IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 20, 2018, accepted February 14, 2018, date of publication March 5, 2018, date of current version April 4, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2812144

Task Offloading in Heterogeneous Mobile Cloud Computing: Modeling, Analysis, and Cloudlet Deployment

HYUN-SUK LEE AND JANG-WON LEE, (Senior Member, IEEE)
Department of Electrical and Electronic Engineering, Yonsei University, Seoul 03722, South Korea

Corresponding author: Jang-Won Lee (jangwon@yonsei.ac.kr)

This work was supported in part by the Midcareer Researcher Program through NRF grant from MSIT, South Korea, under Grant NRF-2017R1A2 B4006908, in part by the Institute for Information and communications Technology Promotion grant through by the Korea Government under Grant B0717-16-0024, and in part by the development on the core technologies of transmission, modulation and coding with low-power and low complexity for massive connectivity in the IoT environment. The work of H.-S. Lee was supported by the Graduate School of YONSEI University Research Scholarship Grants in 2017.

ABSTRACT In this paper, we consider a heterogeneous mobile cloud computing (HMCC) system that consists of remote cloud servers, local cloudlets, task offloading mobile devices (TMDs), non-task offloading MDs (NTMDs), and radio access networks such as cellular networks and WLANs. TMDs have the capability of task offloading to remote cloud servers or cloudlets, whereas NTMDs are conventional cellular users that do not have such capability. By using stochastic geometry, we analyze the outage probability of task offloading in the MCC system with only remote cloud servers and that in the HMCC with both remote cloud servers and cloudlets. The analysis provides useful information, i.e., how the varying system parameters affect the outage probability. From the analysis, we show that there is an intrinsic limitation in reducing the outage probability in the MCC system due to the outage when accessing remote cloud servers. In addition, we show that the use of cloudlets is a promising solution to overcome this limitation. However, a tradeoff exists in using cloudlets due to their deployment and operation costs. Thus, to address this tradeoff, we also study the optimal cloudlet deployment to maximize the cloud service provider's profit while guaranteeing maximum outage probability requirements.

INDEX TERMS Mobile cloud computing, cloudlet, mobile task offloading, stochastic geometry, cloudlet deployment, profit maximization.

I. INTRODUCTION

Recently, the number of mobile devices (MDs) in use has grown significantly, and various types of mobile applications have been emerging [1]. However, some mobile applications, such as natural language processing and image processing, cannot be readily executed on MDs since they require a large amount of computing resources and a large amount of energy, while MDs have limited computing resources and limited battery capacity. To resolve this problem, mobile cloud computing (MCC) is introduced that allows a task for such applications to be offloaded to cloud servers, i.e., *task offloading* [2]. In an MCC system, remote cloud servers having a considerably larger amounts of computing resources than MDs are installed in a data center, and MDs can use the mobile applications that need a large amount of computing resources by using the computing resources of the remote cloud servers. Since the data center is located in a remote place, the MDs connect to the remote cloud servers through radio access networks, such as 3G, LTE, and WiMAX [3]. Hence, due to delays and large transmission power consumption when using the radio access networks, an outage of task offloading to the remote cloud servers can occur [4]. To mitigate such an outage problem, recently, in addition to remote cloud servers, cloud servers installed in a local place, such as a cafe or a library, are considered as a part of the MCC system [5]. These cloud servers in local places are called *cloudlets*, and MDs are allowed to connect to them through built-in WLAN access points (APs) within cloudlets. In general, cloudlets have larger amounts of computing resources than MDs, but have much smaller amounts of computing resources than remote cloud servers. An MCC system that consists of both remote cloud servers and cloudlets is called a heterogeneous mobile cloud computing (HMCC) system [6], [7].

In this paper, we consider an HMCC system, where an MD offloads its task to either the computing resources of remote cloud servers through cellular networks or that of cloudlets through their WLAN APs. We analyze the outage probability of task offloading in the HMCC by using stochastic geometry. We first model the MCC system with cellular networks that has only remote cloud servers, and analyze its outage probability of task offloading to the remote cloud servers. Then, we also model the MCC system that has only cloudlets, and analyze its outage probability of task offloading to the cloudlets. By using both outage probabilities of task offloading, we derive the outage probability of task offloading in the HMCC system that consists of both remote cloud servers and cloudlets. We then study the optimal cloudlet deployment to maximize the cloud service provider's (CSP's) profit by considering the deployment cost of local cloudlets, the operation costs of remote cloud servers and local cloudlets, and the revenue from the task offloading of the MDs.

A. RELATED WORKS AND MOTIVATION

Task offloading is one of the most important research topics related to both MCC and HMCC systems [2], [3], [8]. When the tasks of MDs require a large amount of computing resources, task offloading can help the MDs to reduce their energy consumption due to their tasks [9]. However, task offloading is not always helpful to reduce the energy consumption since the energy consumption from communications might be large [9]. In addition, task offloading is not helpful when it violates the delay requirement of the task [10]. From these backgrounds, most of the existing studies on task offloading focus on the strategy of an individual MD for its task offloading under a given system [11]–[18]. They model the utility of an MD by considering its benefit and cost from task offloading. Besides, they develop algorithms for the MD to decide its task offloading such that its utility is maximized without the outages of task offloading. In other words, they address task offloading from the viewpoint of an individual MD. On the other hand, some other studies consider task offloading from the viewpoint of CSPs, such as the operation and design of the systems. Specifically, in [18] and [19], pricing algorithms that can be used to maximize the profit of the CSP are proposed. They determine the price of cloud service taking into account the network conditions. In [20] and [21], the algorithms for designing a HMCC system with low costs are provided. They determine where to install cloudlets among the available candidate sites by considering the installation costs and the average access delay to offload tasks.

The algorithms for designing the HMCC system in [20] and [21] can be used only when the design parameters, such as the number of cloudlets and the amount of computing resources of cloudlets, are given. In other words, before designing the HMCC system by using the algorithms, a planning process for the HMCC system to determine such design parameters is required. Hence, an analysis that allows us to easily obtain useful information for planning the system is also needed and valuable when designing a system. For example, many researches have been conducted on the analyses for cellular networks such as coverage analyses of downlink (DL) [22], [23] and uplink (UL) [24], [25]. They model and analyze cellular networks by using stochastic geometry. In the analyses, the characteristics of the cellular networks are derived in simple expressions. From the analyses, we can easily obtain the useful characteristics in various network settings, which can be used to plan and design cellular networks [26]. As in the case of cellular networks, the analysis on HMCC systems with regard to various design parameters, such as the intensity of APs, intensity of cloudlets, power consumption of task offloading, and required data rate for task offloading, can provide useful information for planning and designing the HMCC system. Moreover, in the analysis on HMCC systems, various network settings, such as the intensity of BSs, number of channels, and transmission power, can also provide useful information and should be also considered since radio access networks, such as cellular networks and WLAN APs, are used to access the cloud computing resources in HMCC systems. However, to the best of our knowledge, there is no such analytic research on HMCC systems yet.

In particular, in this paper, we focus on the outage probability of task offloading that is the one of the most important performance metric of task offloading in HMCC systems. Moreover, as mentioned above, mitigating the outage of task offloading in MCC systems without cloudlets is the one of the major reasons of deploying the cloudlets in such systems. By using the analysis of the outage probability, the performance of task offloading and the effectiveness of deploying cloudlets with regard to given design parameters and network settings can be easily obtained. Hence, in these aspects, such an analysis involving various design parameters and network settings is important for planning and designing HMCC systems. In addition, the analysis provides a guideline regarding how many tasks could be successfully offload, which is highly relevant to the profit of the CSP. Thus, from an economic view, the analysis allows the CSP to check the marketability of its HMCC system and to find the optimal number of cloudlets to be deployed in its HMCC system.

B. OUR CONTRIBUTIONS

The main contributions of this paper are summarized as follows:

- We provide the analysis for the outage probability of task offloading in the HMCC system, and this has never been tried before, to the best of our knowledge. By using stochastic geometry, we derive the analysis in a simple expression. This allows us to obtain useful information regarding the impacts of the various characteristics of the

HMCC system on the outage probability that are hard to predict in general, with low complexity. To address the outage, in the analysis, the characteristics of radio access networks including UL and DL transmissions such as their energy consumption, scheduling, and data rates are incorporated. Moreover, we also consider the correlation between using remote cloud servers and cloudlets.

- In this work, we specifically address the following system dynamics related to the UL transmission that are not considered in the analysis on the UL transmission in conventional cellular networks based on stochastic geometry: two types of MDs having different statistical properties on the UL transmission power, UL scheduling with a given number of UL channels, and idle UL channels that represent the UL channels not scheduled to any MD.
- The analysis leads to insights for the purpose of using local cloudlets. In the MCC system with only remote cloud servers, an intrinsic limitation exists in reducing the outage probability of task offloading to remote cloud servers due to the outages that occur when MDs try to connect to the remote cloud servers through the cellular network. Since such outages are out of the control of CSPs, the CSPs can deploy cloudlets to reduce the outage probability of task offloading in addition to resolving the drawbacks of cloud servers, i.e., delay and large transmission power consumption.
- As the application of the analysis, we consider the economic tradeoff in using local cloudlets by studying the optimal cloudlet deployment that maximizes the profit of a CSP while considering the maximum outage probability requirement. From it, we show that cloudlets can be used in order that CSPs obtain more revenue.

C. PAPER STRUCTURE

The rest of the paper is organized as follows. Section II presents the HMCC system model. In Section III, the outage probability of task offloading in the MCC system having only remote cloud servers is analyzed. In Section IV, we first analyze the outage probability of task offloading in the MCC system having only local cloudlets. By using both outage probabilities, we then extend the outage probabilities to that in an HMCC system, and study the optimal cloudlet deployment maximizing the CSP's profit. We provide results and discussions in Section V and finally conclude this paper in Section VI.

II. SYSTEM MODEL

We consider an HMCC system consisting of remote cloud servers, local cloudlets, and MDs.[1] Cloud servers are located in a data center with large amounts of computing resources that MDs can use at any time they want. On the other hand, cloudlets are located in the places near MDs, such as a cafe or a library, where only a limited number of servers can be located. Hence, their computing resources are limited and finite. In addition, we also consider a cellular network and the WLAN APs, which are used to access the computing resources of cloud servers and cloudlets, respectively. The WLAN APs are installed in cloudlets.

Since MDs that do not use the cloud computing resources also exist in the cellular network, we consider two types of MDs: task offloading MDs (TMDs) and non-task offloading MDs (NTMDs). The TMDs have the capability of task offloading, while the NTMDs do not have the capability, i.e., the NTMDs are conventional cellular users. When a TMD executes its task, it decides how to execute the task: either execution by using its own computing resources, i.e., *mobile execution*, or execution by using the cloud computing resources of cloud servers or cloudlets, i.e., *cloud execution*.

A. TASK OFFLOADING MODEL

We define a task model that considers the amount of the required energy for mobile execution, E_M(J), delay requirement, T_{req}(sec), execution time for mobile execution, T_M(sec), execution time for cloud execution, T_C(sec), size of the input data for cloud execution, S_{in}(bits), and size of the result data from cloud execution, S_{res}(bits). We assume that the tasks in our system are identical.[2] Then, an *outage* of task offloading occurs if

- the amount of required energy for cloud execution is larger than that for mobile execution or
- cloud execution cannot satisfy the delay requirement of the task.

In general, the amount of required energy for cloud execution of a TMD varies according to conditions such as its location and channel condition. On the other hand, the amount of required energy for mobile execution of a TMD is not affected by such conditions. The amount of required energy for mobile execution, E_M(J), is given by $E_M = P_M \cdot T_M$, where P_M is the constant power consumption of mobile execution. Note that this derivation of E_M is widely used in many researches on MCC [13]–[15], [17]. Then, E_M becomes constant from the assumption of identical tasks, implying identical T_M.

A TMD can conduct a cloud execution by using either cloud servers or cloudlets. For cloud execution using cloud servers, first the TMD transmits the input data for executing its task to the cloud servers through the cellular network with UL transmission. After transmitting the input data, the cloud servers execute the task and transmit the result data to the TMD, i.e., the TMD receives the result data through the cellular network with DL reception. The amount of the required energy for the cloud execution using cloud servers, E_C(J), is given by $E_{UL} + E_{DL}$, where E_{UL} and E_{DL} are the energy

[1] In the rest of this paper, we omit "remote" from remote cloud servers and "local" from local cloudlets for convenience.

[2] It is worth noting that we assume identical tasks for the sake of simple presentation. By considering k-type TMDs having different types of tasks, the identical tasks can be generalized to k-types of tasks having different characteristics. The way to consider k-type TMDs is described in footnote 3.

consumption for UL transmission and DL reception, respectively. Note that E_{UL} depends on the time duration of UL transmission, $T_{UL,\ trans}$(s) and the channel condition of the TMD due to the UL power control scheme. On the other hand, E_{DL} can be modeled by using the time duration of DL reception, $T_{DL,\ trans}$(s) and the RF modem power consumption, which is a constant [27]. Then, when $E_C > E_M$, the outage of task offloading to cloud servers occurs. In our model, we assume that for each cloud execution, $T_{UL,\ trans}$ and $T_{DL,\ trans}$ have fixed values satisfying the delay requirement of the task, i.e., $T_{UL,\ trans}+T_{DL,\ trans}+T_C \leq T_{req}$. Then, satisfying the delay requirement is equivalent that both UL transmission and DL reception should satisfy certain levels of data rates given by $R_{UL} = S_{in}/T_{UL,\ trans}$(bps) and $R_{DL} = S_{res}/T_{DL,\ trans}$(bps), respectively.

