User Association with Unequal User Priorities in Heterogeneous Cellular Networks

Youjia Chen, Jun Li, *Member, IEEE*, Zihuai Lin, *Senior Member, IEEE*, Guoqiang Mao, *Senior Member, IEEE*, and Branka Vucetic, *Fellow, IEEE*

Abstract-In heterogeneous networks (HetNet), the load between macro-cell base stations (MBS) and small-cell base stations (SBS) is imbalanced due to their different transmission powers and locations. This load imbalance significantly impacts the system performance and affects the experience of mobile users (MU) with different priorities. In this paper, we aim to distributively optimize the user association in HetNets with various user priorities to solve the load balancing problem. Since the user association is a binary matching problem, which is NP-hard, we propose a distributed belief propagation (BP) algorithm to approach the optimal solution. We first develop a factor graph model using the network topology to represent this user association problem. With this factor graph, we propose a novel distributed BP algorithm by adopting the proportional fairness as the objective. Next, we theoretically prove the existence of the fixed point in our BP algorithm. To be more practical, we develop an approximation method to significantly reduce the computational and communication complexity of the BP algorithm. Furthermore, we analyze some properties of the factor graph relevant to the performance of the BP algorithm using the stochastic geometry. Simulation results show that 1) the proposed **BP** algorithm well approaches the optimal system performance and achieves a much better performance compared with other association schemes, and 2) the analytical results on the average degree distribution and sparsity of the factor graph match with those obtained from the Monte-Carlo simulations.

Index Terms—Heterogeneous networks, user association, user priority, belief propagation, stochastic geometry

I. INTRODUCTION

Wireless data traffic is expected to increase by a factor of 40 over the next five years, from current 93 to 3600 Petabytes per month, driven by vast demands from bandwidthhungry mobile applications. To cope with the data avalanche, an efficient solution is to enhance the network capacity by embedding small cells with low-power base stations (BS) into existing macro-cell based networks so obtain the so-called heterogeneous networks (HetNet) to significantly boost the area spectrum efficiency [1, 2].

Jun Li is with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Jiangsu, CHINA. Email: jun.li@njust.edu.cn. His research is partially supported by of "Specially Appointed Professor Program" in Jiangsu Province, 2015.

Youjia Chen, Zihuai Lin, and Branka Vucetic are with School of Electrical and Information Engineering, The University of Sydney, NSW, AUSTRALIA. E-mail:{youjia.chen,zihuai.lin,branka.vucetic}@sydney.edu.au.

Guoqiang Mao is with the School of Computing and Communications, The University of Technology Sydney, NSW, AUSTRALIA. He is also with Beijing University of Posts and Telecommunications and Huazhong University of Science and Technology, CHINA. E-mail: g.mao@ieee.org.

This work is supported by Chinese NSF Projects 61501238 and 61428102, by Jiangsu Provincial Science Foundation Project *BK*20150786, and by Australian Research Council Programs DP120100405 and LP110100110.

In LTE-Advanced, HetNets contain regularly deployed macro-cell BSs (MBS) and overlapping small-cell BSs (SBS), e.g., Pico-cells and Femto-cells [3, 4]. The aim of these low-power and flexibly deployed SBSs is to eliminate the coverage holes and increase the capacity in hot-spots. Usually, the locations of MBSs are carefully chosen, and properly configured to minimize the interference among them, while the SBSs are deployed in a relatively unplanned manner.

1

Due to imbalanced power and random locations among various BSs in HetNets, a major issue is how to associate each MU with a proper BS, namely, user association, to achieve the optimal trade-off between load balancing and network throughput. In conventional homogeneous networks, user association is typically based on the maximum signalto-interference-plus-noise (SINR) received at MUs. However, applying this method to HetNets will lead to a severely imbalanced load among the BSs because of the disparity in their transmission power. Many MUs tend to connect to MBSs, though they are located closer to SBSs. In this case, MBSs may have difficulty to support too many MUs, while SBSs only serve a small portion of MUs and become under-utilized. This imbalanced load will lead to a performance loss and uneven user experience. Therefore, many efforts have been made on better user association schemes.

From industrial perspective, an effective association scheme is to adapt the coverage of small cells to control the number of MUs connecting to them. For instance, in wireless local area networks, cell breathing technique is proposed to balance the load of access points (AP) by tuning transmission power [5, 6]. However, this technique is not suitable for HetNets, since transmission power is quite different between MBSs and SBSs. Alternatively, cell range expansion (CRE) is proposed in LTE-Advanced to enable the offloading of MUs from MBSs to SBSs by setting a bias value [7, 8]. That is, a positive bias is added to the received power from SBSs before each MU associates with a BS, which is equivalent to expanding the coverage of SBSs [9, 10].

From academic perspective, the user association problem has typically been formulated as an optimization problem. Different kinds of objective functions are adopted for optimizations. For instance, the max-min fairness is adopted in [11], which distributes resources among all users as equally as possible, the network throughput is considered in [12], the topological potential is used in [13], and several different utility functions are encompassed in [14]. In [15–17], the proportional fairness is chosen as the objective function for optimization, which achieves a good trade-off between the load

balancing, user fairness, and system throughput.

Generally, user association is a binary matching problem, which is proved to be NP-hard [18]. Since in real systems, it is very difficult to implement the case that an MU associates with multiple BSs, most of the works have the constraint that an MU only associates with one BS. Aiming to solve this problem with polynomial complexity, many algorithms relax the binary constraint by assuming that each MU can associate with multiple BSs simultaneously. This converts the binary matching problem to a continuous optimization problem. Then a rounding method is adopted to obtain the final solution [15, 16]. These algorithms will lead to a performance loss due to the relaxation of the binary constraint. In addition, some works on this problem are based on centralized algorithms, which require information of all the BSs and MUs [16, 17]. Authors in [19] perform localized and distributed optimizations, but central tuning is still required. Furthermore, the above papers have not taken into account the user priority when performing the optimization, while in practice, it is common that MUs have different priorities. The unequal user priority (weight) reflects different characteristics among the users. For instance, in practical cellular network management, users with high QoS requirements or users who experience a low long-term transmission rate, are given higher priorities. Therefore, unequal user priority has been widely considered in many scenarios of wireless communications [20, 21].

In this paper, we aim to distributively optimize the user association in HetNets where MUs have different priorities. By utilizing the weighted proportional fairness as the objective, we propose a distributed belief propagation (BP) algorithm to solve the optimization problem without relaxing the binary constraint in order to approach the optimal performance. The BP algorithm is widely used in artificial intelligence, signal processing, and digital communications, for instance, the decoding of LDPC codes, and the interference coordination in HetNets. To our best knowledge, our work is the first application of the BP algorithm in the user association problem. Generally, it can compute various marginal functions (distribution) derived from a global function (distribution), after decomposing the complicated global function of many variables into the product of multiple local functions, each of which only depends on a subset of the variables [22].

We first develop a factor graph model to represent this user association problem. Then with this factor graph, a distributed BP algorithm is proposed to solve the formulated problem by iterative message passing between the MUs and BSs. We theoretically prove that the fixed point exists and show that convergence can be achieved in our BP algorithm. To be more practical, we also develop an approximated method to dramatically reduce the complexity of the BP algorithm. Furthermore, since the factor graph is closely related to the performance of the BP algorithm, we analyze the average degree distribution and sparsity of the factor graph based on the stochastic geometry theory.

Our simulation results show that 1) the proposed BP algorithm can almost achieve the optimal solution via exhaustive search and that the approximate BP algorithm approaches the optimal solution with a very small gap, 2) our BP algorithm outperforms the existing schemes, such as maximum-SINR and CRE, 3) analytical results on the average degree distribution and sparsity of the factor graph match well with the Monte-Carlo simulations.

The rest of this paper is organized as follows. Section II describes the system model and formulates the user association problem. Section III presents the BP algorithm. The approximate method is given in Section IV. The degree distribution and sparsity of the factor graph are analyzed in Section V. Section VI presents the simulations and Section VII draws the conclusions.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. System Model

We focus on the downlink of a HetNet, which consists of multiple MBSs and SBSs with overlapping coverage. Furthermore, we assume that 1) the HetNet is saturated, where the BSs transmit all the time, 2) each MU can only associate with one of the BSs, and 3) channels between MUs and BSs are considered to be static during the optimization process of the association.

Let \mathcal{M} denote the set of MUs and \mathcal{B} denote the set of BSs in the HetNet. Also, we define two sets $\mathcal{I} \triangleq \{1, \dots, |\mathcal{M}|\}$ and $\mathcal{J} \triangleq \{1, \dots, |\mathcal{B}|\}$, where $|\cdot|$ is the cardinality of a set. We assume that each BS allocates its transmission power equally in its entire bandwidth. Then the received SINR of the *i*-th MU $\mathcal{M}_i \in \mathcal{M}$ from the *j*-th BS $\mathcal{B}_j \in \mathcal{B}$ can be written as

$$\rho_{ij} = \frac{P_j d_{ij}^{-\alpha} h_{ij}}{\sum\limits_{q \in \mathcal{J} \setminus \{j\}} P_q d_{iq}^{-\alpha} h_{iq} + \sigma^2}, \ \forall i \in \mathcal{I}, j \in \mathcal{J},$$
(1)

where P_j is the transmission power of \mathcal{B}_j , σ^2 is the power of additive white Gaussian noise. The path loss from \mathcal{B}_j to \mathcal{M}_i is formulated as $d_{ij}^{-\alpha}$, where d_{ij} and α represent their distance and the path loss exponent, respectively. In addition, h_{ij} denotes the fading power, where the random channel is modeled as Rayleigh fading. Thus, the spectral efficiency between \mathcal{M}_i and \mathcal{B}_j can be obtained according to the following formula $\gamma_{ij} = \log_2 (1 + \rho_{ij})$.

