkashlan, A. Nallanathan and J. A. McCann, "Non-orthogonal multiple access in large-scale heterogeneous networks, " *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2667-2680, Dec 2017.

[4] L. Lv, J. Chen, Q. Ni, and Z. Ding, "Design of cooperative non-orthogonal multicast cognitive multiple access for 5G systems: User scheduling and performance analysis, " *IEEE Trans. Commun.*, vol. 65, no. 6, pp. 2641-2656, June 2017.

[5] 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, July 2017.

[6] Z. Ding, P. Fan, G. K. Karagiannidis, R. Schober, and H. V. Poor "NOMA assisted wireless caching: Strategies and performance analysis," [Online]. Available: https://arxiv.org/abs/1709.06951.

[7] Z. Yang, Z. Ding, P. Fan, and G. K. Karagiannidis, "On the performance of non-orthogonal multiple access systems with partial channel information, " *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 654-667, Dec. 2015.

[8] 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, Sep. 2015.

[9] L. Yang, J. Chen, Q. Ni, J. Shi, and X. Xue, "NOMA-enabled cooperative unicast-multicast: Design and outage analysis, " *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 7870-7889, Sep. 2017.

[10] H. A. David and H. N. Nagaraja, *Order Statistics*, 3rd ed. New Jersey, USA: John Wiley. 2003.