We assume that one WLAN AP is installed on each cloudlet, and for each task offloading to cloudlets, a TMD uses a single cloudlet. Each cloudlet can serve at most N_{cl}^{max} TMDs due to its limited computing resources. For cloud execution using a cloudlet, the TMD should be located within the connection range of the WLAN AP installed in the cloudlet. Then, the TMD transmits the input data to the cloudlet and receives the result data from the cloudlet. Similar to the cloud execution using cloud servers, the amount of the required energy for the cloud execution using cloudlets, E_C^{cl}(J), is given by $E_{UL}^{cl} + E_{DL}^{cl}$, where E_{UL}^{cl} and E_{DL}^{cl} are the energy consumption for UL transmission and DL reception, respectively. The required data rates for UL transmission and DL reception are also given by R_{UL}^{cl}(bps) and R_{DL}^{cl}(bps), respectively.

For the convenience of the analysis in Section III, in the following of this paper, we use power consumptions instead of energy consumptions, e.g., $P_{UL} = \frac{E_{UL}}{T_{UL,\ trans}}$ instead of E_{UL}. However, we cannot directly compare the power consumptions instead of comparing the energy consumptions since for each energy consumption, the time duration in which energy is consumed is different. To resolve this, we normalize each power consumption by the time duration of UL transmission for cloud execution using cloud servers, $T_{UL,\ trans}$. For example, let the time duration of mobile execution be T_M(s). Then, the power consumption of mobile execution is given by $\frac{E_M}{T_M}$(watt). However, for the convenience, we use the normalized power consumption of mobile execution obtained by $P_M = \frac{E_M}{T_M} \cdot \frac{T_M}{T_{UL,\ trans}}$(watt) instead of $\frac{E_M}{T_M}$(watt).

B. NETWORK MODEL

We consider a network model consisting of a single-tier cellular network and WLAN APs installed in cloudlets. The BSs of the cellular network are spatially distributed in $\mathbb{R}^2$ according to a homogeneous Poisson point process (PPP) $\mathbf{\Psi} = \{m_i; i = 1, 2, 3, \ldots\}$ with intensity λ, where $m_i \in \mathbf{\Psi}$ is the location of the i-th BS. The cloudlets, i.e., the WLAN APs, are also spatially distributed in $\mathbb{R}^2$ according to a homogeneous PPP $\mathbf{\Psi}_{cl} = \{c_i; i = 1, 2, 3, \ldots\}$ with intensity λ_{cl}. The MDs are spatially distributed in $\mathbb{R}^2$ according to a homogeneous PPP $\mathbf{\Phi} = \{u_i; i = 1, 2, 3, \ldots\}$ with intensity λ_u. Each MD associates to the base station (BS) that provides the best average link quality to it, i.e., the nearest BS to it. We assume that the probability that an MD is a TMD is p_T. Then, the TMDs are spatially distributed in $\mathbb{R}^2$ according to a homogeneous PPP $\mathbf{\Phi}_T$ with intensity $\lambda_u^T = p_T\lambda_u$, i.e., a thinning PPP from the homogeneous PPP $\mathbf{\Phi}$ with probability p_T.[3] The remaining MDs from the homogeneous PPP $\mathbf{\Phi}$ are the NTMDs, and they constitute a homogeneous PPP $\mathbf{\Phi}_{NT}$ with intensity $\lambda_u^{NT} = p_{NT}\lambda_u$, where $p_{NT} = 1 - p_T$. The path loss is modeled as $d^{-\alpha}$, where d is the distance between an MD and a BS, and α is the path-loss exponent. We consider the Rayleigh fading for the channel gain h, and thus, an independent exponential random variable with a unit mean.

For the cellular network, we consider N UL channels and N DL channels. We assume that all NTMDs have the data to transmit and all TMDs also have the input data to transmit for task offloading.[4] We also assume that each MD can use at most one UL channel and conducts the truncated channel inversion power control [25]. The MD controls its UL transmission power such that the received power at the BS becomes a certain constant, called a *cutoff threshold*. For the generality of the model, we denote the cutoff thresholds for TMDs and NTMDs by ρ^T and ρ^{NT}, respectively. If an MD requires its UL transmission power to exceed its usable UL transmission power due to power control, then the MD gives up its UL transmission. In our system, the usable UL transmission power of an NTMD is determined by its maximum UL transmission power. On the other hand, the usable UL transmission power of a TMD is determined by considering the power consumption for its task offloading in order that the task offloading outage does not occur. The details for determining the usable UL transmission power of a TMD is explained in Section III. We denote the usable UL transmission power of NTMDs and TMDs by P_u^{NT} and P_u^T, respectively.

Each WLAN AP has a connection range with radius d_{cl} and the TMDs within the range can access the corresponding cloudlet. Note that the NTMDs do not use the WLAN APs.

III. ANALYSIS OF TASK OFFLOADING IN MCC WITH ONLY REMOTE CLOUD SERVERS

In this section, we present the modeling and analysis of the outage probability of task offloading in the MCC system where cloudlets are not deployed. For task offloading to the cloud servers, a TMD should be connected to the cloud servers through the cellular network. The outage of its task offloading to the cloud servers does not occur if and only if all the following conditions are satisfied.

[3] We can generalize the TMDs as k-type TMDs having different types of tasks by considering k thinning PPPs from the PPP $\mathbf{\Phi}$ with corresponding probabilities, i.e., $\{p_T^l\}_{l=1,\ldots,k}$ such that $p_{NT} + \sum_{l=1,\ldots,k} p_T^l = 1$. Then, the analyses in this paper can be easily extended for k-type of TMDs.

[4] We can easily generalize that the NTMDs and TMDs have the data with probabilities by thinning the PPPs of the NTMDs and TMDs with the probabilities.

1) The total transmission power consumption of the TMD, i.e., $P_{UL} + P_{DL}$, is lower than the power consumption of mobile execution, P_M.[5]
2) The TMD is scheduled for UL transmission to transmit its input data.
3) The UL data rate of the TMD is higher than the required UL data rate, R_{UL}.
4) The TMD is scheduled for DL transmission to receive its result data.
5) The DL data rate of the TMD is higher than the required DL data rate, R_{DL}.

We assume that if the UL transmission for the task offloading succeeds, i.e., conditions 2 and 3 are satisfied, then the DL transmission for the task offloading also succeeds, i.e., conditions 4 and 5 are satisfied. This assumption is reasonable since the cellular network considers a quality-of-service (QoS) class identifier (QCI) of MDs to satisfy the QoS requirements of the MDs such as delay requirements and minimum data rates [28]. In addition, the transmission power of the BS for DL transmission is much higher than that of the MD for UL transmission and the interference coordination schemes can be conducted. That is, we only consider the outage due to conditions 1, 2, and 3. Note that the analyses for the outage probability due to conditions 1, 2, and 3 are presented in Sections III-A, III-B, and III-C, respectively. Next, the analysis of the outage probability of task offloading to cloud servers is provided in Section III-D. Note that in this section, we provide the analyses on NTMDs if they are related with the outage probability of task offloading.

A. ANALYSIS OF THE TRUNCATION OUTAGE PROBABILITY

For UL transmission, MDs control their transmission power using truncated channel inversion as described in Section II. Then, the MDs are divided into two groups on the basis of whether the required transmission power for the channel inversion is within their usable UL transmission power or not. The group of the MDs whose required transmission power is within the usable UL transmission power is called an *active* group, and the group of the other MDs is called an *inactive* group. The MDs in the active group, called the active MDs, can transmit their UL data, while the MDs in the inactive group, called the inactive MDs, give up their UL transmission, i.e., a *truncation outage* occurs.

The usable UL transmission power of TMDs is different from that of NTMDs due to task offloading. A TMD does not offload its task if its UL transmission power from channel inversion power control, P_{UL}, is larger than $P_M - P_{DL}$. Thus, the usable UL transmission power of the TMD is determined as

$$P_u^T = \min\left[P_M - P_{DL}, P_{max,UL}^{MD}\right], \tag{1}$$

where $P_{max,UL}^{MD}$ is the maximum UL transmission power of MDs. On the other hand, the usable UL transmission power of NTMDs is decided as $P_u^{NT} = P_{max,UL}^{MD}$. Then, with the usable UL transmission power and the cutoff threshold of each of TMDs and NTMDs, we define the cutoff distances of TMDs and NTMDs, d^T and d^{NT}, as $d^T = \left(\frac{P_u^T}{\rho^T}\right)^{\frac{1}{\alpha}}$ and $d^{NT} = \left(\frac{P_u^{NT}}{\rho^{NT}}\right)^{\frac{1}{\alpha}}$, respectively. A TMD becomes inactive, i.e., the truncation outage occurs, if its distance from the nearest BS is larger than d^T, whereas an NTMD becomes inactive if its distance from the nearest BS is larger than d^{NT}. Thus, the truncation outage probability for a typical TMD, $\mathcal{O}_p^T$, is obtained by a void probability that depends on the intensity of BSs, λ, i.e.,

$$\mathcal{O}_p^T = e^{-\pi\lambda {d^T}^2}. \tag{2}$$

Similarly, the truncation outage probability for a typical NTMD, $\mathcal{O}_p^{NT}$, is obtained as

$$\mathcal{O}_p^{NT} = e^{-\pi\lambda {d^{NT}}^2}. \tag{3}$$

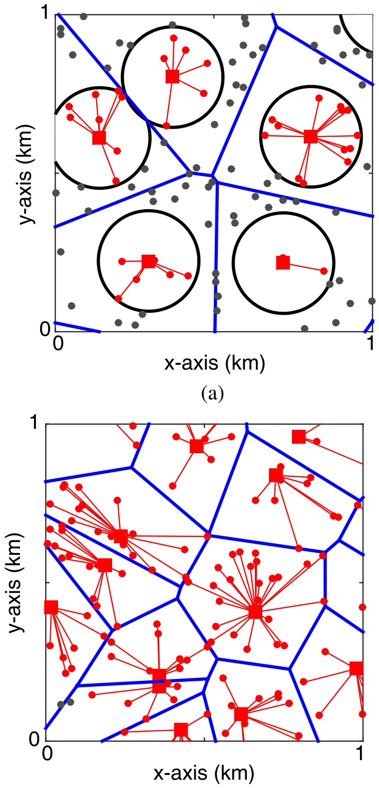
Note that the truncation outage probability for an NTMD is not related to the outage probability of task offloading. However, it will be used for the analysis of outage probability of task offloading presented in the following sections. From (2), we see that the truncation outage probability for a TMD decreases as the intensity of BSs, λ, increases and its cutoff distance becomes longer, i.e., when its usable UL transmission power becomes larger or its cutoff threshold becomes lower.

B. ANALYSIS OF THE SCHEDULING OUTAGE PROBABILITY

For the UL transmission of an active MD, it should be scheduled to use an UL channel. That is, a *scheduling outage* occurs when no UL channel is available to the MD due to the limited number of the UL channels. For each UL channel, a BS chooses one MD to be scheduled among its active MDs that are not only within its Voronoi cell but also within their cutoff distances from it. We assume that when the number of the active MDs is larger than the number of UL channels, the BS chooses the MDs to be scheduled randomly with the same probability. Thus, the scheduling outage probabilities for active TMDs and NTMDs are same.

We first derive the probability distribution of the number of the active MDs in the tagged BS with which a typical active MD is associated, i.e., p^{act}. Note that in the number of the active MDs in the tagged BS, the typical active MD is also included. In general, the number of the MDs in a BS is proportional to its Voronoi cell size. Thus, previous works [22], [23], [29], where the DL data rate of the cellular network is analyzed, utilize the size distribution of Voronoi cells provided in [30]. However, in our system, to be the active MDs in a BS, the MDs should be not only within the Voronoi cell but also within their cutoff distances from it. Thus, to derive p^{act}, we require a new approach considering both Voronoi cells and the cutoff distance.

[5] Note that we can compare the power consumptions instead of the energy consumptions since we use the normalized power consumptions as described in Section II.



(a)

(b)

FIGURE 1. **Part of the network model in 9 km^2 with $\lambda = 5$, $\lambda_u = 100$, and different values of ρ_0. The squares represent the BSs, the dots connected to their BS represent the active MDs, and the remaining dots denote the inactive MDs. (a) Case for $\rho_0 = -65$ dBm. The circle represents the region from the cutoff distance. (b) Case for $\rho_0 = -70$ dBm.**

According to the cutoff distance of MDs and the intensity of BSs, we have two different cases to approximate the number of active MDs, as in Fig. 1. When the cutoff distance is small or the intensity of BSs is sparse, the number of active MDs in a BS is proportional to the size of a circle with a radius of the cutoff distance, as in Fig. 1a. In this case, we can approximate the probability distribution of the number of active TMDs in the tagged BS of a typical active MD as the probability distribution of the number of TMDs within a circle with a radius of their cutoff distance, d^T, conditioning that the typical active MD is in the circle and the probability distribution is obtained as

$$p_C^{act,T}(k) = e^{-\hat{\lambda}_T}\frac{\left(\hat{\lambda}_T\right)^{k-1}}{(k-1)!}, \qquad (4)$$

where $\hat{\lambda}_T = \lambda_u^T \pi {d^T}^2$. It is worth noting that the probability distribution $p_C^{act,T}$ depends on the cutoff distance d^T since the size of the circle considered in the probability distribution is determined by the cutoff distance. In a similar way, the probability distribution of the number of NTMDs within a circle with a radius of their cutoff distance, d^{NT}, conditioning that the typical active MD is in the circle, $p_C^{act,NT}$, is obtained by substituting $\hat{\lambda}_T$ in (4) with $\hat{\lambda}_{NT} = \lambda_u^{NT}\pi {d^{NT}}^2$.

