THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Capacity and Delay Tradeoff of Secondary Cellular Networks With Spectrum Aggregation

Citation for published version:
Chen, L, Liu, C, Hong, X, Wang, C-X, Thompson, J & Shi, J 2018, 'Capacity and Delay Tradeoff of Secondary Cellular Networks With Spectrum Aggregation' IEEE Transactions on Wireless Communications, vol. 17, no. 6, pp. 3974-3987. DOI: 10.1109/TWC.2018.2818730

Digital Object Identifier (DOI):
10.1109/TWC.2018.2818730

Link:
Link to publication record in Edinburgh Research Explorer

Published In:
IEEE Transactions on Wireless Communications

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 05. Apr. 2019

Capacity and Delay Tradeoff of Secondary Cellular Networks with Spectrum Aggregation

Lingyu Chen, Chen Liu, Xuemin Hong, Cheng-Xiang Wang, *Fellow, IEEE,* John Thompson, *Fellow, IEEE,* and Jianghong Shi

***Abstract*—Cellular communication networks are plagued with redundant capacity, which results in low utilization and cost-effectiveness of network capital investments. The redundant capacity can be exploited to deliver secondary traffic that is ultra-elastic and delay-tolerant. In this paper, we propose an analytical framework to study the capacity-delay tradeoff of elastic/secondary traffic in large scale cellular networks with spectrum aggregation. Our framework integrates stochastic geometry and queueing theory models and gives analytical insights into the capacity-delay performance in the interference limited regime. Closed-form results are obtained to characterize the mean delay and delay distribution as functions of per user throughput capacity. The impacts of spectrum aggregation, user and base station (BS) densities, traffic session payload, and primary traffic dynamics on the capacity-delay tradeoff relationship are investigated. The fundamental capacity limit is derived and its scaling behavior is revealed. Our analysis shows the feasibility of providing secondary communication services over cellular networks and highlights some critical design issues.**

***Index Terms*—Capacity-delay tradeoff, secondary traffic, elastic traffic, cellular network, spectrum aggregation.**

I. Introduction

THE capacity of a cellular radio access network (RAN) is fundamentally limited by the density of base stations (BSs), system bandwidth, and spectrum efficiency. Once a particular network is rolled out, its maximum capacity is relatively stable. The traffic load, on the other hand, changes dynamically across space and time. Because the capacity of a cellular network is planned to accommodate peak traffic demand, redundant capacity is unavoidable due to traffic fluctuations. Measurements campaigns (e.g., [1]) have shown that redundant capacity is a pervasive problem, which results in low utilization and cost-effectiveness of network capital investments.

Measurement also revealed that the major cause of mobile traffic is multi-media consumption [2], which includes different types of communication services. The first type is streaming services that are delay-sensitive but loss-tolerant. Typical applications include voice over IP and video conferencing. The second type is elastic traffic services that are delay-tolerant but loss-sensitive. Typical applications include web browsing and file transfer. In practice, the above two types of traffic have no crucial difference in the delay constraints, which are measured by mini-seconds. However, the emergence of new applications such as proactive caching [3]–[5] brings a third type of traffic that has crucial difference from the first two types. Proactive caching systems are able to push content and cache it closer to end users, exploiting the fact that content demand is predictable and that large cache space is becoming affordable. The traffic generated by proactive caching has two distinct characteristics. First, the delay constraint is very relaxed. This is because a piece of content can be pushed to a user device hours or minutes ahead before the user requests actually happen. The delay constraint for such traffic is several orders larger than the constraints of conventional traffic. Second, the traffic volume is very flexible (i.e., can be arbitrarily small) because proactive caching is opportunistic and transparent to users. Due to these two distinct characteristics, we call such a new type of traffic as ultra-elastic traffic.

This paper is motivated by an envision that the redundant capacity in cellular networks can be exploited to deliver the ultra-elastic traffic as secondary traffic, which coexists with other higher priority traffic in the same cellular network [6]. This could allow the redundant capacity to be commercialized to offer a new type of communication service. This paper aims to investigate the performance of the secondary traffic in a context of heterogeneous cellular networks (HCNs) [7]. The HCN represents the future trend of cellular networks, where cell densification and spectrum aggregation are prominent features [8], [9]. Cell densification means heterogeneous BSs will be densely deployed, while spectrum aggregation allows the BSs and/or users to dynamically operate on multiple non-overlapping frequency bands.

Capacity and delay are the two most important performance metrics of a communication service. Given resource constraints, maximizing capacity and reducing delay are conflict-

L. Chen and C. Liu are with the Department of Communications Engineering, School of Information Science and Technology, Xiamen University, Xiamen 361005, Fujian, P.R.China. Email: {chenly, chenl}@xmu.edu.cn.

X. Hong and J. Shi are with the Key Laboratory of Underwater Acoustic Communication and Marine Information Technology, Ministry of Education of China, Xiamen University, Xiamen 361005, Fujian, P.R. China. Email: {xuemin.hong, shijh}@xmu.edu.cn.

C. Wang is with the Institute of Sensors, Signals and Systems, School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh, EH14 4AS, UK. Email: cheng-xiang.wang@hw.ac.uk.

J. Thompson is with the Institute for Digital Communications, School of Engineering, University of Edinburgh, Edinburgh, EH9 3JL, UK. Email: john.thompson@ed.ac.uk.

Corresponding author: Xuemin Hong, Email: xuemin.hong@xmu.edu.cn, Tel: +86–592-2580150.

This work is supported by the National Natural Science Foundation of China (NSFC) (Grant No.: 61571378, 61601388), the Engineering and Physical Sciences Research Council of UK (EPSRC)(Grant No. EP/L020009/1, EP/P000703/1, EP/L026147/1), the EU H2020 5G Wireless project (Grant No. 641985), the EU H2020 RISE TESTBED Project (Grant No. 734325), the EU FP7 QUICK project (Grant No. PIRSES-GA-2013-612652), and the International S&T Cooperation Program of China (No. 2014DFA11640).

Manuscript received XXXX, XXXX; revised XXXX, XXXX.

ing objectives. This is known as the capacity-delay tradeoff, which characterizes the fundamental performance bound of a communication service. The capacity-delay tradeoff has attracted significant research attentions for a variety of communication systems. Multiple analytical frameworks have been proposed to study the tradeoff, including the frameworks of scaling law analysis [10]–[14], interference approximation [15]–[17], and timely throughput [18]–[20]. Scaling law analysis [10]–[14] is a novel framework that can characterize how the mean capacity and delay scale with the network size, but is not able to give an exact quantification on the capacity or delay. The framework of interference approximation [15]–[17] focused on the session level performance of multi-cell networks, but can only provide loose bounds for the estimation of mean delay. The framework of timely throughput [18]–[20] assumed that a queuing packet will be dropped if the packet passes a critical delay. This framework is better suited for the study of loss-tolerant traffic instead of loss-sensitive traffic. Moreover, it does not provide a detailed characterization of the delay distribution.

For performance study of the secondary traffic, it is important to understand the delay distribution. This is because the secondary traffic is loss-sensitive and delay-tolerant, so that a good indicator of the user quality-of-experience is the "outage delay", which gives the probability of having large delays that surpass certain delay-tolerance threshold. Unfortunately, the above-mentioned frameworks for capacity-delay tradeoff study [10]–[20] do not offer the capability to analytically characterize the delay distribution of loss-sensitive traffic.

In this paper, we propose a new framework that integrates stochastic geometry and queueing models to study the session level capacity-delay tradeoff of secondary traffic. Our framework can complement existing ones by offering a means to pinpoint the delay distribution analytically. The merit of our framework comes from the fact that the stochastic geometry and queueing models are the most tractable models in describing the complex spatial and temporal behaviors of a cellular network, respectively. In the spatial domain, stochastic geometry models can yield elegance analytical results [7], [21]–[25] while keeping the same level of accuracy compared with the conventional hexagon models. In the temporal domain, two other widely used models are the discrete/continuous-time Markov chain model [26]–[29] and local delay model [30], [31]. The former can produce delay distribution by numerical computation, but fall short in providing closed-form insights. The latter model focuses on the average delay and reveals no information about the delay distribution. None of these two models can offer the same tractability as the well-established queueing model [32], [33].

How to integrate the stochastic geometry models and queueing models into a coherent framework has long been recognized as a challenging task [34]. The challenge lies in capturing the complex coupling of network behavior in the spatial and temporal domains while preserving the analytical tractability of the model. To this end, some recent attempts were reported in [34]–[36]. In [34], the spatial-temporal dependence of a cellular system is captured by some cell-load equations and eventually resolved via static simulations. Although this framework is mathematically rigorous, it lacks the analytical tractability to reveal closed-form insights. In our previous work [35], [36], stochastic geometric and queueing models are combined to study the uplink capacity of hybrid ad-hoc networks with user collaboration. However, these works focused on a different type of network and did not fully address the issue of multi-user access, which is a critical feature of cellular systems. To our best knowledge, full integration of stochastic geometry and queueing models for the study of cellular networks is still an open problem [34].

This paper proposes a new approach of integrating stochastic geometry models and priority queuing models for the performance study of loss sensitive, delay-tolerant secondary traffic in large scale cellular systems. The main advantage of our approach lies in its analytical tractability to pinpoint delay distributions. Specifically, the following contributions are made.

- Analytical results are derived to characterize the mean delay and delay distribution as functions of per user throughput capacity.
- Analytical results are derived to characterize the capacity limit in some special cases.
- A concise analytical approximation is obtained to describe how the per user capacity scales with user-BS density ratio.

The remainder of this paper is organized as follows. Section II describes the system model. The overall methodology and some useful approximations are introduced in Section III. The capacity-delay tradeoff and fundamental capacity limit are studied in Sections IV and V, respectively. Section VI provides numerical results and discussions. Conclusions are drawn in Section VII.

II. System Model

A. Secondary access protocol

We consider the downlink of a large scale cellular network that aggregates N non-overlapping frequency bands. We assume that these bands are all usable during the considered time frame. BSs operating in the same band are assumed to have homogeneous bandwidth and transmit power denoted by W_n and P_n, respectively, where n ($1 \leq n \leq N$) is the band index. A user can operate in one band at a time, but can handover between different bands. Over the top of such a multi-band physical layer, two general types of services are offered: the primary service that is delay-sensitive and has a higher priority to access the physical layer resource, and the secondary service that is delay-tolerant and only use vacant physical layer resource after the primary service. Secondary users are assumed to comply with the following access protocol as illustrated in Fig. 1.

