An Analytical Model for Flow-level Performance of Large, Randomly Placed Small Cell Networks

George Arvanitakis, Thrasyvoulos Spyropoulos, Florian Kaltenberger
Eurecom Sophia Antipolis, France
{George.Arvanitakis, Thrasyvoulos.Spyropoulos, Florian.Kaltenberger}@eurecom.fr

***Abstract*—In this paper, we develop a flexible and accurate analytical model of large networks with random base station (BS) placement, in order to understand the impact of key network parameters like BS density and load on the network performance. The main goal is to understand the flow level dynamics of such a system, assuming non-saturated users and studying the congestion statistics for BSs and the per flow delay. To achieve this, we base our analysis on two main tools: (a) stochastic geometry, to understand the impact of topological randomness and coverage maps and (b) queueing theory, to model the competition between concurrent flows within the same BS. Our model is then applied the populars Radio Access Technologies (RATs), such as LTE and WIFi. Our results provide some interesting qualitative and quantitative insights about the performance of those networks.**

***Index Terms*—Stochastic Geometry; Queueing; LTE; WiFi; Performance Analysis; Flow-level;**

I. Introduction

The trend of modern networks is to become denser, irregular placed, and more heterogeneous, due to the often unplanned and incremental deployment of new (small cell) BSs. As a result, analyzing such networks, e.g., for protocol comparison or network planning, becomes increasingly challenging. What is more, the usually considered metrics in such analyses, like SINR or capacity, often fail to capture the actual user experience, because flow-level performance (delay, congestion probability, e.t.c.) strongly depends on the network's load and not only the channel conditions [1], [2]. A better metric is latency, which is one of the performance indicators of 5G technologies [3], [4].

To this end, in this paper we present a flexible and accurate model that analyses the performance of random placed networks, in order to understand the impact of important network parameters (BS density, load) on the network's performance. Our model consists of randomly located Base Stations as well as randomly placed users. Users are assumed to be non-saturated, randomly generating requests for new file/flow downloads of varying sizes, and they perceive performance in terms of the average delay to finish such a download.

Our analysis is based on the combination of two key theoretical tools that have recently provided many insights on cellular network performance: (i) We use *queueing theory* to model the performance of dynamic flow arrival and service via the respective scheduler, at the level of a single BS; (ii) We utilize *stochastic geometry*, in order to understand the impact of topological randomness and interaction/competition between BSs at the network level, in order to derive statistics about the *number of users associated with a base station*, and the *modulation and coding schemes (MCS)* offered at each BS. Both these quantities serve as key inputs to the BS queueing model: the former to define the total traffic intensity (in terms of flow arrivals) a given BS has to serve, and the latter to define the average service rate (in terms of flow departures) that a BS is able to offer.

There are a number of works that examine the performance of a network using tools from stochastic geometry: [5] provides distribution of the coverage areas and [6] that derives the distribution of the interference assuming that all neighboring BS are saturated. Additionally, flow-level dynamics of cellular networks have been studied in [2], [7]–[10], some of those focus on spectral efficiency and BS instantaneous throughput, while the rest assume simple cellular topologies (e.g., line networks, or small hexagonal topologies). Compared to these related works, to our best knowledge this is the first work jointly considering the stochastic geometry of the network and flow-level dynamics. Summarizing, our main contributions, are:

I) We present a new analytical result deriving the probability mass function (pmf) of users' cardinality at an arbitrary BS, if both, users and BSs distributed as homogeneous Poisson point process (PPP);

II) We propose an analytical model that captures both physical and MAC layers performance, providing statistics for coverage maps and MCS distributions, as well as flow-level performance as perceived by the user (flow delay) and the network operator (congestion probability);

III) We derive a semi-analytical model that computes the coverage probability of a random placed network, considering the fact that neighboring BSs are not fully loaded (non-saturated) and thus create dynamic interference proportional to their load.

The rest of the paper is organized as follows. In Section II we model performance at the BS level. In Section III, we are modeling the PHY layer. In Section IV, we derive the users cardinality distribution for our topology and we compute the arrival rate. Section V presents the steps in order to specify the service rate, which includes both pure analytical formulas and technical details for each one of the chosen RAT. Section VI, validates our theoretical model and analyses the networks of

This work was funded by SOLDER of the EC Seventh Framework (FP7-CNECT-ICT-619687).

interest. Section VII presents the future steps of our work.

II. Performance at the BS level

We assume that each BS experiences a *dynamic* traffic load and we would like to study the performance at *flow-level*. We state here our assumptions regarding a single randomly chosen BS, and comment where necessary.

A.1: Each *connected* user to a BS generates new *flow* requests randomly, and independently of other users, according to a Poisson Process with density λ_f.

A.2: A flow is a sequence of packets corresponding to the same user or application request (e.g., a file or web page download). Each flow has a random size, in terms of bits, drawn from a *generic* distribution with mean value $\langle s \rangle$.

A.3: The number of users n associated with a BS is a *random* variable with probability mass function (pmf) $f_N(n)$ that depends on the density of the BSs, the density of users, and the association criteria. This pmf will be derived in Section IV.