We denote by $x_{ij} \in \{0, 1\}$ the user association indicator, where $x_{ij} = 1$ if \mathcal{M}_i is associated with \mathcal{B}_j , and $x_{ij} = 0$ otherwise. Assuming that \mathcal{M}_i is associated with \mathcal{B}_j , we denote by β_{ij} the fraction of the resource allocated to \mathcal{M}_i by \mathcal{B}_j , and by R_{ij} the rate that \mathcal{M}_i obtains from \mathcal{B}_j . We have $R_{ij} = W\gamma_{ij}\beta_{ij}$, where W is the bandwidth of \mathcal{B}_j . Since \mathcal{M}_i is possible to associate with any BS, the effective transmission rate of \mathcal{M}_i can be written as

$$R_i = \sum_{j \in \mathcal{J}} x_{ij} R_{ij} = \sum_{j \in \mathcal{J}} x_{ij} W \gamma_{ij} \beta_{ij}.$$
 (2)

B. Problem Formulation

We aim at the network-wide optimization of the user associations x_{ij} , $\forall i, j$, with different user priorities, where the priority of \mathcal{M}_i is denoted by ω_i , namely, priority-based user association (PUA) optimization. Here, ω_i is a positive constant for \mathcal{M}_i , reflecting its physical feature. We choose the

3

objective function based on the proportional fairness, which has been proved to be associated with the logarithmic utility function [23]. The utility of M_i can be given as

$$U_i(R_i) = \log R_i = \log \left(\sum_{j \in \mathcal{J}} x_{ij} R_{ij}\right).$$
(3)

Noting that ω_i denotes the priority of \mathcal{M}_i and U_i denotes the utility achieved by \mathcal{M}_i , it follows that $\omega_i U_i$ can be viewed as the weighted utility. Therefore, the overall system utility is formulated as the sum of each user's weighted utility [16, 20, 24], i.e., $\sum_i \omega_i U_i$, which is used as our objective function. Hence, our optimization problem can be formulated as maximizing the objective function, i.e.,

$$\max_{\{x_{ij},\beta_{ij}\}} \sum_{i \in \mathcal{I}} \omega_i \log R_i = \max_{\{x_{ij},\beta_{ij}\}} \sum_{i \in \mathcal{I}} \omega_i \log \left(\sum_{j \in \mathcal{J}} x_{ij} W \gamma_{ij} \beta_{ij} \right)$$
(4)

s.t.
$$\sum_{j \in \mathcal{J}} x_{ij} = 1, \forall i \in \mathcal{I},$$
 (5a)

$$\sum_{i\in\mathcal{I}} x_{ij}\beta_{ij} = 1, \ \forall j\in\mathcal{J},$$
(5b)

$$x_{ij} \in \{0, 1\}, \forall i \in \mathcal{I}, j \in \mathcal{J},$$
 (5c)

$$\beta_{ij} \in (0,1], \ \forall i \in \mathcal{I}, j \in \mathcal{J}.$$
 (5d)

The constraint (5a) means that each MU can only associate with one BS, (5b) indicates that the MUs associated with the same BS share the resource of the BS, (5c) makes sure that the association indicator must be binary, and (5d) specifies the range of the resource fraction allocated to each MU.

Given x_{ij} , $\forall i \in \mathcal{I}$, $\forall j \in \mathcal{J}$, the optimization problem (4) is equivalent to maximizing $\prod_{i \in \mathcal{I}} (x_{ij}W\gamma_{ij}\beta_{ij})^{\omega_i}$, $\forall j$. Since $\sum_{i \in \mathcal{I}} x_{ij}\beta_{ij} = 1$, the optimal fraction of the resource assigned to \mathcal{M}_i by \mathcal{B}_j can be obtained as $\beta_{ij} = \frac{\omega_i x_{ij}}{\sum_{k \in \mathcal{I}} \omega_k x_{kj}}$. For further details, please refer to Theorem 3 in [16]. Thus, the formulation of our PUA optimization can be rewritten as

$$\max_{\{x_{ij}\}} \sum_{i \in \mathcal{I}} \omega_i \log \left(\sum_{j \in \mathcal{J}} x_{ij} W \gamma_{ij} \frac{\omega_i}{\sum_{k \in \mathcal{I}} x_{kj} \omega_k} \right)$$
(6)

s.t.
$$\sum_{j \in \mathcal{J}} x_{ij} = 1, \ \forall i \in \mathcal{I},$$
 (7a)

$$x_{ij} \in \{0,1\}, \ \forall i \in \mathcal{I}, j \in \mathcal{J}.$$
 (7b)

This problem is proved to be a NP-hard problem [16], due to the nonlinear utility function and the binary association indicator. Note that if each MU has equal priority, i.e., $\omega_i = \omega_k, \forall i, k \in \mathcal{I}, i \neq k$, we can obtain that $\beta_{ij} = x_{ij} / \sum_{k \in \mathcal{I}} x_{kj}$. Then the PUA optimization is reduced to the basic user association problem in [15].

III. FACTOR GRAPH MODEL AND DISTRIBUTED BP Algorithm

In this section, we propose a BP-based distributed algorithm to solve the formulated PUA optimization problem in Eq. (6). To proceed, we first develop a factor graph model according to



Fig. 1. Factor graph model for user association based on the HetNet.

the topology of the HetNet. Then we propose a distributed BP algorithm to efficiently solve the PUA optimization problem with a near-optimal performance. At the end of this section, we prove the existence of the fixed point in our BP algorithm.

A. Factor Graph Model

Based on the topology of the HetNet, we develop a factor graph model $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, as shown in Fig. 1. The vertex set \mathcal{V} consists of factor nodes and variable nodes, where each factor node is related to a BS, and each variable node is related to an MU. To simplify the notations, we use $j \in \mathcal{J}$ to denote the *j*-th factor node and use $i \in \mathcal{I}$ to denote the *i*-th variable node. Hence, the vertex set \mathcal{V} is composed of \mathcal{I} and \mathcal{J} , i.e., $\mathcal{V} = {\mathcal{I}, \mathcal{J}}.$

An edge in the set \mathcal{E} connecting \mathcal{M}_i and \mathcal{B}_j , denoted by (i, j), exists if and only if the received SINR at \mathcal{M}_i from \mathcal{B}_j is no less than a predetermined threshold δ . The node j is called a neighboring node of i if there is an edge (i, j). We use $\mathcal{H}(v)$ to denote the set of neighboring nodes of a node $v, v \in \mathcal{V}$. Thus, $\mathcal{H}(i)$ is the set of neighboring nodes of \mathcal{M}_i , and $\mathcal{H}(j)$ is the set of neighboring nodes of \mathcal{B}_j .

Given a factor graph, our BP algorithm intends to find out the optimal BS for \mathcal{M}_i to associate with from the set $\mathcal{H}(i)$. The messages passing in the BP only occurs between a node and its neighbors, i.e., \mathcal{M}_i forward messages to its neighbors $\mathcal{B}_{\hbar}, \forall \hbar \in \mathcal{H}(i)$, and \mathcal{B}_j forward messages to its neighbors $\mathcal{M}_h, \forall h \in \mathcal{H}(j)$. Now we introduce the details of factor graph and its relationship with the PUA optimization problem.

1) Factor Nodes: We rewrite the optimization problem in Eq. (6) as

$$\max_{\{x_{ij}\}} \sum_{i \in \mathcal{I}} \omega_i \log R_i \stackrel{(a)}{=} \max_{\{x_{ij}\}} \sum_{i \in \mathcal{I}} \omega_i \log \left(\sum_{j \in \mathcal{J}} x_{ij} R_{ij} \right)$$
$$\stackrel{(b)}{=} \max_{\{x_{ij}\}} \sum_{j \in \mathcal{J}} \sum_{i \in \mathcal{I}} x_{ij} \omega_i \log R_{ij}$$
$$\stackrel{(c)}{=} \max_{\{x_{ij}\}} \sum_{j \in \mathcal{J}} \sum_{i \in \mathcal{H}(j)} x_{ij} \omega_i \log R_{ij}$$
$$= \max_{\{x_{ij}\}} \sum_{j \in \mathcal{J}} f_j,$$
(8)

^{0018-9545 (}c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

where the function

$$f_j \triangleq \sum_{i \in \mathcal{H}(j)} x_{ij} \omega_i \log \left(W \gamma_{ij} \frac{\omega_i}{\sum\limits_{k \in \mathcal{H}(j)} \omega_k x_{kj}} \right)$$
(9)

is defined as the local utility function at \mathcal{B}_j . In Eq. (8), step (a) uses the formulation of R_i in Eq. (2). In step (b), we use the constraints $x_{ij} \in \{0,1\}$ and $\sum_j x_{ij} = 1$. For step (c), since only the users in the set of neighboring nodes of \mathcal{B}_j are possible to associate with it, $i \in \mathcal{I}$ can be narrowed to $i \in \mathcal{H}(j)$. From (8), the network-wide system utility in (6) can be decomposed into multiple individual local utility functions f_j at \mathcal{B}_j , $\forall j$. Thus, in the factor graph, we use the factor node j to represent the utility f_j .

2) Variable Nodes: We define a vector \mathbf{x}_i for \mathcal{M}_i consisting of the user association indicators $x_{i\hbar}$, $\forall \hbar \in \mathcal{H}(i)$. The length of \mathbf{x}_i is $|\mathcal{H}(i)|$. Due to the association constraint in (5a), i.e., \mathcal{M}_i only associates with one BS, there is only one element in \mathbf{x}_i equal to 1, and all the other elements are 0. Thus, \mathbf{x}_i has $|\mathcal{H}(i)|$ possible values according to $|\mathcal{H}(i)|$ different locations of 1. In the factor graph, we use the variable node *i* to represent \mathbf{x}_i , namely, user association variable.

