Backhauling in Heterogeneous Cellular Networks: Modeling and Tradeoffs

Daniel C. Chen, *Student Member, IEEE*, Tony Q. S. Quek, *Senior Member, IEEE*, and Marios Kountouris, *Senior Member, IEEE*

***Abstract*—Consumer demand for data has increased tremendously over the last years, and small cell networks are increasingly being considered as one of the key technologies to cope with this demand. Small cell network deployments within the conventional macro cellular networks are creating a significant amount of heterogeneity compared to traditional cellular networks. Nevertheless, the backhaul link is often the bottleneck in terms of system performance and cost. In this paper, we consider the backhaul issue in heterogeneous cellular networks and propose a hierarchical network model using superimposed independent homogeneous Poisson point processes. We derive the total expected delay by taking into account retransmissions over the wireless link, as well as the backhaul delay incurred from both wired and wireless backhaul. For the total expected deployment cost, we take into account infrastructure cost and the construction cost of wired backhaul deployment. Specifically, we are able to characterize the behavior of delay and deployment cost using our simple and tractable model. Furthermore, we propose a delay-based access control policy, which can provide better latency performance especially in dense deployments. Our theoretical framework provides a fundamental understanding of the tradeoff between wired and wireless backhaul and the effect on deployment cost and system performance in heterogeneous cellular networks.**

***Index Terms*—Heterogeneous cellular network, small cell networks, backhaul, deployment cost, mobility, stochastic geometry.**

I. Introduction

CONSUMER demand for cellular data has been growing tremendously, fueled by the ubiquity of wireless devices, such as smartphones, tablets, and laptops. This has placed a greater strain on conventional cellular network infrastructure, and as a result, network densification by deploying small cells has been widely embraced by operators as a promising solution for meeting the galloping demand for mobile traffic [1]–[3]. Small cell access points (SAPs) are low-power base stations that can operate in licensed spectrum and have a range of few meters, compared to macrocell base stations (MBSs) that may have a range of few kilometers. Small cells encompass femtocells, picocells, metrocells, and microcells; they are deployed by either operators or users to improve capacity and coverage, both indoors and outdoors. Despite the demonstrated necessity for small cells, there are still many technical challenges that need to be resolved [1]–[5]. One of them is the incursion of inter-tier and intra-tier interference due to aggressive frequency reuse and dense network deployment [3]. Thus, the fundamental understanding of interference remains an important topic and stochastic geometry has recently gained a lot of popularity as a useful tool to model and quantify network interference in wireless networks [6]–[11]. In small cell networks, SAPs are widely modeled as a homogeneous Poisson point process (PPP) because of their often random, decentralized and uncoordinated deployment [12]–[16].

Besides interference, another key challenge is the need for providing extensive backhaul connectivity economically and reliably to small cells [17], [18]. This problem is further exacerbated by the large number of SAPs, which may eventually become an order of magnitude more than existing MBSs. Furthermore, compared to macrocell backhaul, it is much more challenging to provide backhaul for small cells since first, small cells are typically in hard-to-reach locations near street level rather than above rooftops in the open, and second, reliable connectivity has to be provided at a much lower cost per bit. Many different wired and wireless technologies have been proposed as backhaul solutions for small cells [19]–[21]. Wired backhaul has the advantage of high reliability, high data rates, and high availability. However, cost is an important issue since it is estimated that leased lines currently account for roughly 15% of the network operating expenditure (OPEX) and owning dedicated fiber can be expensive due to high installation expenses. In terms of wired backhaul medium, this can be copper or fiber. The xDSL family of modem technology utilizes legacy copper twisted pair telephone infrastructure and can achieve data rates on the order of a few Mbps to hundreds of Mbps, depending on the copper line length. Fiber offers much higher data rates than copper and can go up to a few Gbps. On the other hand, wireless backhaul is more cost-effective and allows operators to retain end-to-end control of their data and network, without going through a third party. However, the choice of wireless backhaul solutions depends on a number of system considerations such as network capacity, deployment density,

Manuscript received October 9, 2013; revised April 5, 2014 and August 1, 2014; accepted December 17, 2014. Date of publication February 12, 2015; date of current version June 6, 2015. This research was supported by in part by the SRG ISTD2012037, the SUTD-MIT International Design Centre under Grant IDSF1200106OH, and the A⋆STAR SERC under Grant 1224104048, and the project HIATUS of the FET programme within the 7th Framework Programme for Research of the European Commission under FET-Open grant 265578. This paper was presented, in part, at the IEEE Vehicular Technology Conference Spring, Seoul, Korea, May 2014. The associate editor coordinating the review of this paper and approving it for publication was A. A. Abouzeid.

D. C. Chen is with the Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: danielchen1987@gmail.com).

T. Q. S. Quek is with the Singapore University of Technology and Design and the Institute for Infocomm Research, Singapore 487372 (e-mail: tonyquek@sutd.edu.sg).

M. Kountouris is with the Mathematical and Algorithmic Sciences Lab, France Research Center, Huawei Technologies Co. Ltd., 92100 Boulogne-Billancourt, France (e-mail: marios.kountouris@huawei.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2015.2403321

required data rate, infrastructure cost, electromagnetic interference, operating carrier frequency, and the availability of radio spectrum. For instance, wireless backhaul solutions include millimeter wave technologies of 60 GHz and 70–80 GHz, microwave technologies between 6 GHz and 60 GHz, sub 6 GHz radio wave technologies in both licensed and unlicensed bands, TV white space technologies, and satellite technologies.

In this paper, we investigate the challenges associated with backhaul connectivity in heterogeneous cellular networks (HetNets). We build on [22], [23] and we propose a hierarchical network model, where each node connects to the nearest node in the next higher rank. Nodes in each rank are assumed to be distributed according to a homogeneous PPP independent of other ranks. With these assumptions, we ensure that our model is mathematically tractable compared to the traditional grid model. The tractable approach was proposed in [23] to analyze coverage and rate and [24] provides some additional theoretical justification for the use of this model. Based on this network model, we analyze the effect of wired and wireless backhaul delay and the associated deployment cost in HetNets under two extreme cases of mobility: no mobility (static users) and very high mobility (infinite mobility). These simplified mobility models have been extensively used to obtain analytical tractability for gaining insights [25].[1] The main attractiveness of this model is that it allows us to have a simple and tractable way to derive analytical expressions for the average backhaul delay and the average delay experienced by a typical user in the downlink for both wired and wireless backhaul. Specifically, we compare wired and wireless backhaul for small cells deployed within a macrocell network and show that when small cell backhaul is not taken into consideration, mobile users always experience lower delay by connecting to the small cells and densification of small cells is beneficial. However, when backhaul delay is considered, the mobile user may experience lower delay by connecting directly to the macrocell network instead. A key result of this work is that network densification with simply deploying small cells without considering the effect of backhaul may lead to unrealistic conclusions about the performance gains and benefits of HetNets. Moreover, our analysis shows that there is a tradeoff between wired and wireless backhaul, in terms of reliability versus flexibility, which mainly depends on the small cell density.

The remainder of this paper is organized as follows. In Section II, we present our system model and in Section III, the delay performance analysis for both wired and wireless backhaul is provided. In Section IV, we investigate the aggregate network delay and compare two user association policies. In Section V, we provide some numerical results and concluding remarks are given in Section VI.

II. System Model

We consider a macrocell network overlaid with small cells. Mobile users can connect wirelessly to either a MBS or a SAP. The MBSs are connected via dedicated fiber to the core network aggregators (CNAs). The SAPs require wired or wireless backhaul to be connected to small cell gateways (SGs), which are then linked to the CNAs via dedicated fiber. In our model, we consider the heterogeneous cellular network as a collection of nodes and lines. The lines may represent any form of communication links, either wireless links between the mobile users and the base stations or backhaul links. The nodes can represent mobile users, SAPs, MBSs, or SGs, and are assumed to be points drawn from a homogeneous PPP [26].

In our model, we employ a hierarchical network structure, where each node connects to the nearest node in the next highest rank. Nodes in each rank are assumed to be distributed according to a homogeneous PPP independent of other ranks. In particular, the i-th hierarchy of nodes is modeled as a homogeneous PPP Π_i with intensity λ_i, such that $\lambda_1 > \ldots > \lambda_N$ and Π_i is referred to as having *rank* i. For each point x in Π_i and $i < N$, we connect x to the nearest point y, such that $y \in \Pi_{i+1}$, that is, we connect each point to the nearest point from the rank immediately above. Note that the nearest point is unique with probability one [27]. Equivalently, we can say that x is connected to y if x is in the Voronoi cell of y in Π_{i+1}. For instance, we have a heterogeneous cellular network with four tiers comprising of mobile users, SAPs, SGs, and CNAs (in increasing rank with respective densities such that $\lambda_{\rm u} > \lambda_{\rm s} > \lambda_{\rm g} > \lambda_{\rm c}$), overlapping with another heterogeneous network with three tiers comprising of mobile users, MBSs and CNAs (in increasing rank with respective densities such that $\lambda_{\rm u} > \lambda_{\rm m} > \lambda_{\rm c}$).

In the sequel, we consider two cases of mobility:

- Static case: The mobile users are modeled as a static PPP so their locations stay fixed over time. This models stationary users.
- High-mobility case: in each time slot, the positions of the users are modeled as a PPP and are independent of their locations in previous time slots, i.e. a new realization of the PPP of the user locations is drawn in each time slot, and the user locations are independent over time. This models highly mobile users.

The above mobility models will be used to consider the effect of mobility in our delay analysis. Since the links between the MBS and CNA and between the SG and CNA are dedicated and high capacity backhaul links, we assume that there is negligible delay compared to the delay incurred in the other backhaul links. Furthermore, we assume that the delay incurred in the dedicated backhaul for MBS and SG is the same. As such, we do not need to consider this dedicated backhaul delay to CNA in our subsequent delay analysis. In other words, we focus on the delay in wireless access and the wired backhaul between SAPs and SGs. Intermediate mobility cases and analysis under random walk and random waypoint models can be performed using tools from [28], [29] and is left for future work.

1) Delay in Wireless Access: Both MBS and SAP communicate with their intended mobile users via the wireless medium. Transmission through a wireless medium incurs a delay when retransmissions are required due to failure in having successful reception. This may happen for various reasons, but here we mainly focus on the interference from concurrent transmissions and channel fading. In addition, we consider a simple

[1] Note that any practical level of mobility will fall in between these two extreme cases of mobility so we can expect that the results obtained are upper and lower bounds for all levels of mobility.

retransmission protocol in which each packet is repeatedly transmitted until the packet is successfully received, up to a pre-defined maximum number of attempts M. If the packet is successfully received, the mobile user sends a one-bit acknowledgement message to the MBS or SAP, indicating successful reception of the packet. Otherwise, a one-bit negative acknowledgement message is sent. Note that we assume that there is negligible delay and error in these one-bit messages. If after M attempts the packet remains undelivered, an outage is declared. We assume that transmission is successful if the signal-to-interference ratio (SIR) is above a specific threshold γ. We focus on downlink transmissions and consider a Rayleigh fading model with distance dependent path loss attenuation with path loss exponent α. In the following, the SAP and MBS transmit with power $P_{\rm s}$ and $P_{\rm m}$ respectively, such that $P_{\rm s} < P_{\rm m}$.

2) Delay in Wired and Wireless Backhaul: We consider that the SAP-SG links can be either wired or wireless backhaul. For wired backhaul, we model the delay to be an exponentially distributed random variable with mean proportional to the product of the average link distance and the average number of SAPs connecting to a single SG [30]. To derive the mean wired backhaul delay, we introduce the next two lemmas [22], [31] for the hierarchical network structure we introduced earlier.