The following Lemma follows easily, by using a simple Poisson merging argument [11].

Lemma 2.1: If n users are associated with a given BS, the aggregate flow arrival process to that BS is Poisson$(n\lambda_f)$.

Remark: While a Poisson arrival model is pretty standard in related literature, note that if the number of users n at a BS is relatively large, assumption (A.1) can be relaxed to more general traffic arrivals, and we can then use the Palm-Khintchine theorem [11] to support Lemma 2.1 as an approximation.

A.4: In the absence of other flows, *a single flow will be served at full rate*, with the maximum Modulation and Coding Scheme (MCS) that the BS can offer to that UE, which in turns depends on the SINR-BLER specifications for that RAT. The rate of the arbitrary user could be assumed as a random variable and the corresponding pmf, $f_{\mathrm{R}}(r)$, is derived in Section V.

We will assume a single MIMO layer and a single carrier in our analysis [12]. Increased rates due to spatial multiplexing and carrier aggregation can be easily included in the model with a proper physical abstraction models.

A. Queueing Model for BS Schedulers

When more than one flows are served in parallel by a BS, the BS operates as a *queueing system*. The service rate for a flow is generally smaller than what assumption (A.4) predicts. It depends on the number of active flows (BS load), and the centralized scheduler (e.g., in the case of 3G/4G) or distributed media access control (MAC) protocol (in the case of WiFi) which decide how the available resources will be distributed between flows. While a number of different scheduling algorithms exist, we assume for simplicity only the resource-fair one.

Resource Fair Scheduler: Assume all flows are allocated the same amount of resources by the BS, and are served simultaneously, e.g., in a round-robin, TDMA-like manner. If the service time slot is small (e.g., of packet size) compared to the total size of a flow, the flow level performance at that BS can be approximated by a multi-class M/G/1 Processor Sharing (PS) system. This model has already been used to analyzed 3G/3G+ BS performance [2], [9].

LTE schedulers are significantly more complex, allocating competing flows both time and frequency resources (Resource Blocks), possibly taking into account the queue backlog of each flow and flow priority, and also attempting to take advantage of instantaneous SINR variations in time and frequency to achieve further multi-user diversity [12]. While a large number of algorithms have been proposed [13], in the lack flow priority, most implemented schedulers lead to a proportionally fair throughput allocation between flows [12].

The following is a direct application of the multi-class M/G/1/PS result [14].

Lemma 2.2: For a BS with n users generating flows of mean size $\langle s \rangle$, with instantaneous transmission rates drawn from distribution $f_{\mathrm{R}}(r)$, and allocated resources by a resource fair scheduler, the effective service rate of the cell is

$$\langle \mu \rangle = \left(\sum_r \frac{f_R(r) \cdot \langle s \rangle}{r} \right)^{-1} \text{ flows/sec,} \tag{1}$$

and the mean flow delay is given by

$$Delay = \frac{1}{\langle \mu \rangle - n\lambda_f} \ , \tag{2}$$

we define the BS's load as $\rho = \frac{\text{input job rate}}{\text{service job rate}} = \frac{n\lambda_f}{\langle \mu \rangle}$ when the system is stable $\rho < 1$.

Performance gains from opportunistic scheduling can be included in the above equation as a multiplicative factor in front of $\langle \mu \rangle$.

Another often studied scheduler (and good approximation for the 802.11 [15]) is the throughput fair, which equalizes the per flow throughput for all nodes. We ignore it here and we assume that 802.11 performs as resource fair scheduler (which asymptotically the best case, as the load goes to zero), for the following reasons: i) assuming 802.11n characteristics the difference between those two schedulers is negligible, for small average flow size ($\approx$ 1 Mb) and utilization less than 70%, [16], ii) with minor modifications 802.11 is able to operates almost as a resource fair scheduler [17].

B. Network-wide Performance

Our goal in this paper is to understand the network's performance along two main dimensions:

- *Congestion Probability*: We would like to know the percentage of BS whose input load $n \cdot \lambda_f$ exceeds the available service capacity $\langle \mu \rangle$ thus exhibiting per flow delays that grow to infinity. Using the pmf of users' cardinality, congestion probability is

$$P_{cong} = P(n > N_{max}) \ , \tag{3}$$

where $N_{max} = \frac{\langle \mu \rangle}{\lambda_f}$ is the maximum number of users that a BS could serve.

- *Per flow delay*: we would like to know the expected network-wide delay for a randomly chosen user flow, when this flow is served by a stable BS.

These metrics depend on the same two key parameters:

1) The cardinality n of the users associated at a BS, which is a random variable with pmf $f_N(n)$ that depends on the topology of BS and user density.
2) The probability that each of these users is served with a given rate r, namely the rate distribution $f_R(r)$ for this BS that depends on the topology and mutual interference between nearby BSs.

We derive $f_N(n)$ in Section IV and then derive $f_R(r)$ in Section V.

III. PHY Layer Modeling

Before we proceed with the derivation of the cardinality and rate probability distributions, we state here our assumptions about the network topology and physical layer model.

A.5: Users are distributed according to an independent Poisson Point Process with density λ_u.