To illustrate \mathbf{x}_i , we take \mathcal{M}_3 in Fig. 1 as an example. In Fig. 1, \mathcal{M}_3 can receive signals from \mathcal{B}_2 and \mathcal{B}_3 . Due to the constraints (5a) and (5c), the association variable \mathbf{x}_3 for \mathcal{M}_3 has two possible values, i.e., 1) $\mathbf{x}_3 = [1,0]$, that is, \mathcal{M}_3 associates with \mathcal{B}_2 and does not associate with \mathcal{B}_3 ; 2) $\mathbf{x}_3 = [0,1]$, that is, \mathcal{M}_3 associates with \mathcal{B}_3 .

We define the set of the variable nodes in the factor graph as $\mathcal{X} \triangleq \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{|\mathcal{M}|}\}$. Based on the above analysis of the factor graph model, the network-wide optimization of the system utility in (8) can be rewritten as

$$\max_{\boldsymbol{\mathcal{X}}} F(\boldsymbol{\mathcal{X}}), \ F(\boldsymbol{\mathcal{X}}) \triangleq \sum_{j \in \boldsymbol{\mathcal{J}}} f_j(\boldsymbol{\mathcal{X}}_{\boldsymbol{\mathcal{H}}(j)}),$$
(10)

where $\mathcal{X}_{\mathcal{H}(j)} = {\mathbf{x}_h, \forall h \in \mathcal{H}(j)}$ represents the set of the user association variables corresponding to *j*'s neighbors. In the following, we will discuss the maximization of $F(\mathcal{X})$ via distributed BP algorithm.

B. Transformation of Utility Optimization

Generally, BP algorithm converts the optimization problem to a marginal distribution estimation problem [25]. In our optimization problem, we define a probability mass function (PMF) based on the global utility $F(\mathcal{X})$, i.e., $p(\mathcal{X}) \triangleq \frac{1}{Z} \exp(\mu F(\mathcal{X}))$, where μ is a positive number, and Z is a normalization constant. According to [25], the large deviation theory shows that when $\mu \to \infty$, $p(\mathcal{X})$ concentrates around the maxima of $F(\mathcal{X})$, i.e., $\lim_{\mu\to\infty} \mathbb{E}(\mathcal{X}) = \arg \max_{\mathcal{X}} F(\mathcal{X})$, where $\mathbb{E}(\cdot)$ denotes the expectation. From this equation, once we obtain $\mathbb{E}(\mathcal{X})$, we will have a good estimation for the maximization of $F(\mathcal{X})$.

The calculation of $\mathbb{E}(\mathcal{X})$ can be decomposed into the calculation of the expectation of each element in \mathcal{X} , i.e., $\mathbb{E}(\mathbf{x}_i), \forall \mathbf{x}_i \in \mathcal{X}$. Therefore, the optimization problem is transformed into estimating the marginal PMF of each variable node \mathbf{x}_i , i.e., $p(\mathbf{x}_i)$, which can be solved by BP algorithm.

C. Iterative Message Passing

The PMF $p(\mathbf{x}_i)$ is the message updated and exchanged between BSs and MUs. In each iteration, \mathcal{M}_i updates $p(\mathbf{x}_i)$ and forwards it to \mathcal{B}_{\hbar} , $\forall \hbar \in \mathcal{H}(i)$ while \mathcal{B}_j updates $p(\mathbf{x}_h)$ and forwards it to \mathcal{M}_h , $\forall h \in \mathcal{H}(j)$. There are $|\mathcal{H}(i)|$ probabilities in $p(\mathbf{x}_i)$ corresponding to $|\mathcal{H}(i)|$ values of \mathbf{x}_i , representing the probabilities that \mathcal{M}_i associates with its $|\mathcal{H}(i)|$ neighbor BSs. Since \mathcal{B}_{\hbar} only considers whether \mathcal{M}_i is associated with it or not, the message passing along the edge (i, \hbar) only carries the probability that \mathcal{M}_i associates with \mathcal{B}_{\hbar} , i.e., $\Pr(x_{i\hbar} = 1)$, which is one of the $|\mathcal{H}(i)|$ probabilities in $p(\mathbf{x}_i)$.

4

In the following, to simplify the notations, we assume the edge (i, j) exists. We index the iteration by t, and let $m_{i \to j}^{(t)}(x_{ij})$ and $m_{j \to i}^{(t)}(x_{ij})$ denote the messages from \mathcal{M}_i to \mathcal{B}_j and from \mathcal{B}_j to \mathcal{M}_i in the t-th iteration, respectively. The steps for the distributed BP are given as follows.

1) Initialization: The MU \mathcal{M}_i measures the SINR from all the BSs and finds out possible serving BSs \mathcal{B}_{\hbar} , $\forall \hbar \in \mathcal{H}(i)$, according to the SINR threshold and access policy. Instantaneous channel state information (CSI) $g_{i\hbar}$ is estimated at \mathcal{M}_i , which is then fed back to all the \mathcal{B}_{\hbar} . At \mathcal{B}_{\hbar} , the CSI $g_{i\hbar}$ is needed for the update of the messages in each iteration.

2) Message from MU to BS: In the iteration t, \mathcal{M}_i sends its possible serving \mathcal{B}_i the probability of choosing it, i.e.,

$$m_{i \to j}^{(t)} \left(x_{ij} = 1 \right) = \varphi \left(\mathbf{x}_i \right) \prod_{k \in \mathcal{H}(i) \setminus \{j\}} m_{k \to i}^{(t-1)} \left(x_{ik} = 0 \right).$$
(11)

From (11), we can see that the probability of $x_{ij} = 1$ is based on the probabilities of $x_{ik} = 0$, $\forall k \in \mathcal{H}(i) \setminus \{j\}$. This calculation comes from the constraint that \mathcal{M}_i can only associate with one of the BSs in $\mathcal{H}(i)$. Also in (11), $\varphi(\mathbf{x}_i)$ is the normalization function which ensures $\sum_{h \in \mathcal{H}(i)} m_{i \to h}^{(t)} (x_{ih} = 1) = 1$. Obviously, $m_{i \to j}^{(t)} (x_{ij} = 0)$ is not needed to be transmitted by \mathcal{M}_i since we have $m_{i \to j}^{(t)} (x_{ij} = 0) = 1 - m_{i \to j}^{(t)} (x_{ij} = 1)$.

Note that before iterations begin, MUs does not have any *a* priori information about associations. Thus, initial messages are set uniformly. For example, if the \mathcal{M}_i has 3 possible serving BSs, i.e., $|\mathcal{H}(i)| = 3$, the initial messages can be set as $m_{i\to\hbar}^{(1)}(x_{i\hbar} = 1) = 1/3$, $\forall \hbar \in \mathcal{H}(i)$.

3) Message from BS to MU: Now, we consider the message passing from \mathcal{B}_j to \mathcal{M}_i . Note that f_j is only related to x_{hj} in \mathbf{x}_h . We define a vector $\mathbf{x}_{\mathcal{H}(j)}$ as consisting of x_{hj} , $\forall h \in \mathcal{H}(j)$, i.e., $\mathbf{x}_{\mathcal{H}(j)} = [x_{hj}, \forall h \in \mathcal{H}(j)]$. We have

$$m_{j \to i}^{(t)}(x_{ij}) = \sum_{\mathbf{x}_{\mathcal{H}(j) \setminus \{x_{ij}\}}} \left(\exp\left(\mu f_j\left(\mathbf{x}_{\mathcal{H}(j)}\right)\right) \prod_{k \in \mathcal{H}(j) \setminus \{i\}} m_{k \to j}^{(t)}\left(x_{kj}\right) \right)$$
$$= \sum_{\mathbf{x}_{\mathcal{H}(j) \setminus \{x_{ij}\}}} \left(\exp\left(\sum_{k \in \mathcal{H}(j) \setminus \{i\}} \mu x_{kj} \omega_k \log\left(R_{kj}\right) + \mu x_{ij} \omega_i \log\left(R_{ij}\right)\right) \prod_k m_{k \to j}^{(t)}\left(x_{kj}\right) \right),$$
(12)

where $\sum_{\mathbf{x}_{\mathcal{H}(j)} \setminus \{x_{ij}\}}$ represents the summation over all possible values of $\mathbf{x}_{\mathcal{H}(j)}$ given x_{ij} .

