Scalable Transmission over Heterogeneous Network: A Stochastic Geometry Analysis

Liang Wu, Yi Zhong, Wenyi Zhang, *Senior Member, IEEE,*
and Martin Haenggi, *Fellow, IEEE*

***Abstract*—This paper focuses on the transmission of layered source information, such as scalable video coding (SVC), over heterogeneous cellular networks. Scalable transmission enables dynamic adaption of source information to the condition of user equipments (UEs), and thus is suitable for cellular networks in which the transmission link quality varies substantially over space and time. Two novel transmission schemes are proposed, Layered Digital (LD) transmission and Layered Hybrid Digital-Analog (LHDA) transmission. Leveraging tools from stochastic geometry, a comprehensive analysis is conducted focusing on three key performance metrics: outage probability, High-Definition (HD) probability and average distortion. The results show that both proposed transmission schemes can provide a scalable video experience for UEs. For LHDA transmission, the optimal power allocation between digital and analog transmissions is also analyzed. When the proportion of frequency resource allocated to the femto tier exceeds a certain threshold, LHDA transmission is preferable by enabling continuous quality scalability thus avoiding the cliff effect.**

***Index Terms*—Heterogeneous cellular networks, hybrid digital-analog, rate distortion, stochastic geometry, scalable video coding.**

I. Introduction

A. Motivation

The advent of mobile communication and computing keeps driving the data traffic to grow explosively, among which a substantial portion is attributed to multimedia such as mobile video. According to the Cisco Visual Networking Index, mobile video is expected to grow at an average growth rate of 66% until 2019, and within the 24.3 exabytes of data per month crossing mobile networks by 2019, 17.4 exabytes will be video related, such as video on demand, realtime streaming video, video conferencing, and so on. With the release of different types of UEs, the requirements on data rate of video transmission vary in a wide range.

Advanced source coding techniques, such as Scalable Video Coding (SVC), provide a new dimension of dynamically provisioning wireless resources for the varying requirements and the varying link conditions of UEs, thus creating the possibility of extracting video scaled in multiple dimensions, e.g., spatial, temporal, and quality. SVC is an extension of the H.264/MPEG-4 AVC video compression standard [2], in which the bitstream is encoded into multiple layers, namely a Base Layer (BL) and at least one Enhancement Layer (EL). The quality of reconstructed video depends on the number of layers decoded and stays the same until a higher enhancement layer is successfully decoded. The number of layers and their code rates may be determined by the requirement and the link condition of the subscribing UE.

On the other hand, cellular networks are evolving from a homogenous architecture to a composition of heterogeneous networks, comprised of various types of base stations (BSs) [3], [4]. Each type of BSs has its characteristic transmit power and deployment intensity: for example, macro BSs (MBSs) have larger transmit power, aiming at providing global coverage; Femto Access Points (FAPs) are small BSs targeted for home or small business usages. As the distance between a UE and its serving FAP is small, the UE enjoys a high quality link and achieves power savings. Furthermore, the reduced transmission range also enhances spatial reuse and alleviates multiuser interference.

Therefore, when putting together the above two paradigm shifts, namely, shifting from a single-layer video to SVC with multiple layers and shifting from a single-tier cellular network to heterogeneous cellular networks (HCNs) with multiple tiers, these two technologies appear to be inherently compatible and thus can be symbiotically exploited for an improved user experience. The macro cells aim at providing global coverage and therefore are suitable for supporting the BL video content for a majority of UEs, thus enabling the UEs to enjoy basic video (e.g., standard definition 240p) experience; the small cells (e.g., femto cells) aim at providing small-area high-rate service enhancement for hot spots and therefore are suitable for supporting the EL video content for those UEs subscribing to services in their vicinity, thus enabling the UEs to enjoy enhanced video (e.g., high definition 720p) experience. In this paper, we study the problem of scalable transmission over heterogeneous networks and demonstrate that the combination of multi-layer video transmission and multi-tier cellular networks can indeed be beneficially exploited.

L. Wu, Y. Zhong and W. Zhang are with the Key Laboratory of Wireless-Optical Communications, Chinese Academy of Sciences, and the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230027, China (email: {tuohai,geners}@mail.ustc.edu.cn, wenyizha@ustc.edu.cn). M. Haenggi is with the Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN 46556, USA (email: mhaenggi@nd.edu).

The paper was presented in part at the 2015 International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt) [1].

The work of L. Wu, Y. Zhong and W. Zhang was supported in part by the National Basic Research Program of China (973 Program) through grant 2012CB316004, by the National Natural Science Foundation of China through grant 61379003, and by the Fundamental Research Funds for the Central Universities through grant WK3500000003. The work of M. Haenggi was supported by the US NSF through grants CCF 1216407 and 1525904.

B. Related Work

The prior works that consider scalable transmission over wireless networks mainly use digital schemes, consisting of digital source coding (e.g., quantization and entropy coding), digital modulation (e.g., QPSK, 64QAM) and digital channel coding (e.g., turbo or LDPC). The analysis usually focuses on homogeneous networks, and the common feature of the layered structure of SVC and HCNs is not exploited. In [5], an overview of SVC and its relationship to mobile content delivery are discussed focusing on the challenges due to the time-varying characteristics of wireless channels. In [6], a per-subcarrier transmit antenna selection scheme is employed to support multiple scalable video sequences over a downlink cognitive network, and the outage probability is reduced because of video scalability. In [7], real-time use cases of mobile video streaming are presented, for which a variety of parameters like throughput, packet loss ratio and delay are compared with H.264 single-layer video under different degrees of scalability. In [8], the proposed scheme employs WiFi: the BL is always transmitted over a reliable network such as cellular, whereas the EL is opportunistically transmitted through WiFi. Technical issues associated with the simultaneous use of multiple networks are discussed. In [9], HCNs with storage-capable small-cell BSs are studied: versions and layers of video have different impacts on the delay-servicing cost tradeoff, depending on the user demand diversity and the network load.

Besides the above literature based on digital transmission, the recently revitalized analog transmission has shown promising potential in handling channel variations and user heterogeneities for wireless video communication. The analog scheme consists of analog source coding and analog modulation that directly maps a source signal into a linearly transformed channel signal without channel coding. SoftCast [10] is an analog video broadcast scheme that transmits a linear transform of the video signal without quantization, entropy coding, or channel coding. It is claimed to realize continuous quality scalability. However, information-theoretic studies (such as [11], [12]) show that analog schemes with linear mapping (from source signals to channel signals) are relatively inefficient for video transmission while hybrid digital-analog transmission is asymptotically optimal under matched channel conditions for optimally chosen power allocations between the analog and digital parts. The hybrid digital-analog scheme combines digital with analog schemes, transmitting digital and analog signals simultaneously using TDMA, FDMA, or superposition transmission. The authors in [13] propose a hybrid digital-analog scheme for broadcasting, showing a substantial performance gain. However, these works did not consider the impacts of HCNs and the spatial distribution of wireless networks, let alone the design of scalable transmission algorithms utilizing the structure of HCNs.

Considering scalable video transmission over HCNs, we propose two transmission schemes, which adopt digital transmission and hybrid digital-analog transmission, respectively. The system performance is analyzed using stochastic geometry, which has been utilized as an effective tool for modeling and analyzing cellular networks; see, e.g., [14]–[16] and references therein. Generally, the spatial distribution of BSs is modeled as a spatial point process, such as the homogeneous Poisson point process (PPP) for single-tier networks, for which the coverage probability is derived in [17]. For HCNs, the spatial distribution of heterogeneous BSs is often modeled as multiple independent tiers of PPPs, and several key statistics are analyzed in [18], [19]. A comprehensive treatment of the application of stochastic geometry in wireless communication and content can be found in [20], [21].

C. Contributions

In this work, we focus on an analytical performance assessment of SVC transmission over two-tier HCNs utilizing tools from stochastic geometry. The contributions of this work are:

1) An analytical framework is proposed for scalable video transmission exploiting the common feature of a layered structure of SVC and HCNs, which can improve the reception of High-Definition (HD) content and reduce the video distortion, while maintaining an acceptable basic transmission quality for the majority of UEs.
2) A digital and a hybrid digital-analog transmission scheme are proposed and studied. The hybrid digital-analog scheme can further improve the system performance by avoiding the cliff effect[1] and realizing continuous quality scalability when the proportion of frequency resource allocated to the femto tier exceeds a certain threshold.
3) A distortion analysis is provided for different transmission schemes for both orthogonal and non-orthogonal spectrum allocation methods. The power allocation between the digital BL signal and the analog EL signal is also analyzed to minimize the average distortion. The impact of UE load, i.e., the number of UEs served in a cell, is also considered.

The remaining part of this paper is organized as in Fig. 1. Section II describes the system model, including the transmission schemes and spectrum allocation methods. Section III derives the distributions of the number of UEs per cell, subband occupancy probabilities, and SINR distributions. Section IV employs the results obtained in Section III to evaluate the performance metrics, namely outage probability, HD probability, and average distortion. Section V presents numerical results and related discussions. Section VI concludes this paper.

II. System Model

A. Layered Video Model

We consider the downlink performance of SVC over a two-tier HCN. The SVC video content is split into two layers, BL and EL. Two transmission schemes are proposed, Layered

[1]The cliff effect refers to the drastic degradation in video quality when the signal strength fades below the decoding threshold (as opposed to a graceful degradation). There exist certain SINR thresholds at which the video quality changes drastically; in between these thresholds, the quality stays approximately constant. This effect is commonly observed in digital transmission.

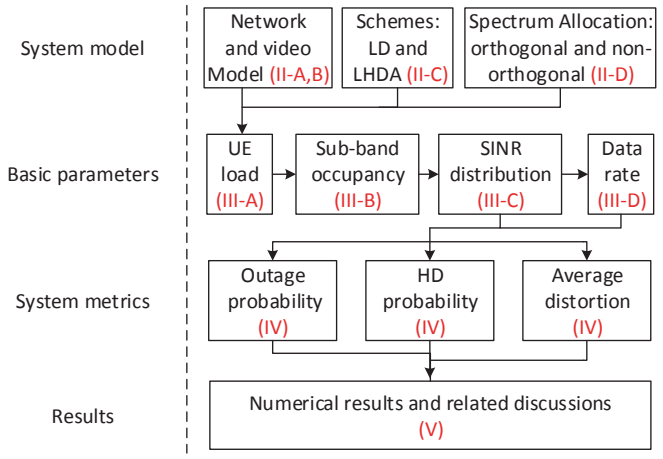


Fig. 1. Paper organization.

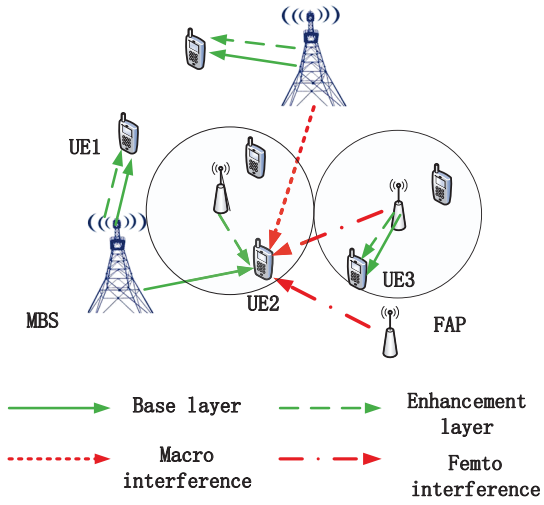


Fig. 2. Illustration of the system model. For the macro UE, UE1 obtains the BL and the EL from the MBS based on the channel quality. There are two cases for a femto UE to receive its signal: UE2 obtains the BL from the MBS and obtains the EL from the FAP; UE3 obtains both the BL and the EL from the FAP.

Digital (LD) and Layered Hybrid Digital-Analog (LHDA). The BL is always modulated into a digital signal and the data rate is R_{B}, while the EL is modulated into a digital signal or an analog signal in the two transmission schemes. If the EL is modulated into a digital signal, then the data rate is R_{E}. Here we focus on the streaming video service, the video can be decoded successfully when the data rate requirements of the BL and the EL are met.

Actually, the proposed analytical framework can be extended to video signals that are encoded to J layers using a fine granularity, and the BS chooses the first J_1 layers for the BL and the following J_2 layers for the EL based on the channel quality for each UE, where $J_1 + J_2 \leq J$.

Here we clarify that SVC allows three types of scalable encoding (spatial, temporal, SNR quality) to be combined and create a single layer [2] [5]. Our layered video model is generic, and we are not concerned with the specifications of the layered encoding and the optimal selection of scalability combinations. Each layer is generated by some combinations of video scalabilities, and the required data rates are the main parameters from the view of networking.

B. Network Model

The two-tier HCN consists of two types of BSs, namely, MBSs and FAPs (see Fig. 2). These two types of BSs are modeled by two independent tiers of homogeneous PPPs, Φ_{mb} and Φ_{fb}, whose intensities are λ_{mb} and λ_{fb}, respectively. FAPs aim at providing network access to UEs in their vicinity within a coverage radius R_{f}. Suppose that there exist N sub-bands each of bandwidth W. The transmit powers of an MBS and an FAP over each sub-band are set as P_{m} and P_{f}, respectively. The path loss model is $r^{-\alpha}$,[2] and the small-scale fading distribution is exponential with mean unity in squared magnitude, i.e., Rayleigh fading. The fading is assumed to be frequency-flat within each sub-band and independent among different sub-bands. We denote the noise variance at each UE by σ^2.