A.6: The number of BSs inside an area S follows a homogeneous Poisson Point Process (PPP), Φ_{BS}, with density λ_{BS}. Therefore, the number of BSs in an are S

$$P(N = n \mid S) = \frac{(\lambda_{\text{BS}}S)^n e^{-\lambda_{\text{BS}}S}}{n!}, \quad n = 0, 1, \dots \quad . \tag{4}$$

A.7: A standard power loss propagation model is used. We assume a path loss exponent $\alpha > 2$ (for $\alpha \leq 2$ the denominator of SINR goes to infinity), Rayleigh fading at the channel with mean 1 and constant transmit power of P_{tx}. So, the received power at distance d from the BS is given by $P_{\text{rx}} = hd^{-\alpha}$ where h follows an exponential distribution, $h \sim \exp(P_{\text{tx}})$. Hence, the SINR is given by

$$SINR_i = \frac{P_{\text{rx}_i}}{\sum\limits_{n \neq i} P_{\text{rx}_n} + \sigma^2} , \tag{5}$$

where sigma is the thermal noise.

A.8: We assume that all BSs have equal transmit power and implement the same scheduling policy.

Assuming that on average, the received power is monotonic in respect to distance, our criterion is simplified to the closest distance criterion, so, the BSs's coverage areas could be represented by Voronoi Regions (Tessellations).

IV. Cardinality of Associated Users

We are now ready to consider the pmf of the users' cardinality for an arbitrary BS, $f_N(n)$, which as explained earlier decides the total input traffic to each BS. Observe that the size of an arbitrary cell is a random variable, depending on the random BS topology, and the number of users given a specific cell size is also a random variable. The proof for the following theorem and a useful and accurate approximation could be found at our technical report [18].

Theorem 4.1: Consider BSs distributed in 2D as a homogeneous PPP with density λ_{BS}, and offering coverage to a set of users distributed as another PPP with density λ_{u}. Assume further that user association within this tier is done using the closest-distance rule, as explained in Section III. Then, the probability of having exactly n users in an arbitrary cell, $f_N(n)$, is given by:

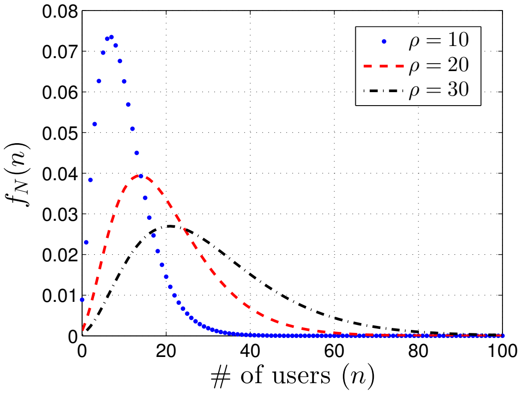


Fig. 1: Pmf of number of users per BS for different values of ratio $\rho = \frac{\lambda_{\text{u}}}{\lambda_{\text{BS}}}$

$$f_N(n) = \frac{343}{n!15}\sqrt{\frac{7}{2\pi}}\frac{\rho^n}{(\rho+\frac{7}{2})^{n+\frac{7}{2}}}\Gamma(n+\frac{7}{2}) , \tag{6}$$

where $\rho = \frac{\lambda_{\text{u}}}{\lambda_{\text{BS}}}$ and Γ is the gamma distribution. Figure 1 depict the user cardinality pmf for different values of ratio ρ.

V. MCS Distribution for each RAT

We are interested in the maximum rate (or equivalently maximum MCS) a user can receive data from the BS it is associated with, given a desired BLER. Our goal is to derive the rate distribution $f_R(r)$ in order to calculate the service rate $\langle\mu\rangle$ in terms of flows/sec for the average BS. This rate depends on the SINR for that user. A given SINR is mapped to an offered MCS [16]. The SINR in turn depends on both the distance of the user to the serving BS and the interference from other nearby BS. Furthermore, a nearby BS might not interfere if it is actually not transmitting at that time, which further complicates analysis. For this reason, we will first consider a "saturated" scenario where interfering BS are assumed to always be ON and interfering. We will then consider the case of load-based interference, where a BS only interferes if it is currently *active* serving at least one user.

A. Rate Distribution for Always ON Interference

We will assume again that BSs and users are distributed according to independent homogeneous PPPs. In [6], the authors present an approach to derive the "coverage probability" of a randomly located user, i.e., the probability that the user's SINR is above a certain threshold. In doing so, it is assumed that interfering BSs always transmit with a power P_{tx}. This assumption is a good approximation when the load of the system is high, in which case the utilization of most BS is close to 1 (i.e., are serving users most of the time). It can also be a valid assumption if the SINR at the user is measured with respect to Reference Signals (i.e., "pilots") that are transmitted at specific times slots by all BS, regardless of whether a BS is serving users or not at that time [12]. Nevertheless, this is not always the case. As a result, in scenarios where BS utilization is lower, this assumption might lead to fairly pessimistic results. We consider this in Section V-B.