- Step 1: Periodically check the buffer of secondary traffic. If the buffer is empty, remain in idle mode. Otherwise turn into active mode and proceed to Step 2.
- Step 2: Randomly select one band and associate with the nearest BS operating in the chosen band. This implies the widely used Poisson-Voronoi cellular network model.

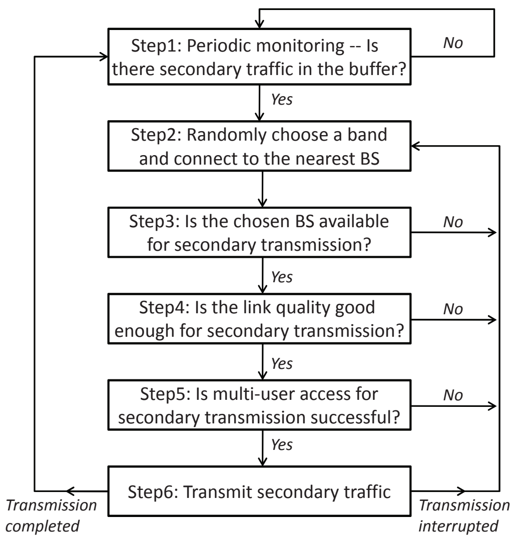
Step1: Periodic monitoring -- Is there secondary traffic in the buffer?
No
Yes
Step2: Randomly choose a band and connect to the nearest BS
Step3: Is the chosen BS available for secondary transmission?
No
Yes
Step4: Is the link quality good enough for secondary transmission?
No
Yes
Step5: Is multi-user access for secondary transmission successful?
No
Yes
Step6: Transmit secondary traffic
Transmission completed
Transmission interrupted

Fig. 1. Flow chart of the secondary multi-user access protocol.

- Step 3: Evaluate whether the associated BS is vacant (i.e., not occupied by primary traffic) and available for secondary services. If yes, proceed to Step 4, otherwise return to Step 2. The probability that a typical BS in the nth band is vacant is called "vacant probability" and is denoted by Ω_n. This parameter indicates the average load of primary traffic.
- Step 4: Evaluate the link quality with respect to the associated BS. If the signal-to-noise-and-interference ratio (SINR) is large enough to support a transmission rate of R bits/s, proceed to Step 5, otherwise return to Step 2. Here, R is the minimum rate requirement of secondary transmission. Such a requirement is imposed to restrict secondary services only to users with high quality links, otherwise the secondary services may become inefficient due to excessive interference and energy consumption. The probability that a typical user in the nth band has good link quality is called "coverage probability" and denoted by $p_{n,v}$.
- Step 5: Compete with other in-coverage users for multiple access to the same BS. We assume a time-division multiple access (TDMA) scheme for multi-user access, where a band is fully allocated to one user at a time and multiple contending users have equal opportunities to access the band through time sharing. If contention is successful, proceed to Step 6, otherwise return to Step 2. The probability that an in-coverage secondary user in the nth band is granted access is called "access probability" and denoted by $p_{n,a}$.
- Step 6: Transmit secondary traffic with a fixed rate R until the buffer is empty. If the buffer is empty, proceed to Step 1. Otherwise if an outage (caused by primary traffic interruption or coverage outage) occurs during transmission, return to Step 2.

For a user to receive secondary service in the nth band, he should firstly be associated with a vacant BS, secondly have a good coverage, and finally be granted access after multi-user contention. It follows that the service probability ε_n is the product of vacant probability Ω_n, coverage probability $p_{n,v}$, and access probability $p_{n,a}$, i.e.,

$$\varepsilon_n = \Omega_n \cdot p_{n,v} \cdot p_{n,a}. \tag{1}$$

The flow chart of the above protocol is illustrated in Fig. 1. We note that this protocol is not a standard-defined protocol. However, it is simple yet sufficient to capture the essence of secondary multi-user access procedure and can represent a wide range of practical access schemes.

B. Spatial interference model

The spatial layout of BSs operating in the nth band is modeled by a stationary Poisson Point Process (PPP) in $\mathbf{R}^2$ with intensity $\lambda_{b,n}$. This is a commonly used model in the literature. For analytical tractability, we ignore the case of co-located BSs and assume that the spatial layout of BSs in different bands are independent. The spatial distribution of secondary users are also assumed to follow a stationary PPP in $\mathbf{R}^2$ with intensity λ_u. Let us consider a typical user in the nth band, the downlink SINR is a random variable, whose cumulative density function (CDF) has been derived for different types of fading channels [38]. For purposes of clarity and tractability, we consider a representative case in which the path loss exponent is 4. The complementary CDF of the user SINR is then given by [38]

$$F_{\gamma,n}(x) = \frac{\pi^{\frac{3}{2}}\lambda_{b,n}}{\sqrt{x/P_n}} e^{\frac{a^2}{\sqrt{2b}}} \mathrm{Q}\left(\frac{a}{\sqrt{2x/P_n}}\right) \tag{2}$$

where $\mathrm{Q}(\cdot)$ denotes the Q-function and

$$a = \lambda_{b,n}\pi\left[1 + \sqrt{x}\arctan(\sqrt{x})\right]. \tag{3}$$

If the system is interference limited, which implies that P_n is sufficiently large and the noise is negligible, (2) can be further simplified to [38]

$$F_{\gamma}^{\lim}(x) = \frac{1}{1 + \sqrt{x}\arctan(\sqrt{x})}. \tag{4}$$

According to the secondary access protocol, a user is in coverage of secondary services if $W_n \log_2(1 + \gamma_n) \geq R$, where γ_n denotes the SINR perceived by the user. The coverage probability in the nth band is therefore given by

$$p_{n,v} = F_{\gamma,n}(2^{R/W_n} - 1). \tag{5}$$

C. Temporal queuing model

As illustrated in Fig. 2, we model the secondary traffic dynamic as a preemptive priority queue, where the transmission of secondary traffic may be preempted (i.e., immediately interrupted) by outages. An outage can be caused by multiple factors such as primary traffic interruption, bad coverage, and failure in multi-user contention. We assume users can handover between bands with negligible time, hence an outage only occurs when no band is available for secondary services. We propose to model the composite outage effect as a stream of higher priority traffic in the priority queue. The arrival of outage events follows a Poisson process with mean interval $\bar{\alpha}_o$.

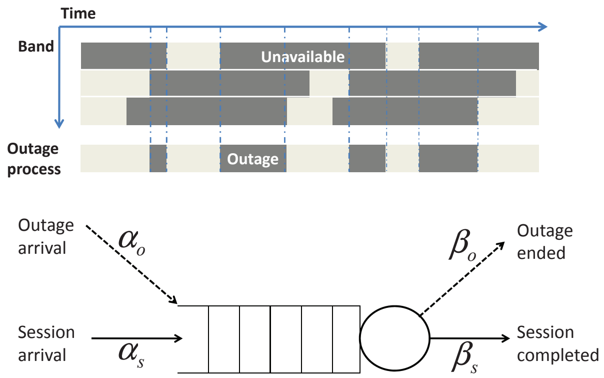

Fig. 2. Priority queuing model of a typical user with secondary traffic and random outage.

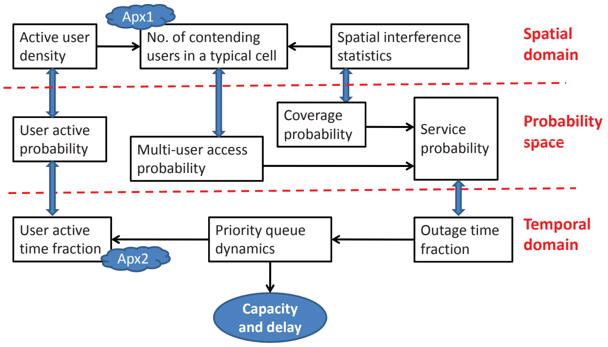

Fig. 3. Connections among parameters in the spatial and temporal domains.

Each outage event contributes to an additive random outage duration denoted by β_o, the mean of which is $\bar{\beta}_o$. Let us define

$$\rho_o = \bar{\beta}_o/\bar{\alpha}_o. \qquad (6)$$

This parameter represents the fraction of time that a user is in outage and cannot be served by a BS in all bands. It is worth noting that we do not make any particular assumption on the distribution of β_o, i.e., it can follow an arbitrary form of continuous distribution. This gives our model the flexibility to represent a wide range of outage phenomenons. We note that in practice, schemes such as packet-wise vertical handover [42] can be used to reduce the handover time to a negligible level. In addition, our model can be refined to account for non-negligible handover time by making the random outage duration β_o to be dependent on the handover rates.

We consider the secondary traffic behavior at the session level. Users are assumed to have homogeneous incoming traffic of sessions that follow i.i.d. Poisson arrival process with mean interval $\bar{\alpha}_s$. Each session carries a file of random size L to be delivered from the BS to the user. The file size L follows a general distribution with mean $\bar{L}$. The mean throughput capacity of a user is given by

$$C = \bar{L}/\bar{\alpha}_s. \qquad (7)$$

Under the assumption of constant transmission rate R, the transmission time of a session is a random variable $\beta_s = L/R$. Let us define

$$\rho_s = \bar{\beta}_s/\bar{\alpha}_s = \bar{L}/(R\bar{\alpha}_s) = C/R. \qquad (8)$$

This parameter represents the fraction of time that a user receives transmission from a BS. The file size L is assumed to follow a general distribution.

The transmission of a secondary session is forced to stop immediately once an outage occurs. Once the secondary service is available again, a session may adapt a ‘resume’ policy to transmit from where it stopped, or adapt a ‘repeat’ policy to retransmit from the beginning. Our paper is restricted to the resume policy, noting that an extension to the repeat policy is straightforward. Based on the above modeling assumptions, the queuing process at a typical secondary user can be captured by a M/G/1 two-level priority queuing model with a preemptive resume policy [43]. The queuing model can be fully characterized by the four random variables shown in Fig. 2.