There are two types of UEs, macro UEs and femto UEs. The locations of macro UEs form a homogeneous PPP Φ_{mu} with intensity λ_{mu}, and each macro UE connects to the nearest MBS. The locations of the femto UEs form a Matern cluster process Φ_{fu} [21] with parent process Φ_{fb} (the FAPs), i.e., the UEs in each cluster form a finite PPP of intensity λ_{fu} on the disk of radius R_{f} centered at each FAP, implying that the mean number of users per cluster is $\bar{U}_{\mathrm{f}} = \lambda_{\mathrm{fu}} \pi R_{\mathrm{f}}^2$. Each femto UE connects to the FAP located at the parent point of the corresponding cluster, called the parent FAP. The access mechanism is as follows: a femto UE always connects to its parent FAP when accessing a femto BS and connects to the MBS closest to its parent FAP when accessing a macro BS; a macro UE can only connect to the nearest MBS, even if it is situated within the coverage of an FAP.[3]

C. Transmission Schemes

Macro UEs can only connect to their serving MBS and attempt to obtain the BL and the EL based on the channel quality. It is assumed that the channel quality can be estimated perfectly.

Femto UEs attempt to obtain their EL contents from their serving FAPs and they attempt to obtain their BL contents from their serving MBSs with probability p or from their serving FAPs with probability $1-p$, independently. If a femto UE attempts to obtain the BL contents from the MBS, the femto UE simultaneously connects to the MBS and the FAP by employing the multi-flow technique [22], which has been proposed in 3GPP enabling a UE to simultaneously connect to two BSs, with the two links using the same or different frequency sub-bands. The probability p is an important tunning

[2]Here for simplicity we assume the path loss exponent to be the same for MBS and FAP and ignore the effect of shadowing; the extended case of heterogenous path loss exponents and shadowing can be similarly treated following our analytical approach but with more tedious derivations.

[3]This corresponds to a closed-access femto network, in which only subscribers are allowed to be served by an FAP.

parameter for load balancing between macro tier and femto tier.

Here we clarify that for macro UEs, both the BL and the EL are modulated into digital signals in LD and LHDA. For a macro UE, the data stream is received from the serving MBS based on the channel quality. The macro UE receives both the BL and the EL when the channel can support a data rate larger than $R_{\mathrm{B}}+R_{\mathrm{E}}$ and receives only the BL when the channel can support a data rate between R_{B} and $R_{\mathrm{B}}+R_{\mathrm{E}}$, while an outage occurs when the channel cannot even support the data rate R_{B}[4].

Based on the different modulations and transmissions of the BL and the EL for femto UEs, we propose the following two transmission schemes:

1) LD transmission: Both the BL and the EL are modulated into digital signals. For a femto UE, the data stream of encoded BL signals for small SINR or jointly encoded signals of both the BL and the EL for large SINR is transmitted from the serving FAP when $p=0$; the digital BL data stream is transmitted from its serving MBS, while the digital EL data stream is transmitted from its serving FAP when $p=1$; a mixed transmission is adopted when $0<p<1$.

2) LHDA transmission: The BL is modulated into a digital signal, while the EL is modulated into an analog signal. For a femto UE, the superposition of the digital BL signal and the analog EL signal is transmitted from the serving FAP when $p=0$; the digital BL data stream is transmitted from its serving MBS, while the analog EL data stream is transmitted from its serving FAP when $p=1$; a mixed transmission is adopted when $0<p<1$.

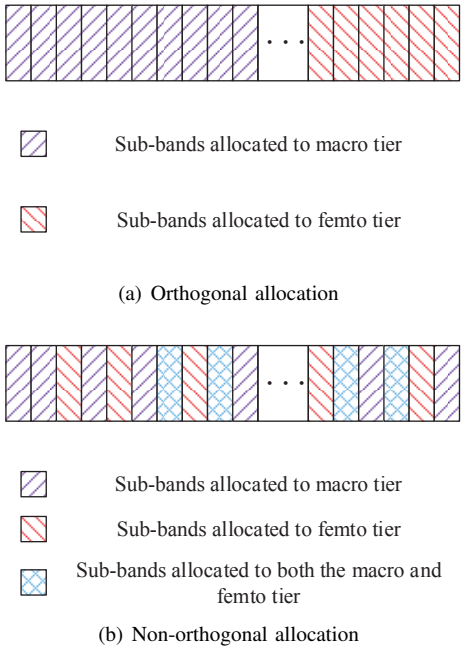
Since the video source is encoded into multiple layers, different layers are transmitted to the UE based on the channel quality, thus providing scalable video quality. Specifically, for those UEs in less favorable conditions, only the BL with relatively low data rate is received in order to ensure basic video experience. When the channel quality improves, the EL is also received for enhanced video experience. Thus, the LD transmission can provide two-level scalable video for the UEs, and LHDA can provide a continuous quality scalability.

D. Spectrum Allocation Methods

Of the N sub-bands, let N_{m} sub-bands be allocated to the macro tier and N_{f} sub-bands to the femto tier. Each UE requires one sub-band for each transmission. We consider the following two spectrum allocation methods [23] (see Fig. 3):

1) Orthogonal Case: The N sub-bands are split as $N=N_{\mathrm{m}}+N_{\mathrm{f}}$, where the N_{m} sub-bands used by all the MBSs of the macro tier are orthogonal to those N_{f} sub-bands used by all the FAPs of the femto tier. So there is no inter-tier interference.

[4]Here we do not consider hybrid digital-analog transmission when macro UEs request both the BL and the EL in LHDA transmission due to its inferior performance (see Fig. 7). Since macro cells aim at providing coverage, the number of served UEs is usually large. From the latter analysis, hybrid digital-analog transmission is beneficial when the UE load is low and the amount of frequency resource exceeds a certain threshold. Thus, digital transmission for macro UEs is preferred.



(a) Orthogonal allocation

(b) Non-orthogonal allocation

Fig. 3. Spectrum allocation methods.

2) Non-orthogonal Case: Compared with the orthogonal case, here the two sets of sub-bands may overlap: each MBS (resp. FAP) independently randomly selects N_{m} (resp. N_{f}) sub-bands from the N sub-bands. The values of both N_{m} and N_{f} can be chosen from 1 to N flexibly and need not add to N. So there is inter-tier interference, while the available spectrum will be abundant as N_{m} and N_{f} grow large.

III. UE Load, Sub-band Occupancy, and SINR Distribution

In this section, we establish several auxiliary results for our derivation of the key performance metrics in Sec. IV. We first provide an approximate characterization of the distribution of the number of UEs connected to a BS and then obtain the probability of a sub-band being occupied. The SINR distribution is subsequently derived, which is used to derive the achievable data rate.

A. UE Load

Since the distribution of femto UEs in an FAP coverage disk is a PPP with intensity λ_{fu}, the number of femto UEs connected to an FAP is a Poisson random variable (r.v.) with mean $\bar{U}_{\mathrm{f}}$,

$$\mathbb{P}\{U_{\mathrm{f}}=i\}=\frac{(\bar{U}_{\mathrm{f}})^{i}}{i!}e^{-\bar{U}_{\mathrm{f}}},\quad i=0,1,\cdots. \tag{1}$$

An MBS not only serves the macro UEs situated in its Voronoi cell but also the femto UEs that belong to the FAPs in this Voronoi cell and connect to the MBS to receive the BL contents. We denote the number of macro UEs in the Voronoi cell as U_{MBS} and the total number of femto UEs served by the MBS as U_{FAP}, which is given by $U_{\mathrm{FAP}}=\sum_{i=1}^{N_{\mathrm{c}}}N_{\mathrm{f},i}$, where N_{c} denotes the number of the FAPs in the Voronoi cell and

$N_{\mathrm{f},i}$ denotes the number of femto UEs which belong to the ith FAP but connect to the MBS to receive the BL contents. The total number of UEs served by an MBS is thus

$$U_{\mathrm{m}} = U_{\mathrm{MBS}} + U_{\mathrm{FAP}}. \tag{2}$$

U_{MBS} is conditionally independent of U_{FAP} given the area of the Voronoi cell. Denote the area of a Voronoi cell by S, the probability generating function (pgf) of U_{m} conditioned on S, denoted by $G_{\mathrm{m}}(z \mid S)$, is

$$G_{\mathrm{m}}(z \mid S) = G_{\mathrm{MBS}}(z \mid S) G_{\mathrm{FAP}}(z \mid S), \tag{3}$$

where $G_{\mathrm{MBS}}(z \mid S)$ and $G_{\mathrm{FAP}}(z \mid S)$ are the pgfs of U_{MBS} and U_{FAP} conditioned on S, respectively.

U_{MBS} is a Poisson r.v. with mean $\lambda_{\mathrm{mu}} S$, and the conditional pgf of U_{MBS} is

$$G_{\mathrm{MBS}}(z \mid S) = e^{\lambda_{\mathrm{mu}} S(z-1)}. \tag{4}$$

Since a femto UE attempts to connect to its serving MBS with probability p, a thinning occurs, i.e., $N_{\mathrm{f},i}$ is a Poisson random variable with mean $p\bar{U}_{\mathrm{f}}$. Meanwhile, N_{c} is also a Poisson r.v. with mean $\lambda_{\mathrm{fb}} S$ because of the PPP distribution of the FAP locations. U_{FAP} is a compound Poisson r.v. with conditional pgf

$$G_{\mathrm{FAP}}(z \mid S) = e^{\lambda_{\mathrm{fb}} S(e^{p\bar{U}_{\mathrm{f}}(z-1)}-1)}. \tag{5}$$

There is no known closed form expression of the probability density function (pdf) of the area S of the typical Poisson Voronoi cell, but the following approximation [24]

$$f_S(x) \approx \frac{(\lambda_{\mathrm{mb}} c)^c}{\Gamma(c)} x^{c-1} e^{-c\lambda_{\mathrm{mb}} x}, \tag{6}$$

where $c = \frac{7}{2}$ and $\Gamma(c) = \int_0^\infty t^{c-1} e^{-t} \mathrm{d}t$, has been known to be handy and sufficiently accurate (see, e.g., [25]). Aided by this approximation, with some manipulations, the pgf of U_{m} is

$$G_{\mathrm{m}}(z) = c^c \Big(c - \frac{\lambda_{\mathrm{mu}}}{\lambda_{\mathrm{mb}}}(z-1) + \frac{\lambda_{\mathrm{fb}}}{\lambda_{\mathrm{mb}}}(1 - e^{p\bar{U}_{\mathrm{f}}(z-1)}) \Big)^{-c}, \tag{7}$$

and the distribution of U_{m} follows as

$$\mathbb{P}\{U_{\mathrm{m}} = i\} = \frac{G_{\mathrm{m}}^{(i)}(0)}{i!}, \quad i = 0, 1, \cdots, \tag{8}$$

where $G_{\mathrm{m}}^{(i)}(0)$ is the i-th derivative of $G_m(z)$ evaluated at $z = 0$.

B. Sub-band Occupancy

Since the number of served UEs for each BS is random, the sub-band frequency resource will be under-utilized in some BSs and over-utilized in some other BSs. As the UE loads in the MBS and the FAP are different under the orthogonal and non-orthogonal spectrum allocations, the sub-band occupancy is calculated for the MBS and the FAP respectively. It is assumed that the available sub-bands are uniformly and independently allocated to the UEs by the BS.

1) Orthogonal Spectrum Allocation: There are N_{m} available sub-bands for the MBS, and each sub-band is equally likely to be chosen. If the number of UEs is smaller than that of sub-bands, the MBS randomly chooses U_{m} out of the total N_{m} sub-bands. Otherwise, all the sub-bands are chosen. The probability that a sub-band is used by an MBS is

$$P_{\mathrm{busy}}^{\mathrm{m},\perp} = \frac{1}{N_{\mathrm{m}}} \sum_{i=0}^{\infty} \min\{i, N_{\mathrm{m}}\} \mathbb{P}\{U_{\mathrm{m}} = i\}, \tag{9}$$

and similarly the probability that a sub-band is used by an FAP is

$$P_{\mathrm{busy}}^{\mathrm{f},\perp} = \frac{1}{N_{\mathrm{f}}} \sum_{i=0}^{\infty} \min\{i, N_{\mathrm{f}}\} \mathbb{P}\{U_{\mathrm{f}} = i\}. \tag{10}$$

2) Non-orthogonal Spectrum Allocation: For the non-orthogonal case, both the MBS and the FAP choose a sub-band randomly from N sub-bands, so the probability that a sub-band is used by an MBS is

$$P_{\mathrm{busy}}^{\mathrm{m},\not\perp} = \frac{1}{N} \sum_{i=0}^{\infty} \min\{i, N_{\mathrm{m}}\} \mathbb{P}\{U_{\mathrm{m}} = i\}, \tag{11}$$

and similarly the probability that a sub-band is used by an FAP is

$$P_{\mathrm{busy}}^{\mathrm{f},\not\perp} = \frac{1}{N} \sum_{i=0}^{\infty} \min\{i, N_{\mathrm{f}}\} \mathbb{P}\{U_{\mathrm{f}} = i\}. \tag{12}$$

The spatial point process of BSs that use a given sub-band is an approximately independent thinning of the original point process Φ_{mb} (resp. Φ_{fb}) by the probability $P_{\mathrm{busy}}^{\mathrm{m,s}}$ (resp. $P_{\mathrm{busy}}^{\mathrm{f,s}}$), denoted by $\tilde{\Phi}_{\mathrm{mb}}$ (resp. $\tilde{\Phi}_{\mathrm{fb}}$) with the intensity $\tilde{\lambda}_{\mathrm{mb}} = \lambda_{\mathrm{mb}} P_{\mathrm{busy}}^{\mathrm{m,s}}$ (resp. $\tilde{\lambda}_{\mathrm{fb}} = \lambda_{\mathrm{mb}} P_{\mathrm{busy}}^{\mathrm{m,s}}$) [25], where the superscript $\mathrm{s} \in \{\perp, \not\perp\}$ indicates whether the orthogonal or the non-orthogonal spectrum allocation method is used. For each sub-band, the event that it is used by an MBS is assumed independent of the event that it is used by all other MBSs. Such an approximation essentially neglects the fact that the areas of adjacent Poisson Voronoi cells are correlated; however, our numerical results reveal that the discrepancy between this approximation (along with others) and simulation experiments is rather slight (see Fig. 9 in Sec. V).