Lemma 1: The probability distribution function of the length of a link between any two successive ranks, say Π_{i-1} and Π_i, is given by

$$f_i(r) = 2\lambda_i \pi r \exp(-\pi\lambda_i r^2). \tag{1}$$

Suppose that we can assign a cost function $c(r)$ to the link as a function of r. Then, the expected cost per link $\tilde{c}$ and the total expected cost $\tilde{C}$ are given by

$$\tilde{c} = \mathbb{E}_{\Pi_i}\left[c(r)\right] = \int_0^\infty c(r) f_i(r)\,\mathrm{d}r \tag{2}$$

$$\tilde{C} = \mathbb{E}_{\Pi_{i-1},\Pi_i} \sum_{\Pi_{i-1}} c(r) = \lambda_{i-1}\int_0^\infty c(r) f_i(r)\,\mathrm{d}r. \tag{3}$$

When $c(r)$ takes the form sr^η for constants s and η, the expected cost per link and the total expected cost become

$$\tilde{c} = s\frac{\Gamma\left(\frac{\eta}{2}+1\right)}{(\pi\lambda_i)^{\eta/2}}, \quad \tilde{C} = \lambda_{i-1} s\frac{\Gamma\left(\frac{\eta}{2}+1\right)}{(\pi\lambda_i)^{\eta/2}}. \tag{4}$$

Lemma 2: In the hierarchical model, points in Π_i are connected to the nearest node in Π_{i+1} and the average number of Π_i nodes connected to each node in Π_{i+1} is given by

$$\lambda_i/\lambda_{i+1}. \tag{5}$$

Using Lemmas 1 and 2, the mean wired backhaul delay is given by

$$\mathbb{E}\left[D_{\rm s}^{\rm (WdB)}\right] = \beta\frac{\lambda_{\rm s}}{\lambda_{\rm g}}\int_0^\infty 2\lambda_{\rm g}\pi r^2 e^{-\pi\lambda_{\rm g} r^2}\mathrm{d}r = \frac{1}{2}\beta\lambda_{\rm s}\lambda_{\rm g}^{-3/2} \tag{6}$$

where β is the constant of proportionality, $\lambda_{\rm s}$ is the intensity of SAPs, and $\lambda_{\rm g}$ is the intensity of SGs. β is a scaling factor that relates the backhaul delay to the non-backhaul delay, and represents how significant the backhaul effect is. Better backhaul infrastructure and techniques correspond to a lower value of β.

For wireless backhaul, the delay incurred is similar to the wireless access delay, where the time required for one transmission is T_2 and the maximum number of attempts employed by the SG to transmit is N. Since there may not be a dedicated spectrum for the wireless backhaul, especially in the case of sub 6 GHz radio wave technologies, we consider two scenarios for the wireless backhaul, namely in-band and out-of-band wireless backhaul. For in-band wireless backhaul, the wireless backhaul transmissions interfere with the SAP and MBS wireless transmissions. On the other hand, there will be no interference between backhaul and wireless access transmissions for the case of out-of-band wireless backhaul.

Besides backhaul delay, another important factor that affects the successful operation of small cell networks is the deployment cost. Based on our hierarchical network model, we define $B_{\rm m}$, $B_{\rm g}$ and $B_{\rm s}$ to be the cost of installing an MBS, SG and SAP respectively, and $L_{\rm m}$, $L_{\rm g}$ and $L_{\rm s}$ to be the cost of constructing a link from an MBS to a CNA, from an SG to a CNA and from an SAP to an SG respectively. Let r represent the length of a link. Then suppose the cost of link construction is given by

$$L_i = l_i r^{k_i} \tag{7}$$

where the index i denotes m, g or s and l_i, k_i are constants. Therefore, by Lemma 1, we can express the total expected deployment cost for a conventional macrocell network as

$$\begin{aligned}\mathbb{E}[C_{\rm m}] &= \lambda_{\rm m}B_{\rm m} + \lambda_{\rm m}\int_0^\infty L_{\rm m} 2\pi\lambda_{\rm c} r e^{-\pi\lambda_{\rm c} r^2}\,\mathrm{d}r \\ &= \lambda_{\rm m}B_{\rm m} + \lambda_{\rm m} l_{\rm m}\frac{\Gamma\left(\frac{k_{\rm m}}{2}+1\right)}{(\pi\lambda_{\rm c})^{k_{\rm m}/2}}\end{aligned} \tag{8}$$

where $\lambda_{\rm m}$ is the intensity of MBSs and $\lambda_{\rm c}$ is the intensity of CNAs. Furthermore, the total expected deployment cost for a small cell network is given by

$$\begin{aligned}\mathbb{E}[C_{\rm s}] &= \lambda_{\rm g}B_{\rm g} + \lambda_{\rm g}\int_0^\infty L_{\rm g} 2\pi\lambda_{\rm c} r e^{-\pi\lambda_{\rm c} r^2}\,\mathrm{d}r + \lambda_{\rm s}B_{\rm s} \\ &\quad + \lambda_{\rm s}\int_0^\infty L_{\rm s} 2\pi\lambda_{\rm g} r e^{-\pi\lambda_{\rm g} r^2}\,\mathrm{d}r \\ &= \lambda_{\rm g}B_{\rm g} + \lambda_{\rm g} l_{\rm g}\frac{\Gamma\left(\frac{k_{\rm g}}{2}+1\right)}{(\pi\lambda_{\rm c})^{k_{\rm g}/2}} + \lambda_{\rm s}B_{\rm s} \\ &\quad + \lambda_{\rm s} l_{\rm s}\frac{\Gamma\left(\frac{k_{\rm s}}{2}+1\right)}{(\pi\lambda_{\rm g})^{k_{\rm s}/2}}.\end{aligned} \tag{9}$$

III. Backhaul Analysis

For notational convenience, we denote a base station or a user by its location. For transmitter x and receiver u, the SIR from x to u is given by

$$\mathsf{SIR}(x \to u) = \frac{P_x F_x(t) g(x-u)}{\sum_{y\in\Omega(x)} P_y F_y(t) g(y-u)} \tag{10}$$

where $\Omega(x)$ is the set of nodes in Ω that interferes with x, P_y is the power of the transmitting node at y, and $F_x(t)$ is the power fading coefficient from x to u. In our model, we assume that $F_x(t)$ is exponentially distributed with mean 1 (Rayleigh fading) and we ignore background thermal noise since we consider an interference-limited network where interference is the dominating noise effect. The distance dependent path loss function is denoted as $g(x) = \|x\|^{-\alpha}$, where α is the path loss exponent. We define the success probability from x to u as $\mathbb{P}(\mathsf{SIR}(x \to u) > \gamma)$, with γ being a pre-determined threshold. Due to the stationarity of the point processes, the success probability for each pair of Tx-Rx in each tier is the same. As a consequence of Slivnyak's Theorem, the distribution of the point process is unaffected by conditioning on a receiver being at the origin (typical receiver).

From [12], [23], we obtain the probability of a successful downlink transmission (ignoring any backhaul effects for now) from the nearest SAP at a distance r away as

$$\mathbb{P}_{\mathrm{s}}^{(\mathrm{Wl,S})}(r) = \exp\left\{-\pi r^2\left[\lambda_{\mathrm{s}}\rho(\gamma,\alpha)+(P_{\mathrm{m}}/P_{\mathrm{s}})^{2/\alpha}\lambda_{\mathrm{m}}\gamma^{2/\alpha}C(\alpha)\right]\right\} \quad (11)$$

where $C(\alpha) = \frac{2\pi/\alpha}{\sin(2\pi/\alpha)}$ and $\rho(\gamma,\alpha) = \gamma^{2/\alpha}\int_{\gamma^{-2/\alpha}}^{\infty}\frac{1}{1+u^{\alpha/2}}\mathrm{d}u$. Similarly, the probability of a successful downlink transmission from the nearest MBS at a distance r away is given by

$$\mathbb{P}_{\mathrm{m}}^{(\mathrm{Wl,S})}(r) = \exp\left\{-\pi r^2\left[\lambda_{\mathrm{m}}\rho(\gamma,\alpha)+(P_{\mathrm{s}}/P_{\mathrm{m}})^{2/\alpha}\lambda_{\mathrm{s}}\gamma^{2/\alpha}C(\alpha)\right]\right\}. \quad (12)$$

Using (11) and (12), we can derive the expected delay in the wireless access as follows:

Theorem 1: In the static case, the expected wireless delay for a typical user connecting to a SAP is given by (13), shown at the bottom of the page, where T_1 is the time taken for a single packet transmission from an SAP to its user. Assuming that the time taken for a single packet transmission from the MBS is also T_1, the expected wireless delay for a typical user connecting to a MBS is given by (14), shown at the bottom of the page. On the other hand, in the high-mobility case, the expected wireless delay for a typical user connecting to an SAP is given by (15), shown at the bottom of the page, and the expected wireless delay for a typical user connecting to an MBS is given by (16), shown at the bottom of the page.

Proof: See Appendix A. □

To obtain further insight from Theorem 1, we introduce the following lemma, which allows us to compare the expected delay when connecting to an SAP with that when connecting to an MBS.

Lemma 3: The functions

$$F_1(x) = \sum_{i=0}^{M-1}(-1)^i\binom{M}{i+1}\frac{1}{1+ix\gamma^{2/\alpha}C(\alpha)} \quad (17)$$

$$F_2(x) = \frac{\left(\rho+1+\gamma^{2/\alpha}C(\alpha)x\right)^M-(\rho+\gamma^{2/\alpha}C(\alpha)x)^M}{\left(\rho+1+\gamma^{2/\alpha}C(\alpha)x\right)^{M-1}} \quad (18)$$

are both increasing in x.

Proof: See Appendix B. □

Remark 1: For network deployments for which the following condition holds

$$\frac{\lambda_{\mathrm{s}}}{\lambda_{\mathrm{m}}} > \left(\frac{P_{\mathrm{m}}}{P_{\mathrm{s}}}\right)^{2/\alpha} \quad (19)$$

we can use Theorem 1 and Lemma 3 to conclude that the typical user experiences a lower expected delay by connecting to the nearest SAP rather than the nearest MBS. This result is independent of whether users are static or highly mobile. The intuition behind this result is that with a high density of SAPs, we can always connect to an SAP with relatively high SIR due to shorter distance. As a result, the expected delay incurred by retransmissions is significantly reduced. This is essentially one of the main motivations why there is vivid interest in network densification using small cells as a solution to provide uniform user experience and QoS [3].

A. Wired Backhaul

First, we analyze the expected delay for wired backhaul.

Theorem 2: In the static case, the total expected delay with wired backhaul is given by (20) and (21), shown at the bottom of the next page, for the small cell user and macrocell user, respectively. For the high-mobility case, the corresponding expressions are given by (22) and (23), shown at the bottom of the next page, respectively.