For the sake of completeness, we mention here again the results from [6] that are applicable to our problem: Given a BS density λ_{BS}, and path loss constant α, the coverage probability for an SINR threshold T is

$$\begin{aligned} p_c\left(T,\lambda_{BS},\alpha\right) &\triangleq \mathbb{P}\left[SINR > T\right] \\ &= \pi\lambda_{BS}\int_0^\infty e^{-\pi\lambda_{BS}u(1+\beta(T,\alpha))-\frac{1}{\mu}T\sigma^2u^{\alpha/2}}du \ , \end{aligned} \qquad (7)$$

where $\beta\left(T,\alpha\right) = T^{2/\alpha}\int_{T^{-2/\alpha}}^{\infty}\frac{1}{1+u^{\alpha/2}}du$.

If we assume that additive noise is negligible w.r.t. interference (a reasonable assumption for the dense modern networks) Eq. (7) can be significantly simplified as $p_c\left(T,\lambda_{BS},\alpha\right) = 1/\left(1+\beta\left(T,\alpha\right)\right)$. Furthermore, if we assume that $\alpha = 4$, we obtain an elegant closed form solution

$$p_c\left(T,\lambda_{BS},4\right) = \frac{1}{1+\sqrt{T}\left(\pi/2-\arctan\left(1/\sqrt{T}\right)\right)} \ . \qquad (8)$$

Finally, assuming and SINR threshold τ_i for each MCS (mcs_i), the pmf of the MCS $f_{MCS}(mcs)$ can be obtained at Eq. (9) through the coverage probability.

$$f_{\text{MCS}}(mcs_i) = p_c\left(\tau_i,\lambda,\alpha\right) - p_c\left(\tau_{(i+1)},\lambda,\alpha\right) \ . \qquad (9)$$

Given the MCS, the actual rate can be easily calculated based on the total bandwidth of the system in question. Existence of multiple antenna ports and resulting MIMO layers can easily be added in this calculation. Similarly for independent carriers, by deriving the respective MCS for each.

B. Rate Distribution for Load-based Interference

As mentioned earlier, the previous results assume that all BS are interfering all the time. In practice, when the load ρ of a BS A is low, e.g., $\rho = 0.5$, then BS A would be transmitting and causing interference only 50% of the time [1]. This implies that another nearby BS B will be actually serving users at higher rates than the ones predicted in the saturated case. This, in turn, means that BS B will also have a higher $\langle\mu\rangle$ and thus lower utilization $\rho = \frac{\lambda}{\langle\mu\rangle}$ than the one predicted, which in turn creates less interference for BS A.

At flow level, this creates a system of dependent PS queues, which is notoriously hard to analyze at Markov chain level (see e.g. [8] for an attempt to derive some performance bounds). We choose to take here a different approach and use an iterative algorithm in order to calculate $\langle\mu\rangle$ of those dependent BSs. Before the algorithm, we have to present the new coverage probability which takes into account the load of interfering BSs and will be one of the components of the algorithm. The following lemma extends the previous analysis based on stochastic geometry, in order to approximate the coverage probability of the load-based interference scenario.

Lemma 5.1: The coverage probability of an arbitrary user in a random cellular network, assuming that BSs are interfering with each other only for the amount of time that they are serving a user is

$$\begin{aligned} p_c^{lb}\left(T,\lambda,\alpha\right) = \sum_{n=0}^{N_{max}-1}&\left(f_N(n)\frac{1}{1+\mathcal{A}_\rho}\right) \\ &+\overline{F_N}\left(N_{max}\right)\frac{1}{1+\mathcal{A}_{\rho=1}} \ . \end{aligned} \qquad (10)$$

Where $\mathcal{A}_\rho = \left(T\rho\right)^{2/\alpha}\int_{(T\rho)^{2/\alpha}}^{\infty}\frac{1}{1+u^{\alpha/2}}du$ and $\overline{F_N}$ is the ccdf of users' cardinality, Section IV.

Assuming $\alpha = 4$, Eq. (10) could further simplified by replacing $\mathcal{A}_\rho$ and $\mathcal{A}_{\rho=1}$ with

$$\begin{aligned} \mathcal{A}_\rho &= \sqrt{\frac{T}{N_{max}}n}\cdot arccot\left(\frac{1}{\sqrt{\frac{T}{N_{max}}n}}\right) \\ \mathcal{A}_{\rho=1} &= \sqrt{T}\cdot arccot\left(\frac{1}{\sqrt{T}}\right) \ . \end{aligned} \qquad (11)$$

The proof of this lemma can be found in Appendix A.

Iterative Algorithm

The calculation of $\langle\mu\rangle$ in load-based scenario is not trivial, but could be estimated iteratively. The steps are:

1) Initially, we calculate the average service rate $\langle\mu\rangle$, Eq. (1), supposing rate distribution of Always ON Interference (worst case).
2) With the given $\langle\mu\rangle$ the MCS distribution is calculated using Eq. (10).
3) With the given MCS distribution we re-calculate the $\langle\mu\rangle$ and go back to step 2.