C. SINR Distribution

The complementary cumulative distribution function (ccdf) of the SINR is defined as $\mathcal{P}(\theta) = \mathbb{P}\{\mathrm{SINR} > \theta\}$, where θ is the SINR threshold. The SINR distributions of a UE connected to the MBS and the FAP are derived under two transmission schemes.

1) LD transmission: For analytical tractability, we assume that both the BL and the EL are modulated into digital signals according to a Gaussian codebook.

For the typical UE which is assumed to be located at the origin and connected to its MBS, the received signal denoted by Y can be written as

$$Y = P_{\mathrm{m}}^{1/2} \|x_0\|^{-\alpha/2} h_{x_0} X_{x_0} + \sum_{x \in \tilde{\Phi}_{\mathrm{mb}} \setminus \{x_0\}} P_{\mathrm{m}}^{1/2} \|x\|^{-\alpha/2} h_x X_x + \kappa \sum_{y \in \tilde{\Phi}_{\mathrm{fb}}} P_{\mathrm{f}}^{1/2} \|y\|^{-\alpha/2} h_y X_y + Z, \tag{13}$$

where the first item of right side of the equation denotes the received signal symbol, the second and the third items denote the interference symbols from the macro and the femto tier, respectively, and Z denotes the Gaussian noise with zero mean and variance σ^2. We use x_0 to denote the location of the serving MBS. Note that if the typical UE is a macro UE, the MBS transmits the encoded BL signals only or the jointly encoded signals of both the BL and the EL based on the link SINR. If the typical UE is a femto UE, the MBS transmits the encoded BL signals only. Actually, the MBS does not need to classify the UE type, it just responds to the different requests by macro UEs and femto UEs. X_{x_0} is the signal symbol, while X_x is the interference symbol transmitted by the interfering MBS x. $X_{x_0}, X_x \sim \mathbb{CN}(0,1)$. X_y is the interference symbol transmitted by the interfering FAP y, and $X_y \sim \mathbb{CN}(0,1)$. The indicator $\kappa \in \{0,1\}$ indicates the orthogonal and non-orthogonal spectrum allocation methods, respectively.

Thus, the received SINR is

$$\gamma_{\mathrm{LD}}^{\mathrm{m}} = \frac{P_{\mathrm{m}}\|x_0\|^{-\alpha}|h_{x_0}|^2}{I_{\mathrm{m}} + \kappa I_{\mathrm{f}} + \sigma^2}, \tag{14}$$

where $I_{\mathrm{m}} = \sum_{x\in\tilde{\Phi}_{\mathrm{mb}}\setminus\{x_0\}} P_{\mathrm{m}}\|x\|^{-\alpha}|h_x|^2$ is the interference from the macro tier, and $I_{\mathrm{f}} = \sum_{y\in\tilde{\Phi}_{\mathrm{fb}}} P_{\mathrm{f}}\|y\|^{-\alpha}|h_y|^2$ is the interference from the femto tier.

For the typical femto UE which is assumed to be located at the origin and connected to its FAP, the received signal can be written as

$$Y = P_{\mathrm{f}}^{1/2}\|y_0\|^{-\alpha/2}h_{y_0}X_{y_0} + \sum_{y\in\tilde{\Phi}_{\mathrm{fb}}\setminus\{y_0\}} P_{\mathrm{f}}^{1/2}\|y\|^{-\alpha/2}h_yX_y + \kappa\sum_{x\in\tilde{\Phi}_{\mathrm{mb}}} P_{\mathrm{m}}^{1/2}\|x\|^{-\alpha/2}h_xX_x + Z, \tag{15}$$

where y_0 denotes the location of the serving FAP. Note that the FAP transmits the encoded EL signals only or the jointly encoded signals of both the BL and the EL to the typical UE based on user request. X_{y_0} is the signal symbol transmitted by the serving FAP, and X_y is the interference symbol transmitted by the interfering FAP y.

Thus, the received SINR is

$$\gamma_{\mathrm{LD}}^{\mathrm{f}} = \frac{P_{\mathrm{f}}\|y_0\|^{-\alpha}|h_{y_0}|^2}{I_{\mathrm{f}} + \kappa I_{\mathrm{m}} + \sigma^2}, \tag{16}$$

where $I_{\mathrm{f}} = \sum_{y\in\tilde{\Phi}_{\mathrm{fb}}\setminus\{y_0\}} P_{\mathrm{f}}\|y\|^{-\alpha}|h_y|^2$ denotes the interference from the femto tier and $I_{\mathrm{m}} = \sum_{x\in\tilde{\Phi}_{\mathrm{mb}}} P_{\mathrm{m}}\|x\|^{-\alpha}|h_x|^2$ denotes the interference from the macro tier.

The following theorem gives the ccdf of the SINR for the typical UE,

Theorem 1. *For LD transmission, the ccdf of the SINR for the typical UE connected to its serving MBS is*

$$\begin{aligned}\mathcal{P}_{\mathrm{LD}}^{\mathrm{m}}(\theta) &= \mathbb{P}\{\gamma_{\mathrm{LD}}^{\mathrm{m}} > \theta\}\\ &= \int_0^{\infty} \pi\lambda_{\mathrm{mb}} \exp\Big(-\pi v(\lambda_{\mathrm{mb}} + \tilde{\lambda}_{\mathrm{mb}}\rho(\theta,\alpha))\\ &\quad - \frac{\theta v^{1/\delta}\sigma^2}{P_{\mathrm{m}}} - \kappa\left(\frac{P_{\mathrm{f}}\theta}{P_{\mathrm{m}}}\right)^{\delta} v\tilde{\lambda}_{\mathrm{fb}}\delta\pi^2\mathrm{csc}(\delta\pi)\Big)\mathrm{d}v,\end{aligned} \tag{17}$$

and the ccdf of the SINR for the typical femto UE connected to its serving FAP is

$$\begin{aligned}\mathcal{P}_{\mathrm{LD}}^{\mathrm{f}}(\theta) &= \mathbb{P}\{\gamma_{\mathrm{LD}}^{\mathrm{f}} > \theta\}\\ &= \int_0^{R_{\mathrm{f}}^2} \frac{1}{R_{\mathrm{f}}^2}\exp\Big(-\frac{\theta v^{1/\delta}\sigma^2}{P_{\mathrm{f}}}\\ &\quad - \delta\pi^2\mathrm{csc}(\delta\pi)\theta^{\delta}v\Big(\tilde{\lambda}_{\mathrm{fb}} + \kappa\Big(\frac{P_{\mathrm{m}}}{P_{\mathrm{f}}}\Big)^{\delta}\Big)\tilde{\lambda}_{\mathrm{mb}}\Big)\mathrm{d}v,\end{aligned} \tag{18}$$

where $\delta = 2/\alpha$, $\tilde{\lambda}_{\mathrm{mb}} = \lambda_{\mathrm{mb}}P_{\mathrm{busy}}^{\mathrm{m,s}}$, $\tilde{\lambda}_{\mathrm{fb}} = \lambda_{\mathrm{fb}}P_{\mathrm{busy}}^{\mathrm{f,s}}$, *and* $\rho(\theta,\alpha) = \theta^{\delta}\int_{\theta^{-\delta}}^{\infty}\frac{1}{1+x^{1/\delta}}\mathrm{d}x$. *In orthogonal spectrum allocation,* $\kappa = 0$, *while in non-orthogonal spectrum allocation,* $\kappa = 1$.

Proof: See Appendix A.

2) LHDA transmission: The BL is modulated to a digital signal, while the EL is modulated to an analog signal. The digital modulation is based on a Gaussian codebook, and the EL signal after analog modulation is also modeled as a Gaussian source with zero mean and unit variance [26], [27]. For analog modulation, it is assumed that the source bandwidth is equal to the channel bandwidth [11], [13].

For the typical UE which is assumed to be located at the origin and connected to its serving MBS, the received signal can be written as

$$Y = P_{\mathrm{m}}^{1/2}\|x_0\|^{-\alpha/2}h_{x_0}X_{x_0} + \sum_{x\in\tilde{\Phi}_{\mathrm{mb}}\setminus\{x_0\}} P_{\mathrm{m}}^{1/2}\|x\|^{-\alpha/2}h_xX_x + \kappa\sum_{y\in\tilde{\Phi}_{\mathrm{fb}}} P_{\mathrm{f}}^{1/2}\|y\|^{-\alpha/2}h_yX_y + Z, \tag{19}$$

which is nearly the same as (13) in LD transmission, the difference lies in that X_y is the analog EL interference symbol or the superposition of digital BL and analog EL interference symbol transmitted by the interfering FAP y based on the transmission scheme of y, and $X_y \sim \mathbb{CN}(0,1)$.

Thus the received SINR is

$$\gamma_{\mathrm{LHDA}}^{\mathrm{m}} = \frac{P_{\mathrm{m}}\|x_0\|^{-\alpha}|h_{x_0}|^2}{I_{\mathrm{m}} + \kappa I_{\mathrm{f}} + \sigma^2}, \tag{20}$$

where $I_{\mathrm{m}} = \sum_{x\in\tilde{\Phi}_{\mathrm{mb}}\setminus\{x_0\}} P_{\mathrm{m}}\|x\|^{-\alpha}|h_x|^2$ is the interference from the macro tier and $I_{\mathrm{f}} = \sum_{y\in\tilde{\Phi}_{\mathrm{fb}}} P_{\mathrm{f}}\|y\|^{-\alpha}|h_y|^2$ is the interference from the femto tier.

For the typical femto UE which is assumed to be located at the origin and connected to its FAP, according to the transmission scheme, it receives only the EL, or it receives the superposition of the digital BL signal and the analog EL signal.

- *Case 1:* The typical femto UE connected to its FAP receives only the EL. The received signal for the typical femto UE is

$$Y = P_{\mathrm{f}}^{1/2}\|y_0\|^{-\frac{\alpha}{2}}h_{y_0}X_{y_0}^{\mathrm{E}} + \sum_{y\in\tilde{\Phi}_{\mathrm{fb}}\setminus\{y_0\}} P_{\mathrm{f}}^{1/2}\|y\|^{-\frac{\alpha}{2}}h_yX_y + \kappa\sum_{x\in\tilde{\Phi}_{\mathrm{mb}}} P_{\mathrm{m}}^{1/2}\|x\|^{-\frac{\alpha}{2}}h_xX_x + Z, \tag{21}$$

  where $X_{y_0}^{\mathrm{E}}$ is the EL signal symbol transmitted by the serving FAP and X_y is the interference symbol transmitted by the interfering FAP y.

Thus, the received SINR for the femto UE connected to its FAP to receive the EL is

$$\gamma^{\mathrm{f}}_{\mathrm{LHDA}} = \frac{P_{\mathrm{f}}\|y_0\|^{-\alpha}|h_{y_0}|^2}{I_{\mathrm{f}} + \kappa I_{\mathrm{m}} + \sigma^2}, \tag{22}$$

where $I_{\mathrm{f}} = \sum_{y\in\tilde{\Phi}_{\mathrm{fb}}\setminus\{y_0\}} P_{\mathrm{f}}\|y\|^{-\alpha}|h_y|^2$ is the interference from the femto tier and $I_{\mathrm{m}} = \sum_{x\in\tilde{\Phi}_{\mathrm{mb}}} P_{\mathrm{m}}\|x\|^{-\alpha}|h_x|^2$ is the interference from the macro tier.

- *Case 2:* The typical femto UE connected to its FAP receives the superposition of the digital BL signal and the analog EL signal. The received signal for the typical femto UE is

$$\begin{aligned} Y = {} & \|y_0\|^{-\alpha/2}h_{y_0}(\sqrt{P^{\mathrm{B}}_{\mathrm{f}}}X^{\mathrm{B}}_{y_0} + \sqrt{P^{\mathrm{E}}_{\mathrm{f}}}X^{\mathrm{E}}_{y_0}) \\ & + \sum_{y\in\tilde{\Phi}_{\mathrm{fb}}\setminus\{y_0\}} P_{\mathrm{f}}^{1/2}\|y\|^{-\alpha/2}h_y X_y \\ & + \kappa \sum_{x\in\tilde{\Phi}_{\mathrm{mb}}} P_{\mathrm{m}}^{1/2}\|x\|^{-\alpha/2}h_x X_x + Z, \end{aligned} \tag{23}$$

where $X^{\mathrm{B}}_{y_0}$ is the BL signal symbol transmitted by the serving FAP, and $X^{\mathrm{E}}_{y_0}$ is the EL signal symbol transmitted by the serving FAP.
Thus, the received SINR for the typical femto UE connected to its FAP to receive the BL, denoted by $\gamma^{\mathrm{f,B}}_{\mathrm{LHDA}}$, is

$$\gamma^{\mathrm{f,B}}_{\mathrm{LHDA}} = \frac{P^{\mathrm{B}}_{\mathrm{f}}\|y_0\|^{-\alpha}|h_{y_0}|^2}{P^{\mathrm{E}}_{\mathrm{f}}\|y_0\|^{-\alpha}|h_{y_0}|^2 + I_{\mathrm{f}} + \kappa I_{\mathrm{m}} + \sigma^2}. \tag{24}$$

Successive Interference Cancellation (SIC) [28] is adopted to demodulate the EL signal. Conditioned on the successful reception of the BL, the received SINR for the typical femto UE connected to the FAP to receive the EL signal, denoted by $\gamma^{\mathrm{f,E}}_{\mathrm{LHDA}}$, is

$$\gamma^{\mathrm{f,E}}_{\mathrm{LHDA}} = \frac{P^{\mathrm{E}}_{\mathrm{f}}\|y_0\|^{-\alpha}|h_{y_0}|^2}{I_{\mathrm{f}} + \kappa I_{\mathrm{m}} + \sigma^2}. \tag{25}$$

The following theorem gives the ccdf of the SINR for the typical UE,

Theorem 2. *For LHDA transmission, the ccdf of the SINR for the typical UE connected to its serving MBS is*

$$\mathcal{P}^{\mathrm{m}}_{\mathrm{LHDA}}(\theta_{\mathrm{B}}) = \mathbb{P}\{\gamma^{\mathrm{m}}_{\mathrm{LHDA}} > \theta_{\mathrm{B}}\} = \mathcal{P}^{\mathrm{m}}_{\mathrm{LD}}(\theta_{\mathrm{B}}), \tag{26}$$

the ccdf of the SINR for the typical femto UE connected to its serving FAP to receive the EL is given by

$$\mathcal{P}^{\mathrm{f}}_{\mathrm{LHDA}}(\theta_{\mathrm{E}}) = \mathbb{P}\{\gamma^{\mathrm{f}}_{\mathrm{LHDA}} > \theta_{\mathrm{E}}\} = \mathcal{P}^{\mathrm{f}}_{\mathrm{LD}}(\theta_{\mathrm{E}}), \tag{27}$$

and the joint ccdf of the SINR for the typical femto UE connected to its serving FAP to receive the superposition of the digital BL and the analog EL is given by (28).