Proof: See Appendix C. □

Corollary 1: For $M = 2$, in both static and high-mobility cases, we have

$$\lambda_{\mathrm{s}} = \frac{\sqrt{\frac{2T_1(P_{\mathrm{m}}/P_{\mathrm{s}})^{2/\alpha}\lambda_{\mathrm{m}}\gamma^{2/\alpha}C}{\beta\lambda_{\mathrm{g}}^{-3/2}}}-(P_{\mathrm{m}}/P_{\mathrm{s}})^{2/\alpha}\lambda_{\mathrm{m}}\gamma^{2/\alpha}C}{1+\rho}$$

which gives the optimal λ_{s} for minimal small cell delay.

$$\mathbb{E}\left[D_{\mathrm{s}}^{(\mathrm{Wl,S})}\right] = T_1\sum_{i=0}^{M-1}(-1)^i\binom{M}{i+1}\frac{1}{1+i\left[\rho(\gamma,\alpha)+(P_{\mathrm{m}}/P_{\mathrm{s}})^{2/\alpha}(\lambda_{\mathrm{m}}/\lambda_{\mathrm{s}})\gamma^{2/\alpha}C(\alpha)\right]} \quad (13)$$

$$\mathbb{E}\left[D_{\mathrm{m}}^{(\mathrm{Wl,S})}\right] = T_1\sum_{i=0}^{M-1}(-1)^i\binom{M}{i+1}\frac{1}{1+i\left[\rho(\gamma,\alpha)+(P_{\mathrm{s}}/P_{\mathrm{m}})^{2/\alpha}(\lambda_{\mathrm{s}}/\lambda_{\mathrm{m}})\gamma^{2/\alpha}C(\alpha)\right]} \quad (14)$$

$$\mathbb{E}\left[D_{\mathrm{s}}^{(\mathrm{Wl,M})}\right] = T_1\frac{\left[\lambda_{\mathrm{s}}(\rho+1)+(P_{\mathrm{m}}/P_{\mathrm{s}})^{2/\alpha}\gamma^{2/\alpha}C(\alpha)\lambda_{\mathrm{m}}\right]^M-\left[\lambda_{\mathrm{s}}\rho+(P_{\mathrm{m}}/P_{\mathrm{s}})^{2/\alpha}\gamma^{2/\alpha}C(\alpha)\lambda_{\mathrm{m}}\right]^M}{\lambda_{\mathrm{s}}\left[\lambda_{\mathrm{s}}(\rho+1)+(P_{\mathrm{m}}/P_{\mathrm{s}})^{2/\alpha}\gamma^{2/\alpha}C(\alpha)\lambda_{\mathrm{m}}\right]^{M-1}} \quad (15)$$

$$\mathbb{E}\left[D_{\mathrm{m}}^{(\mathrm{Wl,M})}\right] = T_1\frac{\left[\lambda_{\mathrm{m}}(\rho+1)+(P_{\mathrm{s}}/P_{\mathrm{m}})^{2/\alpha}\gamma^{2/\alpha}C(\alpha)\lambda_{\mathrm{s}}\right]^M-\left[\lambda_{\mathrm{m}}\rho+(P_{\mathrm{s}}/P_{\mathrm{m}})^{2/\alpha}\gamma^{2/\alpha}C(\alpha)\lambda_{\mathrm{s}}\right]^M}{\lambda_{\mathrm{m}}\left[\lambda_{\mathrm{m}}(\rho+1)+(P_{\mathrm{s}}/P_{\mathrm{m}})^{2/\alpha}\gamma^{2/\alpha}C(\alpha)\lambda_{\mathrm{s}}\right]^{M-1}} \quad (16)$$

For $M = 3$, the optimal λ_s can be found by solving a quartic equation for each case.

From Theorem 2, for the static case, if $\lambda_s > 2T_1 \frac{\lambda_g^{3/2}}{\beta} (M - \sum_{i=0}^{M-1} (-1)^i \binom{M}{i+1} \frac{1}{1+i\rho(\gamma,\alpha)})$, then $\mathbb{E}[D_s^{(\mathrm{Wl+WdB,S})}] \geq \mathbb{E}[D_m^{(\mathrm{Wl+WdB,S})}]$. For the high-mobility case, if $\lambda_s > 2T_1 \frac{\lambda_g^{3/2}}{\beta} \left(M - (\rho+1) \left(1 - \left(\frac{\rho}{\rho+1} \right)^M \right) \right)$, then $\mathbb{E}[D_s^{(\mathrm{Wl+WdB,M})}] \geq \mathbb{E}[D_m^{(\mathrm{Wl+WdB,M})}]$.

Remark 2: From Theorem 2, we see that if β is large enough, the user might prefer to connect to the macrocell network even for high values of λ_s. Indeed, if $\lambda_s > 2T_1 \frac{\lambda_g^{3/2}}{\beta} \left(M - \sum_{i=0}^{M-1} (-1)^i \binom{M}{i+1} \frac{1}{1+i\rho(\gamma,\alpha)} \right)$ for the static case or $\lambda_s > 2T_1 \frac{\lambda_g^{3/2}}{\beta} \left(M - (\rho+1) \left(1 - \left(\frac{\rho}{\rho+1} \right)^M \right) \right)$ for the high-mobility case, then the macrocell delay is smaller than the small cell delay. We note also that increasing λ_s past a certain value actually increases the delay in connecting to the small cell network. In particular, this is true when λ_s exceeds the values given above.

Recall that in Theorem 1, we have seen that without taking the backhaul into account, the typical user is better off connecting to the nearest SAP provided that the density of SAPs is high enough. However, we see that ignoring the backhaul delay can be misleading. In particular, when the backhaul delay is significant enough, it is preferred that the typical user remains connected to the nearest MBS even when the density of SAPs is high. Therefore, our analysis shows the importance of taking into account the effect of backhaul in heterogeneous cellular networks.

In the above analysis, we consider that the macro and small cell tiers interfere with each other. In the case where they do not interfere with each other, the probabilities of successful downlink transmission for small cell and macrocell users in the static case are given by

$$P_s^{(\mathrm{NWl,S})}(r) = \exp\left[-\pi r^2 \lambda_s \rho(\gamma,\alpha)\right] \tag{24}$$

$$P_m^{(\mathrm{NWl,S})}(r) = \exp\left[-\pi r^2 \lambda_m \rho(\gamma,\alpha)\right] \tag{25}$$

and the delay in receiving transmissions from the SAP and MBS is given by

$$\mathbb{E}\left[D_s^{(\mathrm{NWl,S})}\right] = T_1 \sum_{i=0}^{M-1} (-1)^i \binom{M}{i+1} \frac{1}{1+i\rho(\gamma,\alpha)} = \mathbb{E}\left[D_m^{(\mathrm{NWl,S})}\right]. \tag{26}$$

On the other hand, for the high-mobility case, we have $P_s^{(\mathrm{NWl,M})} = \frac{1}{1+\rho(\gamma,\alpha)} = P_m^{(\mathrm{NWl,M})}$ and hence, we have $\mathbb{E}[D_s^{(\mathrm{NWl,M})}] = \mathbb{E}[D_m^{(\mathrm{NWl,M})}]$.

In both cases, we observe that the wireless delay is the same for both tiers. However, in the small cell network, there is still the additional wired backhaul delay, which is unchanged from before. Therefore, depending on the how large the backhaul delay is, it may be beneficial from a latency perspective that the typical user remains connected to the macrocell network.

B. Wireless Backhaul

For the wireless backhaul case, the analysis follows similarly to the previous subsection. If we assume that the wireless backhaul transmission does not interfere with the SAP-user and MBS-user transmissions, then the backhaul delay is constant for each SAP user, and we obtain similar results as shown in the next theorem.

Theorem 3: If the backhaul does not interfere with the other transmissions, then the expected backhaul delay is

$$\mathbb{E}\left[D_s^{(\mathrm{WlBN,S})}\right] = T_2 \sum_{i=0}^{N-1} (-1)^i \binom{N}{i+1} \frac{1}{1+i\rho(\gamma,\alpha)} \tag{27}$$

for the static case, and

$$\mathbb{E}\left[D_s^{(\mathrm{WlBN,M})}\right] = T_2 \left[1+\rho(\gamma,\alpha) - \frac{\rho(\gamma,\alpha)^N}{(1+\rho(\gamma,\alpha))^{N-1}} \right] \tag{28}$$

for the high mobility case.

Proof: The proof is omitted since it follows straightforwardly from the derivation of the delay in wireless access. □

Remark 3: Non-interfering wireless backhaul simply adds a constant term to the expected delay of a single typical user on the small cell network, and if it is large enough, the small cell delay can always be higher than the macrocell delay. If it is not too large, then the small cell delay will eventually (at higher SAP densities) be smaller than the macrocell delay.

In the case that the backhaul transmissions interfere with both MBS and SAP transmissions, the non-backhaul delay is also affected but the qualitative behavior is not too different as shown in the next theorem.

$$\mathbb{E}\left[D_s^{(\mathrm{Wl+WdB,S})}\right] = T_1 \sum_{i=0}^{M-1} (-1)^i \binom{M}{i+1} \frac{1}{1+i\left[\rho(\gamma,\alpha) + (P_m/P_s)^{2/\alpha} (\lambda_m/\lambda_s) \gamma^{2/\alpha} C(\alpha)\right]} + \frac{1}{2}\beta\lambda_s\lambda_g^{-\frac{3}{2}} \tag{20}$$

$$\mathbb{E}\left[D_m^{(\mathrm{Wl+WdB,S})}\right] = T_1 \sum_{i=0}^{M-1} (-1)^i \binom{M}{i+1} \frac{1}{1+i\left[\rho(\gamma,\alpha) + (P_s/P_m)^{2/\alpha} (\lambda_s/\lambda_m) \gamma^{2/\alpha} C(\alpha)\right]} \tag{21}$$

$$\mathbb{E}\left[D_s^{(\mathrm{Wl+WdB,M})}\right] = T_1 \frac{\left[\lambda_s(\rho+1) + (P_m/P_s)^{2/\alpha}\gamma^{2/\alpha}C(\alpha)\lambda_m\right]^M - \left[\lambda_s\rho + (P_m/P_s)^{2/\alpha}\gamma^{2/\alpha}C(\alpha)\lambda_m\right]^M}{\lambda_s\left[\lambda_s(\rho+1) + (P_m/P_s)^{2/\alpha}\gamma^{2/\alpha}C(\alpha)\lambda_m\right]^{M-1}} + \frac{1}{2}\beta\lambda_s\lambda_g^{-\frac{3}{2}} \tag{22}$$

$$\mathbb{E}\left[D_m^{(\mathrm{Wl+WdB,M})}\right] = T_1 \frac{\left[\lambda_m(\rho+1) + (P_s/P_m)^{2/\alpha}\gamma^{2/\alpha}C(\alpha)\lambda_s\right]^M - \left[\lambda_m\rho + (P_s/P_m)^{2/\alpha}\gamma^{2/\alpha}C(\alpha)\lambda_s\right]^M}{\lambda_m\left[\lambda_m(\rho+1) + (P_s/P_m)^{2/\alpha}\gamma^{2/\alpha}C(\alpha)\lambda_s\right]^{M-1}} \tag{23}$$

Theorem 4: In the case that the backhaul transmissions interfere with both MBS and SAP transmissions, the conditional success probabilities are given by (29)–(31), shown at the bottom of the page. In the static case, this results in expected single-user delays given by (32) and (33), shown at the bottom of the page. In the high-mobility case, this results in expected single-user delays given by (34) and (35), shown at the bottom of the page, where

$$q_1 = \lambda_s\rho + (P_m/P_s)^{2/\alpha}\gamma^{2/\alpha}C(\alpha)\lambda_m + (P_g/P_s)^{2/\alpha}\lambda_g\gamma^{2/\alpha}C(\alpha) \quad (36)$$

$$q_2 = \lambda_m\rho(\gamma,\alpha) + (P_s/P_m)^{2/\alpha}\lambda_s\gamma^{2/\alpha}C(\alpha) + (P_g/P_m)^{2/\alpha}\lambda_g\gamma^{2/\alpha}C(\alpha). \quad (37)$$

Proof: The proof is omitted since it follows straightforwardly from the derivation in the wired backhaul case. □

A consequence of limiting the number of attempts at transmitting each packet is that each packet faces an outage probability of $(1-P_s)^M$ or $(1-P_m)^M$ when being transmitted from an SAP or an MBS respectively. Additionally, for the wireless backhaul, there is an extra outage probability of $(1-P_g)^N$ when being transmitted from the SG to the SAP. (Note that P_s, P_m, and P_g are generic notations for the success probabilities in various cases.)