For the convenience of readers, Table I summarizes the major symbols in our system model.

III. Methodology and Approximations

Our system model describes a large scale, dynamic system in the spatial and temporal domains. These two domains are inherently coupled and correlated. Existing work resorted to static simulation to yield results without revealing much theoretical insight [34]. In this paper, instead of trying to capture the detailed relationships between the spatial and temporal domains, we propose a methodology that connects these two domains by establishing analytical relationships among the first-order statistic measure (i.e., mean values) of some critical parameters. When higher order statistics are in concern, we can still use the M/G/1 queueing model (with general distributions) to offer sufficient flexibility to match practical measurements. This section will first explain our overall methodology and then introduce some useful approximations as preliminaries.

A. Overall methodology

Fig. 3 illustrates our overall approach to address the connections between spatial and temporal domains. Our analysis implies two underlying assumptions. First, the queueing processes of users are assumed to be independent and homogeneous. This assumption is reasonable because in the macro time-scale, users are assumed to have independent mobility traces; while in the micro time-scale, users are allowed to hop randomly between independent bands. The composite effects of random mobility and band selection renders the queueing process of a user to be independent from others in the long term. In this case, we can consider a typical user with a typical queueing process, at a typical location and associated with a typical BS. A typical user can be understood as an arbitrary user or a randomly selected user. A probability space can then be defined for the typical user for its status. The second assumption is that all BSs constantly transmit with power P_n. This assumption decouples the interference statistics with user behavior and represents the worst-case interfering scenario. It is reasonable because the combined load of primary and secondary traffic is likely to keep BSs busy.

TABLE I
MAJOR SYMBOLS AND THEIR PHYSICAL MEANINGS

Symbol	Physical meaning of the symbol
N	Number of frequency bands
R	Minimum required transmission rate for secondary traffic (bits/s)
L	Random size of the file carried by a secondary session (bits)
C	Throughput of a secondary user (bits/s)
W_n	Bandwidth of the nth band (Hz)
Ω_n	Vacant (i.e., no primary traffic) probability of the nth band
ε_n	Probability for a secondary user to receive service on the nth band
ε, $p_{service}$	Probability for a secondary user to receive service (on any band)
$p_{n,v}$	Probability for a secondary user to have sufficient signal coverage on the nth band
$p_{n,a}$	Probability for a secondary user to successfully content for multiple access on the nth band
p_{active}	Probability that a secondary user is active (i.e., have buffered secondary traffic)
λ_u and λ_b	Densities of all users and all BSs, respectively (m^{-2})
$\lambda_{u,n}$ and $\lambda_{b,n}$	Densities of users and BSs operating on the nth band, respectively (m^{-2})
α_o and α_s	Random intervals between outage arrivals and secondary session arrivals, respectively
β_o and β_s	Random durations of outage and (uninterrupted) secondary session transmission, respectively
ρ_o and ρ_s	Time fractions of the outage process and secondary session queueing process, respectively

According to the ergodic theory, when the queueing process of the typical user has a statistical equilibrium, the queueing process is ergodic [43] and hence the time average of a queueing parameter is identical to the average over the probability space. This allows us to map time-domain parameters to the probability space. Moreover, according to the theory of Palm probability in stochastic geometry, the spatial average of a large scale network is identical to the probabilistic average over the typical user/BS [37]. This allows us to map spatial-domain parameters to the probability space. Based on these mappings, we are able to deduce a chain of relations in Fig. 3 as follows.

Let us consider the *outage time fraction* in a typical queue, which is the average fraction of time that secondary services is not available. The *outage time fraction* affects the queueing dynamics and hence the *user active time fraction*, which is the average fraction of time that there is secondary traffic buffered in the queue. The *user active time fraction* is identical to the *active probability* of a typical user, which affects the *active user density* in the spatial domain. *Active user density* and *spatial interference statistics* both affect the distribution of the *number of contending users in a typical cell*, which determines the *multi-user access probability*. *Spatial interference statistics* also affects the *coverage probability* of a typical user. Moreover, as shown in (1), the *access probability* and *coverage probability* affects the *service probability*, which ultimately determines the *outage time fraction*. In other words, we have

$$\varepsilon = 1 - \rho_o \tag{9}$$

where ε is the service probability, ρ_o is the outage time fraction, and ρ_o can be expressed as a function of ε. The above chain of relations allows us to establish an equilibrium equation that connects first-order statistics of multiple parameters in the spatial and temporal domains. To establish the equation in an analytical form, two approximations are further introduced.

B. Approximation to the number of in-coverage users in a typical cell

The PDF of the size of a typical Poisson Voronoi cell is analytically intractable but can be approximated using the Monte Carlo method. Let λ be the density of the underlying Poisson process and V denote the random size of a typical Voronoi cell normalized by $1/\lambda$. The PDF of V is given by [39]

$$f_V(x) = \frac{3.5^{3.5}}{\Gamma(3.5)} x^{2.5} e^{-3.5x} \tag{10}$$

where $\Gamma(\cdot)$ is the gamma function. Moreover, consider an arbitrary user and the random size U of the Voronoi cell to which the user belongs to. The PDF of U normalized by $1/\lambda$ is given by [40]

$$f_U(x) = \frac{3.5^{4.5}}{\Gamma(4.5)} x^{3.5} e^{-3.5x}. \tag{11}$$

The difference between $f_V(x)$ and $f_U(x)$ comes from the fact that a user has a higher chance to be covered by larger Voronoi cells.

Let us consider a single band of the network with BS density λ_b and user density λ_u. Denoting K_1 as the random number of users in a non-empty Voronoi cell, the probability mass function (PMF) of K_1 is given by

$$f_{K_1}(k) = \int_0^\infty \frac{(\frac{\lambda_u}{\lambda_b} x)^k}{k!} e^{-\frac{\lambda_u}{\lambda_b} x} f_U(x) dx. \tag{12}$$

Let K be the random number of 'in-coverage' users in a Voronoi cell. The distribution of K is related to the size and shape of the cell and it is difficult to obtain its exact PMF. Keeping the basic form of (12), we propose an approximation to the PMF of K given by

$$\begin{aligned} f_K(k) &\approx \int_0^\infty \frac{(p\Lambda \frac{\lambda_u}{\lambda_b} x)^k}{k!} e^{-p\Lambda \frac{\lambda_u}{\lambda_b} x} f_U(x) dx \qquad (13) \\ &= \frac{3.5^{4.5}\Gamma(4.5+k)}{\Gamma(4.5)k!} \frac{(\Lambda\lambda_u p/\lambda_b)^k}{(3.5+\Lambda\lambda_u p/\lambda_b)^{4.5+k}}. \end{aligned}$$

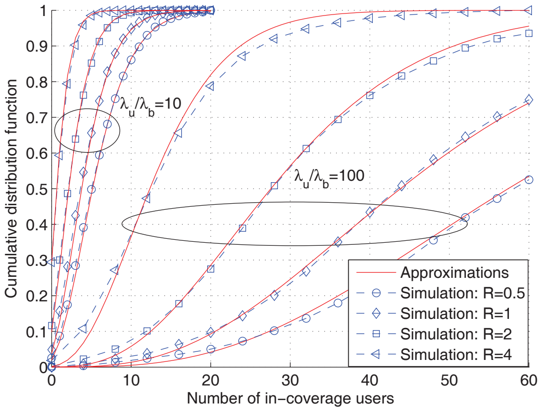

Fig. 4. Approximation on the probability density function of in-coverage users in a typical cell ($\lambda_b = 10^{-6}$).

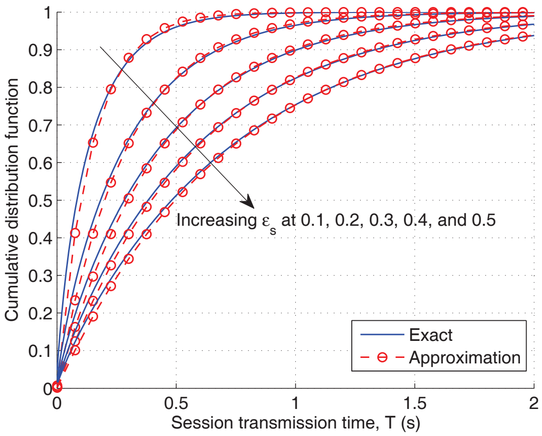

Fig. 5. Exponential approximation for the CDF of session transmission time T ($\bar{\alpha}_o = 0.1$, $\bar{\alpha}_s = 1$, $\rho_o = 0.3$).

where the parameters p and Λ are introduced to capture the effect of colored thinning on the original user point process. Here, p is the probability that an arbitrary user falls within coverage (with target rate R) and can be calculated by (5). The coefficient Λ is an artificial constant to capture the effect of colored thinning. The value of Λ is obtained by searching for the best fit of (13) to the empirical PMF obtained via Monte Carlo simulations. Through extensive simulations, we find that given $\Lambda = 2/3$, the approximation in (13) is valid for a wide range of practical values for λ_u and λ_b. Fig. 4 illustrates the accuracy of this approximation.

C. Approximation to user active time fraction

A user is active when there are sessions buffered or being transmitted in the queue. We are interested in the probability p_{active} that a typical user stays active. This probability also represents the fraction of time for a user to be active. Let T be the total transmission time of a session (including interrupted time). The mean value of T is given by [43]

$$\bar{T} = \frac{\bar{\beta}_s}{1 - \rho_o}. \tag{14}$$

The exact PDF of T is not exponential, but for the purpose of calculating the user active probability, we assume that T follows an exponential distribution with mean $\bar{T}$. The accuracy of this approximation is illustrated in Fig. 5, where we assume exponentially distributed β_o and β_s, set $\bar{\alpha}_o = 0.1$, $\bar{\alpha}_s = 1$, $\varepsilon_o = 0.3$, and let ε_s varies from 0.1 to 0.5. The exact PDF of T is obtained from its Laplace transform $\mathfrak{L}_T(s)$, which is given by [43]

$$\mathfrak{L}_T(s) = \mathfrak{L}_{\beta_s}[K(s)] \tag{15}$$

where $\mathfrak{L}_{\beta_s}(\cdot)$ is the Laplace transform of β_s and

$$K(s) = s + \frac{1 - G(s)}{\bar{\alpha}_o}. \tag{16}$$

Here, $G(s)$ is the solution with the smallest absolute value that satisfies the following equation

$$x - \mathfrak{L}_{\beta_o}\left(s + \frac{1-x}{\bar{\alpha}_o}\right) = 0 \tag{17}$$

where $\mathfrak{L}_{\beta_o}(\cdot)$ is the Laplace transform of β_o.