5

Based on the two values of x_{ij} , \mathcal{B}_j calculates $m_{j \to i}^{(t)}(x_{ij} = 0)$ and $m_{j \to i}^{(t)}(x_{ij} = 1)$, and then forwards the two values to \mathcal{M}_j . Specifically, we have

$$m_{j \to i}^{(t)} (x_{ij} = 0) = \sum_{\mathbf{x}_{\mathcal{H}(j) \setminus \{x_{ij}\}}} \left(\prod_{k \in \mathcal{H}(j) \setminus \{i\}} \left(\frac{W \gamma_{kj} \omega_k}{\sum\limits_{q \in \mathcal{H}(j) \setminus \{i\}} \omega_q x_{qj}} \right)^{\mu x_{kj} \omega_k} m_{k \to j}^{(t)} (x_{kj}) \right),$$
(13)

$$m_{j \to i}^{(t)} (x_{ij} = 1) = \sum_{\mathbf{x}_{\mathcal{H}(j) \setminus \{x_{ij}\}}} \left(\prod_{k \in \mathcal{H}(j) \setminus \{i\}} \left(\frac{W \gamma_{kj} \omega_k}{\sum_{q \in \mathcal{H}(j) \setminus \{i\}} \omega_q x_{qj} + \omega_i} \right)^{\mu x_{kj} \omega_k} \left(\frac{W \gamma_{ij} \omega_i}{\sum_{q} \omega_q x_{qj} + \omega_i} \right)^{\mu \omega_i} m_{k \to j}^{(t)} (x_{kj}) \right). \quad (14)$$

4) Final Decision: We assume there are totally T iterations in our BP algorithm. After T iterations, the probability that the M_i associates with \mathcal{B}_i can be calculated as

$$\Pr(x_{ij} = 1) = \varphi(\mathbf{x}_i) m_{j \to i}^{(T)} (x_{ij} = 1)$$
$$\prod_{k \in \mathcal{H}(i) \setminus \{j\}} m_{j \to i}^{(T)} (x_{ik} = 0). \quad (15)$$

Based on (15), an decision can be made, i.e., \mathcal{M}_i associates with the BS $\mathcal{B}_{\hat{i}}$ such that

$$\hat{j} = \underset{j}{\operatorname{arg\,max}} \operatorname{Pr}(x_{ij} = 1).$$
(16)

D. Fixed Point

The existence of the fixed point is a necessary condition that our distributed BP algorithm can converge. Based on the messages in Eq. (11) and (12), the message for each variable node \mathbf{x}_i in the *t*-th iteration can be obtained from the messages in the (t-1)-th iteration. That is,

$$m_{i \to j}^{(t)}(x_{ij}) = \varphi\left(\mathbf{x}_{i}\right) \prod_{k \in \mathcal{H}(i) \setminus \{j\}} \sum_{\mathbf{x}_{\mathcal{H}(k)} \setminus \{x_{ik}\}} \left(\left(\exp\left(\mu f_{k}\left(\mathbf{x}_{\mathcal{H}(k)}\right)\right) \right) \prod_{q \in \mathcal{H}(k) \setminus \{i\}} m_{q \to k}^{(t-1)}(x_{qk}) \right). \quad (17)$$

Since the message $m_{i \to j}^{(t)}(x_{ij})$ is a probability, we define the probability set $\boldsymbol{\mathcal{S}}^{(t)} \triangleq \left\{ m_{i \to \hbar}^{(t)}(x_{i\hbar}) \right\}$, $\forall i \in \boldsymbol{\mathcal{I}}, \ \hbar \in \boldsymbol{\mathcal{H}}(i)$, and thus $|\boldsymbol{\mathcal{S}}| = \sum_{i \in \boldsymbol{\mathcal{I}}} |\boldsymbol{\mathcal{H}}(i)|$. Define the message mapping function $\boldsymbol{\Gamma} : \mathbb{R}^{|\boldsymbol{\mathcal{S}}|} \to \mathbb{R}^{|\boldsymbol{\mathcal{S}}|}$ based on (17). Then we have $\boldsymbol{\mathcal{S}}^{(t)} = \boldsymbol{\Gamma}(\boldsymbol{\mathcal{S}}^{(t-1)})$.

Lemma 1. The message mapping function Γ is continuous. Proof: Please refer to Appendix A.

With Lemma 1, we have the following theorem.

Theorem 1. A fixed point exists for the proposed distributed BP algorithm.

Proof: Please refer to Appendix B.

Whether the BP algorithm can converge to a fixed point, unfortunately, is not well understood yet [26]. Generally speaking, if the factor graph is sparse and contains no cycles, the BP algorithm can converge to a fixed point exactly and efficiently [27]. In Section V, we will analyze the average sparsity of our factor graph based on the stochastic geometry theory. From the analysis, we show that the sparsity and the loops in the factor graph can be controlled by adapting the SINR threshold δ .

IV. COMPLEXITY REDUCTION VIA APPROXIMATIONS

From the BP algorithm in the previous section, we can see that its computational and communication complexities are relatively high. From communication complexity point of view, the BS \mathcal{B}_j needs to send two messages, i.e., Eq. (13) and (14), to each of its potential MUs in a point-to-point manner, which leads to a heavy communication complexity, especially when the number of the MUs is large.

From computational complexity point of view, the calculation on each of the two messages at \mathcal{B}_j needs to 1) obtain the CSI about the channel from \mathcal{B}_j to \mathcal{M}_h in order to calculate $\gamma_{hj}, \forall h \in \mathcal{H}(j)$, and 2) consider $2^{|\mathcal{H}(j)|-1}$ combination cases to obtain the expectation, which causes a large number of computations at \mathcal{B}_j .

Due to the high complexity of the BP discussed above (referred to as the exact BP), in this section, we will propose an approximate BP to significantly reduce the computational complexity and enable message transmission in a broadcast manner. At the end of this section, we compare the complexity between the exact BP and approximate BP algorithms.

A. Approximations

First, we can rewrite the message from \mathcal{M}_i to \mathcal{B}_j in Eq. (11) as

$$m_{i \to j}^{(t)}(x_{ij} = 1) = \frac{1}{1 + \sum_{k \in \mathcal{H}(i) \setminus \{j\}} \frac{m_{k \to i}^{(t-1)}(x_{ik} = 1)}{m_{k \to i}^{(t-1)}(x_{ik} = 0)}}.$$
 (18)

We can see from (18) that the message $m_{i \to j}^{(t)}(x_{ij} = 1)$ only depends on the likelihood ratios $\frac{m_{k \to i}^{(t-1)}(x_{ik}=1)}{m_{k \to i}^{(t-1)}(x_{ik}=0)}, \forall k \in \mathcal{H}(i) \setminus \{j\}$. Hence, the two messages sent from \mathcal{B}_j to \mathcal{M}_i in Eq. (13) and (14) can be replaced by their likelihood ratio.

Second, after several approximation steps (please refer to Appendix C for details), the likelihood ratio of the two messages from \mathcal{B}_j to \mathcal{M}_i can be approximated as

$$\frac{m_{j \to i}^{(t)} \left(x_{ij} = 1\right)}{m_{j \to i}^{(t)} \left(x_{ij} = 0\right)} \approx \left(\frac{W\gamma_{ij}\omega_i}{\sum\limits_{k \in \mathcal{H}(j) \setminus \{i\}} \omega_k m_{k \to j}^{(t)} \left(x_{kj} = 1\right) + \omega_i}\right)^{\mu\omega_i}$$
(19)

where the denominator of the item in the right hand side can be rewritten as

TABLE I Complexity Comparisons for the Exact BP and approximate BP in Each Iteration

$\sum \omega_k m_{k \to j}^{(t)} \left(x_{kj} = 1 \right) + \omega_i =$	
$k{\in}{oldsymbol{\mathcal{H}}}(j)ackslash\{i\}$	
$\sum \omega_h m_{h \to j}^{(t)} (x_{hj} = 1) - \omega_i m_{i \to j}^{(t)} (x_{ij} = 1) + \omega_i.$	(20)
$h \in \mathcal{H}(j)$	

From Eq. (19), we can see that the likelihood ratio forwarded from \mathcal{B}_j to \mathcal{M}_i can be replaced by $\sum_{h\in\mathcal{H}(j)} \omega_h m_{h\to j}^{(t)} (x_{hj} = 1)$, since the parameters W, γ_{ij} , ω_i , $n\in\mathcal{H}(j)$ and $m_{i\to j}^{(t)} (x_{ij} = 1)$ are all known to \mathcal{M}_i . Then \mathcal{M}_i can obtain the likelihood ratio by calculating Eq. (19) with $\sum_{h\in\mathcal{H}(j)} \omega_h m_{h\to j}^{(t)} (x_{hj} = 1)$. In the approximate BP, we use this $h\in\mathcal{H}(j)$

item as the belief message transmitted by \mathcal{B}_j , denoted by

$$m_{j}^{(t)} = \sum_{h \in \mathcal{H}(j)} \omega_{h} m_{h \to j}^{(t)} \left(x_{hj} = 1 \right).$$
(21)

Intuitively, $m_j^{(t)}$ can be seen as the expectation of the priority of the MUs which associate with \mathcal{B}_j . In the case when all MUs have the same priority configured as 1, $m_j^{(t)}$ is the expected load, i.e., the average number of the associated MUs with \mathcal{B}_j .

It is clear that the message $m_j^{(t)}$ is common to \mathcal{M}_h , $\forall h \in \mathcal{H}(j)$. With this property, \mathcal{B}_j can forward $m_j^{(t)}$ in a broadcast manner to all of its neighboring MUs, rather than in the point-to-point manner to each individual \mathcal{M}_h . Also note that the calculation of $m_j^{(t)}$ at \mathcal{B}_j does not need the channel state information from \mathcal{B}_j to \mathcal{M}_i . As such, the message transmission complexity can be dramatically reduced. Furthermore, the computational complexity at \mathcal{B}_j is reduced significantly: Only $|\mathcal{H}(j)|$ multiplications and $|\mathcal{H}(j)| - 1$ additions are needed on this item in each iteration.

B. Approximate BP Algorithm

Based on the previous approximations, we propose an approximate BP algorithm as follows.

1) Initialization: The initialization of the approximate BP is similar to that in the original BP in Section III except that the instantaneous downlink CSI is not required at the BSs.

2) Message from MU to BS: In the t-th iteration, the MU \mathcal{M}_i calculates the message for \mathcal{B}_j , i.e., $m_{i \to j}^{(t)}(x_{ij} = 1)$, based on Eq. (18), in which, the likelihood ratio $\frac{m_{k \to i}^{(t-1)}(x_{ik}=1)}{m_{k \to i}^{(t-1)}(x_{ik}=0)}$, $\forall k \in \mathcal{H}(i) \setminus \{j\}$, is calculated from (19) based on the message broadcasted by \mathcal{B}_k , i.e., $m_k^{(t-1)}$.

3) Message from BS to MU: The BS \mathcal{B}_j calculates the message $m_j^{(t)}$ based on Eq. (21), and then broadcasts it to $\mathcal{M}_h, \forall h \in \mathcal{H}(j)$.