IV. Optimization

In this section, we move beyond the perspective of a typical user and we optimize the delay from a network perspective. Let r_s and r_m denote the distance from the nearest SAP and MBS, respectively. Suppose each user connects to the nearest SAP if $r_s < \kappa r_m$ for some $\kappa \in \mathbb{R}^+$, and to the nearest MBS otherwise. Then, a typical user connects to an SAP with association probability given by

$$\begin{aligned} p_a &= \mathbb{E}\left[\int_0^{r\kappa} 2\pi\lambda_s t e^{-\pi\lambda_s t^2}\,\mathrm{d}t \,|\, r\right] \\ &= \int_0^{\infty} (1-e^{-\pi\lambda_s\kappa^2 r^2})2\pi\lambda_m r e^{-\pi\lambda_m r^2}\,\mathrm{d}r \\ &= \frac{\kappa^2\lambda_s}{\lambda_m+\kappa^2\lambda_s}, \end{aligned} \quad (38)$$

and the association probability for connecting to an MBS is simply $1-p_a$.

Next, we consider that users can adopt the following two association policies:

1) *Interference-based policy*: each user connects to the node that gives the highest SIR. This means connecting to the nearest SAP if $r_s < \kappa_i r_m$, where

$$\kappa_i = \left(\frac{P_s}{P_m}\right)^{\frac{1}{\alpha}}. \quad (39)$$

2) *Delay-based policy*: each user connects to the node that gives the lowest expected wireless retransmission delay (assuming no other users connect to it). This means connecting to the nearest SAP if $r_s < \kappa_d r_m$, where

$$\kappa_d = \sqrt{\frac{\lambda_m\rho(\gamma,\alpha) + (P_s/P_m)^{2/\alpha}\lambda_s\gamma^{2/\alpha}C(\alpha)}{\lambda_s\rho(\gamma,\alpha) + (P_m/P_s)^{2/\alpha}\lambda_m\gamma^{2/\alpha}C(\alpha)}}. \quad (40)$$

Corollary 2: Under the interference-based policy, the association probability is given by

$$p_{a,i} = \frac{\kappa_i^2\lambda_s}{\lambda_m+\kappa_i^2\lambda_s} \quad (41)$$

and under delay-based policy, the association probability is given by

$$p_{a,d} = \frac{\kappa_d^2\lambda_s}{\lambda_m+\kappa_d^2\lambda_s}. \quad (42)$$

$$P_s^{(\mathrm{WlBI})}(r) = \exp\left\{-\pi r^2\left[\lambda_s\rho(\gamma,\alpha) + (P_m/P_s)^{2/\alpha}\lambda_m\gamma^{2/\alpha}C(\alpha) + (P_g/P_s)^{2/\alpha}\lambda_g\gamma^{2/\alpha}C(\alpha)\right]\right\} \quad (29)$$

$$P_m^{(\mathrm{WlBI})}(r) = \exp\left\{-\pi r^2\left[\lambda_m\rho(\gamma,\alpha) + (P_s/P_m)^{2/\alpha}\lambda_s\gamma^{2/\alpha}C(\alpha) + (P_g/P_m)^{2/\alpha}\lambda_g\gamma^{2/\alpha}C(\alpha)\right]\right\} \quad (30)$$

$$P_g^{(\mathrm{WlBI})}(r) = \exp\left\{-\pi r^2\left[\lambda_g\rho(\gamma,\alpha) + (P_m/P_g)^{2/\alpha}\lambda_m\gamma^{2/\alpha}C(\alpha) + (P_s/P_g)^{2/\alpha}\lambda_s\gamma^{2/\alpha}C(\alpha)\right]\right\} \quad (31)$$

$$\begin{aligned}\mathbb{E}\left[D_s^{(\mathrm{Wl+WlBI,S})}\right] = {} & T_1\sum_{i=0}^{M-1}(-1)^i\binom{M}{i+1}\frac{1}{1+i\left[\rho(\gamma,\alpha)+(P_m/P_s)^{2/\alpha}(\lambda_m/\lambda_s)\gamma^{2/\alpha}C(\alpha)+(P_g/P_s)^{2/\alpha}(\lambda_g/\lambda_s)\gamma^{2/\alpha}C(\alpha)\right]} \\ & + T_2\sum_{i=0}^{N-1}(-1)^i\binom{N}{i+1}\frac{1}{1+i\left[\rho(\gamma,\alpha)+(P_m/P_g)^{2/\alpha}(\lambda_m/\lambda_g)\gamma^{2/\alpha}C(\alpha)+(P_s/P_g)^{2/\alpha}(\lambda_s/\lambda_g)\gamma^{2/\alpha}C(\alpha)\right]}\end{aligned} \quad (32)$$

$$\mathbb{E}\left[D_m^{(\mathrm{Wl+WlBI,S})}\right] = T_1\sum_{i=0}^{M-1}(-1)^i\binom{M}{i+1}\frac{1}{1+i\left[\rho(\gamma,\alpha)+(P_s/P_m)^{2/\alpha}(\lambda_s/\lambda_m)\gamma^{2/\alpha}C(\alpha)+(P_g/P_m)^{2/\alpha}(\lambda_g/\lambda_m)\gamma^{2/\alpha}C(\alpha)\right]} \quad (33)$$

$$\begin{aligned}\mathbb{E}\left[D_s^{(\mathrm{Wl+WlBI,M})}\right] = {} & T_1\frac{(\lambda_s+q_1)^M - q_1^M}{\lambda_s(\lambda_s+q_1)^{M-1}} \\ & + T_2\sum_{i=0}^{N-1}(-1)^i\binom{N}{i+1}\frac{1}{1+i\left[\rho(\gamma,\alpha)+(P_m/P_g)^{2/\alpha}(\lambda_m/\lambda_g)\gamma^{2/\alpha}C(\alpha)+(P_s/P_g)^{2/\alpha}(\lambda_s/\lambda_g)\gamma^{2/\alpha}C(\alpha)\right]}\end{aligned} \quad (34)$$

$$\mathbb{E}\left[D_m^{(\mathrm{Wl+WlBI,M})}\right] = T_1\frac{(\lambda_m+q_2)^M - q_2^M}{\lambda_m(\lambda_m+q_2)^{M-1}} \quad (35)$$

In the sequel, we allow p_{a} to denote either $p_{\mathrm{a,i}}$ or $p_{\mathrm{a,d}}$, depending on the policy used.

A. Wired Backhaul

By approximating the process of SAP users with a thinned PPP with density $p_{\mathrm{a}}\lambda_{\mathrm{u}}$ and the MBS user process as a thinned PPP with density $(1-p_{\mathrm{a}})\lambda_{\mathrm{u}}$, the total expected network delay for the static case is given by (43), shown at the bottom of the page, using Lemma 2 and Theorem 2, where we assume that the users employ TDMA in their multiple access. Similarly, the total expected network delay for the high-mobility case can be derived using Lemma 2 and Theorem 2 as given by (44), shown at the bottom of the page.

From (43) and (44), we can determine whether there is an optimal SAP density to minimize the delay for different association policies. This result is provided in the next theorem, where we define $P=(P_{\mathrm{s}}/P_{\mathrm{m}})^{2/\alpha}$ and $G=\gamma^{2/\alpha}C(\alpha)$ for notational convenience.

Theorem 5: For the static case, if β satisfies

$$\beta<\frac{2PT_1\lambda_{\mathrm{u}}\lambda_{\mathrm{g}}^{3/2}}{3\lambda_{\mathrm{m}}^2}\sum_{i=0}^{M-1}(-1)^i\binom{M}{i+1}\frac{(P-2)iG-1-i\rho}{(1+i\rho+iG)^2}, \quad (45)$$

under the interference-based policy, or β satisfies (46), shown at the bottom of the page, under the delay-based policy, then there is an optimal λ_{s}^* that minimizes the total expected network delay.

For the high-mobility case, if β satisfies (47), shown at the bottom of the page, under the interference-based policy, or β satisfies (48), shown at the bottom of the page, under the delay-based policy, then there is an optimal λ_{s}^* that minimizes the total expected network delay.

Proof: See Appendix D. □

Remark 4: If β is significantly high, the backhaul delay in the small cell network quickly dominates and it is clear that the minimal delay occurs when $\lambda_{\mathrm{s}}=0$. However, if β is sufficiently small, we can determine the optimal λ_{s}^* numerically using the results provided in Theorem 5. Given a fixed deployment cost, the network operator can then decide if this optimal SAP density is feasible to deploy.

B. Wireless Backhaul

Theorem 6: Under the delay-based policy for interfering wireless backhaul, the new value of κ_{d} is

$$\kappa_{\mathrm{d}}^{(\mathrm{WlBI})} = \sqrt{\frac{\lambda_{\mathrm{m}}\rho(\gamma,\alpha)+\left(\frac{P_{\mathrm{s}}}{P_{\mathrm{m}}}\right)^{\frac{2}{\alpha}}\lambda_{\mathrm{s}}\gamma^{2/\alpha}C(\alpha)+\left(\frac{P_{\mathrm{g}}}{P_{\mathrm{m}}}\right)^{\frac{2}{\alpha}}\lambda_{\mathrm{g}}\gamma^{2/\alpha}C(\alpha)}{\lambda_{\mathrm{s}}\rho(\gamma,\alpha)+\left(\frac{P_{\mathrm{m}}}{P_{\mathrm{s}}}\right)^{\frac{2}{\alpha}}\lambda_{\mathrm{m}}\gamma^{2/\alpha}C(\alpha)+\left(\frac{P_{\mathrm{g}}}{P_{\mathrm{s}}}\right)^{\frac{2}{\alpha}}\lambda_{\mathrm{g}}\gamma^{2/\alpha}C(\alpha)}}. \quad (49)$$

With Lemma 2 and Theorems 3 and 4, we can now compute the total expected network delay in the various cases. The total expected network delay for the static case in non-interfering

$$\mathbb{E}\left[D^{(\mathrm{Wl+WdB,S})}\right]=\frac{T_1p_{\mathrm{a}}^2\lambda_{\mathrm{u}}^2}{\lambda_{\mathrm{s}}}\sum_{i=0}^{M-1}(-1)^i\binom{M}{i+1}\frac{1}{1+i\left[\rho(\gamma,\alpha)+(P_{\mathrm{m}}/P_{\mathrm{s}})^{2/\alpha}(\lambda_{\mathrm{m}}/\lambda_{\mathrm{s}})\gamma^{2/\alpha}C(\alpha)\right]}$$
$$+\frac{1}{2}\beta p_{\mathrm{a}}\lambda_{\mathrm{u}}\lambda_{\mathrm{s}}\lambda_{\mathrm{g}}^{-3/2}+\frac{T_1(1-p_{\mathrm{a}})^2\lambda_{\mathrm{u}}^2}{\lambda_{\mathrm{m}}}\sum_{i=0}^{M-1}(-1)^i\binom{M}{i+1}\frac{1}{1+i\left[\rho(\gamma,\alpha)+(P_{\mathrm{s}}/P_{\mathrm{m}})^{2/\alpha}(\lambda_{\mathrm{s}}/\lambda_{\mathrm{m}})\gamma^{2/\alpha}C(\alpha)\right]} \quad (43)$$