On the other hand, when the cutoff distance is large or the intensity of BSs is dense, almost all MDs in a BS are active, as in Fig. 1b. In this case, we can approximate the probability distribution as the probability distribution of the number of TMDs in the Voronoi cell of the tagged BS. Then, by using the size distribution of Voronoi cells in [30], the probability distribution is obtained as [23]

$$p_V^{act,T}(k) = \frac{3.5^{3.5}\Gamma(k+3.5)}{(k-1)!\Gamma(3.5)} \times \left(\frac{\lambda_u^T}{\lambda}\right)^{k-1}\left(3.5+\frac{\lambda_u^T}{\lambda}\right)^{-(k+3.5)}, \qquad (5)$$

where $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$. Note that we can obtain the probability distribution of active NTMDs, $p_V^{act,NT}(k)$, by simply replacing λ_u^T in (5) with λ_u^{NT}.

Let $p^{act,T}$ and $p^{act,NT}$ be the probability distributions of the number of active TMDs and NTMDs in the tagged BS, respectively. As mentioned above, $p^{act,T}$ and $p^{act,NT}$ depend on the corresponding cutoff distance and the intensity of BSs. Thus, we approximate them by appropriately using the four distributions, i.e., $p_C^{act,T}$, $p_C^{act,NT}$, $p_V^{act,T}$, and $p_V^{act,NT}$, according to the cutoff distances and the intensity of BSs. To this end, we propose a threshold-based approximation, in which the approximation for $p^{act,T}$ is chosen among $p_C^{act,T}$ and $p_V^{act,T}$ according to its corresponding cutoff distance d^T and the threshold γ^{act}. Similarly, the approximation for $p^{act,NT}$ is also chosen among $p_C^{act,NT}$ and $p_V^{act,NT}$ according to d^{NT} and γ^{act}.

We now provide the rationale for the derivation of the threshold, γ^{act}. As the cutoff distance becomes larger, the expected number of MDs within a circle with a radius of the cutoff distance, conditioning that the tagged MD is in the circle (e.g., for TMDs, the expected value of $p_C^{act,T}$) increases and will be equal to the expected number of MDs in Voronoi cells (e.g., for TMDs, the expected value of $p_V^{act,T}$). Note that the expected number of MDs in an arbitrary area is proportional to the area. Thus, the same expected number of MDs implies that the circle is large enough such that the probability distribution of the number of active MDs can be approximated by that of MDs in the Voronoi cell. Hence, we determine the threshold, γ^{act}, as the distance that makes the expected number of MDs within the circle equal to that of MDs in Voronoi cells. Then, the threshold γ^{act} is obtained as a function of the intensity of BSs, λ,

$$\gamma^{act}(\lambda) = \sqrt{\frac{4.5}{3.5\pi\lambda}}. \qquad (6)$$

The calculation of the threshold in (6) is provided in Appendix A. Note that the threshold depends only on the intensity of BSs λ, and hence, the same threshold in (6) is used for TMDs and NTMDs.

We then approximate the probability distribution $p^{act,T}$ by adopting the threshold-based approximation method as mentioned earlier. In the approximation method, one of $p_C^{act,T}$ and $p_V^{act,T}$ is chosen as the approximation for $p^{act,T}$ according

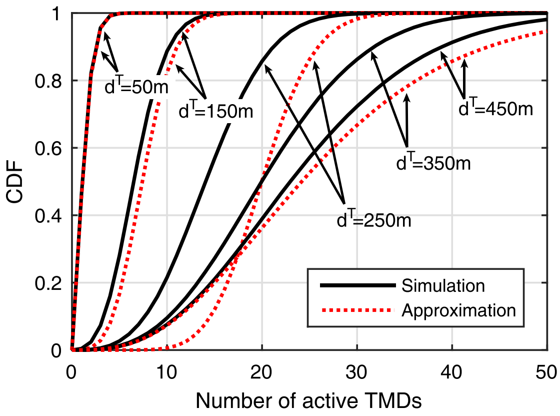


FIGURE 2. Cumulative probability distribution of $p^{act,T}$'s varying cutoff distances, d^T, from simulation and their approximations by using $p_C^{act,T}$ and $p_V^{act,T}$, with $\lambda = 5$ and $\lambda_u^T = 100$. Then, the threshold $\gamma^{act}(\lambda)$ is given by 286.1 m.

to the cutoff distance d^T and the threshold $\gamma^{act}(\lambda)$ in (6). When the cutoff distance d^T is shorter than the threshold, we choose $p_C^{act,T}$, and otherwise, we choose $p_V^{act,T}$ as

$$p^{act,T} = \begin{cases} p_C^{act,T}, & \text{if } d^T < \gamma^{act}(\lambda) \\ p_V^{act,T}, & \text{if } d^T \geq \gamma^{act}(\lambda). \end{cases} \tag{7}$$

For the probability distribution $p^{act,NT}$, we choose it same as in (7) by using $p_C^{act,NT}$, $p_V^{act,NT}$, and d^{NT}. Then, since TMDs and NTMDs are independently distributed, we can approximate the probability distribution of the number of active MDs in the tagged BS, p^{act}, as

$$p^{act}(k) = \sum_{l_T + l_{NT} = k} \left\{ p^{act,T}(l_T) p^{act,NT}(l_{NT}) \right\},$$

where l_T and l_{NT} are positive integer-valued.

In Fig. 2, $p^{act,T}$'s varying cutoff distances, d^T, from simulation and their approximations in (7) are shown. We can see that as the cutoff distance becomes longer or shorter than the threshold, the approximation of $p^{act,T}$ becomes more accurate. This trend is also applicable to the approximation of $p^{act,NT}$. Note that for $d^T < \gamma^{act}(\lambda)$, the approximations are different each other according to d^T since $p_C^{act,T}$ depends on the cutoff distance, d^T, which determines the radius of the circle for $p_C^{act,T}$. On the other hand, for $d^T \geq \gamma^{act}(\lambda)$, the approximations have a same probability distribution, $p_V^{act,T}$, since $p_V^{act,T}$ does not depend on d^T.

We now derive the scheduling outage probability as the following theorem.

Theorem 1: The scheduling outage probability for a typical active MD, $\mathcal{O}_{sc}$, is given by

$$\mathcal{O}_{sc} = \sum_{k=N+1}^{\infty} p^{act}(k) \cdot \frac{k-N}{k}. \tag{8}$$

Proof: The typical active MD is always scheduled if the number of the active MDs in the tagged BS is less than or equal to the number of channels N. On the other hand, when the number of the active MDs is greater than N, the typical MD is scheduled with the same probability with the other active MDs. Thus, the scheduling probability with the given number of the active MDs, n^{act}, is equal to $1/n^{act}$. Then, the scheduling outage probability for the typical MD with the given number of the active MDs, $k \geq N+1$, is obtained as

$$p^{act}(k) \cdot \frac{k-N}{k}.$$

By summing it over all $k \geq N+1$, the scheduling outage probability for a typical active MD is obtained as in (8). ■

From (8) in Theorem 1, we can see that the scheduling outage probability decreases as N becomes large. Suppose that N is given by N', and then, the scheduling outage probability is obtained as

$$\frac{p^{act}(N'+1)}{N'+1} + \frac{2p^{act}(N'+2)}{N'+2} + \frac{3p^{act}(N'+3)}{N'+3} + \cdots.$$

On the other hand, when N is given by $N'+1$, it is obtained as

$$\frac{p^{act}(N'+2)}{N'+2} + \frac{2p^{act}(N'+3)}{N'+3} + \cdots.$$

Then, we can see that compared with the scheduling outage probability when $N = N'+1$, the scheduling outage probability when $N = N'$ has the following additional terms in the summation:

$$\frac{p^{act}(N'+1)}{N'+1} + \frac{p^{act}(N'+2)}{N'+2} + \frac{p^{act}(N'+3)}{N'+3} + \cdots,$$

where all terms are positive. Thus, it is obvious that the scheduling outage probability decreases as N becomes large. On the other hand, it increases when the probability that the number of active MDs in the tagged BS is larger than $N+1$ becomes large. As in (4) and (5), the probability of the number of active MDs in the tagged BS increases in general as the intensity of MDs or the cutoff distance becomes large.

C. ANALYSIS OF THE SINR OUTAGE PROBABILITY

For the UL transmission of a scheduled TMD, the required UL data rate should be satisfied to offload the task within the target delay requirement. Since an MD can use at most one UL channel, in our system model, the required UL data rate can be equivalently expressed as the required UL SINR. Hence, we here derive the *UL SINR outage* probability for an active scheduled TMD.

Without loss of generality, the UL SINR analysis is conducted on a tagged BS located at the origin. According to Slivnyak's theorem [31], the statistical properties on the coexisting PPPs do not change due to conditioning on placing a BS at the origin. Thus, the UL SINR outage analysis for a BS at the origin can be applied to the other BSs. The UL SINR outage probability for a typical active scheduled TMD, $\mathcal{O}_{sinr}^T$, is calculated as follows [25]:

$$\begin{aligned}\mathcal{O}_{sinr}^{T} &= \mathbb{P}\left\{\frac{\rho^T h_o}{n_0+\mathcal{I}} \le \theta\right\} \\ &= \mathbb{P}\left\{h_o \le \frac{\theta}{\rho^T}(n_o+\mathcal{I})\right\} \\ &= \mathbb{E}_{\mathcal{I}}\left[1-e^{-\frac{\theta}{\rho^T}(n_0+\mathcal{I})}\right] \\ &= 1-e^{-\frac{\theta n_o}{\rho^T}}\mathcal{L}_{\mathcal{I}}\left(\frac{\theta}{\rho^T}\right), \end{aligned} \tag{9}$$

where h_o is the channel gain between the BS and the typical TMD at the tagged channel, n_0 is the noise power, θ is the SINR threshold, $\mathcal{I}$ is a random variable representing the aggregate interference at the BS from other active MDs scheduled on the tagged channel, and $\mathcal{L}_{\mathcal{I}}(\cdot)$ is the Laplace transform of its probability density function. The random variable $\mathcal{I}$ is given by $\sum_{u_i\in\tilde{\mathbf{\Phi}}} P_i h_i \|u_i\|^{-\alpha}$, where $\tilde{\mathbf{\Phi}}$ is the point process that consists of interfering MDs on the tagged channel, h_i is the channel gain between the BS and the interfering MD, and P_i is the transmission power of the interfering MD. To model the interference from other active MDs, $\mathcal{I}$, the soft-core process is more appropriate than the PPP since the interfering active MDs have a correlation among them. Nevertheless, many studies ignore the correlation among the interfering active MDs and model the interference by using the PPP since the soft-core process is not tractable and the correlation is weak [24], [25].

We analyze the UL SINR outage probability in a similar way in [25]. However, we consider two types of MDs, i.e., TMDs and NTMDs, which have different statistical properties on the UL transmission power, whereas only one type of MDs is considered in [25]. Thus, we should address their different statistical properties when deriving $\mathcal{I}$, since the interfering MDs consist of both types of MDs. In addition, contrary to [25], we do not assume the *saturation condition* with which each BS is assumed to have at least one MD for each UL channel. In other words, some UL channels in some BSs might not be used by any user in our system. We call those UL channels *idle UL channels*. Thus, when deriving $\mathcal{I}$, we should also consider the different statistical properties due to the idle UL channels.

As mentioned above, in our system, interfering MDs comprise two types of MDs. We obtain the statistical property on the interfering MDs as in the following lemma.

Lemma 1: Let the interfering MDs from the other BSs at the tagged UL channel be modeled as an independent PPP. Then, the probability that a typical interfering MD is a TMD, $\tilde{p}_T$, is obtained as

$$\tilde{p}_T = \frac{p_T(1-\mathcal{O}_p^T)}{p_{NT}(1-\mathcal{O}_p^{NT})+p_T(1-\mathcal{O}_p^T)},$$

where $\mathcal{O}_p^T$ is the truncation outage probability for a typical TMD in (2) *and $\mathcal{O}_p^{NT}$ is that for a typical NTMD in* (3). *Then, the probability that a typical interfering MD is an NTMD, $\tilde{p}_{NT}$, is obtained as $\tilde{p}_{NT} = 1-\tilde{p}_T$.*

Proof: A different type of MDs has a different truncation outage probability, (2) or (3), due to the different usable UL transmit power and cutoff thresholds. Thus, the probability that a typical active MD is an active NTMD is given by $p_{NT}(1-\mathcal{O}_p^{NT})$, and the probability that a typical active MD is an active TMD is given by $p_T(1-\mathcal{O}_p^T)$. Then, $\tilde{p}_T$ and $\tilde{p}_{NT}$ are obtained as this lemma, since each BS randomly chooses an UL scheduled MD among both active TMDs and active NTMDs. ■

In addition, in our system, some idle UL channels exist in some BSs. To address them in deriving $\mathcal{I}$, we assume that the interfering MDs constitute a homogeneous PPP with intensity $(1-p_{idle})\lambda$, where p_{idle} is the probability that an UL channel in a typical BS is idle. The probability p_{idle} is given by the following lemma. We skip the proof of the lemma since it can be easily proved.