C. Rate for each RAT

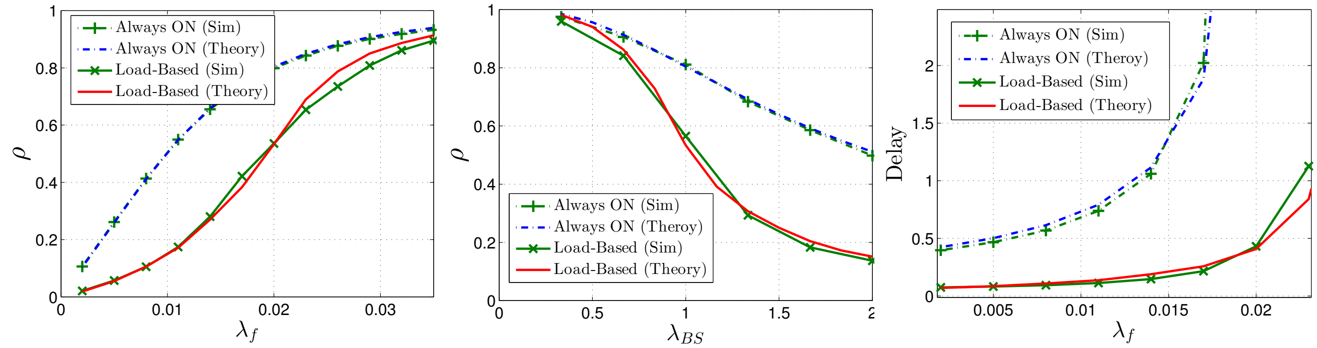
Two parameters are missing in order to derive $\langle\mu\rangle$. Firstly, we need SINR thresholds τ_i for each MCS mode to calculate f_{MCS} from Eq. (9) and secondly, the corresponding rate of each MCS.

The supported MCS are RAT dependent and always are defined at the standard documents [19], [20]. On the other hand, operation threshold for each MCS is not always defined in the protocol since it depends on the receiver implementation characteristics. For example that we demonstrate in this paper we will need one SINR-rate table for the LTE modes and one for WiFi, those tables could be found at our technical report [16].

VI. Simulation Section

The model parameters, for the rest of the simulation section are summarized as: (i) 5 Mbits average flow size, (ii) pathloss $\alpha = 4$, (iii) thermal noise $\sigma^2 = -100$dBm (iv) $BW_{LTE} =$

[1] Even if the SINR estimate is based on the pilot signals, which are always transmitted at the designated LTE resource elements, the *actual* interference experienced during transmission will be lower in practice, leading to better effective rates (e.g., due to fewer HARQ retransmissions required).



a: Load ρ w.r.t flow density λ_f, BS density $\lambda_{BS} = 1$

b: Load ρ w.r.t BS density λ_{BS}, flow density $\lambda_f = 0.02$

c: Delay w.r.t flow density λ_f, BS density $\lambda_{BS} = 1$

Fig. 2: Validation Plots

$BW_{WiFi} = 20$MHz, (v) one antenna per eNodeB and one spatial stream per WiFi AP.

We should mention that if the thermal noise is much smaller than the interference, the value of P_{tx} does not affect the results, as shown in [6].

A. Validation / Performance Analysis

Firstly, we should validate the theoretical models introduced in previous sections. An LTE network is considered in order to compare our theoretical predictions with simulation results. The performance metrics that are used for the comparison are (i) average BS load and (ii) average users' delay. The latter is computed by averaging the users' delay per BS and then taking the median; otherwise even a single congested BS explodes the average delay to infinity.

The simulator generates both BSs and users randomly placed in a large surface with given densities (λ_{BS}, λ_u). Users are associated with the closest BS and generate flows according to Poisson distribution with density λ_f. The flows are forwarded to the corresponding BS which is modeled as a multi-class M/G/1/PS. The service rate of each flow for every time quantum is calculated via SINR. We are interested about two interference scenarios: (1) always ON case, all the neighboring BS are contribute to the interference, (2) load-based case, at the calculation of the interference we taking into consideration only the base stations that are ON at this time quantum. We consider only the users those SINR is at least higher than the threshold of the lowest MCS at always ON case.

Fig. 2 (a) and (b), present the average load ρ (see lemma 2.2) of the system w.r.t. λ_f and λ_{BS} respectively, for both scenarios $\lambda_u = 200$. Two general comments from those plots are i) both of theoretical results is quite accurate, ii) the gap between always ON and load-based interference could be extremely high.

In Fig. 2 (a), for $\lambda_f = 0.02$ the always ON prediction is that the network is 70% loaded instead of 30% of the load-based. That means that the network could be much more robust w.r.t. data traffic than the studies that assume saturated BSs predict.

In Fig. 2 (b), for high density of BS always ON model predicts 50% utilized network, while load-based only 15%. The gap between always ON and load-based prediction increases w.r.t. density of the network. This happens because saturated analysis is able to capture only the gain coming from the fact that an arbitrary BS on average serves less users at a denser network, but not the gain coming from the fact that surroundings BSs will be less loaded, and therefore will cause less interference. Thus, the gain to deploy a denser network is much higher than predicted by an analysis that does not take the load-depended interference into account.