$$\mathbb{E}\left[D^{(\mathrm{Wl+WdB,M})}\right]=\frac{1}{2}\beta p_{\mathrm{a}}\lambda_{\mathrm{u}}\lambda_{\mathrm{s}}\lambda_{\mathrm{g}}^{-3/2}$$
$$+\frac{T_1p_{\mathrm{a}}^2\lambda_{\mathrm{u}}^2}{\lambda_{\mathrm{s}}}\frac{\left[\lambda_{\mathrm{s}}(\rho+1)+(P_{\mathrm{m}}/P_{\mathrm{s}})^{2/\alpha}\gamma^{2/\alpha}C(\alpha)\lambda_{\mathrm{m}}\right]^M-\left[\lambda_{\mathrm{s}}\rho+(P_{\mathrm{m}}/P_{\mathrm{s}})^{2/\alpha}\gamma^{2/\alpha}C(\alpha)\lambda_{\mathrm{m}}\right]^M}{\lambda_{\mathrm{s}}\left[\lambda_{\mathrm{s}}(\rho+1)+(P_{\mathrm{m}}/P_{\mathrm{s}})^{2/\alpha}\gamma^{2/\alpha}C(\alpha)\lambda_{\mathrm{m}}\right]^{M-1}}$$
$$+\frac{T_1(1-p_{\mathrm{a}})^2\lambda_{\mathrm{u}}^2}{\lambda_{\mathrm{m}}}\frac{\left[\lambda_{\mathrm{m}}(\rho+1)+(P_{\mathrm{s}}/P_{\mathrm{m}})^{2/\alpha}\gamma^{2/\alpha}C(\alpha)\lambda_{\mathrm{s}}\right]^M-\left[\lambda_{\mathrm{m}}\rho+(P_{\mathrm{s}}/P_{\mathrm{m}})^{2/\alpha}\gamma^{2/\alpha}C(\alpha)\lambda_{\mathrm{s}}\right]^M}{\lambda_{\mathrm{m}}\left[\lambda_{\mathrm{m}}(\rho+1)+(P_{\mathrm{s}}/P_{\mathrm{m}})^{2/\alpha}\gamma^{2/\alpha}C(\alpha)\lambda_{\mathrm{s}}\right]^{M-1}} \quad (44)$$

$$\beta<\frac{T_1\lambda_{\mathrm{u}}\lambda_{\mathrm{g}}^{3/2}P(\rho+G)}{P\rho+(P+1)G}\sum_{i=0}^{M-1}(-1)^i\binom{M}{i+1}\frac{1}{\lambda_{\mathrm{m}}^2(1+i\rho+iG)^2}\left[\left(4\lambda_{\mathrm{m}}-\frac{2P^2G}{\rho+G}\right)(1+i\rho+iG)+P(P+i\rho P+iG)\right] \quad (46)$$

$$\beta<\frac{PT_1\lambda_{\mathrm{u}}\lambda_{\mathrm{g}}^{3/2}}{6\lambda_{\mathrm{m}}^3}\left[PG+\rho+1+\frac{(M-1)(P-1)G(\rho+G)^{M-1}+(P\rho+(2P-1)G)(\rho+1+G)(\rho+G)^{M-1}}{(\rho+1+G)^M}\right] \quad (47)$$

$$\beta<\frac{4\lambda_{\mathrm{g}}^{3/2}(\rho+G)T_1\lambda_{\mathrm{u}}^2P}{4\lambda_{\mathrm{m}}^3(\rho+2G)}\left[P(\rho+1+2G)-G+\frac{2G}{\rho+G}(\rho+1+G)(1-P)\right.$$
$$\left.+\frac{(\rho+G)^{M-1}}{(\rho+1+G)^M}\left((\rho+G)(4PG-G-GM-P)+4GP-2G+PM\rho\right)\right] \quad (48)$$

wireless backhaul is given by (50), shown at the bottom of the page. The total expected network delay for the high-mobility case in non-interfering wireless backhaul is given by (51), shown at the bottom of the page. The total expected network delay for the static case in interfering wireless backhaul is given by (52), shown at the bottom of the page, and the total expected network delay for the high mobility case in interfering wireless backhaul is given by (53), shown at the bottom of the page.

Corollary 3: In all cases considered under wireless backhaul, the total aggregate delay to be optimized has a derivative that tends to 0 as $\lambda_{\rm s} \to \infty$. So the total delay is not notably reduced by increasing $\lambda_{\rm s}$ past a certain point.

The main difference is that there is no optimal $\lambda_{\rm s}$ that minimize the total aggregate delay. However, we note that the total aggregate delay becomes asymptotically constant as $\lambda_{\rm s} \to 0$. Hence there are diminishing returns from deploying more SAPs. This occurs for both interfering and non-interfering wireless backhaul.

Comparing wired and wireless backhaul from a network perspective, we observe that using wired backhaul would provide higher reliability and an optimal SAP density to aim for, with extremely though poor performance once the SAP density is increased too much beyond this optimal value. Wireless backhaul offers lower reliability, but the delay performance improves and offers greater flexibility in the choice of SAP density, with rapidly though diminishing returns.

TABLE I
SYSTEM PARAMETERS

Parameter	Value
$\lambda_{\rm u}$	1.5×10^{-5}
$\lambda_{\rm s}$	10^{-6} to 2×10^{-3} (Fig 1), 10^{-6} to 4×10^{-3} (Fig 2)
$\lambda_{\rm g}$	10^{-5}
$\lambda_{\rm m}$	10^{-6}
$P_{\rm s}$	23 dBm (0.1995 W)
$P_{\rm g}$	33 dBm (1.995 W)
$P_{\rm m}$	43 dBm (19.95 W)
M	3
N	3
T_1	1
T_2	0.25
γ	4
α	4
β	$10^{-4}, 5 \times 10^{-6}$

V. NUMERICAL RESULTS

In this section, we provide numerical results using the system parameters in Table I unless stated otherwise.

Fig. 1 shows the effect of backhaul on the typical user delay. Comparing Fig. 1(a) and (b), we see that user mobility seems

$$\begin{aligned}\mathbb{E}\left[D^{(\mathrm{Wl+WlBN,S})}\right] &= T_2 p_{\rm a}\lambda_{\rm u}\sum_{i=0}^{N-1}(-1)^i\binom{N}{i+1}\frac{1}{1+i\rho(\gamma,\alpha)}\\ &+\frac{T_1p_{\rm a}^2\lambda_{\rm u}^2}{\lambda_{\rm s}}\sum_{i=0}^{M-1}(-1)^i\binom{M}{i+1}\frac{1}{1+i\left[\rho(\gamma,\alpha)+(P_{\rm m}/P_{\rm s})^{2/\alpha}(\lambda_{\rm m}/\lambda_{\rm s})\gamma^{2/\alpha}C(\alpha)\right]}\\ &+\frac{T_1(1-p_{\rm a})^2\lambda_{\rm u}^2}{\lambda_{\rm m}}\sum_{i=0}^{M-1}(-1)^i\binom{M}{i+1}\frac{1}{1+i\left[\rho(\gamma,\alpha)+(P_{\rm s}/P_{\rm m})^{2/\alpha}(\lambda_{\rm s}/\lambda_{\rm m})\gamma^{2/\alpha}C(\alpha)\right]}\end{aligned}\quad(50)$$

$$\begin{aligned}\mathbb{E}\left[D^{(\mathrm{Wl+WlBN,M})}\right] &= T_2 p_{\rm a}\lambda_{\rm u}\left[1+\rho(\gamma,\alpha)-\frac{\rho(\gamma,\alpha)^N}{(1+\rho(\gamma,\alpha))^{N-1}}\right]\\ &+\frac{T_1p_{\rm a}^2\lambda_{\rm u}^2}{\lambda_{\rm s}}\sum_{i=0}^{M-1}(-1)^i\binom{M}{i+1}\frac{1}{1+i\left[\rho(\gamma,\alpha)+(P_{\rm m}/P_{\rm s})^{2/\alpha}(\lambda_{\rm m}/\lambda_{\rm s})\gamma^{2/\alpha}C(\alpha)\right]}\\ &+\frac{T_1(1-p_{\rm a})^2\lambda_{\rm u}^2}{\lambda_{\rm m}}\sum_{i=0}^{M-1}(-1)^i\binom{M}{i+1}\frac{1}{1+i\left[\rho(\gamma,\alpha)+(P_{\rm s}/P_{\rm m})^{2/\alpha}(\lambda_{\rm s}/\lambda_{\rm m})\gamma^{2/\alpha}C(\alpha)\right]}\end{aligned}\quad(51)$$

$$\begin{aligned}\mathbb{E}\left[D^{(\mathrm{Wl+WlBI,S})}\right] &= T_2 p_{\rm a}\lambda_{\rm u}\sum_{i=0}^{N-1}(-1)^i\binom{N}{i+1}\frac{1}{1+i\left[\rho(\gamma,\alpha)+(P_{\rm m}/P_{\rm g})^{2/\alpha}(\lambda_{\rm m}/\lambda_{\rm g})\gamma^{2/\alpha}C(\alpha)+(P_{\rm s}/P_{\rm g})^{2/\alpha}(\lambda_{\rm s}/\lambda_{\rm g})\gamma^{2/\alpha}C(\alpha)\right]}\\ &+\frac{T_1p_{\rm a}^2\lambda_{\rm u}^2}{\lambda_{\rm s}}\sum_{i=0}^{M-1}(-1)^i\binom{M}{i+1}\frac{1}{1+i\left[\rho(\gamma,\alpha)+(P_{\rm m}/P_{\rm s})^{2/\alpha}(\lambda_{\rm m}/\lambda_{\rm s})\gamma^{2/\alpha}C(\alpha)+(P_{\rm g}/P_{\rm s})^{2/\alpha}(\lambda_{\rm g}/\lambda_{\rm s})\gamma^{2/\alpha}C(\alpha)\right]}\\ &+\frac{T_1(1-p_{\rm a})^2\lambda_{\rm u}^2}{\lambda_{\rm m}}\sum_{i=0}^{M-1}(-1)^i\binom{M}{i+1}\frac{1}{1+i\left[\rho(\gamma,\alpha)+(P_{\rm s}/P_{\rm m})^{2/\alpha}(\lambda_{\rm s}/\lambda_{\rm m})\gamma^{2/\alpha}C(\alpha)+(P_{\rm g}/P_{\rm m})^{2/\alpha}(\lambda_{\rm g}/\lambda_{\rm m})\gamma^{2/\alpha}C(\alpha)\right]}\end{aligned}\quad(52)$$

$$\begin{aligned}\mathbb{E}\left[D^{(\mathrm{Wl+WlBI,M})}\right] &= T_2 p_{\rm a}\lambda_{\rm u}\sum_{i=0}^{N-1}(-1)^i\binom{N}{i+1}\frac{1}{1+i\left[\rho(\gamma,\alpha)+(P_{\rm m}/P_{\rm g})^{2/\alpha}(\lambda_{\rm m}/\lambda_{\rm g})\gamma^{2/\alpha}C(\alpha)+(P_{\rm s}/P_{\rm g})^{2/\alpha}(\lambda_{\rm s}/\lambda_{\rm g})\gamma^{2/\alpha}C(\alpha)\right]}\\ &+\frac{T_1p_{\rm a}^2\lambda_{\rm u}^2}{\lambda_{\rm s}}\frac{(\lambda_{\rm s}+q_1)^M-q_1^M}{\lambda_{\rm s}(\lambda_{\rm s}+q_1)^{M-1}}+\frac{T_1(1-p_{\rm a})^2\lambda_{\rm u}^2}{\lambda_{\rm m}}\frac{(\lambda_{\rm m}+q_2)^M-q_2^M}{\lambda_{\rm m}(\lambda_{\rm m}+q_2)^{M-1}}\end{aligned}\quad(53)$$