We find that the exponential approximation is valid under the condition that the arrival rate of outage is greater than the arrival rate of secondary traffic session. This condition is realistic because our system model considers the secondary traffic delay at the session level, which has a larger time scale than outages caused by packet-level primary traffic.

Now let us consider a discrete-value stochastic process representing the number of sessions staying in the queue. Based on the above mentioned exponential approximation, it is easy to see that this process is a classic birth-death process [43] characterized by an uniform birth rate $1/\bar{\alpha}_s$ and death rate $1/\bar{T}$. Let ϕ_k (k = 0,1,2,3...) denote the steady state probability that there are k sessions in the queue. The equilibrium condition of the birth-death process gives $\phi_k = (\bar{T}/\bar{\alpha}_s)^k \phi_0$. By further considering the constraint of total probability $\Sigma_{k=0}^{\infty}\phi_k = 1$, we have $\phi_0 = 1 - \bar{T}/\bar{\alpha}_s$. It follows that

$$p_{active} = 1 - \phi_0 = \bar{T}/\bar{\alpha}_s = \frac{\bar{\beta}_s}{(1-\rho_o)\bar{\alpha}_s} = \frac{\rho_s}{1-\rho_o} = \frac{\rho_s}{\varepsilon}. \tag{18}$$

IV. Capacity-Delay Tradeoff Analysis

A. Useful Lemmas

Lemma 1: Let ε_n denote the probability that a user can be served with a target rate R by the nearest BS operating in the nth band. We have

$$\varepsilon_n = \frac{\Omega_n}{\Lambda\lambda_n}\left[1 - \left(1 + \frac{\Lambda p_{n,v}\lambda_n}{3.5}\right)^{-3.5}\right] \tag{19}$$

where $p_{n,v}$ is given by (5), $\Lambda = 2/3$, and

$$\lambda_n = \frac{\lambda_u}{\lambda_{b,n}} \cdot \frac{\rho_s}{\varepsilon} \cdot \frac{\Omega_n p_{n,v}}{\sum_{n=1}^{N}\Omega_n p_{n,v}}. \tag{20}$$

Proof: See Appendix I. ■

Lemma 2: When the cellular system independently operates a total number of N bands, the probability that a user can be served by at least one band with a targeted rate R is

$$\varepsilon = 1 - \prod_{n=1}^{N}(1 - \varepsilon_n). \tag{21}$$

Proof: It is straightforward to see that $(1-\varepsilon)$ equals the joint probability that all bands fail to provide service to a user with the target rate R. ■

According to Lemma 1, ε_n is itself a function of ε. Therefore Lemma 2 gives a non-linear equation of ε, based on which the value of ε can be calculated by solving the non-linear equation via numerical methods. In the special case that all bands have the same characteristics in terms of bandwidth, transmit power, and availability, (21) can be simplified to

$$\varepsilon_N = 1-(1-\varepsilon_n)^N. \tag{22}$$

In case when $N = 1$, (21) can be solved to give ε as an explicit function related to capacity C and target rate R as follows

$$\varepsilon_1 = \frac{p_{n,v}}{3.5}\frac{\Lambda\lambda_u}{\lambda_{b,n}}\frac{C}{R}\left[1-\left(1-\frac{\Lambda\lambda_u C}{\Omega_n\lambda_{b,n}R}\right)^{-2/7}\right]^{-1}. \tag{23}$$

B. General results for capacity-delay tradeoff

Once the value of ε is obtained, we can evaluate the mean delay and delay distribution of a session. Established results for two-class M/G/1 priority queues with preemptive-resume policy [43] can be directly applied to give the following two propositions.

Proposition 1: The mean delay of a session is given by

$$\bar{D} = \frac{1}{2\varepsilon(\varepsilon-\frac{C}{R})}\left(\frac{\hat{\beta}_s}{\bar{\alpha}_s}+\frac{\hat{\beta}_o}{\bar{\alpha}_o}\right)+\frac{\bar{L}}{R\varepsilon} \tag{24}$$

where $\hat{\beta}_s$ and $\hat{\beta}_o$ are the second-order moments of random variables β_s and β_o, respectively.

The delay of a session is the total time the session spends in the queue and consists of two parts. The first part is waiting time W, which is the duration from the moment of arrival to the moment when the transmission starts. The second part is transmission time T, which is the duration from the moment when transmission starts to the moment when the transmission ends. It follows that $D = W + T$, where W and T are independent RVs [43]. The PDF of D cannot be obtained directly. However, the Laplace transforms of the PDFs of W and T can be evaluated. Let $\mathfrak{L}_X(\cdot)$ denote the Laplace transform to the PDF of random variable X, we have the following proposition.

Proposition 2: The Laplace transform of the random delay D of a typical session is given by

$$\mathfrak{L}_D(s) = \mathfrak{L}_T(s)\mathfrak{L}_W(s). \tag{25}$$

Here, $\mathfrak{L}_T(s)$ is given by (15). The second term $\mathfrak{L}_W(s)$ in (25) is given by

$$\mathfrak{L}_W(s) = (1-\rho_o-\rho_s)\bar{\alpha}_s\frac{K(s)}{\mathfrak{L}_{\beta_s}[K(s)]+\bar{\alpha}_s s-1}. \tag{26}$$

C. Capacity-delay tradeoff in special cases

1) Exponential distribution: Propositions 1 and 2 are applicable when both the file size L and outage duration β_o follow general distributions. In the special case where both L and β_o follow exponential distributions, we have $\hat{\beta}_s = 2(\bar{\beta}_s)^2$ and $\hat{\beta}_o = 2(\bar{\beta}_o)^2$. The mean delay becomes

$$\bar{D} = \frac{1}{\varepsilon(\varepsilon-\frac{C}{R})}\left(\frac{C\bar{L}}{R^2}+(1-\varepsilon)^2\bar{\alpha}_o\right)+\frac{\bar{L}}{R\varepsilon}. \tag{27}$$

Moreover, given an exponential random variable $X \sim \exp(\bar{X})$, its Laplace transform is

$$\mathfrak{L}_{exp}(s) = \frac{1}{1+s\bar{X}}. \tag{28}$$

Based on (28), closed-form Laplace transforms of $\beta_s = L/R$ and β_o can be obtained in (15) and (17). It follows that Eqn. (17) can be solved explicitly to give

$$G(s) = \frac{\left(1+\varepsilon_o+s\bar{\beta}_o\right)-\sqrt{\left(1+\varepsilon_o+s\bar{\beta}_o\right)^2-4\varepsilon_o}}{2\varepsilon_o}. \tag{29}$$

2) Gamma distribution: A more general distribution we can consider for L and β_o is Gamma distribution, which provides more flexibility to model a variety of practical scenarios. The PDF of Gamma distribution is given by

$$\Gamma(k,\theta) = \frac{1}{\theta^k}\frac{1}{\Gamma(k)}t^{k-1}e^{-\frac{t}{\theta}} \tag{30}$$

where k and θ are the shape and scale parameters, respectively. The first and second moments of the Gamma distribution are $k\theta$ and $k(k+1)\theta^2$, respectively. Let $L \sim \Gamma(k_L, \bar{L}/k_L)$ and $\beta_o \sim \Gamma(k_{\beta_o}, \bar{\beta}_o/k_{\beta_o})$. Here we introduce two new parameters k_L and k_{β_o} to characterize the shape of distributions of L and β_o, respectively. It follows that $\beta_s = L/R \sim \Gamma(k_L, \bar{L}/(k_L R))$, and the mean delay in (24) becomes

$$\bar{D} = \frac{1}{2\varepsilon(\varepsilon-\frac{C}{R})}\left(\frac{C\bar{L}}{R^2}\frac{k_L+1}{k_L}+(1-\varepsilon)^2\bar{\alpha}_o\frac{k_{\beta_o}+1}{k_{\beta_o}}\right)+\frac{\bar{L}}{R\varepsilon}. \tag{31}$$

It is easy to see that when $k_L = 1$ and $k_{\beta_o} = 1$, the Gamma distribution is reduced to exponential distribution and (31) is reduced to (27).

To evaluate the delay distribution, we have the Laplace transform of $G \sim \Gamma(k,\theta)$ given by

$$\mathfrak{L}_{gamma}(s) = (1+\theta s)^{-k}. \tag{32}$$

Based on (32), closed-form Laplace transforms of $\beta_s = L/R$ and β_o can be obtained according to (15) and (17). It follows that when k is an integer or a rational fraction, Eqn. (17) yields a polynomial form. Therefore the function $G(s)$ in (17) can be easily solved using existing root-finding algorithms for polynomials.