Proof: See Appendix B.

D. Data Rate

The instantaneous data rate that a sub-band channel of bandwidth W can accommodate is $R = W\log_2(1+\mathrm{SINR})$. For LD transmission, since both the MBS and the FAP transmit digital signals, the channel from the typical UE to its serving MBS can accommodate the data rate $R_{\mathrm{m}} = W\log_2(1+\gamma^{\mathrm{m}}_{\mathrm{LD}})$, and the channel from the typical UE to its serving FAP can accommodate the data rate $R_{\mathrm{f}} = W\log_2(1+\gamma^{\mathrm{f}}_{\mathrm{LD}})$. For LHDA transmission, only the BL is modulated to a digital signal, so the data rate is defined only for the BL, the channel from the typical UE to its serving MBS can accommodate the data rate $R_{\mathrm{m}} = W\log_2(1+\gamma^{\mathrm{m}}_{\mathrm{LHDA}})$, and the channel from the typical UE to its serving FAP can accommodate data rate $R_{\mathrm{f}} = W\log_2(1+\gamma^{\mathrm{f,B}}_{\mathrm{LHDA}})$.

The actually achieved UE data rates, after taking into consideration the UE load and sub-band occupancy, are given below. Without loss of generality, we take an MBS as an example. When the number of UEs in a macro cell does not exceed the total number of sub-bands (i.e., $U_{\mathrm{m}} \le N_{\mathrm{m}}$), each UE can exclusively occupy a sub-band, and its achieved data rate is R_{m}; when $U_{\mathrm{m}} > N_{\mathrm{m}}$, the U_{m} UEs share the N_{m} sub-bands, and the data rate is thus discounted into $\frac{N_{\mathrm{m}}}{U_{\mathrm{m}}}R_{\mathrm{m}}$, assuming a round-robin sharing mechanism. So the average achieved data rate of a UE served by an MBS is given by

$$R_{\mathrm{mu}} = \xi_{\mathrm{m}}R_{\mathrm{m}}, \tag{29}$$

where ξ_{m} is the scheduling index denoting the probability that a UE is scheduled by the MBS,

$$\xi_{\mathrm{m}} = \frac{\sum_{i=1}^{N_{\mathrm{m}}}\mathbb{P}\{U_{\mathrm{m}}=i\} + \sum_{i=N_{\mathrm{m}}+1}^{\infty}\mathbb{P}\{U_{\mathrm{m}}=i\}\frac{N_{\mathrm{m}}}{i}}{1-\mathbb{P}\{U_{\mathrm{m}}=0\}}. \tag{30}$$

Similarly, the average achieved data rate of a UE served by an FAP is given by

$$R_{\mathrm{fu}} = \xi_{\mathrm{f}}R_{\mathrm{f}}, \tag{31}$$

where ξ_{f} is the scheduling index denoting the probability that a UE is scheduled by the FAP,

$$\xi_{\mathrm{f}} = \frac{\sum_{i=1}^{N_{\mathrm{f}}}\mathbb{P}\{U_{\mathrm{f}}=i\} + \sum_{i=N_{\mathrm{f}}+1}^{\infty}\mathbb{P}\{U_{\mathrm{f}}=i\}\frac{N_{\mathrm{f}}}{i}}{1-\mathbb{P}\{U_{\mathrm{f}}=0\}}. \tag{32}$$

IV. System Performance

In this section we evaluate several important performance metrics, namely, the outage probability, the HD probability, and the average distortion. The outage probability is the probability that a UE cannot receive the BL, namely, the UE data rate is less than R_{B}. The HD probability is the probability that a UE can receive high-definition content, i.e., both the BL and the EL, namely, the UE data rate is greater than $R_{\mathrm{B}}+R_{\mathrm{E}}$. The average distortion evaluates the difference between the received video and source video, which is measured using the distortion-rate function. Note that, for LHDA transmission, the HD probability for the femto UE is not defined since the EL is transmitted as an analog signal and the data rate for an analog signal is undefined.

Table I gives a map of system performance metrics for different transmission schemes.

$$\mathcal{P}_{\mathrm{LHDA}}(\theta_{\mathrm{B}},\theta_{\mathrm{E}})=\mathbb{P}\{\gamma^{\mathrm{f,B}}_{\mathrm{LHDA}}>\theta_{\mathrm{B}},\gamma^{\mathrm{f,E}}_{\mathrm{LHDA}}>\theta_{\mathrm{E}}\}$$
$$=\mathbf{1}\left(\theta_{\mathrm{B}}>\frac{\theta_{\mathrm{E}}P^{\mathrm{B}}_{\mathrm{f}}}{(1+\theta_{\mathrm{E}})P^{\mathrm{E}}_{\mathrm{f}}}\right)\int_0^{R^2_{\mathrm{f}}}\frac{1}{R^2_{\mathrm{f}}}e^{-\frac{\theta_{\mathrm{B}}v^{1/\delta}\sigma^2}{P^{\mathrm{B}}_{\mathrm{f}}-\theta_{\mathrm{B}}P^{\mathrm{E}}_{\mathrm{f}}}-\delta\pi^2\csc(\delta\pi)\theta_{\mathrm{B}}^{\delta}v\left(\tilde{\lambda}_{\mathrm{fb}}(\frac{P_{\mathrm{f}}}{P^{\mathrm{B}}_{\mathrm{f}}-\theta_{\mathrm{B}}P^{\mathrm{E}}_{\mathrm{f}}})^{\delta}+\kappa\tilde{\lambda}_{\mathrm{mb}}(\frac{P_{\mathrm{m}}}{P^{\mathrm{B}}_{\mathrm{f}}-\theta_{\mathrm{B}}P^{\mathrm{E}}_{\mathrm{f}}})^{\delta}\right)}\mathrm{d}v$$
$$+\mathbf{1}\left(\theta_{\mathrm{B}}\leq\frac{\theta_{\mathrm{E}}P^{\mathrm{B}}_{\mathrm{f}}}{(1+\theta_{\mathrm{E}})P^{\mathrm{E}}_{\mathrm{f}}}\right)\int_0^{R^2_{\mathrm{f}}}\frac{1}{R^2_{\mathrm{f}}}e^{-\frac{\theta_{\mathrm{E}}v^{1/\delta}\sigma^2}{P^{\mathrm{E}}_{\mathrm{f}}}-\delta\pi^2\csc(\delta\pi)\theta_{\mathrm{E}}^{\delta}v\left(\tilde{\lambda}_{\mathrm{fb}}(\frac{P_{\mathrm{f}}}{P^{\mathrm{E}}_{\mathrm{f}}})^{\delta}+\kappa\tilde{\lambda}_{\mathrm{mb}}(\frac{P_{\mathrm{m}}}{P^{\mathrm{E}}_{\mathrm{f}}})^{\delta}\right)}\mathrm{d}v. \quad (28)$$

TABLE I
SYSTEM PERFORMANCE METRICS

performance \ transmission	LD	LHDA
macro outage probability	$P^{\mathrm{LD,m}}_{\mathrm{out}}$ in (33)	$P^{\mathrm{LHDA,m}}_{\mathrm{out}}$ in (38)
FAP outage probability	$P^{\mathrm{LD,f}}_{\mathrm{out}}$ in (35)	$P^{\mathrm{LHDA,f}}_{\mathrm{out}}$ in (40)
macro HD probability	$P^{\mathrm{LD,m}}_{\mathrm{HD}}$ in (34)	$P^{\mathrm{LHDA,m}}_{\mathrm{HD}}$ in (39)
femto HD probability	$P^{\mathrm{f}}_{\mathrm{HD}}$ in (36)	—-
average distortion	D_{LD} in (37)	D_{LHDA} in (47)

A. LD Transmission

For a macro UE, both the BL and the EL are transmitted via its serving MBS, so the outage probability, denoted by $P^{\mathrm{LD,m}}_{\mathrm{out}}$, is

$$P^{\mathrm{LD,m}}_{\mathrm{out}}=\mathbb{P}\{R_{\mathrm{mu}}<R_{\mathrm{B}}\}$$
$$=\mathbb{P}\{\gamma^{\mathrm{m}}_{\mathrm{LD}}<2^{\frac{R_{\mathrm{B}}/\xi_{\mathrm{m}}}{W}}-1\}$$
$$=1-\mathcal{P}^{\mathrm{m}}_{\mathrm{LD}}\Big(2^{\frac{R_{\mathrm{B}}/\xi_{\mathrm{m}}}{W}}-1\Big). \quad (33)$$

The HD probability for a macro UE, denoted by $P^{\mathrm{LD,m}}_{\mathrm{HD}}$, is

$$P^{\mathrm{LD,m}}_{\mathrm{HD}}=\mathbb{P}\{R_{\mathrm{mu}}>R_{\mathrm{B}}+R_{\mathrm{E}}\}$$
$$=\mathbb{P}\{\gamma^{\mathrm{m}}_{\mathrm{LD}}>2^{\frac{(R_{\mathrm{B}}+R_{\mathrm{E}})/\xi_{\mathrm{m}}}{W}}-1\}$$
$$=\mathcal{P}^{\mathrm{m}}_{\mathrm{LD}}\Big(2^{\frac{(R_{\mathrm{B}}+R_{\mathrm{E}})/\xi_{\mathrm{m}}}{W}}-1\Big). \quad (34)$$

For a femto UE, it either connects to its serving MBS with probability p or its serving FAP with probability $1-p$ to receive the BL, so the outage probability, denoted by $P^{\mathrm{LD,f}}_{\mathrm{out}}$, is

$$P^{\mathrm{LD,f}}_{\mathrm{out}}=p\mathbb{P}\{R_{\mathrm{mu}}<R_{\mathrm{B}}\}+(1-p)\mathbb{P}\{R_{\mathrm{fu}}<R_{\mathrm{B}}\}$$
$$=p\mathbb{P}\{\gamma^{\mathrm{m}}_{\mathrm{LD}}<2^{\frac{R_{\mathrm{B}}/\xi_{\mathrm{m}}}{W}}-1\}+(1-p)\mathbb{P}\{\gamma^{\mathrm{f}}_{\mathrm{LD}}<2^{\frac{R_{\mathrm{B}}/\xi_{\mathrm{f}}}{W}}-1\}$$
$$=p\Big(1-\mathcal{P}^{\mathrm{m}}_{\mathrm{LD}}\Big(2^{\frac{R_{\mathrm{B}}/\xi_{\mathrm{m}}}{W}}-1\Big)\Big)$$
$$+(1-p)\Big(1-\mathcal{P}^{\mathrm{f}}_{\mathrm{LD}}\Big(2^{\frac{R_{\mathrm{B}}/\xi_{\mathrm{f}}}{W}}-1\Big)\Big). \quad (35)$$

To receive the high-definition video content, a femto UE receives the BL from the MBS and receives the EL from the FAP with probability p, or it receives both the BL and the EL from the FAP with probability $1-p$. Thus, the HD probability for a femto UE, denoted by $P^{\mathrm{f}}_{\mathrm{HD}}$, is

$$P^{\mathrm{f}}_{\mathrm{HD}}=p\mathbb{P}\{R_{\mathrm{mu}}>R_{\mathrm{B}},R_{\mathrm{fu}}>R_{\mathrm{E}}\}$$
$$+(1-p)\mathbb{P}\{R_{\mathrm{fu}}>R_{\mathrm{B}}+R_{\mathrm{E}}\}$$
$$\overset{(a)}{=}p\mathbb{P}\{R_{\mathrm{mu}}>R_{\mathrm{B}}\}\mathbb{P}\{R_{\mathrm{fu}}>R_{\mathrm{E}}\}$$
$$+(1-p)\mathbb{P}\{R_{\mathrm{fu}}>R_{\mathrm{B}}+R_{\mathrm{E}}\}$$
$$=p\mathbb{P}\{\gamma^{\mathrm{m}}_{\mathrm{LD}}>2^{\frac{R_{\mathrm{B}}/\xi_{\mathrm{m}}}{W}}-1\}\mathbb{P}\{\gamma^{\mathrm{f}}_{\mathrm{LD}}>2^{\frac{R_{\mathrm{E}}/\xi_{\mathrm{f}}}{W}}-1\}$$
$$+(1-p)\mathbb{P}\{R_{\mathrm{fu}}>2^{\frac{(R_{\mathrm{B}}+R_{\mathrm{E}})/\xi_{\mathrm{f}}}{W}}-1\}$$
$$=p\mathcal{P}^{\mathrm{m}}_{\mathrm{LD}}\Big(2^{\frac{R_{\mathrm{B}}/\xi_{\mathrm{m}}}{W}}-1\Big)\mathcal{P}^{\mathrm{f}}_{\mathrm{LD}}\Big(2^{\frac{R_{\mathrm{E}}/\xi_{\mathrm{f}}}{W}}-1\Big)$$
$$+(1-p)\mathcal{P}^{\mathrm{f}}_{\mathrm{LD}}\Big(2^{\frac{(R_{\mathrm{B}}+R_{\mathrm{E}})/\xi_{\mathrm{f}}}{W}}-1\Big), \quad (36)$$

where (a) follows from the tier independence approximation. For a femto UE, in the orthogonal case its connection to its serving MBS and its connection to its serving FAP are independent since these two connections use two different sub-bands and they are subject to independent interferences; in the non-orthogonal case, such an independence does not hold, since these two connections may use the same sub-band, thus the same interference from other interfering BSs leads to a certain amount of dependence. Here we make use of a tier independence approximation; that is, for a femto UE, its rate from the macro tier, R_{mu}, and its rate from the femto tier, R_{fu}, are independent r.v.s. Such an approximation is partially motivated by the randomized sub-band selection in the non-orthogonal spectrum allocation method and is found to be accurate via simulation experiments, see Fig. 9.