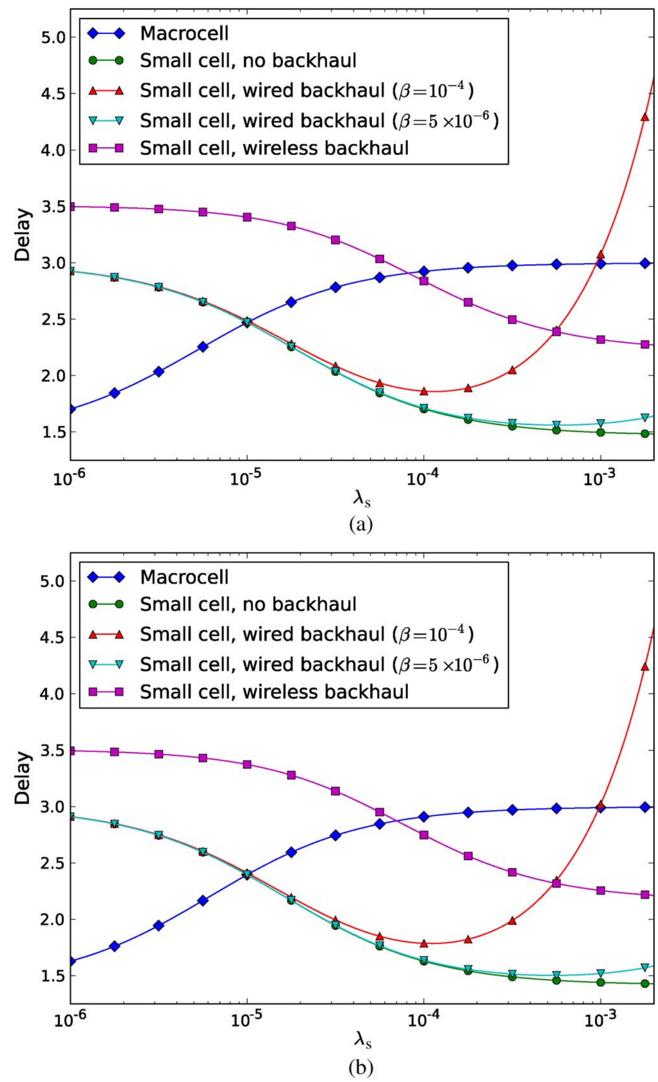

Fig. 1. Effect of backhaul on the typical user delay for different SAP densities. (a) Static case; (b) high-mobility case.

to have negligible effect on the delay. The macrocell delay asymptotically approaches M as $\lambda_s \to \infty$ reaching a ceiling as $\lambda_s \to 0$. Without backhaul, small cell delay asymptotically approaches M as $\lambda_s \to 0$ and has a limit as $\lambda_s \to \infty$. At $\left(\frac{P_s}{P_m}\right)^{2/\alpha} \frac{\lambda_s}{\lambda_m} = 1$, we have the threshold above which small cell delay is less than macrocell delay. For wired backhaul, two values of β are used to compare the effect of the varying quality of backhaul. With higher value of $\beta = 10^{-4}$, the typical small cell user delay quickly exceeds the macrocell delay. With lower value of $\beta = 5 \times 10^{-6}$, the effect of backhaul is far less pronounced and small cell delay remains lower than macrocell delay for the range of λ_s used here. In both cases, the presence of a minimum point indicates that there is an optimal SAP density for a typical small cell user. Wireless delay adds a constant to small cell delay, thus shifting the transition point for which typical small cell delay becomes lower than typical macrocell delay. Overall, we see that depending on the type and the quality of the backhaul used, backhaul can significantly reverse the advantage that small cell users have at sufficiently high SAP densities.

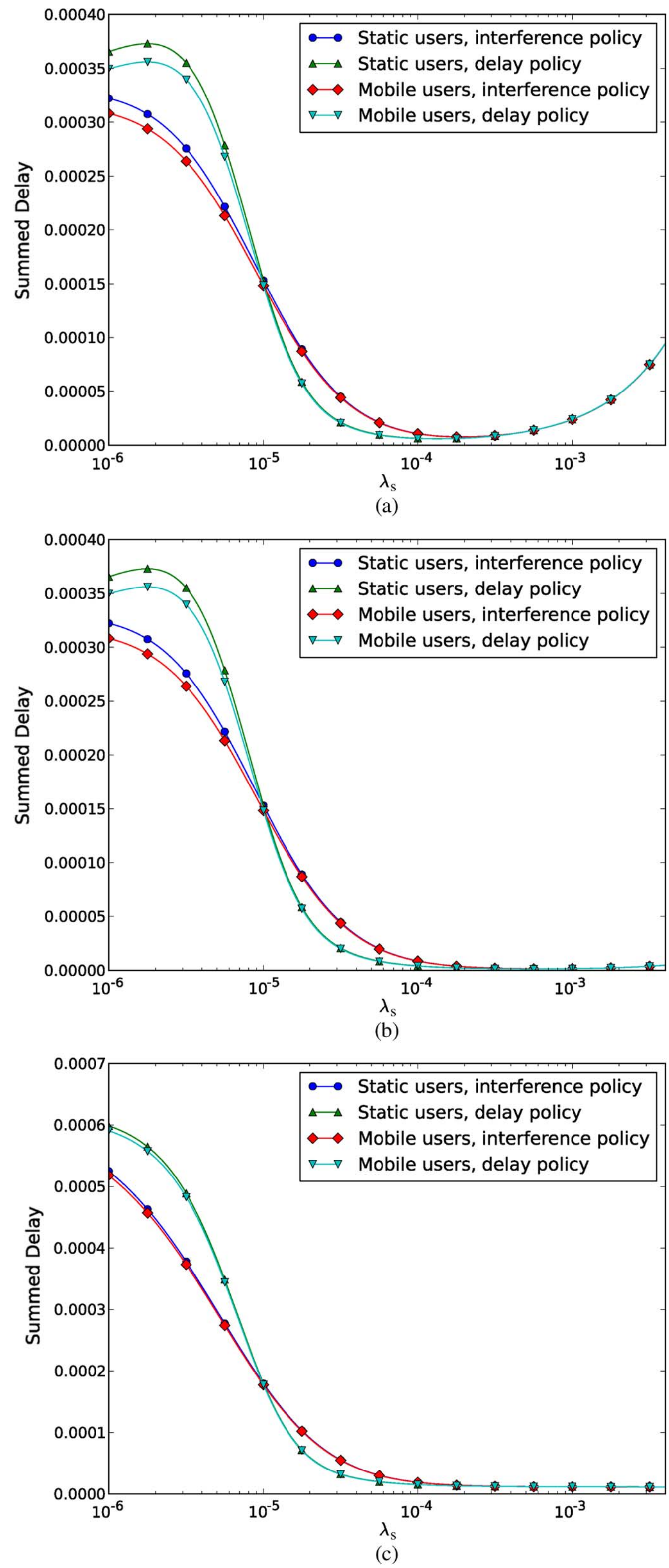

Fig. 2. Total expected network delay vs. SAP density. (a) Wired backhaul with $\beta = 10^{-4}$; (b) wired backhaul with $\beta = 5 \times 10^{-6}$; (c) wireless backhaul.

Fig. 2 shows the total expected network delay under various conditions. Comparing Fig. 2(a) and (b) with Fig. 2(c), we see that the main difference between wired and wireless backhaul in terms of total expected network delay is the existence of an optimal SAP density when there is wired backhaul, which is not the case for the wireless backhaul. The effect of having a lower value of β is again to dampen the effect of increasing delay at high SAP densities. The main takeaway, however, is that in all cases considered here, a sharp improvement in delay experienced takes place around the point where $\left(\frac{P_\mathrm{s}}{P_\mathrm{m}}\right)^{2/\alpha}\frac{\lambda_\mathrm{s}}{\lambda_\mathrm{m}}=1$, and that when λ_s is about an order of magnitude above that, we are close to the optimal point (for wired backhaul) or a point at which increasing SAP density results in minimal improvement in overall delay (wireless backhaul). Additionally, we note that compared to the static user model, the high-mobility user model seems to achieve slightly lower delay, especially for lower SAP density. This may be due to the fact that users being further from SAPs have the chance to relocate to somewhere nearer an access point when retransmitting rather than being stuck at a position that is far from any access point. This would have more effect on the total expected network delay at lower SAP densities, as seen in the numerical results.

VI. Conclusion

In this paper, the effect of wired and wireless backhaul on the performance of heterogeneous cellular networks is studied. The main takeaway is that if we account for the backhaul delay, simplified and unrestricted network densification by deploying SAPs is not necessarily beneficial in terms of the delay experienced by the users.

This is especially true for wired backhaul, where there exists an optimum SAP density given certain operating parameters. For wireless backhaul, we observe that it is not cost-effective to increase SAP density beyond a certain point. Furthermore, we see that wired backhaul provides higher reliability and an optimal SAP density to aim for, but with extremely poor performance once this SAP density is significantly increased beyond this point, whereas wireless backhaul offers greater flexibility in the choice of a near-optimal SAP density, resulting though in lower reliability. Our results shed light to the significant role that backhaul connection plays in the quality of service experienced by users. It is especially worth noting that increasing unrestrictedly and unboundedly the density of small cell networks may not provide a viable long-term solution to mobile data capacity crunch. Moreover, as seen from the effect of the value of β in wired backhaul, improving the capacity, speed or reliability of the backhaul connection is likely to have a greater impact in the medium term. Deploying dense small cell networks may not be effective without similar investment in the backhaul connections.

Since this paper is a first step in the analysis of the backhaul in heterogeneous cellular networks, there remain many interesting issues to be pursued in future work. More elaborated backhaul models need to incorporate the queueing aspect and to provide more insight on the dominant delay. The impact of the backhaul on other performance metrics can also be analyzed.

Appendix

A. Proof of Theorem 1

For the static model, if T_1 is the time taken for a single transmission from the SAP to the user, the expected delay given r is then

$$\begin{aligned}\mathbb{E}&\left[D_\mathrm{s}^{(\mathrm{Wl,S})}|r\right]\\ &=T_1P_\mathrm{s}(r)+2T_1P_\mathrm{s}(r)\left(1-P_\mathrm{s}(r)\right)+3T_1P_\mathrm{s}(r)\left(1-P_\mathrm{s}(r)\right)^2\\ &\quad+\cdots+(M-1)T_1P_\mathrm{s}(r)\left(1-P_\mathrm{s}(r)\right)^{M-2}\\ &\quad+MT_1P_\mathrm{s}(r)\left(1-P_\mathrm{s}(r)\right)^{M-1}+MT_1\left(1-P_\mathrm{s}(r)\right)^{M}. \end{aligned}\tag{54}$$

We can compute the expected delay in another way. The expected delay is at least T_1 with probability 1. The first failure occurs with probability $1-P_\mathrm{s}(r)$ and requires additional time T_1. Given a first failure, the second failure occurs with probability $1-P_\mathrm{s}(r)$ and requires additional time T_1. Continuing in this way, the expected delay is equivalent to

$$\begin{aligned}\mathbb{E}&\left[D_\mathrm{s}^{(\mathrm{Wl,S})}|r\right]\\ &=T_1+\left(1-P_\mathrm{s}(r)\right)\left[T_1+\left(1-P_\mathrm{s}(r)\right)\left(T_1+\cdots\right)\right]\\ &=T_1\left[1+\left(1-P_\mathrm{s}(r)\right)+\left(1-P_\mathrm{s}(r)\right)^2\right.\\ &\qquad\left.+\cdots+\left(1-P_\mathrm{s}(r)\right)^{M-1}\right]\\ &=T_1\left[1-\left(1-P_\mathrm{s}(r)\right)^M\right]/P_\mathrm{s}(r).\end{aligned}\tag{55}$$

By the law of total expectation, the expected wireless delay for a typical user connecting to an SAP is given by (56), shown at the bottom of the page. Similarly, assuming that the time taken for one transmission from the MBS is the same as that from the SAP, the expected wireless delay for a typical user connecting to an MBS is given by (57), shown at the bottom of the page. Note that the temporal correlation of the interference is not taken into account here. Since the success probability is a decreasing event, using the FKG inequality, we can show that our result is a lower bound on the expected delay. For complete characterization of the delay using the joint SIR distribution, tools from [32] can be used.

$$\mathbb{E}\left[D_\mathrm{s}^{(\mathrm{Wl,S})}\right]=T_1\sum_{i=0}^{M-1}(-1)^i\binom{M}{i+1}\frac{1}{1+i\left[\rho(\gamma,\alpha)+(P_\mathrm{m}/P_\mathrm{s})^{2/\alpha}(\lambda_\mathrm{m}/\lambda_\mathrm{s})\gamma^{2/\alpha}C(\alpha)\right]}\tag{56}$$

$$\mathbb{E}\left[D_\mathrm{m}^{(\mathrm{Wl,S})}\right]=T_1\sum_{i=0}^{M-1}(-1)^i\binom{M}{i+1}\frac{1}{1+i\left[\rho(\gamma,\alpha)+(P_\mathrm{s}/P_\mathrm{m})^{2/\alpha}(\lambda_\mathrm{s}/\lambda_\mathrm{m})\gamma^{2/\alpha}C(\alpha)\right]}\tag{57}$$

For the high-mobility case, the SAP-user delay follows a geometric distribution with probability of success as follows:

$$P_{\mathrm{s}}^{(\mathrm{Wl,M})} = \frac{\lambda_{\mathrm{s}}}{\lambda_{\mathrm{s}}(\rho+1) + (P_{\mathrm{m}}/P_{\mathrm{s}})^{2/\alpha}\gamma^{2/\alpha}C(\alpha)\lambda_{\mathrm{m}}}. \quad (58)$$

As such, the expected wireless delay for a typical user connecting to an SAP is given by (59), shown at the bottom of the page. Similarly, following the same procedure, we can straightforwardly obtain the expected wireless delay for a typical user connecting to an MBS.