D. Simulation validation

This subsection aims to validate the previously derived theoretical results via Monte Carlo simulations. We note that a sufficient characterization of the spatial interference requires a large number of BSs and users (tens of thousands) to be simulated. The computational burden prohibits a full-scale, dynamic simulation of the queuing processes of all the users. We therefore adapt a methodology to simulate a typical user in a typical cell following the structure illustrated in Fig. 3. Our simulation includes two Monte-Carlo engines: a

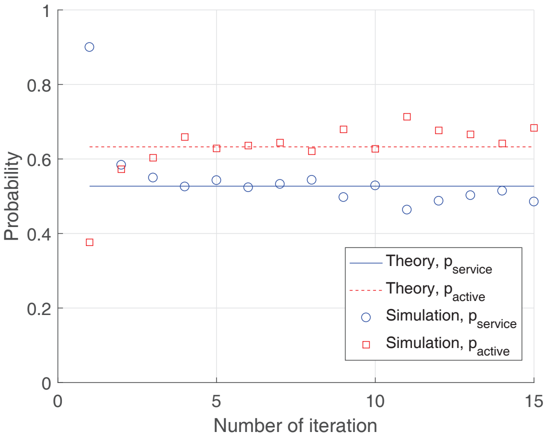

Fig. 6. Theoretical and simulation values of the active user probability p_{active} and service probability $p_{service}$ (λ_b = 10^{-6}/m^2, λ_u/λ_b=50, C=1 bit/s/Hz, R=3 bits/s, N=5, $\bar{L}$=10, $\bar{\alpha}_o$=1, L and α_o follow exponential distributions.)

temporal engine that simulates a preemptive-resume M/G/1 priority queueing process at a typical user, and a spatial engine that simulates the perceived interference and multi-user access process of a typical user located in a random cell. In each simulation run, both engines will conduct a large number of Monte-Carlo experiments until the performances are converged. The temporal engine then outputs the empirical user active probability (i.e., p_{active}) as an input of the spatial engine, while the spatial engine outputs the empirical service probability (i.e., $p_{service}$ or ε) as an input of the temporal engine. Such an iteration process stops after a few simulation runs when the engine outputs are converged. Without loss of generality, Fig. 6 illustrates the iteration process with a typical parameter setting and compares the two empirical probabilities with their theoretical counterparts calculated by Eqns. (18) and (22). It the observed that the empirical probabilities are able to converge to the theoretical ones. The fluctuations of the empirical probabilities are caused by random deviations of the Monte-Carlo simulation engines. The simulation validates that given the modeling assumptions described in Section II, our approximations in Section III and theoretical derivations in Section IV are accurate.

V. Capacity Limit and Scaling

This section studies the fundamental capacity limit at the interference limited regime and investigates how the capacity limit scales with bandwidth and user-BS density ratio. The capacity limit is defined as the maximum capacity that permits a stable queue at a typical user. It is also the capacity that gives infinite mean delay. Interference-limited regime means that power P_n is sufficiently large to justify the closed-form SINR CCDF in (4). For simplicity, we assume that the N bands have homogeneous characteristics in terms of bandwidth and BS density. Two different cases are considered. The first case assumes a fixed bandwidth of each band, which means the system bandwidth scales linearly with N. This case is useful when we want to investigate the impact of spectrum aggregation on the system capacity. The second case assumes a fixed system bandwidth, which means the bandwidth per band is inversely proportional to N. This case is relevant when we are interested in the impacts of spectrum sharing and channelization on the system capacity. Throughout this section, we use the capital letter 'N' as the footnote of parameters to emphasize that we consider homogeneous bands. For example, W_n, ε_n and λ_n are replaced by W_N, ε_N and λ_N, respectively.

A. Fixed bandwidth per band

Proposition 3: In the case of fixed bandwidth per band, the capacity limit $C_I^{\lim}$ is a function of R, λ_u, λ_b, and N given by

$$C_I^{\lim} = R\left[1 - (1 - \varepsilon_N)^N\right] \tag{33}$$

where

$$\varepsilon_N = \frac{\Omega_N}{\Lambda \lambda_N}\left[1 - \left(1 + \frac{\Lambda \lambda_N p_N^I}{3.5}\right)^{-3.5}\right]. \tag{34}$$

Here, p_N^I is given by

$$p_N^I = \left(1 + \sqrt{2^{R/W_N} - 1}\arctan\sqrt{2^{R/W_N} - 1}\right)^{-1} \tag{35}$$

and $\lambda_N = \lambda_u/(\lambda_{b,n}N)$.

Proof: A stable queue requires $1 - \rho_o - \rho_s > 0$, which gives $\varepsilon > \rho_s = C/R$. The capacity limit is achieved when the equality holds, i.e., $\varepsilon = C/R$ or $\rho_d/(1 - \rho_o) = 1$. Substituting this equation into Lemma 1 yields ε_N in (34). ■

We note that by considering the limiting condition, ε_N can be expressed as an explicit function of other parameters (as opposed to numerically solving a non-linear equation in Lemma 2). This allows us to express the capacity limit as a closed-form function of R, N, λ_u, and λ_b, as shown in (33). In the case of fixed bandwidth per band, we are interested in the following optimization problem: given N and the network environment λ_u and λ_b, how can we choose a proper target rate R to maximize the capacity limit? This optimization problem can be formally stated as $C_I^{\max} = \max\limits_R(C_I^{\lim})$. To better understand the nature of this optimization problem, representative numerical examples are presented in Fig. 7 to show $C_I^{\lim}$ as a function of R. We see that there is an unique maximum value of $C_I^{\lim}$, which is achieved when the first-order derivative $dC_I^{\lim}/dR$ equals zero. According to Proposition 3, the derivative function $dC_I^{\lim}/dR$ can be obtained in closed-form to give the following corollary.

Corollary 1: The optimum value R for the optimization problem $C_I^{\max} = \max\limits_R(C_I^{\lim})$ is given by the root of the following non-linear equation:

$$\frac{dC_I^{\lim}}{dR} = Rf_o^{'}(R) + f_0(R) = 0 \tag{36}$$

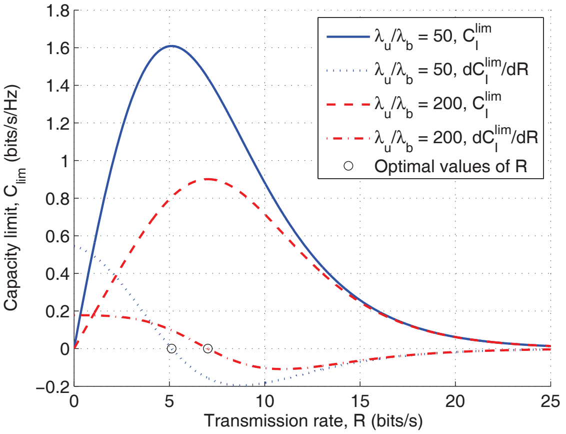

Fig. 7. Capacity limit C_I^{lim} and its first-order derivative as a function of R (fixed bandwidth per band, N=5).

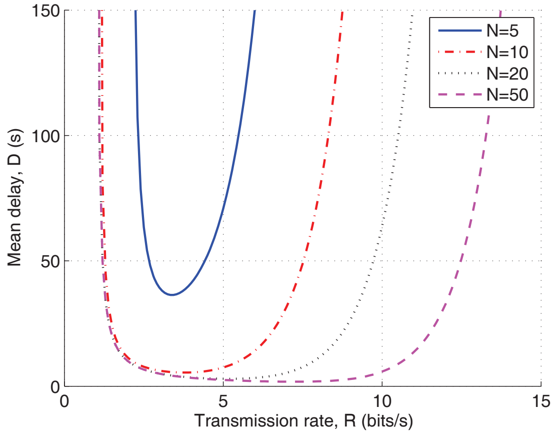

Fig. 8. Mean delay $\bar{D}$ as a function of R with varying N (fixed bandwidth per band, C=1 bit/s/Hz, λ_u/λ_b=50, $\bar{L}$=10, $\bar{\alpha}_o$=10).

where

$$f_0(R) = 1 - (1-\varepsilon_N)^N \tag{37}$$

$$f_0^{'}(R) = f_1(R)\cdot f_2(R)\cdot f_3(R)\cdot f_4(R) \tag{38}$$

$$f_1(R) = N(1-\varepsilon_N)^{N-1} \tag{39}$$

$$f_2(R) = \left(1 + \frac{\lambda_N p_N^I}{3.5}\right)^{-4.5} \tag{40}$$

$$f_3(R) = -\frac{\arctan(\chi_N) + (1+\chi_N^2)^(-1)}{(1+\chi_N \arctan(\chi_N))^2} \tag{41}$$

$$f_4(R) = \frac{\ln 2}{2} 2^R \left(2^R - 1\right)^{-1/2} \tag{42}$$

$$\chi_N = \sqrt{2^R - 1}. \tag{43}$$

In the above equations, p_N^I is defined in (35) and ε_N is defined in (34).

Based on the above corollary, the first-order derivative function dC_I^{lim}/dR is calculated and shown in Fig. 7. The root is obtained by solving the non-linear equation and shown to be accurate for achieving the maximum value of C_I^{lim}.

B. Fixed system bandwidth

In this case, the total system bandwidth is normalized to 1 and the bandwidth of each band becomes $1/N$. Define the capacity limit C_{II}^{lim} as the maximum achievable capacity for a stable queue given R, N, λ_u, and λ_b. Further define the maximum capacity as $C_{II}^{\max} = \max_R(C_{II}^{\mathrm{lim}})$. We have the following two propositions.

Proposition 4: The capacity limit C_{II}^{lim} can be calculated according to Proposition 3 by replacing p_N^I with p_N^{II}, where

$$p_N^{II} = \left(1 + \sqrt{2^{RN/W_N} - 1}\arctan\sqrt{2^{RN/W_N} - 1}\right)^{-1}. \tag{44}$$

Proof: The proof is straightforward by following the proof of Proposition 3 and setting the channel bandwidth to $1/N$. ■

Proposition 5: The maximum capacity is given by

$$C_{II}^{\max} = C_I^{\max}/N \tag{45}$$

where $C_I^{\max}$ can be calculated from Corollary 1.

Proof: According to Propositions 3 and 4, we can write $C_{II}^{\mathrm{lim}}(R) = C_I^{\mathrm{lim}}(RN)/N$. Further considering the fact that adding a scaling on R will not change the maximum value of C_I^{lim}, i.e., $\max_R C_I^{\mathrm{lim}}(R) = \max_R C_I^{\mathrm{lim}}(RN) = C_I^{\max}$, Proposition 5 can be proved. ■

VI. Numerical Results and Discussions

This section presents numerical results and discusses their implications. First, we aim to understand the impacts of various parameters on the capacity-delay tradeoff (Fig. 8 to Fig. 12). Second, we want to investigate how the fundamental capacity limit scales with the number of bands N and user-BS density ratio (Fig. 13 and Fig. 14). For illustration purpose, we consider an interference-limited system and homogeneous bands with $W_N = 1$ and $\Omega_N = 1$.

A. Capacity-delay tradeoff

Due to page limits, we restrict our discussions to the mean delay and the case of fixed bandwidth per band. Except when otherwise mentioned, the default parameter values are set to be $N = 5$, $\lambda_u/\lambda_b = 50$, $\bar{L} = 10$, and $\bar{\alpha}_o = 10$. Moreover, the distributions of L and α_o are treated as exponential. Therefore, our subsequent discussions are primarily based on Eqn. (27).