4) *Final Decision:* The final decision is made similar to that in the exact BP.

C. Complexity Comparisons

1) Communication Complexity: At the MU side, in the exact BP, \mathcal{M}_i needs to forward the instantaneous CSI to \mathcal{B}_{\hbar} , $\forall \hbar \in \mathcal{H}(i)$ before iterations begin, while in the approximate

Number of Distinct Messages		
	Exact BP	Approximate BP
MU \mathcal{M}_i	$ \mathcal{H}(i) $	$ \mathcal{H}(i) $
BS \mathcal{B}_j	$ \mathcal{H}(j) $	1
Instantaneous CSI at BSs	Needed	Not Needed
Computational Complexity		
	Exact BP	Approximate BP
MU \mathcal{M}_i	$O\left(\mathcal{H}(i) \right)$	$O\left(\mathcal{H}(i) ight)$
BS \mathcal{B}_j	$O(2^{ \mathcal{H}(j) } \mathcal{H}(j))$	$O(\mathcal{H}(j))$

BP, this is not the case. At the BS side, in the exact BP, \mathcal{B}_j needs to send a distinct message, i.e., $\frac{m_{k \to i}^{(t-1)}(x_{ik}=1)}{m_{k \to i}^{(t-1)}(x_{ik}=0)}$, to each individual \mathcal{M}_h , $\forall h \in \mathcal{H}(j)$. Thus, the number of distinct messages sent from \mathcal{B}_j is $|\mathcal{H}(j)|$ in each iteration. In the approximate BP, one message in (21) is broadcast to all \mathcal{M}_h from \mathcal{B}_j .

2) Computational Complexity: At the MU side, the computational complexity is $O(|\mathcal{H}(i)|)$ in each iteration for both the exact and approximate BP. At the BS side, in the exact BP algorithm, \mathcal{B}_j needs to calculate $2^{|\mathcal{H}(j)|}$ combinations in each iteration and each combination includes $O(|\mathcal{H}(j)|)$ multiplications and additions. Therefore, the overall computational complexity for \mathcal{B}_j is $O(2^{|\mathcal{H}(j)|}|\mathcal{H}(j)|)$. In the approximate BP, the computational complexity at \mathcal{B}_j reduces to $O(|\mathcal{H}(j)|)$ in each iteration, i.e., only $|\mathcal{H}(j)|$ multiplications and $|\mathcal{H}(j)|-1$ additions.

Table I shows the detailed comparisons on the computational and communication complexity between the two BP algorithms in each iteration as discussed above. As shown in Table I, the computational and communication complexities of the BP algorithm critically depend on $|\mathcal{H}(i)|$ and $|\mathcal{H}(j)|$. Therefore, in the next section, we will analyze the average degree distribution of the factor graph, i.e., $\mathbb{E}(|\mathcal{H}(i)|)$ and $\mathbb{E}(|\mathcal{H}(j)|)$, to provide an insight on the performance of the algorithm.

We compare the computational and communication complexities of the approximate BP with the scheme proposed in [15], which also relies on the messages iteratively passing between MUs and BSs. The computational complexities of the scheme in [15] at the MU and BS side are $O(|\mathcal{H}(i)|)$ and $O(|\mathcal{H}(j)|)$, respectively, which equal to that of the approximate BP. The numbers of distinct messages of the scheme in [15] at the MU and BS side are both 1. Therefore, the number of distinct messages at the MU side of this scheme is less than that of approximate BP which is $|\mathcal{H}(i)|$. However, in this scheme, the iteration number needed is much larger than the approximate BP, and an extra centralized algorithm is used to calculate the step size in each iteration, which involves the information in all BSs.

D. BP for Dynamic Scenarios

In this subsection, we extend the approximate BP algorithm to dynamic scenarios, namely, dynamic BP. The wireless network changes due to the time-varying channels, arrivals of new MUs and so on. In these scenarios, new association

0018-9545 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

7

decisions are needed. We denote by 'new MUs' those MUs in the HetNet who need to make or renew their association decisions, and by 'existing MUs' those MUs who keep their existing association decisions. It is practical by making proper association decisions for 'new MUs' without changing the associations of 'existing MUs'.

To achieve this goal, our dynamic BP algorithm is designed to involve only 'new MUs' and their possible serving BSs. That is, the belief messages are only exchanged between the BS \mathcal{B}_j and its neighboring 'new MUs', denoted by $\mathcal{H}_n(j)$.

Since the dynamic BP needs to take into account the association results of the 'existing MUs', we assume that each BS has recorded the load from its associated 'existing MUs'. Here, the load of \mathcal{B}_j is defined as $\Omega_j \triangleq \sum_{i \in \mathcal{A}(j)} \omega_i$, where $\mathcal{A}(j)$ is the set of 'existing MUs' associated with \mathcal{B}_j . In the dynamic BP, the message that \mathcal{B}_j broadcasts is

$$m_j^{(t)} = \Omega_j + \sum_{h \in \mathcal{H}_n(j)} \omega_h m_{h \to j}^{(t)} (x_{hj} = 1).$$
 (22)

The messages from 'new MU' M_i to B_j are calculated in the same way as in the approximate BP algorithm.

V. FACTOR GRAPH ANALYSIS BASED ON STOCHASTIC GEOMETRY

From the analysis in the previous section, the complexity of the proposed BP algorithm depends on the degree of variable nodes and factor nodes, i.e., $|\mathcal{H}(i)|$ and $|\mathcal{H}(j)|$. Aiming to provide a general understanding of the complexity of the proposed BP algorithm, in this section, we analyze the average degree distribution of the factor graph using the stochastic geometry theory, i.e., $\mathbb{E}(|\mathcal{H}(i)|)$ and $\mathbb{E}(|\mathcal{H}(j)|)$. Also, we investigate the sparsity of the factor graph with the aim to provide a good indication on the performance of the proposed BP algorithm.

To be general, we consider a *L*-tier HetNet. Here, "tier" represents the set of BSs with the same properties. That is, the BSs in the *l*-th tier, $\forall l \in \mathcal{L} \triangleq \{1, \dots, L\}$, have transmission power P_l and the path loss exponent α_l . Also, the distribution of the BSs in the *l*-th tier is modeled as an independent homogeneous Poisson point process (HPPP) Φ_l with the intensity λ_l [28]. Thus, the *l*-th tier is defined by $\{\Phi_l, \lambda_l, P_l, \alpha_l\}$. Also, the distribution of the MUs in the HetNet is modeled as an HPPP Φ_u with the intensity λ_u .

Note that the tier structure is transparent to our BP algorithm, since in the BP all the BSs are treated as independent factor nodes, and in the factor graph, the existence of the edge (i, j) depends on whether the received SINR at \mathcal{M}_i from \mathcal{B}_j is larger than the SINR threshold δ . Now, we investigate the average degree distribution of the factor graph. Note that the degree of a variable node i is defined as the number of its neighboring factor nodes, i.e., $|\mathcal{H}(i)|$. The degree of a factor node j is defined as the number of a variable node i. Now, we can formulate the following theorem.

Theorem 2. The variable nodes in the factor graph have the average degree

$$D_{u} \triangleq \mathbb{E}(|\mathcal{H}(i)|) = \sum_{l=1}^{L} 2\pi \lambda_{l} Z\left(\lambda_{l}, P_{l}, \alpha_{l}, \delta\right), \quad (23)$$

and the factor nodes corresponding to the BSs in the *l*-th tier have the average degree

$$D_{l} \triangleq \mathbb{E}(|\mathcal{H}(j)|) = 2\pi\lambda_{u}Z(\lambda_{l}, P_{l}, \alpha_{l}, \delta), \qquad (24)$$

where

$$Z\left(\lambda_{l}, P_{l}, \alpha_{l}, \delta\right) = \int_{0}^{\infty} \exp\left\{-\sum_{k=1}^{L} \frac{2\lambda_{k}\pi}{\alpha_{k}} \left(\frac{\delta P_{k}}{P_{l}}\right)^{\frac{2}{\alpha_{k}}} \\ B\left(\frac{2}{\alpha_{k}}, 1 - \frac{2}{\alpha_{k}}\right) r^{\frac{2\alpha_{l}}{\alpha_{k}}} - \frac{\delta\sigma^{2}}{P_{l}}r^{\alpha_{l}}\right\} r \mathrm{d}r \quad (25)$$

and the Beta function $B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1} dt$. Proof: Please refer to Appendix D.

Based on Theorem 2, we have the following two corollaries.

Corollary 1. Assuming all the tiers have the equal path loss exponents, i.e., $\alpha_1 = \cdots = \alpha_L = \alpha$, and neglecting the noise, the function $Z(\lambda_l, P_l, \alpha_l, \delta)$ in (25), can be rewritten as

$$Z\left(\lambda_{l}, P_{l}, \alpha, \delta\right) = \frac{\alpha}{4\pi} \left(\frac{P_{l}}{\delta}\right)^{\frac{2}{\alpha}} \frac{1}{B\left(\frac{2}{\alpha}, 1 - \frac{2}{\alpha}\right) \sum_{k=1}^{L} \lambda_{k} P_{k}^{\frac{2}{\alpha}}}.$$
(26)

Then we simplify the average degree of the variable nodes as

$$D_u = \frac{\alpha}{2\delta^{\frac{2}{\alpha}} B\left(\frac{2}{\alpha}, 1 - \frac{2}{\alpha}\right)},\tag{27}$$

and the average degree of the factor nodes related to the BSs in the l-th tier as

$$D_{l} = \frac{\lambda_{u}\alpha}{2} \left(\frac{P_{l}}{\delta}\right)^{\frac{2}{\alpha}} \frac{1}{B\left(\frac{2}{\alpha}, 1 - \frac{2}{\alpha}\right) \sum_{k=1}^{L} \lambda_{k} P_{k}^{\frac{2}{\alpha}}} = \frac{P_{l}^{\frac{2}{\alpha}}}{\sum_{k=1}^{L} \lambda_{k} P_{k}^{\frac{2}{\alpha}}} \lambda_{u} D_{u}.$$
(28)

Proof: Please refer to Appendix E.