B. Proof of Lemma 3

We need to prove that $F_1(x) = \sum_{i=0}^{M-1}(-1)^i\binom{M}{i+1}\frac{1}{1+ixL}$ is increasing in x, where $L = \gamma^{2/\alpha}C(\alpha)$. It is sufficient to show that $F_1'(x) > 0,\ \forall M \geq 2, L \geq 0, x \geq 0$.

$$\begin{aligned} F_1'(x) &= \sum_{i=1}^{M-1}(-1)^{i+1}\binom{M}{i+1}\frac{iL}{(1+ixL)^2} \\ &= \sum_{i=1}^{M-1}(-1)^{i+1}\frac{M(M-1)}{i+1}\binom{M-2}{i-1}\frac{L}{(1+ixL)^2} \\ &= M(M-1)(-1)^M\Delta^{M-2}[f](0) \end{aligned} \quad (60)$$

where $f(i) = \frac{L}{(i+2)[1+(i+1)xL]^2}$ and $\Delta^n[f]$ represents the n-th order forward difference. This is always positive since the sign of $f^{(M-2)}(y)$ is $(-1)^M\ \forall y > -2$, as we show by induction below, so the sign of $\Delta^{M-2}[f](0)$ is $(-1)^M$.

Claim: $f^{(n)}(y)$ is of the form $(-1)^n\sum_{j=0}^{n}\frac{c_{n,i}}{(y+2)^{1+j}[1+(y+1)xL]^{2+n-j}}$ $\forall n \geq 0$, where $c_{n,i} \geq 0$.

For $n = 0$, the claim is obviously true. If it is true for some n, then using Leibniz's rule to differentiate gives the claim for $n+1$. Hence, the claim is true by induction. Therefore, we have proved that $F_1(x)$ is increasing in x. For $F_2(x)$, it is straightforward to find $F_2'(x)$ and show that it is always positive. Hence, $F_2(x)$ is also increasing in x.

C. Proof of Theorem 2

This comes from adding the backhaul delay in (6) to the wireless transmission delays found in Theorem 1.

D. Proof of Theorem 5

1) Static Case, Interference-Based Policy: Let $D_{\mathrm{S,i}}$ denote $\mathbb{E}[D^{(\mathrm{Wl+WdB,S})}]$ under the interference-based policy; we first find a solution for $\frac{dD_{\mathrm{S,i}}}{d\lambda_{\mathrm{s}}} = 0$. As $\lambda_{\mathrm{s}} \to \infty$, $\frac{dD_{\mathrm{S,i}}}{d\lambda_{\mathrm{s}}} \to \frac{\beta\lambda_{\mathrm{u}}}{2\lambda_{\mathrm{g}}^{3/2}} > 0$. When $\lambda_{\mathrm{s}} = \frac{\lambda_{\mathrm{m}}}{P}$, we have

$$\frac{dD_{\mathrm{S,i}}}{d\lambda_{\mathrm{s}}} = \frac{1}{4}\sum_{i=0}^{M-1}(-1)^i\binom{M}{i+1}\frac{(P-2)i\gamma^{2/\alpha}C(\alpha) - 1 - i\rho}{(1+i\rho+i\gamma^{2/\alpha}C(\alpha)^2} + \frac{3\beta\lambda_{\mathrm{m}}^2}{8PT_1\lambda_{\mathrm{u}}\lambda_{\mathrm{g}}^{3/2}} < 0 \quad (61)$$

if β satisfies (45). Hence, there exists $z > 0$ such that $D_{\mathrm{S,i}}(\lambda_{\mathrm{s}})$ is decreasing on $(-\infty, -z)$ and increasing on (z, ∞). $D_{\mathrm{S,i}}(\lambda_{\mathrm{s}})$ is continuous on $[-z, z]$, so by the extreme value theorem, $D_{\mathrm{S,i}}(\lambda_{\mathrm{s}})$ attains its minimum on $[-z, z]$, which is also a minimum on $(-\infty, \infty)$.

2) Static Case, Delay-Based Policy: Let $D_{\mathrm{S,d}}$ denote $\mathbb{E}[D^{(\mathrm{Wl+WdB,S})}]$ under the delay-based policy; we first find a solution for $\frac{dD_{\mathrm{S,d}}}{d\lambda_{\mathrm{s}}} = 0$. As $\lambda_{\mathrm{s}} \to \infty$, $\frac{dD_{\mathrm{S,d}}}{d\lambda_{\mathrm{s}}} \to \frac{\beta\lambda_{\mathrm{u}}}{2\lambda_{\mathrm{g}}^{3/2}} > 0$. When $\lambda_{\mathrm{s}} = \frac{\lambda_{\mathrm{m}}}{P}$, we have

$$\begin{aligned} &\frac{dD_{\mathrm{S,d}}}{d\lambda_{\mathrm{s}}} \\ &= \frac{T_1\lambda_{\mathrm{u}}^2PG}{2(\rho+G)\lambda_{\mathrm{m}}}\sum_{i=0}^{M-1}(-1)^{i+1}\binom{M}{i+1}\frac{iPG}{\lambda_{\mathrm{m}}^2(1+i\rho+iG)^2} \\ &+ \frac{T_1\lambda_{\mathrm{u}}^2}{4}\sum_{i=0}^{M-1}(-1)^{i+1}\binom{M}{i+1}\frac{(1+i\rho)P^2}{\lambda_{\mathrm{m}}^2(1+i\rho+iG)^2} \\ &+ \frac{\beta\lambda_{\mathrm{u}}}{4\lambda_{\mathrm{g}}^{3/2}}\left(\frac{G}{P(\rho+G)}+1\right) \\ &- T_1\lambda_{\mathrm{u}}^2\sum_{i=0}^{M-1}(-1)^i\binom{M}{i+1}\frac{1}{\lambda_{\mathrm{m}}(1+i\rho+iG)} \\ &+ \frac{T_1\lambda_{\mathrm{u}}^2}{4}\sum_{i=0}^{M-1}(-1)^{i+1}\binom{M}{i+1}\frac{iPG}{\lambda_{\mathrm{m}}^2(1+i\rho+iG)^2)} < 0 \end{aligned} \quad (62)$$

if β satisfies (46). Hence, there exists $z > 0$ such that $D_{\mathrm{S,d}}(\lambda_{\mathrm{s}})$ is decreasing on $(-\infty, -z)$ and increasing on (z, ∞). $D_{\mathrm{S,d}}(\lambda_{\mathrm{s}})$ is continuous on $[-z, z]$, so by the extreme value theorem, $D_{\mathrm{S,d}}(\lambda_{\mathrm{s}})$ attains its minimum on $[-z, z]$, which is also a minimum on $(-\infty, \infty)$.

3) High-Mobility Case, Interference-Based Policy: Let $D_{\mathrm{M,i}}$ denote $\mathbb{E}[D^{(\mathrm{Wl+WdB,M})}]$ under the interference-based policy; we first find a solution for $\frac{dD_{\mathrm{M,i}}}{d\lambda_{\mathrm{s}}} = 0$. As $\lambda_{\mathrm{s}} \to \infty$, $\frac{dD_{\mathrm{M,i}}}{d\lambda_{\mathrm{s}}} \to \frac{\beta\lambda_{\mathrm{u}}}{2\lambda_{\mathrm{g}}^{3/2}} > 0$. When $\lambda_{\mathrm{s}} = \frac{\lambda_{\mathrm{m}}}{P}$, we have (63), shown at the top of the next page, if β satisfies (47). Hence, there exists $z > 0$ such that $D_{\mathrm{M,i}}(\lambda_{\mathrm{s}})$ is decreasing on $(-\infty, -z)$ and increasing on (z, ∞). $D_{\mathrm{M,i}}(\lambda_{\mathrm{s}})$ is continuous on $[-z, z]$, so by the extreme value theorem, $D_{\mathrm{M,i}}(\lambda_{\mathrm{s}})$ attains its minimum on $[-z, z]$, which is also a minimum on $(-\infty, \infty)$.

$$\begin{aligned} \mathbb{E}\left[D_{\mathrm{s}}^{(\mathrm{Wl,M})}\right] &= T_1\left[1 - \left(1 - P_{\mathrm{s}}^{(\mathrm{Wl,M})}\right)^M\right]/P_{\mathrm{s}}^{(\mathrm{Wl,M})} \\ &= \frac{\left[\lambda_{\mathrm{s}}(\rho+1) + (P_{\mathrm{m}}/P_{\mathrm{s}})^{2/\alpha}\gamma^{2/\alpha}C(\alpha)\lambda_{\mathrm{m}}\right]^M - \left[\lambda_{\mathrm{s}}\rho + (P_{\mathrm{m}}/P_{\mathrm{s}})^{2/\alpha}\gamma^{2/\alpha}C(\alpha)\lambda_{\mathrm{m}}\right]^M}{\lambda_{\mathrm{s}}\left[\lambda_{\mathrm{s}}(\rho+1) + (P_{\mathrm{m}}/P_{\mathrm{s}})^{2/\alpha}\gamma^{2/\alpha}C(\alpha)\lambda_{\mathrm{m}}\right]^{M-1}} \end{aligned} \quad (59)$$

$$\frac{dD_{\mathrm{M,i}}}{d\lambda_{\mathrm{s}}} = \frac{3\beta\lambda_{\mathrm{u}}\lambda_{\mathrm{m}}}{2\lambda_{\mathrm{g}}^{3/2}} - \frac{PT_1\lambda_{\mathrm{u}}^2}{4\lambda_{\mathrm{m}}^2}\left[PG+\rho+1+\frac{(M-1)(P-1)G(\rho+G)^{M-1}+(P\rho+(2P-1)G)(\rho+1+G)(\rho+G)^{M-1}}{(\rho+1+G)^M}\right]<0 \quad (63)$$

$$\frac{dD_{\mathrm{M,d}}}{d\lambda_{\mathrm{s}}} = \frac{\beta\lambda_{\mathrm{m}}}{4\lambda_{\mathrm{g}}^{3/2}}\frac{\rho+2G}{\rho+G} + \frac{T_1\lambda_{\mathrm{u}}^2 P}{4\lambda_{\mathrm{m}}^2}\left[-P(\rho+1+2G)+\frac{P(\rho+G)^M}{(\rho+1+G)^M}-\frac{MP\rho(\rho+G)^{M-1}}{(\rho+1+G)^M}+G+\frac{2G}{\rho+G}(\rho+1+G)(P-1)+\frac{2G(\rho+G)^{M-1}}{(\rho+1+G)^{M-1}}(1-P)+\frac{G(\rho+G)^M(M-1-2P)}{(\rho+1+G)^M}-\frac{2PG(\rho+G)^{M-1}}{(\rho+1+G)^M}\right]<0 \quad (64)$$