Lemma 2: The probability that an UL channel in a typical BS is idle, p_{idle}, is obtained as

$$p_{idle} = \sum_{k=0}^{N} \hat{p}^{act}(k)\cdot\frac{N-k}{N}, \tag{10}$$

where $\hat{p}^{act}$ is the probability distribution of the number of active MDs in the typical BS.

For Lemma 2, the probability distribution of the number of the active MDs in a typical BS, $\hat{p}^{act}$, should be derived. Since it can be derived in a similar way to the probability distribution of the number of active MDs in the BS with which a typical active MD is associated, p^{act}, as in Section III-B, we provide the details of the derivation of $\hat{p}^{act}$ in Appendix B.

From Lemmas 1 and 2, the UL SINR outage probability of a typical active scheduled TMD is derived as the following theorem.

Theorem 2: In a cellular network with two types of MDs, let us assume that the interfering MDs constitute a PPP with intensity $(1-p_{idle})\lambda$ and their transmit powers are independent, where p_{idle} is given by (10). *Then, the UL SINR outage probability for a typical active scheduled TMD, $\mathcal{O}_s^T$, is given by*

$$\begin{aligned}\mathcal{O}_{sinr}^{T} = 1-\exp\Bigg[-\frac{\theta n_0}{\rho^T}-2\theta^{\frac{2}{\alpha}}\Bigg(\tilde{p}_{NT}\left(\frac{\rho^{NT}}{\rho^T}\right)^{\frac{2}{\alpha}}\xi\left(\frac{P_u^{NT}}{\rho^{NT}},\alpha\right) \\ +\tilde{p}_T\xi\left(\frac{P_u^T}{\rho^T},\alpha\right)\Bigg)\int_{\theta^{-\frac{1}{\alpha}}}^{\infty}\frac{y}{y^\alpha+1}dy\Bigg],\end{aligned} \tag{11}$$

where $\tilde{p}_T$ and $\tilde{p}_{NT}$ are given by Lemma 1, and $\xi(x,\alpha)$ is given by

$$\xi(x,\alpha) = \frac{\gamma\left(2,\pi(1-p_{idle})\lambda x^{\frac{2}{\alpha}}\right)}{1-\exp\left(-\pi(1-p_{idle})\lambda x^{\frac{2}{\alpha}}\right)},$$

where $\gamma(a,b) = \int_0^b t^{a-1}e^{-t}dt$ is the lower incomplete gamma function.

Proof: See Appendix C. ■

Theorem 2 provides the UL SINR outage probability for a typical active scheduled TMD in a simple expression. Moreover, when α is an integer, the integral in (11) is reduced in a closed-form expression [25]. For example, for $\alpha = 4$, the UL SINR outage probability is simplified as

$$\mathcal{O}^T_{sinr} = 1 - \exp\left[-\frac{\theta n_0}{\rho^T} - \sqrt{\theta}\left(\tilde{p}_{NT}\sqrt{\frac{\rho^{NT}}{\rho^T}}\xi\left(\frac{P^{NT}_u}{\rho^{NT}}, 4\right) + \tilde{p}_T\xi\left(\frac{P^T_u}{\rho^T}, 4\right)\right)\arctan(\sqrt{\theta})\right].$$

Theorem 2 shows that the UL SINR outage probability decreases when the UL SINR threshold, θ, becomes small. In addition, it also decreases as the cutoff threshold for TMDs, ρ^T, increases. However, note that when ρ^T increases, the truncation outage probability also increases. Thus, to optimize the total outage probability, ρ^T should be properly chosen as will be shown later. Moreover, the UL SINR outage probability decreases as p_{idle} increases. As shown in Lemma 2, p_{idle} increases as N becomes larger, and thus, larger N is always more favorable to reduce the outage probability as will be shown later since the scheduling outage probability also decreases when N becomes large.

D. OUTAGE PROBABILITY OF TASK OFFLOADING TO REMOTE CLOUD SERVERS

As mentioned at the beginning of this section, there are three conditions in order that a TMD connects to cloud servers through the cellular network and successfully offloads its task to the cloud servers: 1. being active, 2. being scheduled, and 3. satisfying the required UL data rate. Thus, by combining these conditions and the outage probability analyses in (2), (8), and (11), the outage probability that a TMD cannot offload its task to cloud servers, $\mathcal{O}_t$, is derived as

$$\mathcal{O}_t = \mathcal{O}^T_p + (1 - \mathcal{O}^T_p)\mathcal{O}_{sc} + (1 - \mathcal{O}^T_p)(1 - \mathcal{O}_{sc})\mathcal{O}^T_{sinr}. \tag{12}$$

IV. ANALYSIS OF TASK OFFLOADING IN HMCC

In this section, we present the analyses of the outage probabilities of task offloading in the MCC system having only cloudlets and in the HMCC system where both cloud servers and cloudlets exist. We also study the cloudlet deployment problem maximizing the profit of a CSP.

A. OUTAGE PROBABILITY OF TASK OFFLOADING TO LOCAL CLOUDLETS

For task offloading to cloudlets, a TMD connects to a cloudlet by using the WLAN AP installed in the cloudlet. The outage of the task offloading to cloudlets does not occur if and only if a TMD satisfies all the following conditions:

1) The TMD is within the connection range of the WLAN AP installed in the cloudlet.
2) The UL data rate of the TMD is higher than the required UL data rate, R^{cl}_{UL}.
3) The DL data rate of the TMD is higher than the required DL data rate, R^{cl}_{DL}.
4) The TMD is scheduled to use the computing resources of the cloudlet.

The last condition comes from the limited computing resources of cloudlets.

We assume that the connection range of the WLAN AP is determined to be short enough such that the energy consumption for the cloud execution using cloudlets, E^{cl}_C (J), is always less than that of mobile execution, and both UL and DL transmissions always satisfy the required data rates for the delay requirement, i.e., R^{cl}_{UL} and R^{cl}_{DL}.[6] Then, the satisfaction of condition 1 implies that conditions 2 and 3 are also satisfied. Thus, we only consider the outage due to conditions 1 and 4, i.e., the outage due to the connection range of the WLAN AP and the limited number of serving TMDs of cloudlets. Note that the analyses for the outage probability due to conditions 1 and 4, are presented in Sections IV-A.1 and IV-A.2, respectively. Then, the analysis of the outage probability of task offloading to cloudlets is provided in Section IV-A.3.

1) ANALYSIS OF THE RANGE OUTAGE PROBABILITY

A WLAN AP is installed in its corresponding cloudlet and has the connection range d_{cl}. Thus, when a TMD is not in the connection range of any WLAN AP, the TMD cannot offload its task to cloudlets, i.e., a *range outage* occurs. Hence, the range outage probability is given by the probability that any cloudlet does not exist within the distance d_{cl} from the TMD. Since the cloudlets constitute a homogeneous PPP, the range outage probability for a typical TMD, $\mathcal{O}^{cl}_r$, can be obtained by the void probability of a PPP with the distance d_{cl} as

$$\mathcal{O}^{cl}_r = e^{-\pi\lambda_{cl}d^2_{cl}}. \tag{13}$$

From (13), we see that the range outage probability decreases as d_{cl} or λ_{cl} becomes large.

2) ANALYSIS OF THE CLOUDLET SCHEDULING OUTAGE PROBABILITY

Cloudlets have a constraint on the number of serving TMDs due to their limited computing resources. Thus, a typical TMD within the connection range of the tagged cloudlet should be scheduled to use the computing resources of the tagged cloudlet for task offloading to cloudlets. According to the number of the TMDs in the connection range of the tagged cloudlet including the typical TMD, there are two cases that the typical TMD is scheduled to use the computing resources of the tagged cloudlet:

- *Case 1*: The number of the TMDs in the connection range is smaller than or equal to N^{max}_{cl}.

[6] These assumptions are reasonable since the power consumption using WLAN APs is lower than that using cellular networks [4], [32]. In addition, each WLAN AP in our system can provide sufficiently high data rate and low latency to serve its TMDs since it is directly connected to the cloudlet and serves only a limited number of TMDs due to the number of maximum serving TMDs of the cloudlet [4].

- *Case 2*: The number of the TMDs in the connection range is larger than N_{cl}^{max}, and the typical TMD is chosen to be scheduled among the TMDs in the connection range.

When the number of TMDs in the connection range is larger than N_{cl}^{max}, the typical TMD cannot use the computing resources if it is not scheduled to use it. We call this a *cloudlet scheduling outage*. The probability distribution of the number of the TMDs in the connection range is given by a Poisson distribution with intensity $\lambda_u^T \pi d_{cl}^2$, since the TMDs constitute a homogeneous PPP. We assume that when the number of the TMDs in the connection range of a cloudlet is greater than N_{cl}^{max}, the cloudlet randomly chooses N_{cl}^{max} TMDs to serve among the TMDs with the same probability for each TMD for the fairness among TMDs. Then, the cloudlet scheduling outage probability for a typical TMD in the connection range of tagged cloudlet, $\mathcal{O}_u^{cl}$, is given by

$$\mathcal{O}_u^{cl} = \sum_{k=N_{cl}^{max}+1}^{\infty} e^{-\lambda_u^T \pi d_{cl}^2} \frac{\left(\lambda_u^T \pi d_{cl}^2\right)^{k-1}}{(k-1)!} \cdot \frac{k - N_{cl}^{max}}{k}. \quad (14)$$

From (14), we can see that the cloudlet scheduling outage probability has the same form with the scheduling outage probability in (8), if we regard $e^{-\lambda_u^T \pi d_{cl}^2} \frac{(\lambda_u^T \pi d_{cl}^2)^{k-1}}{(k-1)!}$ as a function of k. Thus, in a similar way to the scheduling outage probability in (8), we can easily show that the cloudlet scheduling outage probability decreases as N_{cl}^{max} becomes large. On the other hand, it increases when the probability that the number of TMDs in the connection range is larger than $N_{cl}^{max} + 1$ becomes large, i.e., λ_u^T or d_{cl} becomes large.

3) OUTAGE PROBABILITY OF TASK OFFLOADING TO LOCAL CLOUDLETS

There are two conditions in order that a TMD connects to a cloudlet through its included WLAN AP and offloads its task to the cloudlet: 1. being within the connection range and 2. being scheduled. By combining the conditions and the outage probability analyses in (13) and (14), the outage probability that a typical TMD cannot offload its task to cloudlets, $\mathcal{O}_t^{cl}$, is derived as

$$\mathcal{O}_t^{cl} = \mathcal{O}_r^{cl} + (1 - \mathcal{O}_r^{cl})\mathcal{O}_u^{cl}. \quad (15)$$

B. OUTAGE PROBABILITY OF TASK OFFLOADING IN HMCC

We now analyze the outage probability of task offloading in the HMCC system consisting of both cloud servers and cloudlets. The outage of task offloading in the HMCC system implies that a task cannot be offloaded to both cloud servers and cloudlets. In the HMCC system, a correlation exists between the outage probabilities of task offloading to cloud servers and cloudlets, since the TMDs using the cloudlets, i.e., within the connection range of the cloudlets, do not connect to the cellular network.

Considering this correlation, the TMDs that offload tasks to the cloud servers using the cellular network, i.e., the TMDs that do not use the cloudlets, can be modeled by a Poisson hole process (PHP), where its holes represent the region within the connection range of the cloudlets. We can formally define the PHP for the TMDs as follows: Let the homogeneous PPP of TMDs, $\mathbf{\Phi}_T$, be a *baseline* PPP from which the holes will be carved out. Let the homogeneous PPP of cloudlets, $\mathbf{\Psi}_{cl}$, be the locations of holes with a radius d_{cl}. Then, the region covered by the holes, i.e., the region within the connection range of the cloudlets, is given by

$$\Xi_{cl} = \cup_{c_i \in \mathbf{\Psi}_{cl}} B(c_i, d_{cl}),$$

where $B(c_i, d_{cl}) = \{x \in \mathbb{R}^2 : \|x - c_i\| < d_{cl}\}$. By using the region Ξ_{cl}, the TMDs that offload tasks to the cloud servers can be formally expressed as

$$\mathbf{\Phi}_T' = \{u_i \in \mathbf{\Phi}_T : u_i \notin \Xi_{cl}\} = \mathbf{\Phi}_T \setminus \Xi_{cl}.$$

Then, with this PHP, $\mathbf{\Phi}_T'$, we can define a *new* MCC system only with the cloud servers, formed by omitting the TMDs that use the cloudlets from the original HMCC system. This new MCC system does not have the correlation with the cloudlets in the HMCC system any longer, since it does not include any TMDs that use the cloudlets. Thus, the outage probability for the HMCC system, $\mathcal{O}$, is given by

$$\mathcal{O} = \mathcal{O}_t' \times \mathcal{O}_t^{cl}, \quad (16)$$

where $\mathcal{O}_t'$ is the outage probability for the new MCC system. However, the characteristics of the PHP is not tractable in general [33], and thus, in many studies, they are addressed by approximating the PHP. Hence, in this subsection, we first derive the outage probability of task offloading in the new MCC system, $\mathcal{O}_t'$, by approximating the PHP $\mathbf{\Phi}_T'$. By using $\mathcal{O}_t'$, we then derive the outage probability of task offloading in the HMCC system as in (16).