The distortion-rate function $D(R)$ [11], [29] is used to measure the distortion per source sample when the source rate is R bits/sample. As the bandwidth of a sub-band is W and the data rate of the BL (resp. the EL) is R_{B} (resp. R_{E}), the source rate is $\frac{R_{\mathrm{B}}}{W}$ (resp. $\frac{R_{\mathrm{E}}}{W}$). Since the source signal is modeled as a Gaussian signal with zero mean and unit variance, the distortion of the received video signal can be divided into three cases based on the reception. If the BL is not decoded correctly, the distortion is $D_0=1$; if the BL is decoded correctly while the EL is not, then the distortion is $D_{\mathrm{B}}=2^{-2\frac{R_{\mathrm{B}}}{W}}$; if both the BL and the EL are decoded correctly, the distortion is $D_{\mathrm{HD}}=2^{-2\frac{R_{\mathrm{B}}+R_{\mathrm{E}}}{W}}$.

The average distortion for femto UEs, denoted by D_{LD}, is given by

$$D_{\mathrm{LD}}=P^{\mathrm{LD,f}}_{\mathrm{out}}D_0+(1-P^{\mathrm{LD,f}}_{\mathrm{out}}-P^{\mathrm{f}}_{\mathrm{HD}})D_{\mathrm{B}}+P^{\mathrm{f}}_{\mathrm{HD}}D_{\mathrm{HD}}. \quad (37)$$

B. LHDA transmission

For macro UEs, both the BL and the EL are digitally transmitted via its serving MBS, just the same as that in

LD transmission. From the SINR analysis in Section III-C, the outage probability $P_{\rm out}^{\rm LHDA,m}$, and the HD probability $P_{\rm HD}^{\rm LHDA,m}$ for the macro UE are the same as that in LD transmission.

$$P_{\rm out}^{\rm LHDA,m} = \mathbb{P}\{R_{\rm mu} < R_{\rm B}\} = P_{\rm out}^{\rm LD,m} \tag{38}$$
$$P_{\rm HD}^{\rm LHDA,m} = \mathbb{P}\{R_{\rm mu} \geq R_{\rm B} + R_{\rm B}\} = P_{\rm HD}^{\rm LD,m}. \tag{39}$$

Here we do not adopt hybrid digital-analog transmission when macro UEs request both the BL and the EL in LHDA transmission due to its inferior performance (see Fig. 7 and footnote 4).

For a femto UE, since it receives the BL from the MBS with probability p or receives the BL from the FAP with probability $1-p$, the outage probability, denoted by $P_{\rm out}^{\rm LHDA,f}$, is

$$\begin{aligned} P_{\rm out}^{\rm LHDA,f} &= p\mathbb{P}\{R_{\rm mu} < R_{\rm B}\} + (1-p)\mathbb{P}\{R_{\rm fu} < R_{\rm B}\} \\ &= p\mathbb{P}\{\gamma_{\rm LHDA}^{\rm m} < 2^{\frac{R_{\rm B}/\xi_{\rm m}}{W}} - 1\} \\ &\quad + (1-p)\mathbb{P}\{\gamma_{\rm LHDA}^{\rm f,B} < 2^{\frac{R_{\rm B}/\xi_{\rm f}}{W}} - 1\} \\ &= p\Big(1 - \mathcal{P}_{\rm LHDA}^{\rm m}\Big(2^{\frac{R_{\rm B}/\xi_{\rm m}}{W}} - 1\Big)\Big) \\ &\quad + (1-p)\Big(1 - \mathcal{P}_{\rm LHDA}\Big(2^{\frac{R_{\rm B}/\xi_{\rm f}}{W}} - 1, 0\Big)\Big). \end{aligned} \tag{40}$$

The femto UE has two choices to receive the video content, and the average distortion is calculated accordingly.

1) *Case 1:* The femto UE receives the BL from MBS, and receives the EL from FAP. Since the EL signal is analog, an MMSE estimator is employed for the estimation of the EL, and thus we have $\mathrm{MMSE} = \frac{1}{1+\gamma_{\rm LHDA}^{\rm f}}$, where $\gamma_{\rm LHDA}^{\rm f}$ is the received SINR. Since there are multiple femto UEs in a FAP, a round-robin mechanism is used to schedule time slots for each femto UE to transmit the EL. If a UE is scheduled, its distortion for the EL is MMSE; otherwise, its distortion is unity. So the distortion is $e_{\rm LHDA} = \xi_{\rm f} \cdot \frac{1}{1+\gamma_{\rm LHDA}^{\rm f}} + (1-\xi_{\rm f})\cdot 1$. Since the EL is estimated only if the BL is decoded successfully, the cdf of $e_{\rm LHDA}$ conditioned on the successful reception of the BL is given by

$$\begin{aligned} &\mathbb{P}\{e_{\rm LHDA} < T \mid R_{\rm mu} \geq R_{\rm B}\} \overset{(a)}{=} \mathbb{P}\{e_{\rm LHDA} < T\} \\ &= \mathbb{P}\left\{\xi_{\rm f}\frac{1}{1+\gamma_{\rm LHDA}^{\rm f}} + (1-\xi_{\rm f})1 < T\right\} \\ &= \mathbb{P}\left\{\gamma_{\rm LHDA}^{\rm f} > \frac{1-T}{T-1+\xi_{\rm f}}\right\} \\ &= \mathcal{P}_{\rm LHDA}^{\rm f}\left(\frac{1-T}{T-1+\xi}\right), \end{aligned} \tag{41}$$

where (a) follows from the tier independence approximation.
Since for a positive random variable X, $\mathbb{E}\{X\} = \int_{t>0}\mathbb{P}\{X > t\}\mathrm{d}t$, the mean distortion for the EL, denoted by $D_{\rm E}$, is

$$\begin{aligned} D_{\rm E} &= \mathbb{E}\{e_{\rm LHDA} \mid R_{\rm mu} \geq R_{\rm B}\} \\ &= 1 - \xi_{\rm f} + \int_{1-\xi_{\rm f}}^{1}\left(1 - \mathcal{P}_{\rm LHDA}^{\rm f}\Big(\frac{1-T}{T-1+\xi_{\rm f}}\Big)\right)\mathrm{d}T. \end{aligned} \tag{42}$$

Since the EL corresponds to the residual between the BL and the source signal, the distortion when both the BL and the EL are received, denoted by $D_{\rm HD}$, is given by $D_{\rm HD} = D_{\rm B}D_{\rm E}$.
So the average distortion for the femto UE in *Case 1*, denoted by $D_{\rm LHDA}^{(1)}$, is

$$\begin{aligned} D_{\rm LHDA}^{(1)} &= \mathbb{P}\{R_{\rm mu} < R_{\rm B}\}D_0 + \mathbb{P}\{R_{\rm mu} \geq R_{\rm B}\}D_{\rm HD} \\ &= 1 - \mathcal{P}_{\rm LHDA}^{\rm m}(2^{\frac{R_{\rm B}}{\xi_{\rm m}W}} - 1) + \mathcal{P}_{\rm LHDA}^{\rm m}(2^{\frac{R_{\rm B}}{\xi_{\rm m}W}} - 1)2^{-2R_{\rm B}} \\ &\quad \times \Big(1 - \xi_{\rm f} + \int_{1-\xi_{\rm f}}^{1}\Big(1 - \mathcal{P}_{\rm LHDA}^{\rm f}\Big(\frac{1-T}{T-1+\xi_{\rm f}}\Big)\Big)\,\mathrm{d}T\Big). \end{aligned} \tag{43}$$

2) *Case 2:* The femto UE receives both the BL and the EL from the FAP. Since the EL signal is analog and superposed with the digital BL signal, an MMSE estimator is employed for the estimation of the EL conditioned on the correct reception of the BL, thus we have $\mathrm{MMSE} = \frac{1}{1+\gamma_{\rm LHDA}^{\rm f,E}}$, where $\gamma_{\rm LHDA}^{\rm f,E}$ is the received SINR after the cancellation of the BL. The distortion for the EL is $e_{\rm LHDA} = \xi_{\rm f}\frac{1}{1+\gamma_{\rm LHDA}^{\rm E}} + (1-\xi_{\rm f})1$. The cdf of $e_{\rm LHDA}$ conditioned on the successful reception of the BL is given by

$$\begin{aligned} &\mathbb{P}\{e_{\rm LHDA} < T \mid R_{\rm fu} \geq R_{\rm B}\} \\ &= \mathbb{P}\{\xi_{\rm f}\frac{1}{1+\gamma_{\rm hc}^{\rm E}} + (1-\xi_{\rm f})1 < T \;\Big|\; R_{\rm fu} \geq R_{\rm B}\} \\ &= \mathbb{P}\{\gamma_{\rm LHDA}^{\rm f,E} > \frac{1-T}{T-1+\xi_{\rm f}} \;\Big|\; \gamma_{\rm LHDA}^{\rm f,B} > 2^{\frac{R_{\rm B}/\xi_{\rm f}}{W}} - 1\} \\ &= \frac{\mathcal{P}_{\rm LHDA}(2^{\frac{R_{\rm B}/\xi_{\rm f}}{W}} - 1, \frac{1-T}{T-1+\xi_{\rm f}})}{\mathcal{P}_{\rm LHDA}(2^{\frac{R_{\rm B}/\xi}{W}} - 1, 0)}. \end{aligned} \tag{44}$$

Then, we can obtain the distortion of the EL as

$$\begin{aligned} D_{\rm E} &= \mathbb{E}\{e_{\rm LHDA} < T \mid R_{\rm fu} \geq R_{\rm B}\} \\ &= 1 - \xi_{\rm f} + \int_{1-\xi_{\rm f}}^{1}\Big(1 - \frac{\mathcal{P}_{\rm LHDA}(2^{\frac{R_{\rm B}}{W\xi_{\rm f}}} - 1, \frac{1-T}{T-1+\xi_{\rm f}})}{\mathcal{P}_{\rm LHDA}(2^{\frac{R_{\rm B}}{W\xi_{\rm f}}} - 1, 0)}\Big)\mathrm{d}T. \end{aligned} \tag{45}$$

So the average distortion for the femto UE in *Case 2*, denoted by $D_{\rm LHDA}^{(2)}$, is

$$\begin{aligned} D_{\rm LHDA}^{(2)} &= \mathbb{P}\{R_{\rm fu} < R_{\rm B}\}D_0 + \mathbb{P}\{R_{\rm fu} \geq R_{\rm B}\}D_{\rm HD} \\ &= 1 - \mathcal{P}_{\rm LHDA}(2^{\frac{R_{\rm B}/\xi_{\rm f}}{W}} - 1, 0) \\ &\quad + \mathcal{P}_{\rm LHDA}(2^{\frac{R_{\rm B}/\xi_{\rm f}}{W}} - 1, 0)2^{-2R_{\rm B}}\Big(1 - \xi_{\rm f} \\ &\quad + \int_{1-\xi_{\rm f}}^{1}\Big(1 - \frac{\mathcal{P}_{\rm LHDA}(2^{\frac{R_{\rm B}/\xi_{\rm f}}{W}} - 1, \frac{1-T}{T-1+\xi_{\rm f}})}{\mathcal{P}_{\rm LHDA}(2^{\frac{R_{\rm B}/\xi_{\rm f}}{W}} - 1, 0)}\Big)\mathrm{d}T\Big). \end{aligned} \tag{46}$$

Since a femto UE follows *Case 1* with probability p and follows *Case 2* with probability $1-p$, the average distortion for a femto UE, denoted by $D_{\rm LHDA}$, is

$$D_{\rm LHDA} = pD_{\rm LHDA}^{(1)} + (1-p)D_{\rm LHDA}^{(2)}. \tag{47}$$

TABLE II
SYSTEM PARAMETERS

Symbol	Description	Typical Value
N	number of sub-bands	20
W	bandwidth of a sub-band (MHz)	5
P_{m}	MBS transmit power per sub-band (dBm)	39
P_{f}	FAP transmit power per sub-band (dBm)	13
σ^2	noise power (dBm)	-104
λ_{mb}	MBS intensity (m^{-2})	1E-5
λ_{fb}	FAP intensity (m^{-2})	5E-5
λ_{mu}	macro UE intensity (m^{-2})	2E-4
λ_{fu}	femto UE intensity in coverage (m^{-2})	8E-3
R_{f}	coverage radius of FAP (m)	20
α	path loss exponent	4
R_{B}	rate for the BL transmission (Mbps)	0.5
R_{E}	rate for the EL transmission (Mbps) [5]	4.5