4) High-Mobility Case, Delay-Based Policy: Let $D_{\mathrm{M,d}}$ denote $\mathbb{E}[D^{(\mathrm{Wl+WdB,M})}]$ under the delay-based policy; we first find a solution for $\frac{dD_{\mathrm{M,d}}}{d\lambda_{\mathrm{s}}}=0$. As $\lambda_{\mathrm{s}}\to\infty$, $\frac{dD_{\mathrm{M,d}}}{d\lambda_{\mathrm{s}}}\to\frac{\beta\lambda_{\mathrm{u}}}{2\lambda_{\mathrm{g}}^{3/2}}>0$. When $\lambda_{\mathrm{s}}=\frac{\lambda_{\mathrm{m}}}{P}$, we have (64) if β satisfies (48). Hence, there exists $z>0$ such that $D_{\mathrm{M,d}}(\lambda_{\mathrm{s}})$ is decreasing on $(-\infty,-z)$ and increasing on (z,∞). $D_{\mathrm{M,d}}(\lambda_{\mathrm{s}})$ is continuous on $[-z,z]$, so by the extreme value theorem, $D_{\mathrm{M,d}}(\lambda_{\mathrm{s}})$ attains its minimum on $[-z,z]$, which is also a minimum on $(-\infty,\infty)$.

REFERENCES

[1] D. López-Pérez *et al.*, "Enhanced intercell interference coordination challenges in heterogeneous networks," *IEEE Wireless Commun. Mag.*, vol. 18, no. 3, pp. 22–30, Jun. 2011.

[2] J. G. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. C. Reed, "Femtocells: Past, present, future," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 497–508, Apr. 2012.

[3] T. Q. S. Quek, G. de la Roche, İ. Güvenc, and M. Kountouris, *Small Cell Networks—Deployment, PHY Techniques, Resource Management*. Cambridge, U.K.: Cambridge Univ. Press, 2013.

[4] T. Nakamura *et al.*, "Trends in small cell enhancements in LTE advanced," *IEEE Commun. Mag.*, vol. 51, no. 2, pp. 98–105, Feb. 2013.

[5] I. Hwang, B. Song, and S. S. Soliman, "A holistic view on hyper-dense heterogeneous and small cell networks," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 20–27, Jun. 2013.

[6] M. Haenggi and R. K. Ganti, "Interference in large wireless networks," *Found. Trends Netw.*, vol. 3, no. 2, pp. 127–248, Feb. 2009.

[7] J. G. Andrews, R. K. Ganti, M. Haenggi, N. Jindal, and S. Weber, "A primer on spatial modeling and analysis in wireless networks," *IEEE Commun. Mag.*, vol. 48, no. 11, pp. 156–163, Nov. 2010.

[8] M. Z. Win, P. C. Pinto, and L. A. Shepp, "A mathematical theory of network interference and its applications," *Proc. IEEE*, vol. 97, no. 2, pp. 205–230, Feb. 2009.

[9] 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, Sep. 2009.

[10] A. Rabbachin, T. Q. S. Quek, P. C. Pinto, I. Oppermann, and M. Z. Win, "Non-coherent UWB communication in the presence of multiple narrowband interferers," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3365–3379, Nov. 2010.

[11] A. Rabbachin, T. Q. S. Quek, H. Shin, and M. Z. Win, "Cognitive network interference," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 2, pp. 480–493, Feb. 2011.

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

[13] Y. S. Soh, T. Q. S. Quek, M. Kountouris, and H. Shin, "Energy efficient heterogeneous cellular networks," *IEEE J. Select. Areas Commun.*, vol. 31, no. 5, pp. 840–850, May 2013.

[14] M. Wildemeersch, T. Q. S. Quek, C. H. Slump, and A. Rabbachin, "Cognitive small cell networks: Energy efficiency and trade-offs," *IEEE Trans. Commun.*, vol. 61, no. 9, pp. 4016–4029, Sep. 2013.

[15] Y. S. Soh, T. Q. S. Quek, M. Kountouris, and G. Caire, "Cognitive hybrid division duplex for two-tier femtocell networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 10, pp. 4852–4865, Oct. 2013.

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

[17] S. Chia, M. Gasparroni, and P. Brick, "The next challenge for cellular networks: Backhaul," *IEEE Microw. Mag.*, vol. 10, no. 5, pp. 54–66, Aug. 2009.

[18] H. Raza, "A brief survey of radio access network backhaul evolution: Part II," *IEEE Commun. Mag.*, vol. 51, no. 5, pp. 170–177, May 2013.

[19] S. Little, "Is microwave backhaul up to the 4G task?" *IEEE Microw. Mag.*, vol. 10, no. 5, pp. 67–74, Aug. 2009.

[20] O. Tipmongkolsilp, S. Zaghloul, and A. Jukan, "The evolution of cellular backhaul technologies: Current issues and future trends," *IEEE Commun. Mag.*, vol. 13, no. 1, pp. 97–113, 2011.

[21] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, "Robust and efficient distributed compression for cloud radio access networks," *IEEE Trans. Veh. Technol.*, vol. 62, no. 2, pp. 692–703, Feb. 2013.

[22] F. Baccelli, M. Klein, M. Lebourges, and S. Zuyev, "Stochastic geometry and architecture of communication networks," *J. Telecommun. Syst.*, vol. 7, no. 1–3, pp. 209–227, Jun. 1997.

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

[24] B. Błaszczyszyn, M. Karray, and H. Keeler, "Using Poisson processes to model lattice cellular networks," in *Proc. IEEE INFOCOM*, Turin, Italy, Apr. 2013, pp. 773–781.

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

[26] F. Baccelli, M. Klein, M. Lebourges, and S. Zuyev, "Géométrie aléatoire et architecture de réseaux de communications," *Annales Des Télécommun.*, vol. 51, no. 3/4, pp. 158–179, Mar./Apr. 1996.

[27] F. Baccelli and B. Błaszczyszyn, "Stochastic geometry and wireless networks," *Found. Trends Netw.*, vol. 3, no. 3/4, pp. 249–449, 2009.

[28] 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.

[29] X. Lin, R. K. Ganti, P. J. Fleming, and J. G. Andrews, "Towards understanding the fundamentals of mobility in cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 4, pp. 1686–1698, Apr. 2013.

[30] D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed. Englewood Cliffs, NJ, USA: Prentice-Hall, 1992.

[31] F. Baccelli and S. Zuyev, "Poisson-Voronoi spanning trees with applications to the optimization of communication networks," *Oper. Res.*, vol. 47, no. 4, pp. 619–631, Jul./Aug. 1999.

[32] U. Schilcher, C. Bettstetter, and G. Brandner, "Temporal correlation of interference in wireless networks with Rayleigh block fading," *IEEE Trans. Mobile Comput.*, vol. 11, no. 12, pp. 2109–2120, Dec. 2012.

Daniel C. Chen (S'13) received the B.A. degree in mathematics from the University of Cambridge, U.K. in 2012. He is currently working towards the Ph.D. degree in Operations Research at the Massachusetts Institute of Technology. From August 2012 to July 2013, he was a Research Engineer in the Institute for Infocomm Research, Singapore. His current research interests include data driven optimization and its applications to the real world. He received the National Science Scholarship from the Agency for Science, Technology and Research, Singapore in 2006.

Tony Q. S. Quek (S'98–M'08–SM'12) received the B.E. and M.E. degrees in electrical and electronics engineering from Tokyo Institute of Technology, Tokyo, Japan, respectively. At Massachusetts Institute of Technology, he earned the Ph.D. in Electrical Engineering and Computer Science. Currently, he is an Assistant Professor with the Information Systems Technology and Design Pillar at Singapore University of Technology and Design (SUTD). He is also a Scientist with the Institute for Infocomm Research. His main research interests are the application of mathematical, optimization, and statistical theories to communication, networking, signal processing, and resource allocation problems. Specific current research topics include sensor networks, heterogeneous networks, green communications, smart grid, wireless security, compressed sensing, big data processing, and cognitive radio.

Dr. Quek has been actively involved in organizing and chairing sessions, and has served as a member of the Technical Program Committee as well as symposium chairs in a number of international conferences. He is serving as the Co-chair for the PHY & Fundamentals Track for IEEE WCNC in 2015, the Communication Theory Symposium for IEEE ICC in 2015, and the PHY & Fundamentals Track for IEEE EuCNC in 2015. He is currently an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE WIRELESS COMMUNICATIONS LETTERS, and an Executive Editorial Committee Member for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. He was Guest Editor for the IEEE SIGNAL PROCESSING MAGAZINE (Special Issue on Signal Processing for the 5G Revolution) in 2014, and the IEEE WIRELESS COMMUNICATIONS MAGAZINE (Special Issue on Heterogeneous Cloud Radio Access Networks) in 2015.

Dr. Quek was honored with the 2008 Philip Yeo Prize for Outstanding Achievement in Research, the IEEE Globecom 2010 Best Paper Award, the CAS Fellowship for Young International Scientists in 2011, the 2012 IEEE William R. Bennett Prize, the IEEE SPAWC 2013 Best Student Paper Award, and the IEEE WCSP 2014 Best Paper Award.

Marios Kountouris (S'04–M'08–SM'15) received the Diploma in electrical and computer engineering from the National Technical University of Athens, Greece, in 2002 and the M.S. and Ph.D. degrees in electrical engineering from the Ecole Nationale Supérieure des Télécommunications (Télécom ParisTech), France in 2004 and 2008, respectively. His doctoral research was carried out at Eurecom Institute, France, and it was funded by Orange Labs, France. From February 2008 to May 2009, he has been with the Department of ECE at The University of Texas at Austin as a research associate, working on wireless *ad hoc* networks under DARPA's IT-MANET program. From June 2009 to December 2013, he has been an Assistant Professor at the Department of Telecommunications at Supélec (Ecole Supérieure d'Electricité), France, where he is currently an Associate Professor. Since January 2015, he is a Principal Researcher in the Mathematical and Algorithmic Sciences Lab, France Research Center, Huawei Technologies Co. Ltd.

Dr. Kountouris has published several papers and patents all in the area of communications, wireless networks, and signal processing. He has served as technical program committee member for several top international conferences and has served as Workshop Chair for the IEEE Globecom 2010 Workshop on Femtocell Networks, the IEEE ICC 2011 Workshop on Heterogeneous Networks, and the IEEE Globecom 2012 Workshop on Heterogeneous and Small Cell Networks.

He is currently an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the EURASIP Journal on Wireless Communications and Networking, the KSII Transactions on Internet and Information Systems, and the Journal of Communications and Networks (JCN). He is also a founding member and Vice Chair of IEEE SIG on Green Cellular Networks. He received the 2013 IEEE ComSoc Outstanding Young Researcher Award for the EMEA Region, the 2014 EURASIP Best Paper Award for EURASIP Journal on Advances in Signal Processing (JASP), the 2012 IEEE SPS Signal Processing Magazine Award, the IEEE SPAWC 2013 Best Student Paper Award, and the Best Paper Award in Communication Theory Symposium at IEEE Globecom 2009. He is a Professional Engineer of the Technical Chamber of Greece.