We approximate the PHP of the TMDs that use the cloud servers to offload, $\mathbf{\Phi}_T'$, as a homogeneous PPP having the same intensity. The intensity of the PHP was derived in [31], and the intensity of the PHP $\mathbf{\Phi}_T'$ in our system, λ_{PHP}, is given by

$$\lambda_{PHP} = \lambda_u^T e^{-\pi \lambda_{cl} d_{cl}^2}. \quad (17)$$

Then, the PHP $\mathbf{\Phi}_T'$ is approximated as a homogeneous PPP with intensity, λ_{PHP}. With this PPP approximating the PHP and the PPP representing the NTMDs, we can approximate the new MCC system. We then apply the analysis in Section III to the *approximated* MCC system with its parameters that can be derived as follows. The intensity of the total MDs in the approximated MCC system, λ_u' is given by

$$\lambda_u' = \lambda_{PHP} + \lambda_u^{NT} = \lambda_u(p_T e^{-\pi \lambda_{cl} d_{cl}^2} + p_{NT}).$$

Then, the probability that a typical MD is a TMD, p_T', is derived by

$$p_T' = \frac{p_T e^{-\pi \lambda_{cl} d_{cl}^2}}{p_T e^{-\pi \lambda_{cl} d_{cl}^2} + p_{NT}},$$

and the probability that a typical MD is an NTMD, p'_{NT}, is derived by

$$p'_{NT} = \frac{p_{NT}}{p_T e^{-\pi \lambda_{cl} d_{cl}^2} + p_{NT}}.$$

With these parameters, i.e., λ'_u, p'_t, and p'_{NT}, the outage probability for the new MCC system, $\mathcal{O}'_t$, is approximated as in (12). Then, we can approximate the outage probability for the HMCC system as in (16).

From (17), we see that the intensity of the TMDs in the new MCC system decreases as the intensity and/or the connection range of the cloudlets increase. Thus, the scheduling outage probability in the new MCC system decreases as the intensity and/or the connection range of the cloudlets increase.

C. OPTIMAL CLOUDLET DEPLOYMENT

In this subsection, we study an optimal cloudlet deployment that maximizes the profit of a CSP. A tradeoff exists in using cloudlets since the costs occur when the CSP deploys and operates them. Thus, to address the tradeoff, we define an economic model of the HMCC system considering the CSP's expenses and revenue due to the deployment of cloudlets. We then formulate a cloudlet deployment problem to maximize the profit of the CSP and obtain the optimal cloudlet deployment by solving the problem.

In the economic model, the CSP calculates its profit from its HMCC system over a discrete time horizon, where each timeslot has a fixed equal duration, e.g., a week, a day, an hour. Let τ be an index of timeslots and the set of timeslots denoted by $\mathcal{T} = \{1, 2, \ldots, T\}$. We assume that the cloud servers of the CSP are already installed in its data center and their operational cost for each timeslot is constant. Note that with the assumption, maximizing the profit of the CSP is equivalent to maximizing the additional profit due to the cloudlets since the operational costs of the cloud servers are constant, and the revenue from the cloud servers and the deployment of the cloudlet are independent. Thus, we consider only the additional cost and revenue due to the deployment of cloudlets.

When the CSP deploys cloudlets to help more TMDs offload their tasks, the deployment cost, i.e., CAPEX, and operational cost, i.e., OPEX, occur. The CAPEX is the expenses from buying the cloudlets and their placement, and the OPEX consists of the electricity cost and internet connection cost for the cloudlets. Both CAPEX and OPEX depend on the intensity of the cloudlets to be deployed, λ_{cl}. Note that the CAPEX occurs only once when the cloudlets are deployed. On the other hand, the OPEX occurs during each timeslot after the deployment. We denote the functions of the CAPEX and the OPEX during each timeslot by $C_{CAPEX}(\lambda_{cl})$ and $C^{cl}_{OPEX}(\lambda_{cl})$. We assume that they are convex functions of λ_{cl}. The total OPEX by timeslot T is obtained as $\sum_{\tau=1}^{T} C^{cl}_{OPEX}(\lambda_{cl})$. Then, the CSP's total cost, C_{CSP}, is obtained as

$$C_{CSP}(\lambda_{cl}) = C_{CAPEX}(\lambda_{cl}) + \sum_{\tau=1}^{T} C^{cl}_{OPEX}(\lambda_{cl}). \tag{18}$$

The average intensity of the TMDs and that of the NTMDs during timeslot τ are denoted by $\Lambda^T_u(\tau)$ and $\Lambda^{NT}_u(\tau)$, respectively. The number of the tasks that the TMDs want to offload during timeslot τ is denoted by $N_U(\Lambda^T_u(\tau))$, which is a random variable depending on $\Lambda^T_u(\tau)$. We assume that a priori knowledge of $\Lambda^T_u(\tau)$, $\Lambda^{NT}_u(\tau)$, and the probability distribution of $N_U(\Lambda^T_u(\tau))$ are given from the past information of the MCC system. For given $\Lambda^T_u(\tau)$, $\Lambda^{NT}_u(\tau)$, and λ_{cl}, we can obtain the outage probability for the MCC system, $\mathcal{O}_t\left(\Lambda^T_u(\tau), \Lambda^{NT}_u(\tau)\right)$, and that for the HMCC system, $\mathcal{O}\left(\Lambda^T_u(\tau), \Lambda^{NT}_u(\tau), \lambda_{cl}\right)$, by using (12) and (16), respectively. Then, we can obtain the number of the task offloading outage during timeslot τ in the MCC system, i.e., without the cloudlets, as $N_U(\Lambda^T_u(\tau))\mathcal{O}_t\left(\Lambda^T_u(\tau), \Lambda^{NT}_u(\tau)\right)$, and that in the HMCC system, i.e., with the cloudlets, as $N_U(\Lambda^T_u(\tau))\mathcal{O}\left(\Lambda^T_u(\tau), \Lambda^{NT}_u(\tau), \lambda_{cl}\right)$. Then, the number of additional offloaded tasks by the cloudlets during timeslot τ is given by

$$N_U(\Lambda^T_u(\tau))\Big(\mathcal{O}_t(\Lambda^T_u(\tau), \Lambda^{NT}_u(\tau)) - \mathcal{O}(\Lambda^T_u(\tau), \Lambda^{NT}_u(\tau), \lambda_{cl})\Big).$$

When a TMD offloads its task by using the cloud computing resources, it should pay the fee for its task offloading to the CSP. The cost of task offloading for a single task is denoted by C_U. Then, the CSP's additional revenue during timeslot τ from the cloudlets is obtained as

$$\begin{aligned} R_{CSP}(\tau, \lambda_{cl}) = C_U N_U(\Lambda^T_u(\tau)) \Big(&\mathcal{O}_t(\Lambda^T_u(\tau), \Lambda^{NT}_u(\tau)) \\ &- \mathcal{O}(\Lambda^T_u(\tau), \Lambda^{NT}_u(\tau), \lambda_{cl})\Big). \end{aligned} \tag{19}$$

We formulate the cloudlet deployment problem maximizing the profit of the CSP from (18) and (19) while guaranteeing the outage requirement of task offloading, i.e., the maximum outage probability of task offloading, as

$$\begin{aligned} &\underset{0 \le \lambda_{cl} \le \lambda_{cl}^{max}}{\text{maximize}} \; \sum_{\tau=1}^{T} \mathbb{E}\left\{R_{CSP}(\tau, \lambda_{cl})\right\} - C_{CSP}(\lambda_{cl}) \\ &\text{subject to } \mathcal{O}\left(\Lambda^T_u(\tau), \Lambda^{NT}_u(\tau), \lambda_{cl}\right) \le \xi, \quad \forall \tau \in \mathcal{T}, \end{aligned} \tag{20}$$

where λ_{cl}^{max} is the maximum intensity of cloudlets, which the CSP can deploy in practice, ξ is the maximum outage probability of task offloading, and the expectation is taken over N_U. Then, by solving the problem, we can obtain the optimal intensity of the cloudlets to be deployed, which maximizes the profit of the CSP while guaranteeing the outage requirement of task offloading.

Since the problem has only a single decision variable, λ_{cl}, we can find the near-optimal intensity of the cloudlets quite easily. To this end, we first quantize the search interval of the problem, $\mathcal{D}_{cl} = \{\lambda_{cl} | 0 \le \lambda_{cl} \le \lambda_{cl}^{max}\}$. Let N_q denote

Algorithm 1 Algorithm for Solving Cloudlet Deployment Problem

```
$PROFIT_{OPT} = -\infty$
Quantize the search interval $\mathcal{D}_{cl}$ into $\bar{\mathcal{D}}_{cl}$
for each $\lambda_{cl} \in \bar{\mathcal{D}}_{cl}$ do
    if $\mathcal{O}\left(\Lambda_u^T(\tau), \Lambda_u^{NT}(\tau), \lambda_{cl}\right) \leq \xi, \forall\tau \in \mathcal{T}$ then
        Obtain $PROFIT(\lambda_{cl})$ as in (20)
        if $PROFIT_{OPT} < PROFIT(\lambda_{cl})$ then
            $PROFIT_{OPT} \leftarrow PROFIT(\lambda_{cl})$
            $\lambda_{cl}^* \leftarrow \lambda_{cl}$
        end if
    end if
end for
```

the cardinality of quantized search interval set. For example, we can adopt a simple linear quantization method, and then, the quantized search interval set, $\bar{\mathcal{D}}_{cl}$, is defined as

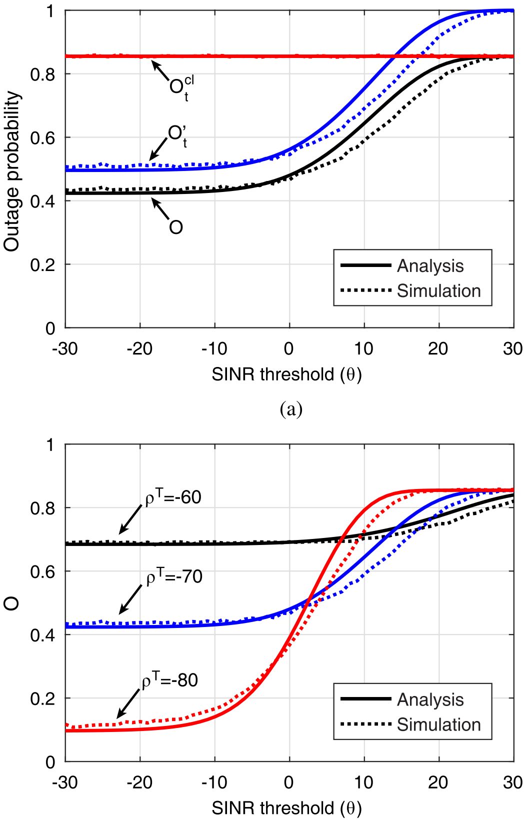
$$\bar{\mathcal{D}}_{cl} = \left\{0, \frac{\lambda_{cl}^{max}}{N_q - 1}, \frac{2\lambda_{cl}^{max}}{N_q - 1}, \cdots, \lambda_{cl}^{max}\right\}.$$

It is worth emphasizing that for quantizing the search interval, any other method other than the linear quantization method above can be used. For each $\lambda_{cl} \in \bar{\mathcal{D}}_{cl}$, we can obtain the profit of the CSP, i.e., the objective value of the problem in (20), and check if the outage requirement of task offloading are satisfied or not. Then, by comparing the profits with λ_{cl}'s that satisfy the outage requirements, we can obtain the near-optimal intensity, λ_{cl}^*, which has the maximum profit among such λ_{cl}'s. This algorithm is presented in Algorithm 1 as a pseudo code.

In the algorithm, as increasing N_q, i.e., quantizing the search interval more finely, the profit gap between the near-optimal intensity and the optimal intensity decreases. However, at the same time, the computational complexity of the algorithm also linearly increases according to N_q, since the algorithm should obtain the profit and check the outage requirements for each $\lambda_{cl} \in \bar{\mathcal{D}}_{cl}$. Nevertheless, the computational complexity is still reasonable owing to the simple expressions of the outage probabilities. Besides, the complexity to solve this problem is not a critical problem since the CSP does not have to solve it in real time.

V. RESULTS AND DISCUSSIONS

In this section, we verify our analysis through simulations and provide results and discussions for the HMCC system. To this end, we develop a dedicated C++-based simulator on which the following system can run. We set the BS intensity $\lambda = 5$ BSs/km^2, the cloudlet intensity $\lambda_{cl} = 20$ cloudlets/km^2, and the MD intensity $\lambda_u = 100$ MDs/km^2. Then, from (6), the threshold for approximation of p^{act}, $\gamma^{act}(\lambda)$, is given by 286.1m. Besides, from (23), the threshold for approximation of $\hat{p}^{act}$, $\hat{\gamma}^{act}(\lambda)$, is given by 252.31m. The probability that a typical MD is a TMD is set to be $p_T = 0.2$. We set the maximum UL transmission power $P_{max,UL}^{MD} = 0.2$ W, which is the maximum UL transmit power in LTE (23 dBm) [28].