Fig. 2 (c), shows the median delay of the simulator as well as the theoretical predictions for saturated and load-based cases. Again the theoretical predictions are quite accurate, on the other hand, always ON interference differs orders of magnitude from the load based due to delay's sensitivity at average service rate.

B. Different RATs

Given that the validation of the theory worked well, in this section we will use directly the theoretical results, in order to avoid figures being too cluttered. Fig. 3 and 4 present compactly the performance (congested probability and delay) for the two networks of interest (LTE, WiFi) for the same density of connected users ($\lambda_u = 100$). Taking into account that we have assumed the MAC performance of WiFi equal with LTE (best case, valid for low load or with small modifications as discussed at Section II) all differences between the RATs are due to the PHY characteristics of RATs (different MCS threshold and different rates).

First, focusing on the saturated case, perhaps is not clear why LTE network performs worst than WiFi, especially if we take into account that for same SINR, LTE operates with higher rate. The reason are the edge users, since LTE is much more robust to low SINR compared to WiF. Users with low SINR, are achieving a low bit rate in LTE, as opposed to WiFi where they would be regarded as "out of service", and therefore are not taken into consideration. We should mention that on the one hand both networks have the same number of

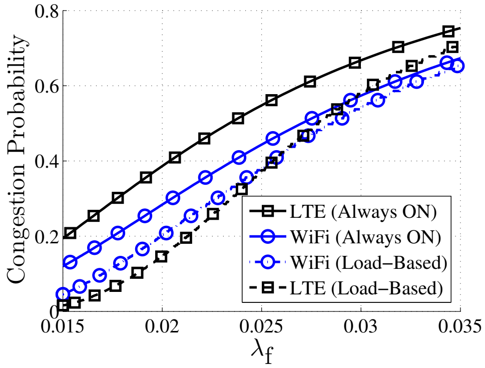


Fig. 3: Congestion probability w.r.t flow density λ_f

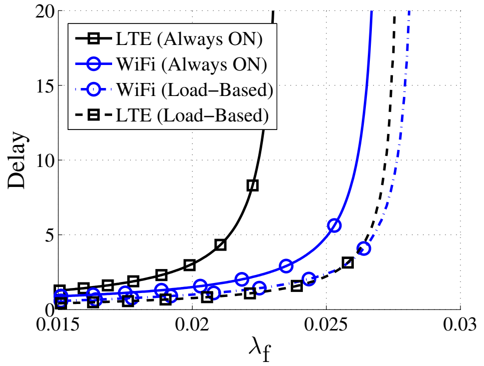


Fig. 4: Delay of each network w.r.t flow density λ_f

connected users, but on the other hand LTE network's coverage area is wider than WiFi. For the always ON case, coverage area is 0.67 and 0.47 of the total area for LTE and WiFi networks respectively.

The low values of coverage area originated from two previous-mentioned worst case assumptions 1) the random BS placement; is possible a BS to ends up asymptotical close to one else and 2) the interference is calculated assuming that neighboring BSs are saturated. By examining the load-based case, we notice how critical the second assumption is, for low values of load the coverage area is almost 1 for both RATs, while for a load around $\rho = 0.5$ coverage areas are 0.9 and 0.7 for LTE and WiFi respectively.

Another interesting remark is that for low or middle-load scenarios in contradiction to always ON case the LTE operates better than WiFi. his happened due to LTE's smaller granularity between the MCS, so, the LTE attain higher SINR improvement. As the load increases the two networks approaching the always ON case which WiFi operates better.

VII. Conclusions / Future Work

We presented an analytical framework to model the flow-level performance of large randomly placed networks assuming saturated BS, as well as a semi-analytical model for the more realistic case of load-based interference. The gap between those two cases could be huge, leading to an underestimation of the network performance. If the BSs do not interfere all the time, the network is much more robust to the total incoming load and the gain of denser deployment is much higher than the saturated case predicts. Additionally, which network's PHY characteristics are perform better turns out that is load-dependent.

As future work, we will apply the same framework in multi-tier HetNet scenarios in order to analyze different inter-tier association and aggregation criteria.

Appendix A
Derivation of Load-Based Coverage Probability

Lemma A.1: Distance of an arbitrary user to the nearest BS is a random variable r with pdf

$$f_r\left(r\right) = e^{-\lambda\pi r^2} 2\pi\lambda r \; . \tag{12}$$

Positions of BS are described by a 2-D homogenous Poisson process, so the cdf is given by $P\left[r \leq R\right] = F_r\left(R\right) = 1 - e^{-\lambda\pi R^2}$ and the pdf can be found as $f_r\left(r\right) = \frac{dF_r(r)}{dr}$.