Fig. 8 shows the mean delay $\bar{D}$ as a function of R with varying N while the capacity is fixed to $C = 1$ bits/s. U-shape curves are observed, indicating that given other parameters, there is an optimal value for R to minimize the mean delay. Because we are interested in the fundamental capacity-delay tradeoff, it is desirable to consider the minimized delay over feasible values of R. Define $\bar{D}_{\min} = \min_R(\bar{D})$, we will subsequently evaluate $\bar{D}_{\min}$ as a function of C. The value of $\bar{D}_{\min}$ is obtained by performing a numerical optimization over R.

Fig. 9 shows the impact of λ_u/λ_b on the capacity-delay tradeoff curve. Two interesting phenomena are observed. First, when the user-BS density ratio is relatively high ($100 \le \lambda_u/\lambda_b \le 1000$), the capacity per user (at a fixed delay) appears

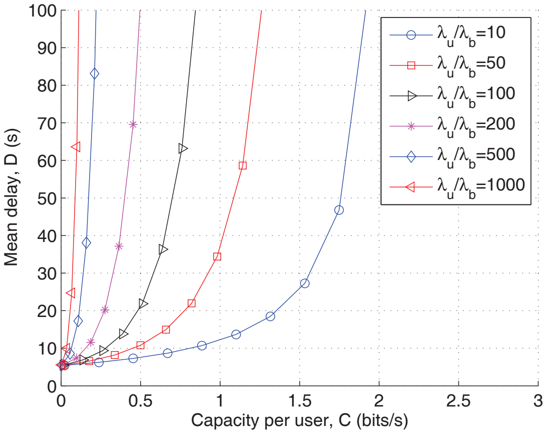

Fig. 9. Mean delay $\bar{D}$ as a function of per user capacity C with varying λ_u/λ_b (fixed bandwidth per band, N=5, $\bar{L}$=10, $\bar{\alpha}_o$=10).

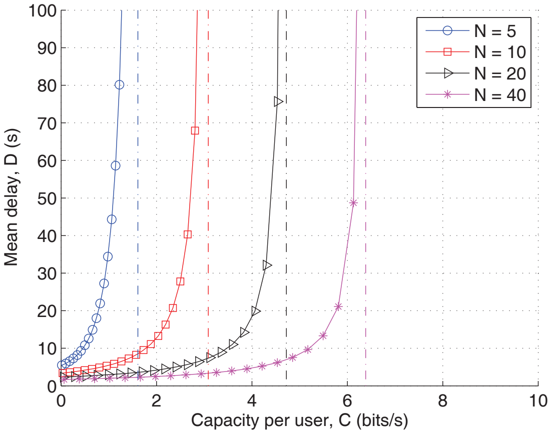

Fig. 10. Mean delay $\bar{D}$ as a function of per user capacity C with varying N (fixed bandwidth per band, λ_u/λ_b=50, $\bar{L}$=10, $\bar{\alpha}_o$=10).

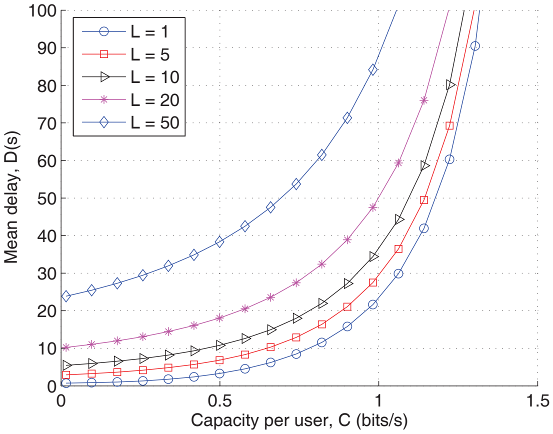

Fig. 11. Mean delay $\bar{D}$ as a function of per user capacity C with varying $\bar{L}$ (fixed bandwidth per band, λ_u/λ_b=50, N=5, $\bar{\alpha}_o$=10).

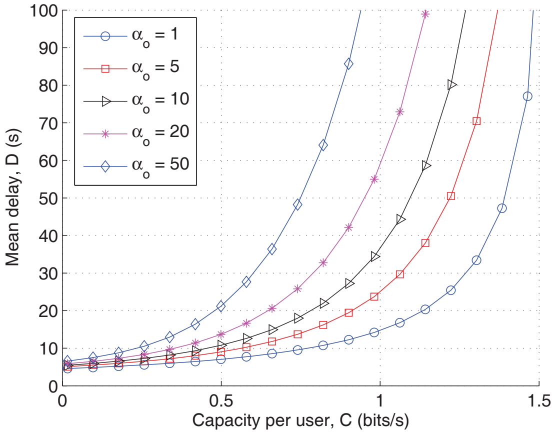

Fig. 12. Mean delay $\bar{D}$ as a function of per user capacity C with varying $\bar{\alpha}_o$ (fixed bandwidth per band, λ_u/λ_b=50, N=5, $\bar{L}$=10).

to scale linearly with λ_b/λ_u. We called this "infrastructure-limited" regime, in which the investment in BS infrastructure yields linear returns on the capacity. In contrast, when the user-BS density is relatively low ($10 \leq \lambda_u/\lambda_b \leq 100$), investment in BS infrastructure only yields sub-linear returns. Second, in the low delay regime, there is minimum delay even when C approaches zero. Such a minimum delay is caused by coverage outage and primary traffic interruption, which caps the secondary service probability.

Fig. 10 shows the impact of the number of channels N on the capacity-delay tradeoff curve. The capacity limits with respect to different values of N are also shown. The delays are shown to rise quickly when C approaches the capacity limits. It is observed that in the medium to high delay regime, capacity at a fixed delay scales linearly with N. In the low delay regime, increasing N contributes slightly to reducing the minimum delay. Fig. 10 indicates that spectrum aggregation is effective for both capacity enhancement and delay reduction.

Fig. 11 shows the impact of average file size $\bar{L}$ on the capacity-delay tradeoff curve. The capacity limit is also shown, which is unrelated to the value of $\bar{L}$. In the low to medium capacity regime, $\bar{L}$ is shown to have a significant effect on the delay. A smaller value of $\bar{L}$ leads to a smaller delay because the file transmission has a lower probability of being interrupted by an outage. In the high delay regime, the impact of $\bar{L}$ diminishes as all delay curves eventually converge to the capacity limit. Fig. 11 suggests that file/session size management is an important factor to consider if a system is designed for low delay performance.

Fig. 12 shows the impact of mean outage arrival interval $\bar{\alpha}_o$ on the capacity-delay tradeoff curve. The capacity limit, which is independent from the values of $\bar{\alpha}_o$, is also shown. In the low delay regime, the curves converge to a minimum delay. In the high delay regime, we can predict that the curves also slowly converge to the capacity limit. However, significant differences are observed in the low to medium delay regimes. A smaller value of $\bar{\alpha}_o$ leads to smaller delays. This is because an interrupted session is less likely to be prolonged for a long period. Fig. 12 implies that introducing extra dynamics into the system (such as dynamic scheduling) can potentially help to reduce the delay.

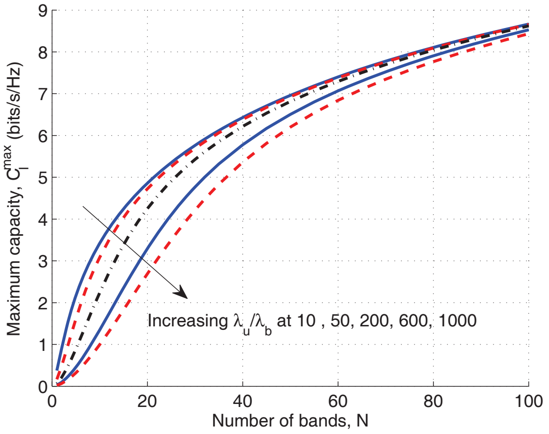

Fig. 13. Maximum capacity $C_I^{\max}$ as a function of N with varying λ_u/λ_b (fixed bandwidth per band).

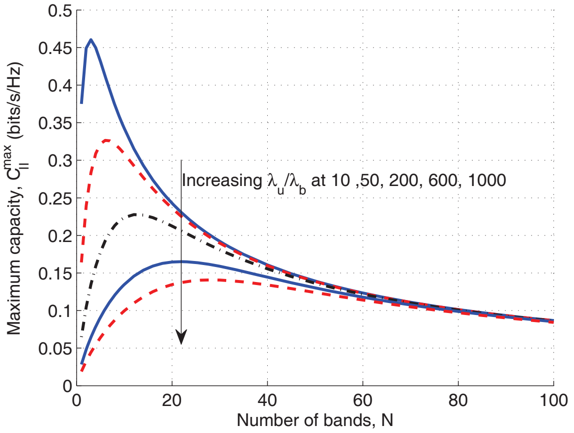

Fig. 14. Maximum capacity $C_{II}^{\max}$ as a function of N with varying λ_u/λ_b (fixed system bandwidth).

B. Capacity limit and scaling

This subsection investigates how the capacity limit scales with N and user-BS density ratio. Consider the case of fixed bandwidth per band, Fig. 13 applies Corollary 1 to show the maximum capacity $C_I^{\max}$ as a function of N with varying λ_u/λ_b. We see that the capacity increases monotonically with increasing N, indicating the benefits of spectrum aggregation. However, increasing N shows diminishing returns on the capacity gain. It is also interesting to observe that the curves with different values of λ_u/λ_b converge to the same value when N becomes large. These observations differ from the common intuition that user capacity scales linearly with system bandwidth (i.e., the number of bands). The reason for this counter-intuitive result is because we assume that a user is allowed to access only one band. The capacity per user depends on the bandwidth per band and the (successful) multi-user access probability. When N is relatively small compared to the average number of users per cell, the multi-user access probability scales roughly linearly with N. However, when N tends large, the multi-user access probability saturates to one and the capacity is limited by the bandwidth per band rather than the number of bands. Fig. 13 suggests that to achieve the full potential of spectrum aggregation, it is important to allow users to access multiple bands simultaneously, although this would introduce extra hardware cost and power consumption.