(a)

(b)

FIGURE 3. **The outage probability varying the SINR threshold θ. (a) $\mathcal{O}$, $\mathcal{O}_t'$, and $\mathcal{O}_t^{cl}$, varying the SINR threshold θ with $\rho_0 = -70$ dBm. (b) $\mathcal{O}$'s varying the SINR threshold θ with $\rho^T = -60, -70, -80$ dBm.**

We assume that the mobile execution power is large enough to make the usable UL transmission power of the TMD be $P_u^T = P_{max,UL}^{MD} = 0.2$ W as in (1). Note that this assumption is reasonable since in general, the TMDs want to offload the tasks that have high computational complexity and consume a lot of power. We set the SINR threshold of TMDs be $\theta = 0$ dB with which the data rate requirements of tasks such as audio and image processing can be satisfied in LTE [34]. The number of UL channels is set to be $N = 25$. All MDs conduct channel inversion power control, and the cutoff threshold of TMDs and NTMDs are set to be $\rho^T = \rho^{NT} = -70$ dBm.[7] The noise spectral intensity is set to be -174 dBm. We set the connection range of a WLAN AP in a cloudlet $d_{cl} = 50$ m and the maximum number of serving TMDs in a cloudlet $N_{cl}^{max} = 5$. We drop BSs, cloudlets, and MDs randomly and uniformly over a 100 km^2 simulation area to realize PPPs. The above simulation parameters are used unless mentioned explicitly.

In our analysis, we approximate p^{act}, $\hat{p}^{act}$ and the correlation between using cloud servers and cloudlets for task offloading. Thus, we should verify our analysis for the outage probability of task offloading in the HMCC system with

[7] Note that this cutoff threshold is not favorable for our analysis since its cutoff distances for TMDs and NTMDs is close to the thresholds for approximation.

different system parameters through the simulation results. In the following results, we not only verify our analysis but also provide discussions on the task offloading in the HMCC system from our analysis.

In Fig. 3, we first compare the simulation results and our analyses varying the SINR threshold θ that represents the required UL data rates for task offloading. From Fig. 3a, we see that our analyses for the outage probabilities, $\mathcal{O}$, $\mathcal{O}'_t$, and $\mathcal{O}^{cl}_t$, closely follow the simulation results even with the approximation of p^{act} and $\hat{p}^{act}$. Note that θ does not affect the outage probability of task offloading to cloudlets, $\mathcal{O}^{cl}_t$, and thus, it is constant. When θ is small, i.e., when the required UL data rate is low, the SINR outage does not occur, and thus, $\mathcal{O}'_t$ is determined only by the truncation outage probability and the scheduling outage probability. As θ becomes larger, i.e., as the required UL data rate is high, the SINR outage probability also increases, and eventually, the task offloading to cloud servers always fails, i.e., $\mathcal{O}'_t \approx 1$. Thus, when offloading the task whose required UL data rate is high, the cloudlets are effective to offload it. In particular, in the case that the required UL data rate of the task is too high, using the cloudlets located close to the TMDs might be only a way to offload it.

In Fig. 3b, the total outage probabilities, $\mathcal{O}$, with different cutoff thresholds of TMDs, ρ^T, are shown. Since ρ^T is the system parameter of the cellular network that determines the received power level at BSs, it directly affects the outage probability. We see that our analyses closely follow the simulation results regardless of the cutoff threshold of TMDs. From the figure, we see that due to the truncation outage probability, when θ is small, the outage probability decreases as ρ^T becomes smaller. On the other hand, due to the SINR outage probability, when θ is large, the outage probability decreases as ρ^T becomes larger. Thus, with given tasks, the outage can be reduced by properly choosing the parameter of the cellular network, i.e., ρ^T. When offloading the tasks whose required UL data rate is high, the cellular network with high ρ^T is more favorable than that with low ρ^T, and vice versa.

In Fig. 4, we now compare the simulation results and our analyses for the outage probabilities varying the intensity of cloudlets, λ_{cl}. From Fig. 4a, we see that as λ_{cl} becomes larger, the total outage probability, $\mathcal{O}$, decreases since the outage probability of task offloading to cloudlets, $\mathcal{O}^{cl}_t$, decreases. This implies that using more cloudlets is effective to reduce the total outage probability of the task offloading. In addition, we can see that our analyses closely follow the simulation results despite the fact that we approximate the PHP of the TMDs that do not use the cloudlets. Note that the correlation between using cloud servers and cloudlets for task offloading becomes stronger as the intensity of cloudlets becomes larger. However, in the figure, the simulation result for the outage probability of task offloading to cloud servers is almost constant regardless of the intensity of cloudlets. From this result, we can infer that the correlation is still weak even though the intensity becomes larger.

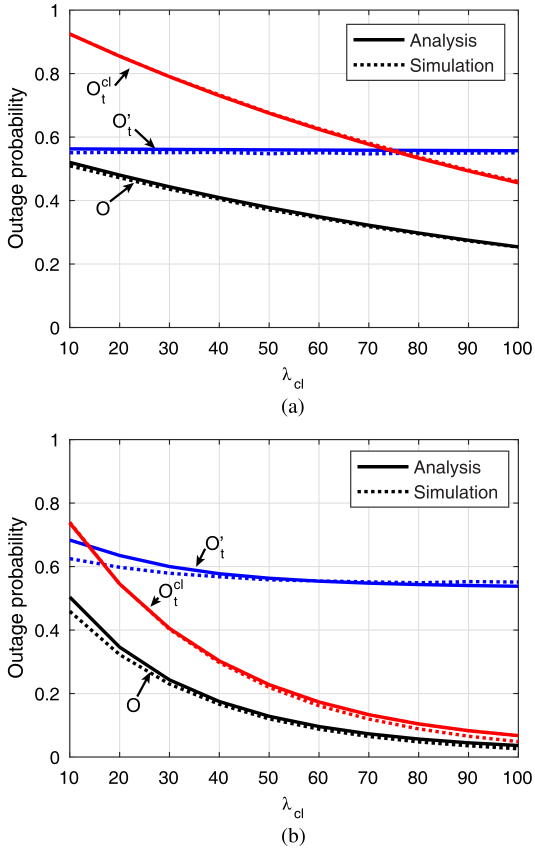


FIGURE 4. The outage probabilities varying the intensity of cloudlets, λ_{cl}. (a) $\mathcal{O}$, $\mathcal{O}'_t$, and $\mathcal{O}^{cl}_t$, varying the intensity of cloudlets, λ_{cl}. (b) $\mathcal{O}$, $\mathcal{O}'_t$, and $\mathcal{O}^{cl}_t$, varying the intensity of cloudlets, λ_{cl} with $p^T = 0.8$, $N = 10$, and $d_{cl} = 100$ m.

With such weak correlation, it is hard to show how the correlation affects the outage probability as in Fig. 4a. Thus, in Fig. 4b, to show the effect from the correlation, we consider a scenario in which the correlation is considerably stronger by setting the parameters to $p_T = 0.8$, $N = 10$, and $d_{cl} = 100$. Then, due to the strong correlation, we can see that the simulation results for the outage probability of task offloading to cloud servers, $\mathcal{O}'_t$, decreases as λ_{cl} increases. We can also see that our analyses closely follow the simulation results even with such strong correlation.

Compared with the previous work [25], we additionally consider the scheduling outage and idle UL channels in our analysis. They are directly affected by the number of UL channels. Thus, we compare the simulation results and our analysis for the outage probabilities varying the number of UL channels, N. Note that N does not affect the outage probability of task offloading to cloudlets, $\mathcal{O}^{cl}_t$, and thus, it is constant. From Fig. 5, we see that our analyses closely follow the simulation results. Naturally, as N increases, the scheduling outage probability decreases. In addition, the SINR outage probability also decreases since the interference becomes weaker due to the large number of idle UL channels. Thus, larger N is always more favorable to

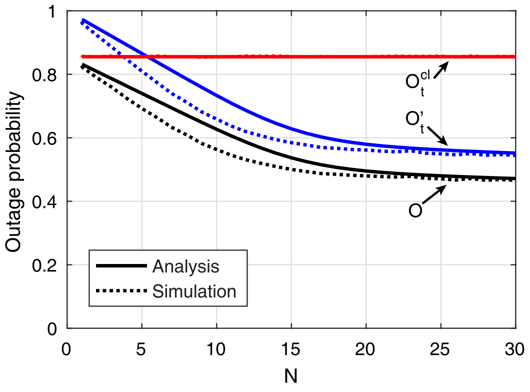


FIGURE 5. The outage probabilities, $\mathcal{O}$, $\mathcal{O}_t'$, and $\mathcal{O}_t^{cl}$, varying the number of UL channels, N.

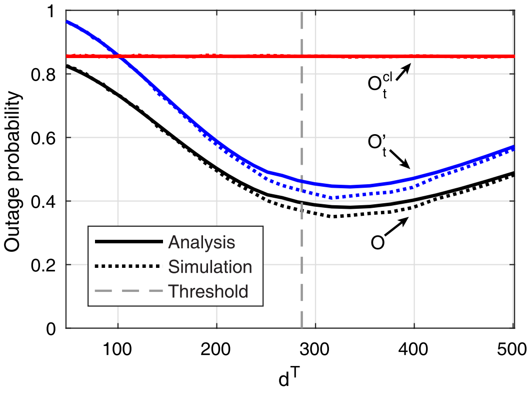


FIGURE 6. The outage probabilities, $\mathcal{O}$, $\mathcal{O}_t'$, and $\mathcal{O}_t^{cl}$, varying the cutoff distance, d^T.

offload the tasks regardless of the characteristics of the tasks such as the require UL data rate and the power consumption of mobile execution.

To consider the scheduling outage and idle UL channels, we approximate the probability distributions of the number of active TMDs and NTMDs, $p^{act,T}$ and $p^{act,NT}$, in the tagged BS using the proposed threshold-based approximation. The approximations become more accurate as the cutoff distances becomes longer or shorter than the threshold as shown in Fig. 2. Thus, to show how much the approximation errors affect the outage probabilities, in Fig. 6, we compare the simulation results and our analyses for the outage probabilities varying the cutoff distance, d^T. From the figure, we can see that despite of the approximations, our analyses closely follow the simulation results. This shows that the impact of the approximation errors for $p^{act,T}$ on the outage probability is not significant. Moreover, similar to the case of the approximations in Fig. 2, our analyses become more accurate as the cutoff distances becomes longer or shorter than the threshold.

In Fig. 7, we provide the total outage probabilities, $\mathcal{O}$, varying the system parameters in the MCC system, i.e., ρ^T, N, and P_u^T, with different intensities of cloudlets, λ_{cl}, to show a limitation in reducing the outage probability in the MCC system with only the cloud servers and also show the effectiveness of using more cloudlets to reduce the total outage probability. In Fig. 7a, the total outage probabilities varying the cutoff threshold of TMDs, ρ^T, are provided. As shown in Fig. 3b, ρ^T is a critical parameter in the outage probability of task offloading to cloud servers. When ρ^T is small, the SINR outage mainly occurs since the received power at the BS is weak. On the other hand, when ρ^T is large, truncation outage mainly occurs since large amount of power is required for the truncated channel inversion power control. Thus, as shown in the figure, to optimize the total outage probability, ρ^T should be properly chosen. However, ρ^T is hard to be controlled by the CSP since it is the system parameter of the cellular network.

In Fig. 7b, the total outage probabilities varying the number of UL channels, N, are shown. As N increases, the total outage probability decreases as also shown in Fig. 5 since both scheduling outage probability and SINR outage probability decrease. However, the total outage probability converges to the truncation outage probability since it does not depend on N. Thus, there is a limitation in reducing the total outage probability by increasing N. Besides, N is also the given parameter of the cellular network that cannot be controlled by the CSP.

In Fig. 7c, the total outage probabilities varying the usable transmission power of TMDs, P_u^T, are shown. When P_u^T is small, the total outage probability is high since the truncation outage frequently occurs. As P_u^T becomes larger, the total outage probability decreases, but it converges to a certain level since the scheduling outage and the SINR outage occur. Thus, as N, P_u^T also has a limitation in reducing the total outage probability. Moreover, as in (1), P_u^T depends on the power consumption of mobile execution, P_M, and the power consumption of DL reception, P_{DL}, and the maximum UL transmission power of MDs, $P_{max,UL}^{MD}$, which are not easily changed.

As shown in the figures in Fig. 7, a better outage performance can be achieved by optimizing the system parameters such as ρ^T, N, and P_u^T in the cellular network.[8] Moreover, optimizing the system parameters commonly has an intrinsic limitation in reducing the total outage probability. On the other hand, from the figures, we see that as the intensity of cloudlets increases, the total outage probability proportionally decreases. This implies that the cloudlets can be also utilized to achieve a better outage performance. In particular, when the QoS requirement such as the maximum outage probability is given, the cloudlets might be necessarily used to satisfy it due to the limitation of optimizing the system parameters. However, when using the cloudlets, the deployment cost and the operation cost occur. Thus, an economic tradeoff exists between the addition revenue and the costs from the cloudlets.

[8] Note that in general, such system parameters are hard to be controlled by the CSP. Nevertheless, the analysis according to the system parameters is useful to the CSP since it should predict the effect of the changes of such system parameters on its HMCC system to cope with them.