C. Power Allocation in LHDA Transmission

For LHDA transmission, the FAP transmits the superposition of the digital BL signal and the analog EL signal when the femto UE claims both the BL and the EL from the FAP. Since the total transmit power is limited in each FAP, if more power is allocated to transmit the BL, not only less UEs encounter video outage, but also less UEs enjoy HD video. Otherwise, if more power is allocated to transmit the EL, the transmission of the BL suffers while the transmission of the EL is enhanced. Hence, the overall performance depends on the power allocation. In order to optimally allocate the transmit power between the digital BL signal and the analog EL signal, we formulate the following optimization problem:

$$\begin{aligned} \min_{P_{\mathrm{f}}^{\mathrm{B}}, P_{\mathrm{f}}^{\mathrm{E}}} \quad & D_{\mathrm{LHDA}} \\ \text{s.t.} \quad & P_{\mathrm{f}}^{\mathrm{B}} + P_{\mathrm{f}}^{\mathrm{E}} \le P_{\mathrm{f}}. \end{aligned} \tag{48}$$

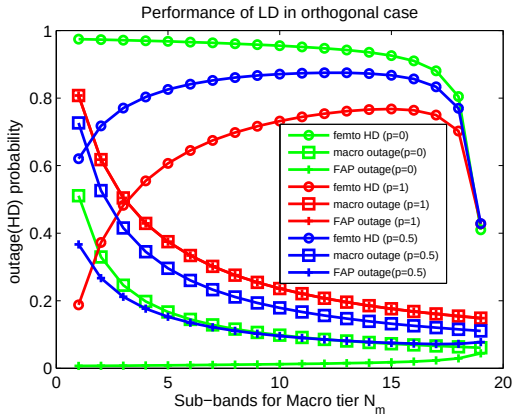
The aim is to find the optimal power allocation for the FAP to minimize the video distortion under the condition that the total transmit power is limited. Since this is a univariate optimization problem, it is practically efficient to find the optimal transmit power allocation among the BL and the EL using a line search.

V. NUMERICAL ILLUSTRATION

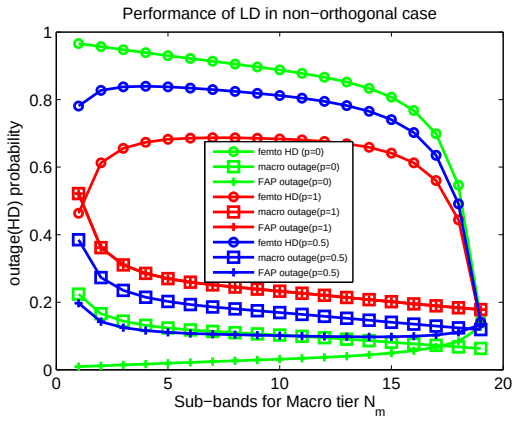
In this section, the outage probabilities, the HD probabilities and the average distortions are evaluated for LD and LHDA transmission schemes considered under orthogonal and non-orthogonal spectrum allocation methods. Meanwhile, the optimal power allocation for the digital BL and the analog EL for LHDA transmission is assessed. We also give the comparison of our analytical results and Monte Carlo simulations to verify the tier independence approximation and the approximate statistics of the Poisson Voronoi cell area in Fig. 9. Unless otherwise specified, the system parameters are listed in Table II.

Fig. 4(a) displays the performance of LD transmission in the orthogonal case. In that case, N_{m} sub-bands for the macro tier and N_{f} sub-bands for the femto tier are orthogonal with $N_{\mathrm{m}}+N_{\mathrm{f}}=N$. As N_{m} increases, more resources are allocated

[5]https://support.google.com/youtube/answer/2853702?hl=en. From Youtube Live encoder settings, we set the basic video (240p) data rate R_{B} as 0.5 Mbps, and the HD video data (720p) rate as 5 Mbps, thus $R_{\mathrm{E}}=4.5$ Mbps.



(a)



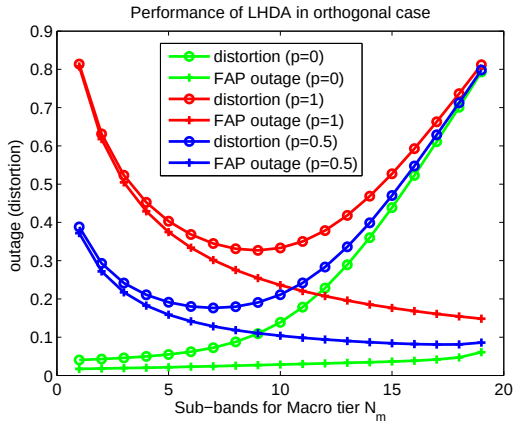
(b)

Fig. 4. Performances of LD in both orthogonal and non-orthogonal cases.

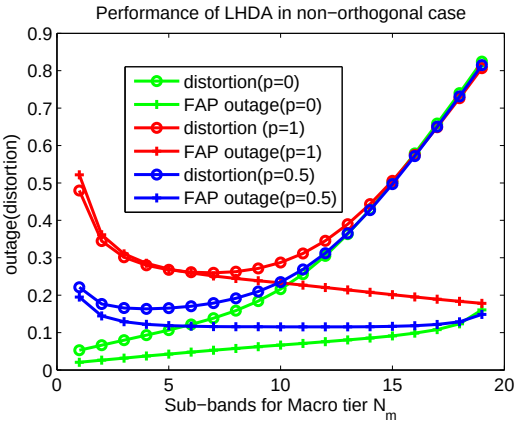
to the macro tier, and the outage probabilities decrease for both macro UEs and femto UEs, except that the femto UE outage probabilities slightly increase for very large values of N_{m}. The HD probability of the femto UE with $p=0$ decreases with N_{m} because the EL transmission via FAPs deteriorates as the resources for the femto tier are reduced. The HD probabilities of the femto UE for $p=0.5$ and $p=1$ increase for small N_{m} and then decrease as N_{m} grows large, reflecting the tension between the resources for the BL transmission and the EL transmission.

Fig. 4(b) displays the performance of LD transmission in the non-orthogonal case. For comparison with Fig. 4(a), we still let $N_{\mathrm{m}}+N_{\mathrm{f}}=N$, but let the sub-bands be selected by each BS independently. The general trend is similar to that in the orthogonal case, but the difference lies in that the curves show less variability with N_{m} (except for those values near to N). The reason for such a practically desirable insensitivity is due to the lessened tension between the resources for macro tiers and femto tiers from randomized sub-band selection.

Note that if p is large, the femto UE tends to connect to an MBS to receive the BL, the outage probability increases and the HD probability decreases, i.e., the performance deteriorates. However, since an MBS can provide continuous coverage while an FAP cannot, if a femto UE is moving, then it may prefer to connect to an MBS to receive the BL, which

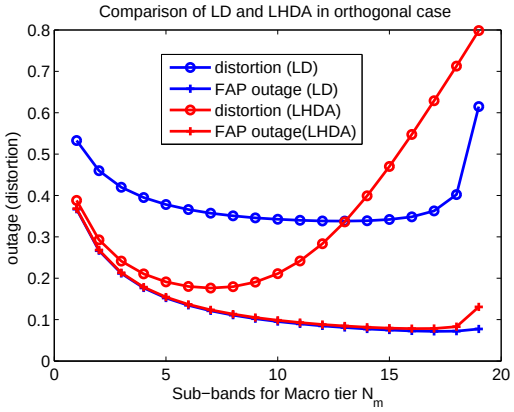


(a)

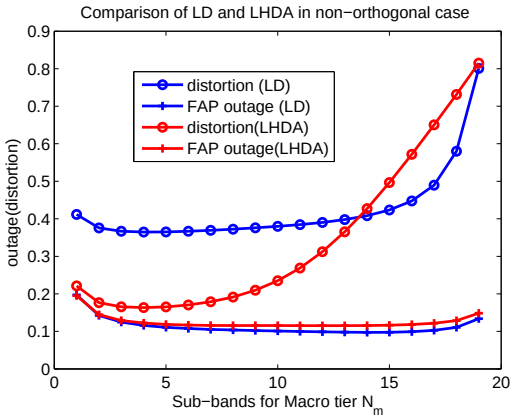


(b)

Fig. 5. Performances of LHDA in both orthogonal and non-orthogonal cases.



(a)



(b)

Fig. 6. Comparisons between LD and LHDA in both orthogonal and non-orthogonal cases.

prevents frequent handover between femto cells and enables uninterrupted reception of the BL video.

Fig. 5(a) displays the performance of LHDA transmission in the orthogonal case. The outage probability for macro UE is the same as that in LD transmission, so we just neglect it in LHDA transmission. Since the frequency resource allocated to the macro tier increases, the resource for the femto tier decreases. The outage probability for the femto UE connected to the FAP (corresponding to $p = 0$) to receive the BL increases while the outage probability for the femto UE connected to the MBS (corresponding to $p = 1$) to receive the BL decreases. The case where $p = 0.5$ shows a tradeoff of these two extreme cases: the outage probability for femto UE first decreases and then slightly increases when the allocated resource for the FAP is small. When N_{m} is small, the performance of the macro tier is poor, and thus the distortion for the UE connected to the MBS to receive the BL is large. When increasing N_{m}, the performance of the macro tier becomes good while that of the femto tier is poor.

Fig. 5(b) displays the performance of LHDA transmission in the non-orthogonal case. The general trends of the curves of the outage and the average distortion are almost the same as that of Fig. 5(a). The difference lies in that the outage probability is lower in the non-orthogonal case than that in the orthogonal case when N_{m} is small.

Fig. 6 displays a comparison between LD transmission and LHDA transmission. Since the comparisons for different p are more or less the same, we set $p = 0.5$ as an example. In both orthogonal and non-orthogonal cases, LHDA outperforms LD when the proportion of frequency resource allocated to the femto tier exceeds a certain threshold, for example, 35% (i.e., $N_{\mathrm{f}} \geq 7$) in the current deployment, as the outage probability is slightly increasing while the average distortion is obviously decreasing when N_{m} is small. The reason is that analog transmission avoids the cliff effect and offers the continuous quality scalability.

Fig. 7 displays the HD probability and distortion for the macro UE. Compared to that of the femto UE, the HD probability is generally low and the distortion is larger, which is owing the large number of UEs served by the MBS. If we adopt the hybrid digital-analog transmission of the digital BL and the analog EL for macro UEs in LHDA transmission, the distortion of video is even deteriorated for a large range of N_{m} compared to that in LD, which verify the conclusion that hybrid digital-analog transmission performs well when the amount of frequency resource exceeds a certain threshold in Fig. 6. Thus we do not adopt the hybrid digital-analog transmission for macro UEs.

Fig. 8 displays the power allocation between the digital BL

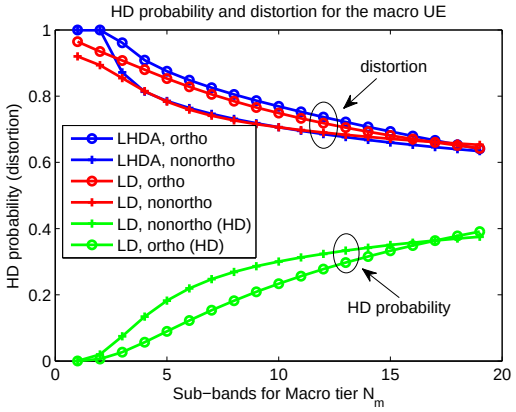


Fig. 7. HD probability and distortion for the macro UE.

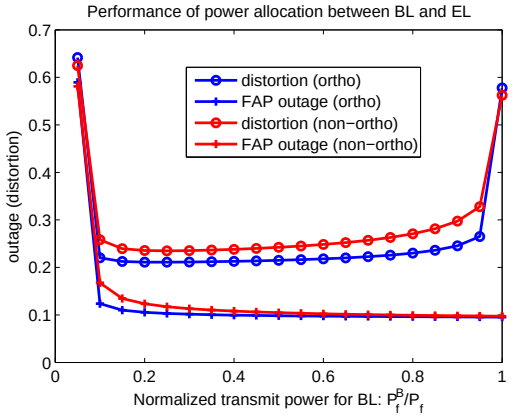
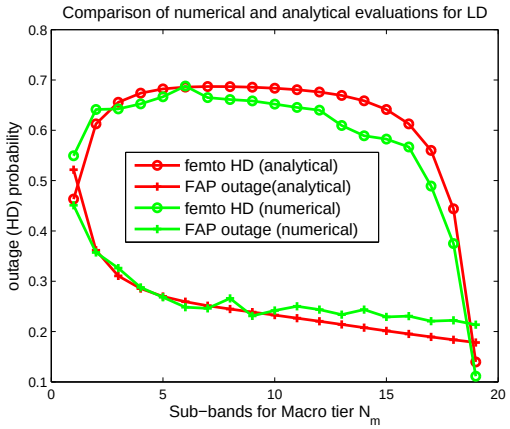
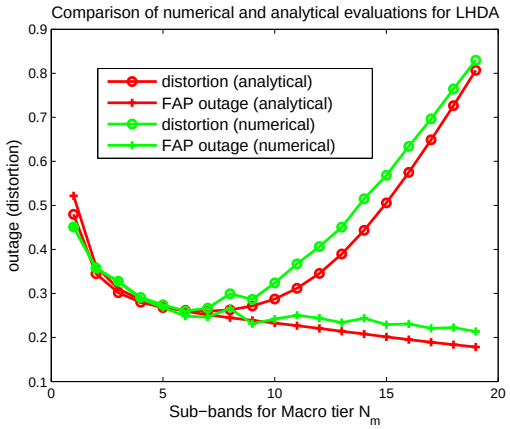


Fig. 8. Power allocation between the BL and the EL in FAPs for LHDA transmission.