Considering the case of fixed system bandwidth, Fig. 14 applies Proposition 5 to show the maximum capacity $C_{II}^{\max}$ as a function of N with varying λ_u/λ_b. It is shown that with increasing N, the capacity increases initially but eventually declines. For each value of λ_u/λ_b, there exists an optimal value of N to maximize the capacity. Fig. 14 reveals a design tradeoff between maximizing single channel capacity and maximizing multi-user access probability. It implies that proper channelization of the available spectrum resource is important, particularly when λ_u/λ_b is small. By performing a numerical search for the optimal value of N based on results in Fig. 14, Table II shows the corresponding maximum values of $C_{II}^{\max}$ as a function of λ_u/λ_b. We find that there exists a convenient approximation given by

$$C_{II}^{*\max} \approx 0.6359 - 0.052 \log_2(\lambda_u/\lambda_b). \tag{46}$$

The actual values obtained from numerical calculation and the approximated values obtained from (46) are compared in Table II. It is shown that the approximation is reasonably accurate for $2 < \lambda_u/\lambda_b < 500$. In addition, we find that a convenient approximation exists to give the optimal value of N as $N = \lceil\sqrt{\lambda_u/\lambda_b}\rceil$, where $\lceil\cdot\rceil$ is the ceiling function. The accuracy of this approximation is also shown in Table II. It shows that the optimal number of channels is roughly proportional to the square-root of the user-BS density ratio. This observation provides a useful guideline for system designers in practice.

C. Discussions and future work

Finally, we would like to address the aspect of modeling accuracy and limitations. The proposed analytical model in this paper is based on an integration of two well-established models: the spatial Poisson Point Process model and the temporal M/G/1 queueing model. The accuracies of these two models have been evaluated against real-world measurement data in [38] and [45]. We note that more realistic models are also available, such as clustered Poisson Point Process [37] and Ginibre point process [25] in the spatial domain, G/G/1 queue [43] and self-similarity traffic models [44] in the temporal domain. Providing analytically tractable results based on these realistic models is challenging and will be considered in our future work. Other directions of future work include considering more advanced secondary access protocols, evaluating the energy-efficiency, and addressing practical aspects such as channel sensing and handover.

VII. Conclusions

An analytical framework has been proposed for the study of the capacity-delay tradeoff in cellular networks with spectrum aggregation. The framework compliments existing ones by focusing on the secondary traffic and offering tractable analytical insights. Analytical results have been derived to characterize

TABLE II
APPROXIMATIONS OF $C_{II}^{\max}$ AND OPTIMAL N AS FUNCTIONS OF USER-BS DENSITY RATIO

User-BS density ratio $\log_2(\lambda_u/\lambda_b)$	3	4	5	6	7	8	9	10
Actual $C_{II}^{\max}$	0.677	0.4194	0.3625	0.3077	0.2575	0.2125	0.1733	0.1397
Approximated $C_{II}^{\max}$	0.4799	0.4279	0.3759	0.3239	0.2719	0.2199	0.1679	0.1159
Actual N	3	4	5	7	10	14	18	26
Approximated N	3	4	6	8	12	16	23	32

the capacity-delay tradeoff and the fundamental capacity limit. Numerical studies have shown that while spectrum aggregation primarily affects the capacity in the high-delay regime, session size management and dynamic scheduling have bigger impacts on the capacity in the low delay regime. Moreover, when different bands have homogeneous configurations, it has been shown that the per user throughput per Hertz is upper bounded by a constant and reduces at a rate proportional to the logarithm of user-BS density ratio. Our analysis offers useful guidelines for providing novel secondary services over cellular networks to improve the overall capacity utilization.

PROOF FOR LEMMA 1

We assume that an active user randomly selects a band for access, in an equilibrium state, the density of users in a band is proportional to the area fraction of coverage of this band. The density of active users in the nth band is then given by

$$\lambda_{u,n} = \lambda_u \cdot p_{active} \cdot \frac{\Omega_n p_{n,v}}{\sum_{n=1}^{N} \Omega_n p_{n,v}} = \frac{\lambda_u \rho_s}{\varepsilon} \cdot \frac{\Omega_n p_{n,v}}{\sum_{n=1}^{N} \Omega_n p_{n,v}}. \quad (47)$$

Now consider an active user in band n, the number of contenting users in the same cell can be evaluated according to (13) with user density $\lambda_{u,n}$ and BS density $\lambda_{b,n}$. When strict fairness is assumed, the access probability of a user is given by

$$\begin{aligned} p_{n,a} &= \sum_{k=0}^{\infty} \frac{1}{k+1} f_K(k) \qquad (48) \\ &= \sum_{k=0}^{\infty} \frac{1}{k+1} \int_{k=0}^{\infty} \frac{(\lambda_n \Lambda p_{n,v} x)^k}{k!} e^{-\lambda_n \Lambda p_{n,v} x} f_U(x) dx \\ &= \sum_{k=0}^{\infty} \frac{1}{\lambda_n \Lambda p_{n,v} x} \left[\int_{k=1}^{\infty} \frac{(\lambda_n \Lambda p_{n,v} x)^k}{k!} \right] e^{-\lambda_n \Lambda p_{n,v} x} f_U(x) dx \\ &= \sum_{k=0}^{\infty} \frac{1}{\lambda_n \Lambda p_{n,v} x} \left(1 - e^{-\lambda_n \Lambda p_{n,v} x}\right) f_U(x) dx \\ &= \frac{3.5^{4.5}}{\Gamma(4.5)} \frac{1}{\lambda_n \Lambda p_{n,v}} \left[\int_0^{\infty} x^{2.5} \left(e^{-3.5x} - e^{(-3.5+\lambda_n \Lambda p_{n,v})x} \right) dx \right] \\ &= \frac{3.5^{4.5}}{\Gamma(4.5)} \frac{1}{\lambda_n \Lambda p_{n,v}} \left[\frac{\Gamma(3.5)}{3.5^{3.5}} - \frac{\Gamma(3.5)}{(3.5 + \lambda_n \Lambda p_{n,v})^{3.5}} \right] \\ &= \frac{1}{\lambda_n \Lambda p_{n,v}} \left[1 - \left(1 + \frac{\Lambda p_{n,v} \lambda_n}{3.5} \right)^{-3.5} \right]. \end{aligned}$$

Finally, Lemma 1 can be obtained by substituting (48) into (1).

REFERENCES

[1] D. Lee, S. Zhou, X. Zhong, Z. Niu, X. Zhou, and H. Zhang, "Spatial modeling of the traffic density in cellular networks," *IEEE Wireless Commun. Mag.*, vol. 21, no. 1, pp. 80-88, Feb. 2014.

[2] "Cisco visual networking index: Global mobile data traffic forecast update 2015-2020," Cisco White Paper, Feb. 2016.

[3] S. Zhou, J. Gong, Z. Zhou, W. Chen, and Z. Niu, "Green delivery: proactive content caching and push with energy-harvesting-based small cells," *IEEE Commun. Mag.*, vol. 53, no. 4, pp. 142–149, Apr. 2015.

[4] X. Wang, M. Chen, T. Taleb, A. Ksentini and V. C. M. Leung, "Cache in the air: exploiting content caching and delivery techniques for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 131–139, Feb. 2014.

[5] N. Niesen, D. Shah, and G. W. Wornell, "Caching in wireless networks," *IEEE Trans. Info. Theory*, vol. 58, no. 10, pp. 6524–6540, Oct. 2012.

[6] Y. Zhang, H. Lu, H. Wang, and X. Hong, "Cognitive cellular content delivery networks: cross-layer design and analysis," *Proc. VTC-Spring'16*, May 2016, Nanjing, China.

[7] H. S. Dhillon, R. K. Ganti, F. Baccelli, and J. G. Andrews, "Modeling and analysis of K-tier downlink heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 550–560, Apr. 2012.

[8] C.-X. Wang, F. Haider, X. Gao, X.-H. You, Y. Yang, D. Yuan, H. Aggoune, H. Haas, S. Fletcher, and E. Hepsaydir, "Cellular architecture and key technologies for 5G wireless communication networks," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 122–130, Feb. 2014.

[9] X. Ge, S. Tu, T. Han, Q. Li, and G. Mao, "Energy efficiency of small cell backhaul networks based on Gauss-Markov mobile models," *IET Netw.*, vol. 4, no. 2, pp. 158–167čňMar. 2015.

[10] M. Neely and E. Modiano, "Capacity and delay tradeoffs for ad-hoc mobile networks," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 1917–1937, June 2005.

[11] A. Gamal, J. Mammen, B. Prabhakar, and D. Shah, "Throughput-delay trade-Off in wireless networks Part I: The fluid model," *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2568–2592, Jun. 2006.

[12] A. Gamal, J. Mammen, B. Prabhakar, and D. Shah, "Throughput-delay trade-off in wireless networks part II: Constant-size packets," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 5111–5116, Nov. 2006.

[13] P. Li, C. Zhang, and Y. Fang, "Capacity and delay of hybrid wireless broadband access networks," *IEEE J. Sel. Area, Commun.*, vol. 27, no. 2, pp. 117-125, Feb. 2009.

[14] X. Ta, G. Mao, and B.D.O. Anderson, "On the giant component of wireless multihop networks in the presence of shadowing," *IEEE Trans. Veh. Technol.*, vol. 58, no. 9, pp. 5152–5163, Nov. 2009.

[15] A. J. Fehske and G. P. Fettweis, "Aggregation of variables in load models for cellular data networks," *Proceedings ICC 2012*, Ottawa, Canada, May 2012, pp. 5102–5107.

[16] I. Siomina and D. Yuan, "Analysis of cell load coupling for LTE network planning and optimization," *IEEE Trans. Wireless Commun.*, vol. 11, no. 6, pp. 2287–2297, June 2012.

[17] A. J. Fehske and G. P. Fettweis, "On flow level modeling of multi-cell wireless networks," *Proc. IEEE 11th Int. Model. Optim. Mobile Ad Hoc Wireless Netw.*, Tsukuba, Japan, May 2013, pp. 572–579.