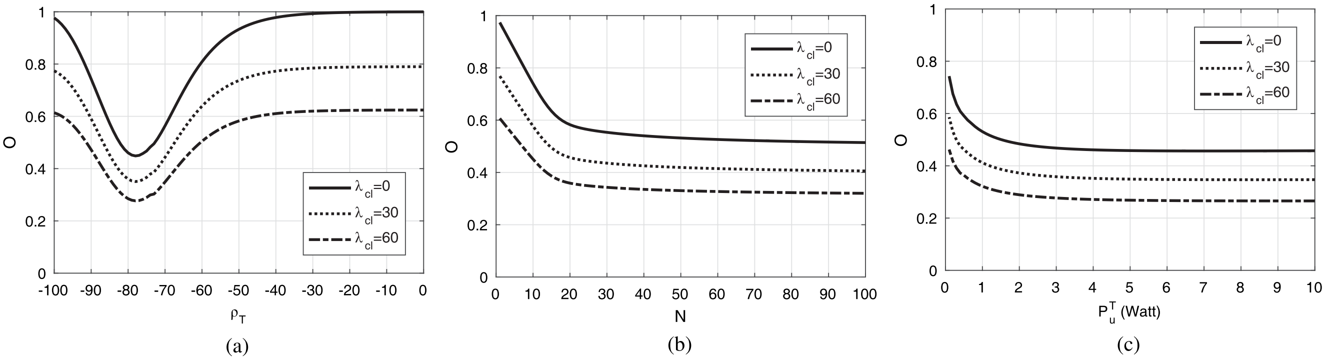


FIGURE 7. The outage probability varying the system parameters with different intensities of cloudlets. (a) $\mathcal{O}$'s varying the cutoff threshold of TMDs ρ^T with $\lambda_{cl} = 0, 30, 60$. (b) $\mathcal{O}$'s varying the number of UL channels N with $\lambda_{cl} = 0, 30, 60$. (c) $\mathcal{O}$'s varying the usable transmission power of TMDs P_u^T with $\lambda_{cl} = 0, 30, 60$.

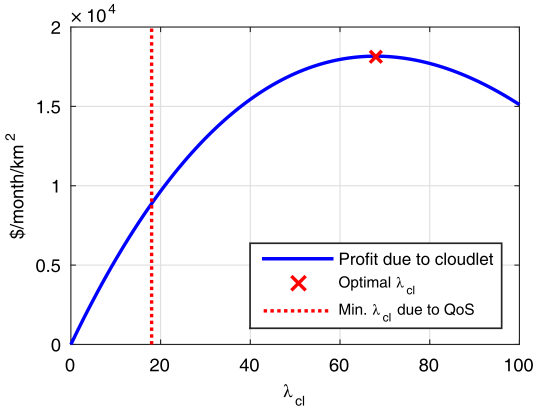


FIGURE 8. The expected profit for the cloudlet deployment problem.

We then show the optimal cloudlet deployment addressing the tradeoff from the economic model in Section IV-C. We set the duration of each timeslot to be a month and consider the time horizon of 5 years, i.e., 60 months. The deployment cost of the cloudlets is defined by $C_{CAPEX}(\lambda_{cl}) = 200\lambda_{cl}$ \$/km^2 and the operating cost of them is defined by $C_{OPEX}(\lambda_{cl}) = 10\lambda_{cl}$ \$/month/km^2. We set the cost of task offloading for a single task $C_U = 0.5$ \$/task. The average intensities of the TMDs and the NTMDs, Λ_u^T and Λ_u^{NT}, is set to be 20 TMDs/km^2 and 80 NTMDs/km^2 over all timeslots, respectively. The random variable for the number of the tasks which TMDs want to offload during each timeslot is generated as a Poisson distribution and its parameter (mean) is set to be $500\times\Lambda_u^T$ tasks/month/km^2. The QoS requirement, i.e., the maximum outage probability of task offloading, is set to be 0.5.

In Fig. 8, the expected profit varying the intensity of cloudlets in the cloudlet deployment problem is shown. Since the total outage probability is a nonincreasing function of the intensity of cloudlets, we can obtain the minimum intensity of cloudlets to satisfy the QoS requirement as shown in Fig. 8. We then find the intensity of cloudlets which maximizes the expected profit of the CSP. The CSP can achieve more expected profit by deploying more number of cloudlets than the minimum number of cloudlets to satisfy the QoS requirement. This result implies that the cloudlets can be used not only to resolve the weaknesses of using the cloud servers or to satisfy the QoS requirement, but also to get more revenue.

VI. CONCLUSION

In this paper, we modeled and analyzed the outage probability of task offloading in the HMCC system by using stochastic geometry. The analysis addresses the main causes of the outage of task offloading such as energy consumption and delay requirements of tasks. In addition, the correlation between the task offloading to cloud servers and cloudlet was incorporated. Then, it was verified through the simulation results. From the analysis, we showed that a lower bound exists on the outage probability of task offloading to remote cloud servers since outages due to power constraint, scheduling, and required data rate occur when TMDs access the cellular network. We also showed that using cloudlets is a promising solution to achieve a better outage probability. However, when using the cloudlets, an economic tradeoff exists between the additional revenue and costs. To address the tradeoff, we proposed an economic model of the HMCC system considering the CSP's costs and revenue. Based on the model, we formulated the cloudlet deployment problem to maximize the CSP's profit, and we then found the optimal deployment of the cloudlet by solving the problem. In the results, it is shown that the cloudlets can be utilized not only to satisfy the QoS requirement, i.e., the maximum outage probability, but also to generate more revenue.

APPENDIX A
CALCULATION OF THE THRESHOLD IN (6)

We determine the threshold, γ^{act}, as the distance that makes the expected number of MDs within a circle equal to that of MDs in Voronoi cells. We first calculate the expected number of MDs within a circle with a radius of γ^{act} conditioning that the typical MD is in the circle as

$$\sum_{k=1}^{\infty} k e^{-\lambda_u \pi (\gamma^{act})^2} \frac{(\lambda_u \pi (\gamma^{act})^2)^{k-1}}{(k-1)!} = \lambda_u \pi (\gamma^{act})^2 + 1. \quad (21)$$

We calculate the expected number of MDs in the Voronoi cell of the tagged BS as

$$\sum_{k=1}^{\infty} k \frac{3.5^{3.5}\Gamma(k+3.5)}{(k-1)!\Gamma(3.5)} \left(\frac{\lambda_u^T}{\lambda}\right)^{k-1} \left(3.5 + \frac{\lambda_u^T}{\lambda}\right)^{-(k+3.5)}$$
$$\overset{(i)}{=} \left(\frac{4.5\frac{\lambda_u}{\lambda}}{3.5+\frac{\lambda_u}{\lambda}}\right)\left(\frac{3.5+\frac{\lambda_u}{\lambda}}{3.5}\right) + 1 = \frac{4.5\lambda_u}{3.5\lambda} + 1, \quad (22)$$

where (i) follows from the regularization of the binomial series [35]. Then, we can determine the threshold in (6) by obtaining γ^{act}, which makes (21) and (22) the same.

APPENDIX B
DERIVATION OF $\hat{p}^{act}$

The number of active MDs in a typical BS and that in the tagged BS with the typical active MD have different statistics, i.e., $\hat{p}^{act}$ is different from p^{act} that was derived in Section III-B. In the tagged BS, at least one active MD exists due to the typical active MD. Moreover, the tagged BS probably has a larger area, which incurs more active MDs, since a BS having a larger area has more chance to cover the typical active MD [36].

As shown in Fig. 1, according to the cutoff distance and the intensity of BSs, we can approximate $\hat{p}^{act}$ by using the probability distributions, $\hat{p}_C^{act,T}$, $\hat{p}_C^{act,NT}$, $\hat{p}_V^{act,T}$, and $\hat{p}_V^{act,NT}$, which will be derived as follows. The probability distribution of the number of TMDs within a circle with a radius of their cutoff threshold, $\hat{p}_C^{act,T}$, is obtained as

$$\hat{p}_C^{act,T}(k) = e^{-\hat{\lambda}_T} \frac{\left(\hat{\lambda}_T\right)^k}{k!},$$

where $\hat{\lambda}_T = \lambda_u^T \pi {d^T}^2$. We can obtain that of NTMDs, $\hat{p}_C^{act,NT}$, by simply substituting $\hat{\lambda}_T$ in $\hat{p}_C^{act,T}$ with $\hat{\lambda}_{NT} = \lambda_u^{NT} \pi {d^{NT}}^2$. The probability distribution of the number of TMDs in a typical BS, $\hat{p}_V^{act,T}$, is obtained as

$$\hat{p}_V^{act,T}(k) = \frac{3.5^{3.5}\Gamma(k+3.5)}{k!\Gamma(3.5)} \left(\frac{\lambda_u^T}{\lambda}\right)^k \left(3.5 + \frac{\lambda_u^T}{\lambda}\right)^{-(k+3.5)}$$

in [37]. We can obtain that of NTMDs, $\hat{p}_V^{act,NT}$, by simply substituting λ_u^T in $\hat{p}_V^{act,T}$ with λ_u^{NT}. Then, we can apply the threshold-based approximation proposed in Section III-B.

To provide a function to choose the threshold for the approximation of $\hat{p}^{act}$, $\hat{\gamma}^{act}$, we first obtain the expected number of MDs in a circle with a radius of $\hat{\gamma}^{act}$ as in Appendix A. We can easily obtain it as $\lambda_u \pi \hat{\gamma}^{act2}$ since the number of MDs in a circle is given by a Poisson random variable. The expected number of MDs in the Voronoi cell of a typical BS is intuitively given by $\frac{\lambda_u}{\lambda}$. Note that we can also obtain the same result by using the regularization of the binomial series. Then, $\hat{\gamma}^{act}(\lambda)$ is given by

$$\hat{\gamma}^{act}(\lambda) = \sqrt{1/\pi\lambda}. \quad (23)$$

By using the threshold, we choose a probability distribution to use for $\hat{p}^{act,T}$ and $\hat{p}^{act,NT}$. For $\hat{p}^{act,T}$, we use $\hat{p}_C^{act,T}$ if $d^T < \hat{\gamma}^{act}(\lambda)$, and $\hat{p}_V^{act,T}$ otherwise. Similarly, for $\hat{p}^{act,NT}$, we use $\hat{p}_C^{act,NT}$ if $d^{NT} < \hat{\gamma}^{act}(\lambda)$, and $p_V^{act,NT}$ otherwise. With $\hat{p}^{act,T}$ and $\hat{p}^{act,NT}$, we can approximate the probability distribution of the number of active MDs in a typical BS, $\hat{p}^{act}(k)$, as

$$\hat{p}^{act}(k) = \sum_{l_T + l_{NT} = k} \left\{\hat{p}^{act,T}(l_T)\hat{p}^{act,NT}(l_{NT})\right\},$$

where l_T and l_{NT} are nonnegative integer-valued.

APPENDIX C
PROOF OF THEOREM 2

We first derive the probability distribution of the UL transmission power of active TMDs and that of active NTMDs. Let P^T be the UL transmission power of an active TMD and P^{NT} be the UL transmission power of an active NTMD. The probability distribution of the UL transmission power of an active TMD, f_{P^T}, is given in the following lemma.

Lemma 3 [25]: In a single-tier Poisson cellular network using truncated channel inversion power control with cutoff threshold ρ^T, the moments of the UL transmission power of an active TMD are obtained as

$$\mathbb{E}[{P^T}^\nu] = \frac{{\rho^T}^\nu \gamma\left(\frac{\nu\alpha}{2}+1, \pi\lambda {d^T}^2\right)}{(\pi\lambda)^{\frac{\nu\alpha}{2}}\left[1-\exp\left\{-\pi\lambda {d^T}^2\right\}\right]}.$$

Moreover, the moments of the transmission of an active NTMD in the UL can be similarly obtained by using its maximum transmit power P_u^{NT} and its cutoff threshold ρ^{NT}.

To obtain the UL SINR outage probability, the Laplace transform of the total interference is needed as in (9). To obtain it, we use the following facts and assumption:

- *Fact 1*: The average received power from the TMDs at any BS is equal to ρ^T.
- *Fact 2*: The average interference received from any single interfering MD is strictly less than $\hat{\rho} = \max\{\rho^{NT}, \rho^T\}$.
- *Fact 3*: Due to the idle UL channels, the intensity of interfering MDs on each channel is given by $(1 - p_{idle})\lambda$. Then, from Lemma 1, the intensities of interfering TMDs and NTMDs on each channel are given by $\tilde{p}_T(1-p_{idle})\lambda$ and $\tilde{p}_{NT}(1-p_{idle})\lambda$, respectively.
- *Assumption*: The interfering MDs constitute a homogeneous PPP and have independent UL transmission powers.