(a)



(b)

Fig. 9. Comparisons between numerical evaluation and analytical evaluation for LD transmission and LHDA transmission.

and the analog EL for LHDA transmission. If the power allocated to the BL is increasing, the outage probability decreases monotonously and then approaches stable as the network is interference-limited. With $P_{\mathrm{f}}^{\mathrm{B}}$ increasing, the distortion for the BL is sharply decreasing while the distortion for the EL is increasing. Thus, the total distortion firstly decreases owing to superior transmission of the BL, then increases owing to inferior transmission of the EL. Because of the tradeoff between the transmissions of the BL and the EL, the average distortion varies little when the power allocation ratio $P_{\mathrm{f}}^{\mathrm{B}}/P_{\mathrm{f}}$ lies in a wide range, thus the power allocation is robust.

Fig. 9 displays a comparison between Monte Carlo simulation and analytical results of the performance. Since the comparisons with different p are more or less the same, we set $p = 1$ as a representative example. The simulation region in Monte Carlo simulation is 10 km*10 km. In order to mitigate the boundary effect, we only use the central [3/4 length * 3/4 width] part of the entire region to analyze. The deployment parameters for the BSs and UEs are listed in the Table II. The statistics of the system parameters are derived from that of the total of UEs in the central area. A small gap exists between the curves of Monte Carlo simulation and the analytical results, suggesting that the approximations we adopt in the analysis are sensible.

Here we consider some practical issues on implementation. Firstly, the cooperation between the MBS and its associated FAPs which reside in its Voronoi cell is relatively easy to manage, since the cooperation relationship is location-determined. Secondly, in order to realize analog transmission, we encode the video source by only linear real codes for both compression and error protection, such as 3D DCT for compression and a scaling matrix to adjust the magnitude of the DCT components for error protection, thus ensuring the final coded samples are linearly related to the original pixels. The details can be found in [10], [30].

VI. Conclusion

In this paper, we proposed an analytical framework for scalable video transmission, which exploits the common feature of a layered structure of SVC and HCNs. Specifically, we presented two scalable transmission schemes, LD and LHDA, which are shown to be an effective means for providing differentiated services for users. Through the analysis and comparison of system performance metrics, i.e., outage probability, HD probability and average distortion, under orthogonal and non-orthogonal spectrum allocation methods, we observe that: 1) Compared to the traditional non-scalable video transmission, our schemes can adaptively provide basic or high-

definition video; 2) The frequency resource should be elaborately allocated between tiers to achieve good performance, and the choice of orthogonal and non-orthogonal spectrum allocation methods depend on the system configuration; 3) The hybrid digital-analog transmission can further improve the system performance by reducing video distortion and providing continuous quality scalability of high-definition video; 4) The performance is quite insensitive to the power allocation between the digital BL and the analog EL.

APPENDIX A

Let $\|x_0\|$ be the distance from the typical UE to its serving MBS, which is the nearest MBS, so the pdf of $\|x_0\|$ is $f_{\|x_0\|}(r) = e^{-\lambda_{\rm mb}\pi r^2} 2\pi\lambda_{\rm mb} r$.

The SINR experienced by the typical UE connected to its serving MBS is given by $\gamma^{\rm m}_{\rm LD} = \frac{P_{\rm m}\|x_0\|^{-\alpha}|h_{x_0}|^2}{I_{\rm m}+\kappa I_{\rm f}+\sigma^2}$, where $I_{\rm m} = \sum_{x\in\tilde{\Phi}_{\rm mb}\setminus\{x_0\}} P_{\rm m}\|x\|^{-\alpha}|h_{x_0}|^2$ is the interference from the macro tier, and $I_{\rm f} = \sum_{y\in\tilde{\Phi}_{\rm fb}} P_{\rm f}\|y\|^{-\alpha}|h_y|^2$ is the interference from the femto tier. $\kappa \in \{0,1\}$ is the indicator that whether the orthogonal or the non-orthogonal spectrum allocation is used. Due to the independent thinning approximation, the set of interfering MBSs is a PPP $\tilde{\Phi}_{\rm mb}$ with intensity $\tilde{\lambda}_{\rm mb}$ and the set of interfering FAPs is a PPP $\tilde{\Phi}_{\rm fb}$ with intensity $\tilde{\lambda}_{\rm fb}$.

The ccdf of the SINR experienced by the typical UE connected to its serving MBS

$$\begin{aligned}\mathcal{P}^{\rm m}_{\rm LD}(\theta) &= \mathbb{P}\{\gamma^{\rm m}_{\rm LD} > \theta\}\\ &= \int_0^\infty 2\pi\lambda_{\rm mb} r e^{-\pi\lambda_{\rm mb} r^2}\mathbb{P}\left\{\frac{P_{\rm m}|h_{x_0}|^2 r^{-\alpha}}{I_{\rm m}+\kappa I_{\rm f}+\sigma^2} > \theta\right\}{\rm d}r\\ &\overset{(a)}{=} \int_0^\infty 2\pi\lambda_{\rm mb} r e^{-\pi\lambda_{\rm mb} r^2 - \frac{\theta r^\alpha \sigma^2}{P_{\rm m}}}\mathcal{L}_{I_{\rm m}+\kappa I_{\rm f}}\left(\frac{\theta r^\alpha}{P_{\rm m}}\right){\rm d}r. \quad (49)\end{aligned}$$

where (a) follows from $|h_{x_0}|^2 \sim {\rm Exp}(1)$.

After excluding the serving BS x_0, $\tilde{\Phi}_{\rm mb}\setminus\{x_0\}$ is still a PPP, so we apply the pgfl of PPP to obtain the Laplace transform of $I_{\rm m}$

$$\begin{aligned}\mathcal{L}_{I_{\rm m}}(s) &= \exp\left(-2\pi\tilde{\lambda}_{\rm mb}\int_r^\infty (1-\frac{1}{1+sP_{\rm m}x^{-\alpha}})x{\rm d}x\right)\\ &= e^{-\pi\tilde{\lambda}_{\rm mb} r^2\rho(\frac{sP_{\rm m}}{r^\alpha},\alpha)}. \quad (50)\end{aligned}$$

Since $\tilde{\Phi}_{\rm fb}$ is a PPP, we apply the pgfl of PPP to obtain the Laplace transform of $I_{\rm f}$

$$\begin{aligned}\mathcal{L}_{I_{\rm f}}(s) &= \exp\left(-2\pi\tilde{\lambda}_{\rm fb}\int_0^\infty\left(1-\frac{1}{1+sP_{\rm f}x^{-\alpha}}\right)x{\rm d}x\right)\\ &= e^{-\delta\pi^2\csc(\delta\pi)\tilde{\lambda}_{\rm fb}(sP_{\rm f})^\delta}. \quad (51)\end{aligned}$$

Substituting (50) and (51) into $\mathcal{P}^{\rm m}_{\rm LD}(\theta)$, we can obtain (17).

Let y_0 be the distance between the typical femto UE and its serving FAP. Since femto UEs are uniformly distributed in the circular coverage area of radius $R_{\rm f}$ of each FAP, the pdf of y_0 is given by $f_{y_0}(r) = \frac{2r}{R_{\rm f}^2}$.

The received SINR for the typical femto UE connected to its serving FAP follows as $\gamma^{\rm f}_{\rm LD} = \frac{P_{\rm f}\|y_0\|^{-\alpha}|h_{y_0}|^2}{I_{\rm f}+\kappa I_{\rm m}+\sigma^2}$, where $I_{\rm f} = \sum_{y\in\tilde{\Phi}_{\rm fb}} P_{\rm f}\|y\|^{-\alpha}|h_y|^2$ is the interference from the femto tier, and $I_{\rm m} = \sum_{x\in\tilde{\Phi}_{\rm mb}} P_{\rm m}\|x\|^{-\alpha}|h_{x_0}|^2$ is the interference from the macro tier.

The ccdf of the SINR experienced by the typical femto UE connected to its serving FAP is

$$\begin{aligned}\mathcal{P}^{\rm f}_{\rm LD}(\theta) &= \mathbb{P}\{\gamma^{\rm f}_{\rm LD} > \theta\}\\ &= \int_0^{R_{\rm f}} \frac{2r}{R_{\rm f}^2}\mathbb{P}\left\{\frac{P_{\rm f}|h_{y_0}|^2 r^{-\alpha}}{I_{\rm f}+\kappa I_{\rm m}+\sigma^2} > \theta\right\}{\rm d}r\\ &= \int_0^{R_{\rm f}} \frac{2r}{R_{\rm f}^2} e^{\frac{-\theta r^\alpha\delta^2}{P_{\rm f}}}\mathcal{L}_{I_{\rm f}+\kappa I_{\rm m}}\Big(\frac{\theta r^\alpha}{P_{\rm f}}\Big){\rm d}r, \quad (52)\end{aligned}$$

which, after expanding the Laplace transform of $I_{\rm m}, I_{\rm f}$ and further manipulations, leads to (18).

APPENDIX B

The received SINR for the typical UE connected to its serving MBS is $\gamma^{\rm m}_{\rm LHDA} = \frac{P_{\rm m}\|x_0\|^{-\alpha}|h_{x_0}|^2}{I_{\rm m}+\kappa I_{\rm f}+\sigma^2}$, where $I_{\rm m} = \sum_{x\in\tilde{\Phi}_{\rm mb}\setminus\{x_0\}} P_{\rm m}\|x\|^{-\alpha}|h_{x_0}|^2$ is the interference from the macro tier, and $I_{\rm f} = \sum_{y\in\tilde{\Phi}_{\rm fb}} P_{\rm f}\|y\|^{-\alpha}|h_y|^2$ is the interference from the femto tier.

Similar to the derivation of $\mathcal{P}^{\rm m}_{\rm LD}(\theta)$, the ccdf of $\gamma^{\rm m}_{\rm LHDA}$ follows as

$$\mathcal{P}^{\rm m}_{\rm LHDA}(\theta) = \mathbb{P}\{\gamma^{\rm m}_{\rm LHDA} > \theta\} = \mathcal{P}^{\rm m}_{\rm LD}(\theta). \quad (53)$$

According to the transmission scheme, the FAP transmits the analog EL signal with probability p or the superposition of the digital BL signal and the analog EL signal with probability $1-p$.

1) Case 1: The received SINR for the typical femto UE connected to the FAP receives the EL follows as $\gamma^{\rm f}_{\rm LHDA} = \frac{P_{\rm f}\|y_0\|^{-\alpha}|h_{y_0}|^2}{I_{\rm f}+\kappa I_{\rm m}+\sigma^2}$. $I_{\rm f} = \sum_{y\in\tilde{\Phi}_{\rm fb}} P_{\rm f}\|y\|^{-\alpha}|h_y|^2$ is the interference from the femto tier, $I_{\rm m} = \sum_{x\in\tilde{\Phi}_{\rm mb}} P_{\rm m}\|x\|^{-\alpha}|h_{x_0}|^2$ is the interference from the macro tier. Similar to the derivation of $\mathcal{P}^{\rm f}_{\rm LD}(\theta)$, the ccdf of $\gamma^{\rm f}_{\rm LHDA}$ follows as

$$\mathcal{P}^{\rm f}_{\rm LHDA}(\theta) = \mathbb{P}\{\gamma^{\rm f}_{\rm LHDA} > \theta\} = \mathcal{P}^{\rm f}_{\rm LD}(\theta). \quad (54)$$

2) Case 2: The received SINR for the typical femto UE connected to the FAP receives the superposition of the digital BL signal and the analog EL signal follows as

$$\gamma^{\rm f,B}_{\rm LHDA} = \frac{P^{\rm B}_{\rm f}\|y_0\|^{-\alpha}|h_{y_0}|^2}{P^{\rm E}_{\rm f}\|y_0\|^{-\alpha}|h_{y_0}|^2 + I_{\rm f}+\kappa I_{\rm m}+\sigma^2}, \quad (55)$$

where $P^{\rm E}_{\rm f}\|y_0\|^{-\alpha}|h_{y_0}|^2$ is the interference of the superposed EL, the interference from the femto tier is $I_{\rm f} = \sum_{y\in\tilde{\Phi}_{\rm fb}} P_{\rm f}\|y\|^{-\alpha}|h_y|^2$ and the interference from the macro tier is $I_{\rm m} = \sum_{x\in\tilde{\Phi}_{\rm mb}} P_{\rm m}\|x\|^{-\alpha}|h_{x_0}|^2$. The ccdf of $\gamma^{\rm f,B}_{\rm LHDA}$ follows as

$$\begin{aligned}\mathbb{P}\{\gamma^{\rm f,B}_{\rm LHDA} > \theta\} &= \int_0^\infty \frac{2r}{R_{\rm f}^2} e^{-\frac{\theta r^\alpha\sigma^2}{P^{\rm B}_{\rm f}-\theta P^{\rm E}_{\rm f}}}\mathcal{L}_{I_{\rm f}+\kappa I_{\rm m}}\left(\frac{\theta r^\alpha}{P^{\rm B}_{\rm f}-\theta P^{\rm E}_{\rm f}}\right){\rm d}r\\ &= \int_0^{R_{\rm f}^2}\frac{1}{R_{\rm f}^2}\exp\Big(-\frac{\theta v^{1/\delta}\sigma^2}{P^{\rm B}_{\rm f}-\theta P^{\rm E}_{\rm f}} - \delta\pi^2\csc(\delta\pi)\theta^\delta v\\ &\quad\times\big(\tilde{\lambda}_{\rm fb}(\frac{P_{\rm f}}{P^{\rm B}_{\rm f}-\theta P^{\rm E}_{\rm f}})^\delta + \kappa\tilde{\lambda}_{\rm mb}(\frac{P_{\rm m}}{P^{\rm B}_{\rm f}-\theta P^{\rm E}_{\rm f}})^\delta\big)\Big){\rm d}v. \quad (56)\end{aligned}$$

SIC is adopted to decode the EL signal. After successful reception of the BL, the received SINR for the EL signal is $\gamma^{\rm f,E}_{\rm LHDA} = \frac{P^{\rm E}_{\rm f}\|y_0\|^{-\alpha}|h_{y_0}|^2}{I_{\rm f}+\kappa I_{\rm m}+\sigma^2}$.