Coverage Probability: is the probability that SINR of an arbitrary user is greater than a given threshold T

$$\begin{aligned} p_c\left(T,\lambda,\alpha\right) &= E_r\left[P\left[SINR > T|r\right]\right] \\ &= \int_{r>0} P\left[SINR > T|r\right] f_r\left(r\right) dr \\ &= \int_{r>0} P\left[h > Tr^{\alpha}\left(\sigma^2 + I_r\right)|r\right] e^{-\lambda\pi R^2} 2\pi\lambda r dr \; . \end{aligned} \tag{13}$$

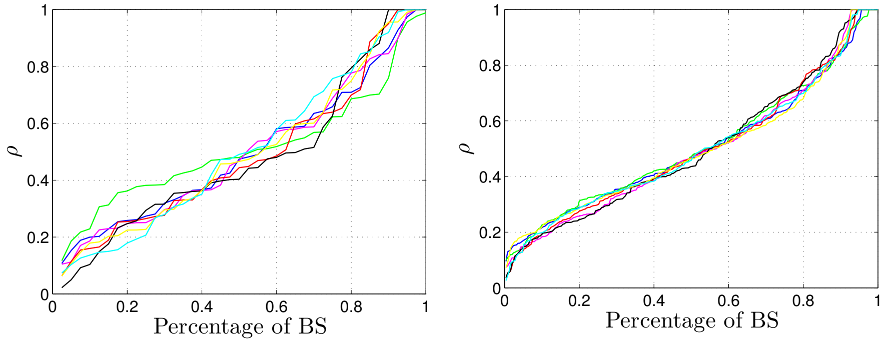
Where I_r is the mean Interference at distance r (the rest of parameters have been defined at *A.7*). Taking into account the channel model $h \sim \exp\left(P_{\text{tx}}\right)$, The probability $P\left[h > Tr^{\alpha}\left(\sigma^2 + I_r\right)|r\right]$ could be re-defined

$$\begin{aligned} P\left[h > Tr^{\alpha}\left(\sigma^2 + I_r\right)|r\right] &= E_{I_r}\left[e^{\left(-P_{\text{tx}}Tr^{\alpha}\left(\sigma^2+I_r\right)\right)}|r\right] \\ &= e^{-P_{\text{tx}}Tr^{\alpha}\sigma^2} E_{I_r}\left[e^{\left(-P_{\text{tx}}Tr^{\alpha}I_r\right)}|r\right] \; . \end{aligned} \tag{14}$$

In the non-saturated case interference is given by $I = \sum_{i\in\Phi n\{b_0\}} \rho_i h_i R_i^{-\alpha}$, where ρ_i is the utility of the i-th BS which is equal to the probability to be ON. So, setting $s = P_{\text{tx}}Tr^{\alpha}$, the expectation of Eq.(14), $E_{I_r}\left[\exp\left(-P_{\text{tx}}Tr^{\alpha}I_r\right)|r\right]$, could be re-written as

$$\begin{aligned} E_{I_r}\left[\exp\left(-sI_r\right)|r\right] &= E_{\rho,\Phi,h}\left[\exp\left(-s\sum_{i\in\Phi n\{b_0\}} h_i\rho_i R_i^{-\alpha}\right)\right] \\ &= E_{\rho,\Phi}\left[\prod_{i\in\Phi n\{b_0\}} \frac{P_{\text{tx}}}{P_{\text{tx}} + s\rho_i R_i^{-\alpha}}\right] \; , \end{aligned} \tag{15}$$

Lemma A.2: As the cardinality of BSs is raising, the distribution of ρ becomes independent from Φ's realizations. Thus, $E_{\rho,\Phi}[\cdot]$ could be treated as two independent expectations $E_{\rho}[E_{\Phi}[\cdot]]$. This happens because of the law of large numbers and the ergodicity of the process and can be illustrated at Fig. 5, where the "variance" of the cdf ρ is decreasing w.r.t. BS cardinality.



a: $\#BSs = 90$

b: $\#BSs = 250$

Fig. 5: different realizations of load for different number of BSs

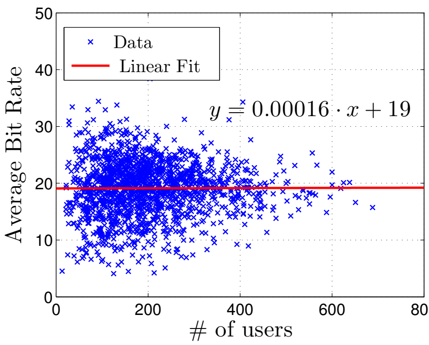


Fig. 6: Linear Interpolation of Average Rate w.r.t. users' cardinality

Additionally, due to the properties of exponential distribution, $E_\Phi\left[\prod_{x\in\Phi} f(x)\right] = \exp\left(-\lambda\int_{R^2}\left(1-f(x)\right)dx\right)$, and after some trivial calculations Eq.(15) is equal to

$$E_\rho\left[\exp\left(-2\pi\lambda\int_r^\infty\left(1-\frac{1}{1+T\rho\left(\frac{r}{R}\right)^\alpha}\right)RdR\right)\right].$$

Finally, by setting $u=\left(\frac{R}{r(\rho T)^{1/\alpha}}\right)^2$ the initial expectation of eq. 14 is equal to

$$E_{I_r}\left[\exp\left(-sI_r\right)|r\right] = E_\rho\left[\exp\left(-2\pi\lambda\mathcal{A}_\rho\right)\right]\ . \tag{16}$$

Where $\mathcal{A}_\rho = (T\rho)^{2/\alpha}\int_{(T\rho)^{2/\alpha}}^\infty \frac{1}{1+u^\alpha/2}du$.