[18] I.-H. Hou, V. Borkar, and P. R. Kumar, "A theory of QoS for wireless," *Proc. IEEE INFOCOM*, Rio de Janeiro, Brazil, Apr. 2009, pp. 486–494.

[19] S. Lashgari and A. S. Avestimehr, "Timely throughput of heterogeneous wireless networks: Fundamental limits and algorithms," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8414–8433, Dec. 2013.

[20] G. Zhang, T. Q. S. Quek, A. Huang and H. Shan, "Delay and reliability tradeoffs in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1101–1113, Feb. 2016.

[21] M. Di Renzo, A. Guidotti, and G. E. Corazza, "Average rate of downlink heterogeneous cellular networks over generalized fading channels: A stochastic geometry approach," *IEEE Trans. Commun.*, vol. 61, no. 7, pp. 3050–3071, Jul. 2013.

[22] A. Guo and M. Haenggi, “Spatial stochastic models and metrics for the structure of base stations in cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 11, pp. 5800–5812, Nov. 2013.

[23] X. Lin, J. G. Andrews, and A. Ghosh, “Modeling, analysis and design for carrier aggregation in heterogeneous cellular networks," *IEEE Trans. Commun.*, vol. 61, no. 9, pp. 4002–4015, Sep. 2013.

[24] X. Zhang and M. Haenggi, “A stochastic geometry analysis of inter-cell interference coordination and intra-cell diversity," *IEEE Trans. Wireless Commun.*, vol. 13, no. 12, pp. 6655–6669, Dec. 2014.

[25] N. Deng, W. Zhou, and M. Haenggi, “The Ginibre point process as a model for wireless networks with repulsion,” *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 107–121, Jan. 2015.

[26] V. Tumuluru, P. Wang, D. Niyato, and W. Song, “Performance analysis of cognitive radio spectrum access with prioritized traffic," *IEEE Trans. Veh. Technol.*, vol. 61, no. 4, pp. 1895–1906, May 2012.

[27] J. Wang, A. Huang, W. Wang, and T. Quek, “Admission control in cognitive radio networks with finite queue and user impatience," *IEEE Wireless Commun. Lett.*, vol. 2, no. 2, pp. 175–178, Apr. 2013.

[28] M. Rashid, M. Hossain, E. Hossain, and V. Bhargava, “Opportunistic spectrum scheduling for multiuser cognitive radio: a queueing analysis,” *IEEE Trans. Wireless Commun.*, vol. 8, no. 10, pp. 5259–5269, Oct. 2009.

[29] S. Lirio Castellanos-Lopez, F. Cruz-Perez, M. Rivero-Angeles, and G. Hernandez-Valdez, “Joint connection level and packet level analysis of cognitive radio networks with VoIP traffic," *IEEE J. Select. Areas Commun.*, vol. 32, no. 3, pp. 601–614, Mar. 2014.

[30] M. Haenggi, “The local delay in Poisson networks," *IEEE Trans. Inf. Theory*, vol. 59, no. 3, pp. 1788–1802, Mar. 2013.

[31] Z. Gong and M. Haenggi, “The local delay in mobile Poisson networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 9, pp. 4766–4777, Sep. 2013.

[32] Q. Liu, S. Zhou, and G. Giannakis, “Queuing with adaptive modulation and coding over wireless links: cross-layer analysis and design,” *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, pp. 1142–1153, May 2005.

[33] L. Le, E. Hossain, and A. Alfa, “Delay statistics and throughput performance for multi-rate wireless networks under multiuser diversity,” *IEEE Trans. Wireless Commun.*, vol. 5, no. 11, pp. 3234–3243, Nov. 2006.

[34] B. Baszczyszyn, M. Jovanovic and M. K. Karray, “Performance laws of large heterogeneous cellular networks,” *Proc. 13th Int’l Symp. Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, Mumbai, 2015, pp. 597–604.

[35] H. Ye, C. Liu, X, Hong, and H. Shi, “Uplink capacity-delay trade-off in hybrid cellular D2D networks with user collaboration,” *Proc. IEEE WPMC*, Nov. 2016, Shenzhen, China.

[36] L. Chen, W. Luo, C. Liu, X. Hong, and J. Shi, “Capacity-delay trade-off in collaborative hybrid ad-hoc networks with coverage sensing,” *MDPI Sensors*, under review.

[37] D. Stoyan, W. S. Kendall, and J. Mecke, *Stochastic Geometry and Its Applications*, 2nd Edition, Wiley, 2008.

[38] J. G. Andrews, F. Bacccelli, 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.

[39] J.-S. Ferenc and Z. Neda, “On the size distribution of Poisson Voronoi cells,” *Physica A: Statistical Mechanics and its Applications*, vol. 385, no. 2, pp. 518–526, 2007.

[40] S. M. Yu and S.-L. Kim, “Downlink capacity and base station density in cellular networks,” *Proc. WiOpt 2013*, Tsukuba Science City, 2013, pp. 119–124.

[41] X. Hong, Y. Jie, C. X. Wang, J. Shi and X. Ge, “Energy-spectral efficiency trade-off in virtual MIMO cellular systems,” *IEEE J. Sel. Areas Commun.*, vol. 31, no. 10, pp. 2128–2140, Oct. 2013.

[42] J.Gambini, O. Simeone, Y. Bar-Ness, U. Spagnolini, and T. Yu, “Packet-wise vertical handover for unlicensed multi-standard spectrum access with cognitive radios,” *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 5172–5176, Dec. 2008.

[43] J. W. Cohen, *The Single Server Queue*, North-Holland Publishing Company, 1982.

[44] A. Feldmann, A. C. Gilbert, P. Huang, and W. Willinger, “Dynamics of IP traffic: A study of the role of variability and the impact of control," *ACM SIGCOMM Computer Communication Review*, vol. 29, no. 4, pp. 301–313, Oct. 1999.

[45] S. B. Fredj, T. Bonald, A. Proutiere, G. Regnie, and J. W. Roberts, “Statistical bandwidth sharing: A study of congestion at flow level," *SIGCOMM’01*, vol. 31, no. 4, Aug. 2001, pp. 111–122.

Lingyu Chen received his B.S. degree in software engineering (2006) and the Ph.D. degree in communication and information systems (2011), both from Xiamen University. He is currently an assistant professor in the School of Information Science and Technology, Xiamen University, China. His research interests include vehicular and wireless sensor networks, acoustic sensor networks, edge computing, and wireless signal processing.

Liu Chen received his B.S. degree and M.Eng. degrees from Wuhan University, China, in 2010 and 2012, respectively. He is currently a Ph.D. candidate in the School of Information Science and Engineering, Xiamen University. His research interests include cognitive radio networks and fifth-generation mobile communications, with emphasis on cross-layer design and resource allocation.

Xuemin Hong (S’05–M’12) received his Ph.D. degree from Heriot-Watt University, UK, in 2008. He is currently a professor at Xiamen University, China. He has published one book chapter and over 60 papers in refereed journals and conference proceedings. His research interests include cognitive radio networks, wireless channel modeling, and fifth-generation mobile communications.

Cheng-Xiang Wang (S’01–M’05–SM’08–F’17) received the B.Sc. and M.Eng. degrees in communication and information systems from Shandong University, Jinan, China, in 1997 and 2000, respectively, and the Ph.D. degree in wireless communications from Aalborg University, Aalborg, Denmark, in 2004.

He was a Research Assistant with the Hamburg University of Technology, Hamburg, Germany, from 2000 to 2001, a Research Fellow at the University of Agder, Grimstad, Norway, from 2001 to 2005, and a Visiting Researcher with Siemens AG-Mobile Phones, Munich, Germany, in 2004. He has been with Heriot-Watt University, Edinburgh, U.K., since 2005, and became a Professor in wireless communications in 2011. He is also an Honorary Fellow at The University of Edinburgh, U.K., a Chair Professor of Shandong University, and a Guest Professor of Southeast University, China. He has co-authored two books, one book chapter, and over 320 papers in refereed journals and conference proceedings. His current research interests include wireless channel measurements/modeling and (B)5G wireless communication networks, including green communications, cognitive radio networks, high mobility communication networks, massive MIMO, millimeter wave communications, and visible-light communications.

Dr. Wang is a fellow of the IET and HEA. He received nine Best Paper Awards from IEEE GLOBECOM 2010, IEEE ICCT 2011, ITST 2012, IEEE VTC 2013-Spring, IWCMC 2015, IWCMC 2016, IEEE/CIC ICCC 2016, and WPMC 2016. He has served as a technical program committee (TPC) member, the TPC chair, and a general chair for over 80 international conferences. He has served as an editor for nine international journals, including the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2007 to 2009, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY since 2011, and the IEEE TRANSACTIONS ON COMMUNICATIONS since 2015. He was the Lead Guest Editor of the IEEE JOURNAL ON SELECTED AREAS IN COM- MUNICATIONS Special Issue on Vehicular Communications and Networks. He was also a Guest Editor of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS Special Issue on Spectrum and Energy Efficient Design of Wireless Communication Networks and Special Issue on Airborne Communication Networks, and a Guest Editor of the IEEE TRANSACTIONS ON BIG DATA Special Issue on Wireless Big Data. He is recognized as a Web of Science 2017 Highly Cited Researcher.

John Thompson is currently a Professor at the School of Engineering in the University of Edinburgh. He specializes in antenna array processing, cooperative communications systems and energy efficient wireless communications. He has published in excess of three hundred papers on these topics, He was coordinator for the recently completed EU Marie Curie Training Network ADVANTAGE, which studies how communications and power engineering can provide future smart grid systems). In 2018, he will be a technical programme co-chair of the IEEE Smartgridcomm conference to be held in Aalborg, Denmark. He currently leads two UK research projects which study new concepts for fifth generation wireless communications. In January 2016, he was elevated to Fellow of the IEEE for contributions to antenna arrays and multi-hop communications. In 2015-2017, he has been recognised by Thomson Reuters as a highly cited researcher.

Jianghong Shi received his PhD from Xiamen University, China, in 2002. He is currently a professor in the School of Information Science and Engineering, Xiamen University. He is also the director of the West Straits Communications Engineering Center, Fujian Province, China. His research interests include wireless communication networks and satellite navigation systems.