Equations (27) and (28) can be seen as approximations of (23) and (24), respectively, when noise power is neglected. These approximations are accurate enough for the HetNets, since interference is dominant due to the dense deployments of the SBSs.

From (27), the average degree of a variable node, i.e., D_u , is only related to δ and α , while independent of λ_l and P_l . In other words, the number of potential serving BSs for an MU is independent of the BSs' deployment intensities and transmission powers. An intuitive explanation is that although increasing the BSs' deployment intensities or transmission powers can enhance the MUs' received signal, the interference increases at the same time. Since D_u keeps constant, the average computational and communication complexity of the proposed BP algorithm at the MU side also remains constant, even when the scale of the network increases.

From (28), the average degree of factor nodes in *l*-th tier, i.e., D_l , increases with λ_u . This means that the number of MUs in the coverage of an BS is proportional to the intensity of MUs. Also, D_l decreases when the deployment intensities of the BSs increases. That is, the coverage area of an BS shrinks when the number of BSs increases. Furthermore, D_l increases with $\frac{P_l}{P_k}$, $k = 1, \dots, L$, and $k \neq l$, which means the

http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

coverage area of the BSs in l-th tier depends on the power ratios between the l-th tier and the other tiers.

Now we focus on the sparsity of the factor graph. Given a factor graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, the sparsity is defined as the ratio between the number of existing edges in the graph, i.e., $|\mathcal{E}|$, and the number of all possible edges. The latter can be calculated as $|\mathcal{I}| \times |\mathcal{J}|$. Generally, the sparsity of a factor graph is important to the performance of the BP algorithm. If the graph is too dense, the complexity will be high, and the highly correlated messages cause a poor performance due to a large number of loops. On the other hand, if the graph is too sparse, messages cannot be effectively conveyed between nodes, causing a performance degradation.

In the following corollary, we give the expression of the average sparsity of our factor graph.

Corollary 2. Given the average degree distribution of the variable nodes, we can express the average sparsity of the factor graph in the area of R^2 as

$$S = \frac{D_u}{R^2 \sum_{l=1}^L \lambda_l}.$$
(29)

Proof: Please refer to Appendix F.

From (29), the sparsity of the factor graph is independent of the intensity of MUs, i.e., λ_u , and inversely proportional to the number of BSs in the network. That is, the factor graph is sparser when the scale of the network increase. Also intuitively, we could control the average sparsity of the factor graph by tuning the threshold δ .

Remark 1. We observe that in (27) the beta function $B\left(\frac{2}{\alpha}, 1-\frac{2}{\alpha}\right) = \pi$ when $\alpha = 4$. Thus, we can have closed forms for D_u , D_l , and S in (27), (28), and (29), respectively, when $\alpha = 4$.

VI. SIMULATION RESULTS

In this section, we first analyze the performance of the proposed BP algorithm using Monte-Carlo simulations. Then we present analytical and simulation results for the average degree distribution and sparsity of the factor graph.

We consider a HetNet consisting of 4 macro cells where MBSs are deployed at the center of macro cells while multiple SBSs are randomly distributed. The inter-site distance (ISD) between MBSs is set to be 500 meters, and the transmission power of MBSs is 20W [29]. The deployment intensity of SBSs are 24 per cell and the transmission power of SBSs is 2W [29]. The MUs are uniformly distributed in the network with the intensity of 30 MUs per cell, and the MUs' priorities are modeled as 3 different levels, i.e., $\omega_i \in \{1, 2, 3\}$. The path loss exponent is set to be 4. The SINR threshold is 0.1, and the noise power is 10^{-13} W [29]. The simulations focus on the spectral efficiency where the bandwidth W = 1. Besides, in the simulation, $\mu = 10$ and the iteration number T = 5.

A. Performance of Distributed BP Algorithms

We consider two BP algorithms proposed in our paper. One is the original BP algorithm without approximation in Section III, which is denoted by 'BP-Exact'. The other is



Fig. 2. Utility comparison between two BP algorithms and the optimal result by exhaustive search.

the complexity reduced BP algorithm with approximation in Section IV, which is denoted by 'BP-Approx'. Also we compare them with the following user association schemes.

1) Max-SINR (denoted by 'Max-SINR'): Each MU chooses to associate with the BS that provides the strongest received SINR.

2) Cell Range Expansion: A positive bias will be added to the received power of SBSs before the MU selects the BS that provides the strongest SINR. Two bias values that are commonly used are selected for simulations, 4dB and 8dB, denoted by 'CRE-4dB' and 'CRE-8dB', respectively [30, 31].

3) Optimal results (denoted by 'Optimal'): The exhaustive search method is used to find out the optimal user-association solution.

Fig. 2 shows the cumulative distribution function (CDF) curves of the utility obtained by 'BP-Exact', 'BP-Approx' and 'Optimal'. Due to the high complexity of the exhaustive search, Fig. 2 considers one macro-cell with 4 overlapping small-cells. From the figure, we can see that 'BP-Exact' almost achieves the performance of 'Optimal' with 0.086% utility loss, and 'BP-Approx' can well approach 'Optimal' with 1.54% utility loss. The reason that 'BP-Exact' cannot achieve the optimal performance is as follows: a) The value of μ is finite; b) There could be loops in the factor graph. Since 'BP-Approx' has a very close performance with 'BP-Exact', we only consider 'BP-Approx' in the following simulations due to its low complexity.

Fig. 3 compares the system utility performance of different user association schemes: 'Max-SINR', 'CRE-4dB', 'CRE-8dB' and 'BP-Approx'. From the figure, we can see that, compared with the traditional 'Max-SINR' scheme, 'CRE-4dB' slightly improves the system utility by 4.39%. However, 'CRE-8dB' degrades the system utility. Among all the schemes, 'BP-Approx' has the best system performance, which increases the system utility of 'Max-SINR' by 28%.



Fig. 3. Utility performance of Max-SINR, CRE and BP-Approx with different user priorities.



Fig. 4. Comparison of System Utility Between Dynamic and Static BP Algorithm with different user priorities.

Generally, the CRE is simple and does not need message overhead once the bias is fixed. However, selecting a universally good bias is a non-trivial optimization problem, since it depends on many factors such as the deployment intensity of BSs and MUs, the transmission powers of BSs, and so on. The performance becomes unpredictable for a fixed bias when the network environment changes. In contrast, our BP is always near-optimal and robust to network environments. Particularly, the approximate BP has dramatically reduced the complexity with little performance loss.

To evaluate the performance of the dynamic BP algorithm, we apply the dynamic BP to two scenarios, i.e., the first



Fig. 5. Average degree of the variable nodes in the factor graph.

scenario: 80% 'existing MUs' with 20% 'new MUs' joining the network, and the second scenario: 60% 'existing MUs' with 40% 'new MUs' joining the network. Denoted by 'DBP-Approx(80% + 20%)' and 'DBP-Approx(60% + 40%)' for the two scenarios, respectively. For comparison, we also consider the start-over user association scheme, in which the approximate BP algorithm is applied to all MUs as a whole. This start-over scenario, denoted by 'BP-Approx(100%)', provides the benchmark with the optimal result.

From Fig. 4, we can see that the performance of the dynamic BP algorithm in the two scenarios approaches the 'BP-Approx(100%)' algorithm with a very small gap. Specifically, the performance gap between 'DBP-Approx(80% + 20%)' and 'BP-Approx(100%)' is only 2.28%, and the gap between 'DBP-Approx(100%)' is 5.59%. This means that we can rely on the dynamic BP to establish associations for the 'new MUs' with a much lower complexity, rather than starting over to perform the original BP algorithm for all the MUs.

B. Degree Distributions of Factor Graph

Fig. 5 plots the average degree of variable nodes in the factor graph. Here, δ is a linear SINR value. First, we can see that the analytical results in Eq. (27) match well with the simulation results. The average degree decreases when δ increases, which reduces the complexity at the MU side. This is because that the number of the BSs, which can provide large enough SINR, decreases when the SINR threshold increases. From the figure, when $\delta > 0.4$, the average degree of a variable node is below one. In this case, the BP algorithm cannot perform well because of the limited association options for each MU. Importantly, the variable node degree only depends on the threshold δ , and is constant relative to the deployment intensities and transmission powers of BSs. This property provides a constant computational and communication complexity at the MU side once δ is determined.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2015.2488039, IEEE Transactions on Vehicular Technology



Fig. 6. Average degree of the factor nodes in Tier-1 (MBSs).



Fig. 7. Average degree of the factor nodes in Tier-2 (Pico-cell BSs).

In Fig. 6 and Fig. 7, we present the analytical and simulation results for the average degree of the factor nodes in different tiers. As shown in the two figures, the analytical results both match well with the simulation ones. The average degree of factor nodes in tier-2 is much less than the those in tier-1 because the transmission power in tie-2 is much smaller. The average degree of factor nodes decreases with the increase of δ . Furthermore, we can see that when we increase the deployment intensities of BSs in tier-2, the degree of factor nodes in both tiers will decrease, as shown in the two figures. The reason is that more BSs are contributing to the inter-cell interference. However, when the transmission power of BSs in tier-2 increases, as well, while the degree of factor nodes in tier-1 decreases.



Fig. 8. Sparsity of the factor graph in an unit area $(1km^2)$.

Also, we can see that the degrees of factor nodes in all tiers will increase as the intensity of MUs increases.