With the above, the total interference is given by

$$\mathcal{I} = \sum_{u_i \in \{\tilde{\Phi}_T \cup \tilde{\Phi}_{NT}\}\setminus\{o\}} \mathbb{I}\left(P_i \|u_i\|^{-\alpha} < \hat{\rho}\right) P_i h_i \|u_i\|^{-\alpha},$$

$$
\begin{aligned}
\mathcal{L}_{\mathcal{I}}(s) &= \mathbb{E}\left[e^{-s\mathcal{I}}\right] = \mathbb{E}\left[e^{-s\left(\sum_{u_i\in\tilde{\mathbf{\Phi}}_{NT}\setminus\{o\}}\mathbb{I}(P_i\|u_i\|^{-\alpha}<\hat{\rho})P_ih_i\|u_i\|^{-\alpha}+\sum_{u_i\in\tilde{\mathbf{\Phi}}_{T}\setminus\{o\}}\mathbb{I}(P_i\|u_i\|^{-\alpha}<\hat{\rho})P_ih_i\|u_i\|^{-\alpha}\right)}\right] \\
&\overset{(i)}{=} \mathbb{E}_{\tilde{\mathbf{\Phi}}_{NT}}\left[\prod_{u_i\in\tilde{\mathbf{\Phi}}_{NT}\setminus\{o\}}\mathbb{E}_{P_i,h_i}\left\{e^{-s\mathbb{I}\left(\|u_i\|>\left(\frac{P_i}{\hat{\rho}}\right)^{\frac{1}{\alpha}}\right)P_ih_i\|u_i\|^{-\alpha}}\right\}\right] \\
&\qquad\times\mathbb{E}_{\tilde{\mathbf{\Phi}}_{T}}\left[\prod_{u_i\in\tilde{\mathbf{\Phi}}_{T}\setminus\{o\}}\mathbb{E}_{P_i,h_i}\left\{e^{-s\mathbb{I}\left(\|u_i\|>\left(\frac{P_i}{\hat{\rho}}\right)^{\frac{1}{\alpha}}\right)P_ih_i\|u_i\|^{-\alpha}}\right\}\right] \\
&\overset{(ii)}{=} e^{-2\pi(1-p_{idle})\lambda\left\{\tilde{p}_{NT}\int_{\left(\frac{P^{NT}}{\hat{\rho}}\right)^{\frac{1}{\alpha}}}^{\infty}\mathbb{E}_{P^{NT},h}\left[\left(1-e^{-sP^{NT}hx^{-\alpha}}\right)\right]xdx+\tilde{p}_T\int_{\left(\frac{P^{T}}{\hat{\rho}}\right)^{\frac{1}{\alpha}}}^{\infty}\mathbb{E}_{P^{T},h}\left[\left(1-e^{-sP^{T}hx^{-\alpha}}\right)\right]xdx\right\}} \\
&\overset{(iii)}{=} e^{-2\pi(1-p_{idle})\lambda\left\{\tilde{p}_{NT}\int_{\left(\frac{P^{NT}}{\hat{\rho}}\right)^{\frac{1}{\alpha}}}^{\infty}\mathbb{E}_{P^{NT}}\left[\left(1-\frac{1}{1+sP^{NT}x^{-\alpha}}\right)\right]xdx+\tilde{p}_T\int_{\left(\frac{P^{T}}{\hat{\rho}}\right)^{\frac{1}{\alpha}}}^{\infty}\mathbb{E}_{P^{T},h}\left[\left(1-\frac{1}{1+sP^{T}x^{-\alpha}}\right)\right]xdx\right\}} \\
&\overset{(iv)}{=} e^{-2\pi(1-p_{idle})\lambda s^{\frac{2}{\alpha}}\left(\tilde{p}_{NT}\mathbb{E}\left[P^{NT\,\frac{2}{\alpha}}\right]+\tilde{p}_T\mathbb{E}\left[P^{T\,\frac{2}{\alpha}}\right]\right)\int_{(s\hat{\rho})^{\frac{-1}{\alpha}}}^{\infty}\frac{y}{y^{\alpha}+1}dy}.
\end{aligned}
\tag{24}
$$

where $\tilde{\mathbf{\Phi}}_T$ is a homogeneous PPP that represents the interfering TMDs, $\tilde{\mathbf{\Phi}}_{NT}$ is a homogeneous PPP that represents the interfering NTMDs, $\hat{\rho} = \max\{\rho^{NT}, \rho^{T}\}$, and $\mathbb{I}(\cdot)$ is an indicator function. Note that the indicator function is used to reflect the fact that the average interference received from any single interfering MD is strictly less than $\hat{\rho}$. By using the above equation, the Laplace transform of the total interference is derived as in (24), where (i) follows from the independence between $\tilde{\mathbf{\Phi}}_{NT}$, $\tilde{\mathbf{\Phi}}_T$, P_i, and h_i, (ii) follows from the probability generation functional of the PPP [31], (iii) follows from the Laplace transform of the channel gain h, and (iv) follows from substituting y with $\frac{x}{(sP)^{\frac{1}{\alpha}}}$. Then, the equation (11) is derived by substituting $\mathcal{L}_{\mathcal{I}}\left(\frac{\theta}{\rho^T}\right)$ in (9). $\mathcal{L}_{\mathcal{I}}\left(\frac{\theta}{\rho^T}\right)$ can be obtained from Lemma 3 and (24).

REFERENCES

[1] D. Huang, T. Xing, and H. Wu, “Mobile cloud computing service models: A user-centric approach,” *IEEE Netw.*, vol. 27, no. 5, pp. 6–11, Sep./Oct. 2013.

[2] N. Fernando, S. W. Loke, and W. Rahayu, “Mobile cloud computing: A survey,” *Future Generat. Comput. Syst.*, vol. 29, no. 1, pp. 84–106, 2013.

[3] Z. Sanaei, S. Abolfazli, A. Gani, and R. Buyya, “Heterogeneity in mobile cloud computing: Taxonomy and open challenges,” *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 369–392, 1st Quart., 2014.

[4] Y. Jararweh, L. Tawalbeh, F. Ababneh, and F. Dosari, “Resource efficient mobile computing using cloudlet infrastructure,” in *Proc. IEEE MSN*, Dec. 2013, pp. 373–377.

[5] H. Qi and A. Gani, “Research on mobile cloud computing: Review, trend and perspectives,” in *Proc. IEEE DICTAP*, May 2012, pp. 195–202.

[6] D. T. Hoang, D. Niyato, and P. Wang, “Optimal admission control policy for mobile cloud computing hotspot with cloudlet,” in *Proc. IEEE WCNC*, Apr. 2012, pp. 3145–3149.

[7] M. R. Rahimi, N. Venkatasubramanian, S. Mehrotra, and A. V. Vasilakos, “MAPCloud: Mobile applications on an elastic and scalable 2-tier cloud architecture,” in *Proc. IEEE UCC*, Nov. 2012, pp. 83–90.

[8] S. Abolfazli, A. Gani, and M. Chen, “HMCC: A hybrid mobile cloud computing framework exploiting heterogeneous resources,” in *Proc. IEEE MobileCloud*, Mar./Apr. 2015, pp. 157–162.

[9] K. Kumar and Y.-H. Lu, “Cloud computing for mobile users: Can offloading computation save energy?” *Computer*, vol. 43, no. 4, pp. 51–56, Apr. 2010.

[10] T. Zhao, S. Zhou, X. Guo, Y. Zhao, and Z. Niu, “A cooperative scheduling scheme of local cloud and Internet cloud for delay-aware mobile cloud computing,” in *Proc. IEEE Globecom Workshops*, Dec. 2015, pp. 1–6.

[11] M. V. Barbera, S. Kosta, A. Mei, and J. Stefa, “To offload or not to offload? The bandwidth and energy costs of mobile cloud computing,” in *Proc. IEEE INFOCOM*, Apr. 2013, pp. 1285–1293.

[12] J. Barrameda and N. Samaan, “A novel application model and an offloading mechanism for efficient mobile computing,” in *Proc. IEEE WiMob*, Oct. 2014, pp. 419–426.

[13] W. Zhang, Y. Wen, K. Guan, D. Kilper, H. Luo, and D. O. Wu, “Energy-optimal mobile cloud computing under stochastic wireless channel,” *IEEE Trans. Wireless Commun.*, vol. 12, no. 9, pp. 4569–4581, Sep. 2013.

[14] X. Chen, “Decentralized computation offloading game for mobile cloud computing,” *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 4, pp. 974–983, Apr. 2015.

[15] Y. Zhang, D. Niyato, and P. Wang, “Offloading in mobile cloudlet systems with intermittent connectivity,” *IEEE Trans. Mobile Comput.*, vol. 14, no. 12, pp. 2516–2529, Dec. 2015.

[16] W. Zhang, Y. Wen, and D. O. Wu, “Collaborative task execution in mobile cloud computing under a stochastic wireless channel,” *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 81–93, Jan. 2015.

[17] T. Liu, F. Chen, Y. Ma, and Y. Xie, “An energy-efficient task scheduling for mobile devices based on cloud assistant,” *Future Generat. Comput. Syst.*, vol. 61, pp. 1–12, Aug. 2016.

[18] H. Shah-Mansouri, V. W. S. Wong, and R. Schober, “Joint optimal pricing and task scheduling in mobile cloud computing systems,” *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 5218–5232, Aug. 2017.

[19] Z. Yin, F. R. Yu, S. Bu, and Z. Han, “Joint cloud and wireless networks operations in mobile cloud computing environments with telecom operator cloud,” *IEEE Trans. Wireless Commun.*, vol. 14, no. 7, pp. 4020–4033, Jul. 2015.

[20] A. Ceselli, M. Premoli, and S. Secci, “Mobile edge cloud network design optimization,” *IEEE/ACM Trans. Netw.*, vol. 25, no. 3, pp. 1818–1831, Jun. 2017.

[21] M. Jia, J. Cao, and W. Liang, “Optimal cloudlet placement and user to cloudlet allocation in wireless metropolitan area networks,” *IEEE Trans. Cloud Comput.*, vol. 5, no. 4, pp. 725–737, Oct./Dec. 2017.

[22] J. G. Andrews, F. Baccelli, and R. K. Ganti, “A tractable approach to coverage and rate in cellular networks,” *IEEE Trans. Commun.*, vol. 59, no. 11, pp. 3122–3134, Nov. 2011.

[23] S. Singh and J. G. Andrews, "Joint resource partitioning and offloading in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, pp. 888–901, Feb. 2014.

[24] T. D. Novlan, H. S. Dhillon, and J. G. Andrews, "Analytical modeling of uplink cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2669–2679, Jun. 2013.

[25] H. ElSawy and E. Hossain, "On stochastic geometry modeling of cellular uplink transmission with truncated channel inversion power control," *IEEE Trans. Wireless Commun.*, vol. 13, no. 8, pp. 4454–4469, Aug. 2014.

[26] W. Guo, S. Wang, X. Chu, J. Zhang, J. Chen, and H. Song, "Automated small-cell deployment for heterogeneous cellular networks," *IEEE Commun. Mag.*, vol. 51, no. 5, pp. 46–53, May 2013.

[27] *DRX Parameter in LTE*, document R2-071285, 3GPP, Nokia, 2007.

[28] A. Ghosh and R. Ratasuk, *Essentials of LTE and LTE-A*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[29] S. Singh, H. S. Dhillon, and J. G. Andrews, "Offloading in heterogeneous networks: Modeling, analysis, and design insights," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 2484–2497, May 2013.

[30] J.-S. Ferenc and Z. Néda, "On the size distribution of Poisson Voronoi cells," *Phys. A, Statist. Mech. Appl.*, vol. 385, no. 2, pp. 518–526, 2007.

[31] M. Haenggi, *Stochastic Geometry for Wireless Networks*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[32] J. Huang, F. Qian, A. Gerber, Z. M. Mao, S. Sen, and O. Spatscheck, "A close examination of performance and power characteristics of 4G LTE networks," in *Proc. 10th Int. Conf. Mobile Syst., Appl., Services*, 2012, pp. 225–238.

[33] Z. Yazdanshenasan, H. S. Dhillon, M. Afshang, and P. H. J. Chong, "Poisson hole process: Theory and applications to wireless networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7531–7546, Nov. 2016.

[34] M. H. Zarei, M. A. Shirsavar, and N. Yazdani, "A QoS-aware task allocation model for mobile cloud computing," in *Proc. ICWR*, Apr. 2016, pp. 43–47.

[35] V. Kowalenko, *The Stokes Phenomenon, Borel Summation and Mellin-Barnes Regularisation*. Emirate of Sharjah, UAE: Bentham Science, 2009.

[36] F. Baccelli and B. Blaszczyszyn, *Stochastic Geometry for Wireless Networks*. Breda, The Netherlands: Now Publishers, 2009.

[37] S. M. Yu and S.-L. Kim, "Downlink capacity and base station density in cellular networks," in *Proc. WiOpt*, May 2013, pp. 119–124.

HYUN-SUK LEE received the B.S. and Ph.D. degrees in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2012 and 2018, respectively. Since 2018, he has been a Post-Doctoral Research Associate with the Institute of BioMed-IT, Energy-IT and Smart-IT Technology (BEST), a Brain Korea 21 Plus Program, Yonsei University. His research interests include communication networks, mobile cloud computing, and smart grid.

JANG-WON LEE (M'04–SM'12) received the B.S. degree in electronic engineering from Yonsei University, Seoul, South Korea, in 1994, the M.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 1996, and the Ph.D. degree in electrical and computer engineering from Purdue University, West Lafayette, IN, USA, in 2004. From 1997 to1998, he was with the Dacom R&D Center, Daejeon, South Korea. From 2004 to 2005, he was a Post-Doctoral Research Associate with the Department of Electrical Engineering, Princeton University, Princeton, NJ, USA. Since 2005, he has been with the School of Electrical and Electronic Engineering, Yonsei University, where he is currently a Professor. His research interests include resource allocation, QoS and pricing issues, optimization, and performance analysis in communication networks, and smart grid.