So, by replacing Eq.(14), (16) to Eq.(13) the coverage probability becomes

$$\begin{aligned} p_c\left(T,\lambda,\alpha\right) &= \int_{r>0} 2\pi\lambda r e^{-\lambda\pi r^2} e^{-P_{\text{tx}}Tr^\alpha\sigma^2} E_\rho\left[e^{-2\pi\lambda\mathcal{A}_\rho}\right] dr \\ &= E_\rho\left[\int_{r>0} 2\pi\lambda r e^{-\lambda\pi r^2(1+\mathcal{A}_\rho)} e^{-P_{\text{tx}}Tr^\alpha\sigma^2} dr\right]\ . \end{aligned}$$

The above equation can be significantly simplified under the assumption that $\sigma^2 << I$ so $\sigma^2 = 0$.

$$p_c\left(T,\lambda,\alpha\right) = E_\rho\left[\frac{1}{1+\mathcal{A}_\rho}\right]\ . \tag{17}$$

Lemma A.3: We assume that cell's average service rate $\langle\mu\rangle$ is independent from the users' cardinality. There is a dependency between cell size and users' cardinality as well as between cell's size and $\langle\mu\rangle$. We state that the dependent of those dependencies is negligible.

A large scale topology is presented in Fig. 6. Each dot represents a BS of a given number of users and average cell rate. We can observe that the linear fit is almost constant w.r.t. the cardinality of users; the linear term is 5 orders of magnitude less than the constant term. So, our assumption that the $\langle\mu\rangle$ and the number of associated users could be treated as independent variables is confirmed.

Applying lemma A.3 we assume that $\langle\mu\rangle$ is the equal to all cells. Thus, we define ρ distribution via users' cardinality $f_N(n)$ see Eq.(18), where $N_{max} = \frac{\langle\mu\rangle}{\lambda_f}$ is the maximum number of users that a BS could serve. Applying Eq.(18) to Eq.(17) we end up to Eq.(10)

$$\rho = \begin{cases} \frac{n\cdot\lambda_f}{\langle\mu\rangle} & , n < N_{max} \\ 1 & , n \geq N_{max}\ . \end{cases} \tag{18}$$

References

[1] J. Andrews, S. Singh, Q. Ye, X. Lin, and H. Dhillon, "An overview of load balancing in hetnets: old myths and open problems," *Wireless Communications, IEEE*, 2014.

[2] A. P. Thomas Bonald, "Wireless downlink data channels: User performance and cell dimensioning," in *ACM MOBICOM*, 2003.

[3] Nokia, "5g use cases and requirements," *White Paper*, 2015.

[4] Ericsson, "5g radio access," *White Paper*, 2015.

[5] F. Baccelli, B. Blaszczyszyn, and F. Tournois, "Spatial averages of downlink coverage characteristics in cdma networks," in *INFOCOM IEEE*, 2002.

[6] J. Andrews, F. Baccelli, and R. Ganti, "A tractable approach to coverage and rate in cellular networks," *Communications, IEEE Transactions on*, 2011.

[7] J. R. Thomas Bonald, "Scheduling network traffic," in *ACM SIGMETRICS*, 2007.

[8] T. Bonald and A. Proutière, "On performance bounds for the integration of elastic and adaptive streaming flows," in *ACM SIGMETRICS*, 2004.

[9] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks," *Networking, IEEE/ACM Transactions on Networking*, 2005.

[10] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "Distributed α -optimal user association and cell load balancing in wireless networks," *IEEE/ACM Transactions on Networking,*, 2012.

[11] M. Harchol-Balter, *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*. Cambridge University Press.

[12] S. Sesia, I. Toufik, and M. Baker, *LTE - the UMTS long term evolution : from theory to practice*. Wiley, 2009.

[13] F. Capozzi, G. Piro, L. Grieco, G. Boggia, and P. Camarda, "Downlink packet scheduling in lte cellular networks: Key design issues and a survey," *Communications Surveys Tutorials, IEEE*, 2013.

[14] G. Fayolle, I. Mitrani, and R. Iasnogorodski, "Sharing a processor among many job classes," *J. ACM*, 1980.

[15] M. Heusse, F. Rousseau, G. Berger-Sabbatel, and A. Duda, "Performance anomaly of 802.11b," in *INFOCOM IEEE*, 2003.

[16] G. Arvanitakis and F. Kaltenberger, "Phy layer modeling of lte and wifi rats," Eurecom, Tech. Rep. RR-16-317, 2016.

[17] Y. Lin and V. Wong, "Wsn01-1: Frame aggregation and optimal frame size adaptation for ieee 802.11n wlans," in *GLOBECOM IEEE*, 2006.

[18] G. Arvanitakis, "Distribution of the number of poisson points in poisson voronoi tessellation," Eurecom, Tech. Rep. RR-15-304, 2014.

[19] *LTE Specifications, http://www.3gpp.org/DynaReport/36-series.htm.*

[20] *802.11 Specifications, http://standards.ieee.org/about/get/802/802.11.html.*