Fig. 8 shows the sparsity of the factor graph within a unit area, i.e., 1 square kilometer. From Eq. (29), we can see that the sparsity is independent of the transmission power of the BSs. This can be verified from this figure, i.e., the sparsity keeps constant when we increase the power of small-cell BSs P_2 from 1W to 2W. However, when the intensities of BSs increases, the sparsity decreases, which can be seen from Eq. (29), and is also verified by Fig. 8. Also, we can see that the sparsity is below 0.1 per area, and will decrease significantly when the area of the HetNet increases. The sparsity also decreases when the threshold δ increases. Thus, if δ is close to 0, the relative high density of the factor graph will affect the performance of the BP algorithm. We can tune the value of δ to achieve a good BP performance. Practically, $\delta = 0.1$ will lead to a reasonably good performance.

VII. CONCLUSION

In this paper, we proposed a distributed BP algorithm to solve the user association problem in HetNets with various user priorities. In addition, we developed an approximate BP algorithm to reduce the computational and communication complexity at the BS side with a slight performance loss. The computational complexity can be reduced to only $|\mathcal{H}(j)|$ multiplications and $|\mathcal{H}(j)| - 1$ additions, and the messages are transmitted in a broadcast manner. This is in contrast with the complexity of the exact BP algorithm, of which the computational complexity is $O(2^{|\mathcal{H}(j)|}|\mathcal{H}(j)|)$, and the messages are transmitted in a point-to-point manner. Also the practical dynamic algorithm is proposed. Furthermore, the average degree distribution and the sparsity of the developed factor graph were analyzed based on the stochastic geometry theory. Simulation results showed that the proposed BP algorithm converged quickly within five iterations. Its performance

11

almost overlapped with the optimal result obtained by an exhaustive search. Compared with the existing schemes, our proposed distributed BP algorithm improves the system utility by nearly 30%. The simulations also showed that with the parameters in 3GPP specification, the average value of |H(i)|, i.e., the variable node degree D_u , is practically small (around two), which indicated a low computational and communication complexity at the MU side. Considering the scenario that each MU can associate with more than one BS, the constraint $\sum_j x_{ij} = 1, \forall i$, will be relaxed and more potential MU-BS associations will be involved in to the BP algorithm. Although the framework of our BP algorithm still works, the detail messages will be reformulated. Hence, we leave this issue as our future work.

APPENDIX A Proof of Lemma 1

To simplify the notation in the proof, we assume that each MU's neighbor set is \mathcal{J} , i.e., $\mathcal{H}(i) = \mathcal{J}$ and $\mathcal{H}(j) = \mathcal{I}$. This will not change the proof. Consider two probability sets $\mathcal{S}^{(t-1)} = \left\{ m_{i \to j}^{(t-1)}(x_{ij}) \right\}$ and $\widetilde{\mathcal{S}}^{(t-1)} = \left\{ \widetilde{m}_{i \to j}^{(t-1)}(x_{ij}) \right\}$. Then we have the supremum norm

$$\begin{aligned} \left\| \mathbf{\Gamma} \left(\mathbf{S}^{(t-1)} \right) - \mathbf{\Gamma} \left(\widetilde{\mathbf{S}}^{(t-1)} \right) \right\|_{\sup} &= \max_{i \in \mathbf{\mathcal{I}}, j \in \mathbf{\mathcal{J}}} \\ \left| m_{i \to j}^{(t)}(x_{ij}) - \widetilde{m}_{i \to j}^{(t)}(x_{ij}) \right| &= \frac{1}{\varphi} \max_{i,j} \left\| \prod_{k \in \mathbf{\mathcal{J}} \setminus \{j\}} \sum_{\mathbf{x} \mathbf{z} \setminus \{x_{ik}\}} \\ \exp(\mu f_k) \left(\prod_{q \in \mathbf{\mathcal{I}} \setminus \{i\}} m_{q \to k}^{(t-1)}(x_{qk}) - \prod_{q \in \mathbf{\mathcal{I}} \setminus \{i\}} \widetilde{m}_{q \to k}^{(t-1)}(x_{qk}) \right) \right\| \\ \stackrel{(a)}{\leq} \frac{1}{\varphi} \max_{j} \prod_{k \in \mathbf{\mathcal{J}} \setminus \{j\}} N \sum_{\mathbf{x} \mathbf{z} \setminus \{x_{ik}\}} \max_{i} \left\| \prod_{q \in \mathbf{\mathcal{I}} \setminus \{i\}} m_{q \to k}^{(t-1)}(x_{qk}) - \prod_{q \in \mathbf{\mathcal{I}} \setminus \{i\}} \widetilde{m}_{q \to k}^{(t-1)}(x_{qk}) \right\| \\ & \int_{q \in \mathbf{\mathcal{I}} \setminus \{i\}} \widetilde{m}_{q \to k}^{(t-1)}(x_{qk}) - \widetilde{m}_{q \to k}^{(t-1)}(x_{qk}) \right\| \\ & \leq \frac{\left(2^{|\mathbf{\mathcal{I}}|-1}(||\mathbf{\mathcal{I}}|-1)N\right)^{|\mathbf{\mathcal{J}}|-1}}{\varphi} \\ & \max_{q,k} \left| m_{q \to k}^{(t-1)}(x_{qk}) - \widetilde{m}_{q \to k}^{(t-1)}(x_{i}) - \widetilde{m}_{i \to k}^{(t-1)}(x_{i}) \right\| \\ & = \frac{\left(2^{|\mathbf{\mathcal{I}}|-1}(||\mathbf{\mathcal{I}}|-1)N\right)^{|\mathbf{\mathcal{J}}|-1}}{\varphi} \\ & \left\| \mathbf{\mathcal{S}}^{(t-1)} - \widetilde{\mathbf{\mathcal{S}}}^{(t-1)} \right\|_{sup}^{t}. \end{aligned}$$
(30)

The inequality (a) comes from the two following two facts: 1) Given μ , the function $\exp(\mu f_k)$ is a bounded value, say upper bounded by a constant N, since the utility f_k is always a bounded value, and 2) $|\sum_s x_s| \leq \sum_s |x_s|$ for arbitrary x_s . The inequality (b) can be obtained from 1) the fact that $\sum_{\mathbf{x}_{\mathcal{I}} \setminus \{x_{ik}\}}$ is the summation of $2^{|\mathcal{I}|-1}$ items, and 2) the following inequality.

$$\max_{i \in \mathcal{I}} \left| \prod_{q \in \mathcal{I} \setminus \{i\}} m_{q \to k} - \prod_{q \in \mathcal{I} \setminus \{i\}} \widetilde{m}_{q \to k} \right| \\ \leq (|\mathcal{I}| - 1) \max_{q \in \mathcal{I} \setminus \{i\}} |m_{q \to k} - \widetilde{m}_{q \to k}|. \quad (31)$$

The proof for (31) is based on the constraint that all the message values are between 0 and 1. The proof follows iterative manner by first considering $|\mathcal{I}| = 2$. Then we plug the result to the case when $|\mathcal{I}| = 3$, and so on. To be concise with the paper, we skipped this iterative process.

From (30), we say that Γ is a continuous mapping since the coefficient $(2^{|\mathcal{I}|-1}(|\mathcal{I}|-1)K)^{|\mathcal{J}|-1}$ is a finite number, and this completes the proof.

APPENDIX B PROOF OF THEOREM 1

Let \mathcal{A} be the collection of the message set $\mathcal{S}^{(t)}$. The mapping function Θ maps \mathcal{A} to \mathcal{A} with the function Γ . According to Lemma 1, Θ is continuous since Γ is continuous. Also, it is clear that the set \mathcal{A} is convex, closed and bounded. According to Schauder fixed point theorem, Θ has a fixed point. This completes the proof.

APPENDIX C APPROXIMATION OF THE LIKELIHOOD

The likelihood ratio $\frac{m_{j \to i}(x_{ij}=1)}{m_{j \to i}(x_{ij}=0)}$ can be approximated as follows. To simplify the notation, we assume $\mu = 1$ here, since it is a constant and does not affect the approximations.

$$\frac{m_{j \to i} (x_{ij} = 1)}{m_{j \to i} (x_{ij} = 0)} \approx \frac{\mathbb{E} \left(\exp \left(f_{j} \right) |_{x_{ij} = 1} \right)}{\mathbb{E} \left(\exp \left(f_{j} \right) |_{x_{ij} = 0} \right)} \approx \frac{\exp \left(\mathbb{E} \left(f_{j} |_{x_{ij} = 1} \right) \right)}{\exp \left(\mathbb{E} \left[f_{j} |_{x_{ij} = 0} \right] \right)} = \frac{\exp \left(\mathbb{E} \left(\sum_{k \in \mathcal{H}(j) \setminus \{i\}} x_{kj} \omega_{k} \log \frac{W \gamma_{kj} \omega_{k}}{\sum_{q \in \mathcal{H}(j) \setminus \{i\}} x_{qj} \omega_{q} + \omega_{i}} \right) \right)}{\exp \left(\mathbb{E} \left(\sum_{k \in \mathcal{H}(j) \setminus \{i\}} x_{kj} \omega_{k} \log \frac{W \gamma_{kj} \omega_{k}}{\sum_{q \in \mathcal{H}(j) \setminus \{i\}} x_{qj} \omega_{q}} \right) \right) \times \exp \left(\mathbb{E} \left(\omega_{i} \log \frac{W \gamma_{ij} \omega_{i}}{\sum_{q \in \mathcal{H}(j) \setminus \{i\}} x_{qj} \omega_{q} + \omega_{i}} \right) \right),$$
(32)

where (a) comes from the approximation that $\mathbb{E}(\exp(x)) \approx \exp(\mathbb{E}(x))$. Since the priority ω_i is uniformed distributed with a relatively small range, e.g., $\omega_i \in [1,3]$, we further make the approximation on the last equation in (32) as $\log\left(\sum_{q \in \mathcal{H}(j) \setminus \{i\}} x_{qj} \omega_q + \omega_i\right) \approx \log\left(\sum_{q \in \mathcal{H}(j) \setminus \{i\}} x_{qj} \omega_q\right)$,

which is accurate enough when $|\mathcal{H}(j)|$ is large. Then (32)