The ccdf of $\gamma_{\mathrm{LHDA}}^{\mathrm{f,E}}$ follows as

$$\mathbb{P}\{\gamma_{\mathrm{LHDA}}^{\mathrm{f,E}}>\theta\}=\int_0^{\infty}\frac{2r}{R_{\mathrm{f}}^2}e^{-\frac{\theta r^{\alpha}\sigma^2}{P_{\mathrm{f}}^{\mathrm{E}}}}\mathcal{L}_{I_{\mathrm{f}}+\kappa I_{\mathrm{m}}}\left(\frac{\theta r^{\alpha}}{P_{\mathrm{f}}^{\mathrm{E}}}\right)\mathrm{d}r$$
$$=\int_0^{R_{\mathrm{f}}^2}\frac{1}{R_{\mathrm{f}}^2}e^{-\frac{\theta v^{1/\delta}\sigma^2}{P_{\mathrm{f}}^{\mathrm{E}}}-\delta\pi^2\csc(\delta\pi)\theta^{\delta}v\left(\tilde{\lambda}_{\mathrm{fb}}(\frac{P_{\mathrm{f}}}{P_{\mathrm{f}}^{\mathrm{E}}})^{\delta}+\kappa\tilde{\lambda}_{\mathrm{mb}}(\frac{P_{\mathrm{m}}}{P_{\mathrm{f}}^{\mathrm{E}}})^{\delta}\right)}\mathrm{d}v. \tag{57}$$

The joint ccdf of $\gamma_{\mathrm{LHDA}}^{\mathrm{f,B}}$ and $\gamma_{\mathrm{LHDA}}^{\mathrm{f,E}}$ is

$$\mathcal{P}_{\mathrm{LHDA}}(\theta_{\mathrm{B}},\theta_{\mathrm{E}})=\mathbb{P}\{\gamma_{\mathrm{LHDA}}^{\mathrm{f,B}}>\theta_{\mathrm{B}},\gamma_{\mathrm{LHDA}}^{\mathrm{f,E}}>\theta_{\mathrm{E}}\}$$
$$=\mathbb{P}\left\{|h_{x0}|^2>\frac{\theta_{\mathrm{B}}I_{\mathrm{total}}}{(P_{\mathrm{B}}-\theta_{\mathrm{B}}P_{\mathrm{E}})\|x_0\|^{-\alpha}},|h_{x0}|^2>\frac{\theta_{\mathrm{E}}I_{\mathrm{total}}}{P_{\mathrm{E}}\|x\|^{-\alpha}}\right\}$$
$$=\mathbb{P}\left\{|h_{x0}|^2>\max\Big(\frac{\theta_{\mathrm{B}}I_{\mathrm{total}}}{(P_{\mathrm{B}}-\theta_{\mathrm{B}}P_{\mathrm{E}})\|x_0\|^{-\alpha}},\frac{\theta_{\mathrm{E}}I_{\mathrm{total}}}{P_{\mathrm{E}}\|x\|^{-\alpha}}\Big)\right\}$$
$$=\mathbb{P}\{\gamma_{\mathrm{LHDA}}^{\mathrm{f,B}}>\theta_{\mathrm{B}}\}\mathbf{1}\Big(\theta_{\mathrm{B}}>\frac{\theta_{\mathrm{E}}P_{\mathrm{f}}^{\mathrm{B}}}{(1+\theta_{\mathrm{E}})P_{\mathrm{f}}^{\mathrm{E}}}\Big)$$
$$+\mathbb{P}\{\gamma_{\mathrm{LHDA}}^{\mathrm{f,E}}>\theta_{\mathrm{E}}\}\mathbf{1}\Big(\theta_{\mathrm{B}}\le\frac{\theta_{\mathrm{E}}P_{\mathrm{f}}^{\mathrm{B}}}{(1+\theta_{\mathrm{E}})P_{\mathrm{f}}^{\mathrm{E}}}\Big), \tag{58}$$

where $I_{\mathrm{total}}=I_{\mathrm{f}}+\kappa I_{\mathrm{m}}+\sigma^2$.

ACKNOWLEDGEMENT

The authors wish to thank the anonymous reviewers for their constructive comments.

REFERENCES

[1] L. Wu, Y. Zhong, W. Zhang, and M. Haenggi, "Scalable transmission over heterogenous networks," in *Proc. International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, 2015, pp. 459–466.

[2] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H. 264/AVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, 2007.

[3] V. Chandrasekhar, J. G. Andrews, and A. Gatherer, "Femtocell networks: a survey," *IEEE Commun. Mag.*, vol. 46, no. 9, pp. 59–67, 2008.

[4] C.-H. Ko and H.-Y. Wei, "On-demand resource-sharing mechanism design in two-tier OFDMA femtocell networks," *IEEE Trans. Veh. Technol.*, vol. 60, no. 3, pp. 1059–1071, 2011.

[5] T. Schierl, T. Stockhammer, and T. Wiegand, "Mobile video transmission using scalable video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1204–1217, 2007.

[6] M. Z. Bocus, J. P. Coon, C. N. Canagarajah, S. Armour, A. Doufexi, and J. P. McGeehan, "Per-subcarrier antenna selection for H. 264 MGS/CGS video transmission over cognitive radio networks," *IEEE Trans. Veh. Technol.*, vol. 61, no. 3, pp. 1060–1073, 2012.

[7] R. Radhakrishnan and A. Nayak, "Cross layer design for efficient video streaming over LTE using scalable video coding," in *Proc. IEEE Intl. Conf. on Communications*, 2012, pp. 6509–6513.

[8] V. Gupta, S. Somayazulu, N. Himayat, H. Verma, M. Bisht, and V. Nandwani, "Design challenges in transmitting scalable video over multi-radio networks," in *Proc. IEEE Globecom Workshops*, 2012, pp. 46–51.

[9] K. Poularakis, G. Iosifidis, A. Argyriou, and L. Tassiulas, "Video delivery over heterogeneous cellular networks: Optimizing cost and performance," in *Proc. IEEE INFOCOM*, 2014, pp. 1078–1086.

[10] S. Jakubczak and D. Katabi, "A cross-layer design for scalable mobile video," in *Proc. ACM Proceedings of Annual Intl. Conf. on Mobile Computing and Networking*, 2011, pp. 289–300.

[11] Y. Gao and E. Tuncel, "New hybrid digital/analog schemes for transmission of a Gaussian source over a Gaussian channel," *IEEE Trans. Inf. Theory*, vol. 56, no. 12, pp. 6014–6019, 2010.

[12] P. Minero, S. H. Lim, and Y.-H. Kim, "A unified approach to hybrid coding," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 1509–1523, 2015.

[13] L. Yu, H. Li, and W. Li, "Wireless scalable video coding using a hybrid digital-analog scheme," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 2, pp. 331–345, 2014.

[14] C. C. Chan and S. V. Hanly, "Calculating the outage probability in a CDMA network with spatial Poisson traffic," *IEEE Trans. Veh. Technol.*, vol. 50, no. 1, pp. 183–204, 2001.

[15] M. Haenggi, J. G. Andrews, F. Baccelli, O. Dousse, and M. Franceschetti, "Stochastic geometry and random graphs for the analysis and design of wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 7, pp. 1029–1046, 2009.

[16] H. ElSawy, E. Hossain, and M. Haenggi, "Stochastic geometry for modeling, analysis, and design of multi-tier and cognitive cellular wireless networks: A survey," *IEEE Commun. Surveys & Tutorials*, vol. 15, no. 3, pp. 996–1019, 2013.

[17] 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, 2011.

[18] 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, 2012.

[19] G. Nigam, P. Minero, and M. Haenggi, "Coordinated multipoint joint transmission in heterogeneous networks," *IEEE Trans. Commun.*, vol. 62, no. 11, pp. 4134–4146, 2014.

[20] F. Baccelli and B. Blaszczyszyn, *Stochastic Geometry and Wireless Networks: Volume 1: THEORY*. Now Publishers Inc, 2009, vol. 1.

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

[22] 3GPP, *3GPP TR 25.872 V11.0.0 (2011-09) High Speed Packet Access (HSDPA) multipoint transmission*. Technical Specification, 2011.

[23] W. C. Cheung, T. Q. Quek, and M. Kountouris, "Throughput optimization, spectrum allocation, and access control in two-tier femtocell networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 561–574, 2012.

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

[25] Y. Zhong and W. Zhang, "Multi-channel hybrid access femtocells: a stochastic geometric analysis," *IEEE Trans. Commun.*, vol. 61, no. 7, pp. 3016–3026, 2013.

[26] V. M. Prabhakaran, R. Puri, and K. Ramchandran, "Hybrid digital-analog codes for source-channel broadcast of Gaussian sources over Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 57, no. 7, pp. 4573–4588, 2011.

[27] Y. Kochman and R. Zamir, "Analog matching of colored sources to colored channels," *IEEE Trans. Inf. Theory*, vol. 57, no. 6, pp. 3180–3195, 2011.

[28] M. Wildemeersch, T. Q. Quek, M. Kountouris, A. Rabbachin, and C. H. Slump, "Successive interference cancellation in heterogeneous networks," *IEEE Trans. Commun.*, vol. 62, no. 12, pp. 4440–4453, 2014.

[29] X. Xu, D. Gunduz, E. Erkip, and Y. Wang, "Layered cooperative source and channel coding," in *Proc. IEEE Intl. Conf. on Communications*, 2005, pp. 1200–1204.

[30] H. Cui, R. Xiong, C. Luo, Z. Song, and F. Wu, "Denoising and resource allocation in uncoded video transmission," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 1, pp. 102–112, 2015.

Liang Wu received his B.S. degree in Electronic Engineering from Jilin University in 2011, Jilin, China. He is now a Ph.D. student in Electronic Engineering at University of Science and Technology of China, Hefei, China. His research interests include heterogeneous cellular networks, scalable video transmission, wireless caching and stochastic geometry.

Yi Zhong (S'11, M'15) received his B.S. and Ph.D. degree in Electronic Engineering from University of Science and Technology of China (USTC) in 2010 and 2015 respectively. From August to December 2012, he was a visiting student in Prof. Martin Haenggi's group at University of Notre Dame. From July to October 2013, he worked as an intern in Qualcomm, Corporate Research and Development, Beijing. Now, he is a PostDoctoral research fellow with Singapore University of Technology and Design in the WNDS group led by Prof. Tony Q.S. Quek. His research interests include heterogeneous and femtocelloverlaid cellular networks, wireless ad hoc networks, stochastic geometry and point process theory.

Wenyi Zhang (S'00, M'07, SM11) is with the faculty of the Department of Electronic Engineering and Information Science, University of Science and Technology of China. Prior to that, he was affiliated with the Communication Science Institute, University of Southern California, as a postdoctoral research associate, and with Qualcomm Incorporated, Corporate Research and Development. He studied at Tsinghua University (Bachelor's degree in Automation, in 2001), and the University of Notre Dame, Indiana, USA (Master's and Ph.D. degrees, both in Electrical Engineering, in 2003 and 2006, respectively).

Martin Haenggi (S-95, M-99, SM-04, F-14) is the Frank M. Freimann Professor of Electrical Engineering and a Concurrent Professor of Applied and Computational Mathematics and Statistics at the University of Notre Dame, Indiana, USA. He received the Dipl.-Ing. (M.Sc.) and Dr.sc.techn. (Ph.D.) degrees in electrical engineering from the Swiss Federal Institute of Technology in Zurich (ETH) in 1995 and 1999, respectively. After a postdoctoral year at the University of California in Berkeley, he joined the University of Notre Dame in 2001. In 2007-2008, he spent a Sabbatical Year at the University of California at San Diego (UCSD). For both his M.Sc. and Ph.D. theses, he was awarded the ETH medal, and he received a CAREER award from the U.S. National Science Foundation in 2005 and the 2010 IEEE Communications Society Best Tutorial Paper award. He served an Associate Editor of the Elsevier Journal of Ad Hoc Networks from 2005-2008, of the IEEE Transactions on Mobile Computing (TMC) from 2008-2011, and of the ACM Transactions on Sensor Networks from 2009-2011, as a Guest Editor for the IEEE Journal on Selected Areas in Communications in 2008-2009 and the IEEE Transactions on Vehicular Technology in 2012-2013, and as a Steering Committee Member for the TMC. Presently he is the chair of the Executive Editorial Committee of the IEEE Transactions on Wireless Communications. He also served as a Distinguished Lecturer for the IEEE Circuits and Systems Society in 2005-2006, as a TPC Co-chair of the Communication Theory Symposium of the 2012 IEEE International Conference on Communications (ICC'12), and as a General Co-chair of the 2009 International Workshop on Spatial Stochastic Models for Wireless Networks (SpaSWiN'09) and the 2012 DIMACS Workshop on Connectivity and Resilience of Large-Scale Networks, and as the Keynote Speaker of SpaSWiN'13. He is a co-author of the monograph "Interference in Large Wireless Networks" (NOW Publishers, 2009) and the author of the textbook "Stochastic Geometry for Wireless Networks" (Cambridge University Press, 2012). His scientific interests include networking and wireless communications, with an emphasis on ad hoc, cognitive, cellular, sensor, and mesh networks.