^{0018-9545 (}c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See

http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

12

can be rewritten as

$$\exp\left(\mathbb{E}\left(\omega_{i}\log\frac{W\gamma_{ij}\omega_{i}}{\sum\limits_{q\in\mathcal{H}(j)\setminus\{i\}}x_{qj}\omega_{q}+\omega_{i}}\right)\right)$$

$$=\frac{\exp\omega_{i}\log\left(W\gamma_{ij}\omega_{i}\right)}{\exp\mathbb{E}\left(\omega_{i}\log\left(\sum\limits_{q\in\mathcal{H}(j)\setminus\{i\}}x_{qj}\omega_{q}+\omega_{i}\right)\right)\right)}$$

$$\stackrel{(a)}{\approx}\frac{(W\gamma_{ij}\omega_{i})^{\omega_{i}}}{\left(\mathbb{E}\left(\sum\limits_{q\in\mathcal{H}(j)\setminus\{i\}}x_{qj}\omega_{q}+\omega_{i}\right)\right)^{\omega_{i}}}$$

$$=\left(\frac{W\gamma_{ij}\omega_{i}}{\sum\limits_{q\in\mathcal{H}(j)\setminus\{i\}}\omega_{q}m_{q\rightarrow j}\left(x_{qj}=1\right)+\omega_{i}}\right)^{\omega_{i}},\quad(33)$$

where (a) comes from the same approximation as in Eq. (32). This completes the approximation process.

APPENDIX D PROOF OF THEOREM 2

A. The Average Degree of Variable Nodes

Without loss of generality, we conduct the analysis on a typical MU that is located at the origin and assume that the potential serving BSs in *l*-th tier locate at the point x_l , $\forall l \in \mathcal{L}$. The fading (power) is denoted by h_{x_l} , which is assumed to be exponential distributed, i.e., $h_{x_l} \sim \exp(1)$. The path loss function is given by $||x_l||^{-\alpha_l}$, where $||\cdot||$ denotes the Euclidian distance.

The average number of edges emanated from the typical MU to the BSs in the l-th tier can be formulated as

$$N_l = \int_{\mathbb{R}^2} \lambda_l \Pr\left(\rho(x_l) > \delta\right) dx_l, \tag{34}$$

where $\rho(x_l)$ represents the received SINR at the typical MU from the *l*-th tier BS located at x_l .

Now, we focus on the probability $\Pr(\rho(x_l) > \delta)$ in (34) as follows.

$$\Pr\left(\rho(x_l) > \delta\right) = \Pr\left(\frac{P_l h_{x_l} \|x_l\|^{-\alpha_l}}{\sum\limits_{k=1}^{L} \sum\limits_{x_k \in \Phi_k} P_k h_{x_k} \|x_k\|^{-\alpha_k} + \sigma^2} > \delta\right)$$
$$= \Pr\left(h_{x_l} > \frac{\delta\left(I + \sigma^2\right)}{P_l \|x_l\|^{-\alpha_l}}\right)$$
$$= \mathbb{E}_I\left(\exp\left(-sI\right)\right) \exp\left(-s\sigma^2\right), \qquad (35)$$

where x_k denotes the locations of interfering BSs, $I \stackrel{\triangle}{=} \sum_{k=1}^{L} \sum_{x_l \in \Phi_k} P_k h_{x_l} ||x_l||^{-\alpha_k}$ represents the aggregate interference, and $s = \frac{\delta ||x_l||^{\alpha_l}}{P_l}$. The last step results due to the exponential distribution of h_{x_l} . Then, we derive $\mathbb{E}_I (\exp(-sI))$ in (35) as follows.

$$\mathbb{E}_{I}\left(\exp\left(-sI\right)\right) \stackrel{(a)}{=} \prod_{k=1}^{L} \mathbb{E}_{I_{k}}\left(\exp\left(-sI_{k}\right)\right) \stackrel{(b)}{=} \prod_{k=1}^{L} \mathbb{E}_{\Phi_{k}}$$

$$\left(\prod_{x_{k}\in\Phi_{k}}\int_{0}^{\infty} \exp\left(-sP_{k}h_{x_{k}} \|x_{k}\|^{-\alpha_{k}}\right) \exp(-h_{x_{k}}) dh_{x_{k}}\right)$$

$$\stackrel{(c)}{=} \prod_{k=1}^{L} \exp\left(-\lambda_{k}\int_{\mathbb{R}^{2}} \left(1 - \frac{1}{1 + sP_{k}} \|x_{k}\|^{-\alpha_{k}}\right) dx_{k}\right)$$

$$= \prod_{k=1}^{L} \exp\left(-2\pi\lambda_{k}\frac{1}{\alpha_{k}}\left(sP_{k}\right)^{\frac{2}{\alpha_{k}}} B\left(\frac{2}{\alpha_{k}}, 1 - \frac{2}{\alpha_{k}}\right)\right),$$
(36)

where $I_k \triangleq \sum_{x_l \in \Phi_k} P_k h_{x_l} ||x_l||^{-\alpha_k}$. In (36), (a) follows the independence of Φ_k , i.e., the point process of one tier is independent of other tiers, (b) is based on the independence of channel fading, and (c) follows $\mathbb{E}\left(\prod_x u(x)\right) = \exp\left(-\lambda \int_{\mathbb{R}^2} (1-u(x)) \, dx\right)$, where $x \in \Phi$ and Φ represents a Poisson point process in \mathbb{R}^2 with the intensity λ [32].

Based on the derivation above, the average number of edges emanated from the typical MU to BSs from all the tiers can be calculated as

$$N = \sum_{l=1}^{L} N_l = \sum_{l=1}^{L} \lambda_l \int_{\mathbb{R}^2} \exp\left(-\sum_{k=1}^{L} 2\pi \frac{\lambda_k}{\alpha_k} \left(\frac{\delta P_k}{P_l}\right)^{\frac{2}{\alpha_k}}\right)$$
$$B\left(\frac{2}{\alpha_k}, 1 - \frac{2}{\alpha_k}\right) \|x_l\|^{\frac{2\alpha_l}{\alpha_k}} - \frac{\delta\sigma^2}{P_l} \|x_l\|^{\alpha_l} dx_l$$
$$= \sum_{l=1}^{L} 2\pi \lambda_l \int_0^\infty \exp\left(-\sum_{k=1}^{L} 2\pi \frac{\lambda_k}{\alpha_k} \left(\frac{\delta P_k}{P_l}\right)^{\frac{2}{\alpha_k}}\right)$$
$$B\left(\frac{2}{\alpha_k}, 1 - \frac{2}{\alpha_k}\right) r^{\frac{2\alpha_l}{\alpha_k}} - \frac{\delta\sigma^2}{P_l} r^{\alpha_l} r^{\alpha_l} dr$$
(37)

It can be seen that the average degree of variable nodes, i.e., D_u , equals to N.

B. The Degree of Factor Nodes in l-th Tier

In this subsection, we assume a typical BS in the l-th tier that is located at the origin, and assume that an MU is located at the point x. The average number of edges emanated from the typical BS to the MUs can be formulated as

$$N_u = \int_{\mathbb{R}^2} \lambda_u \Pr\left(\rho(x) > \delta\right) \mathrm{d}x. \tag{38}$$

where $\rho(x)$ represents the received SINR at the MU located at x from the typical BS, i.e.,

$$\Pr\left(\rho(x) > \delta\right) = \left\{ \frac{P_l h_x \left\|x\right\|^{-\alpha_l}}{\sum\limits_{k=1}^{L} \sum\limits_{x_k \in \Phi_k} P_k h_{x_k, x} \left\|x_k - x\right\|^{-\alpha_k} + \sigma^2} > \delta \right\}, \quad (39)$$

where x_k denotes the location of an interfering BS.

0018-9545 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

13

Since HPPP is a stationary process, the distribution of $||x_k - x||$ is independent of the value of x, i.e., $p(||x_k - x||) = p(||x_k||)$, where $p(\cdot)$ represents the probability density function. Then, we have similar results with Eq. (35). That is,

$$\Pr\left(\rho(x) > \delta\right) = \mathbb{E}_{I}\left(\exp\left(-sI\right)\right)\exp\left(-s\sigma^{2}\right), \qquad (40)$$

where $s = \frac{\delta ||x||^{\alpha_l}}{P_l}$. Then we have

$$N_{u} = 2\pi\lambda_{u} \int_{0}^{\infty} \exp\left(-\sum_{k=1}^{L} 2\pi \frac{\lambda_{k}}{\alpha_{k}} \left(\frac{\delta P_{k}}{P_{l}}\right)^{\frac{2}{\alpha_{k}}}\right) B\left(\frac{2}{\alpha_{k}}, 1 - \frac{2}{\alpha_{k}}\right) r^{\frac{2\alpha_{l}}{\alpha_{k}}} - \frac{\delta\sigma^{2}}{P_{l}} r^{\alpha_{l}}\right) r \mathrm{d}r. \quad (41)$$

We can see that the average degree of factor nodes corresponding to the BSs in the *l*-th tier, i.e., D_l , equals to N_u . Combined with the results Eq. (36), we completes the proof.

APPENDIX E Proof of Corollary 1

By plugging $\alpha_1 = \cdots = \alpha_L = \alpha$ into (25), and ignoring the noise, we have

$$Z(\lambda_{l}, P_{l}, \alpha, \delta)$$

$$= \int_{0}^{\infty} \exp\left(-\sum_{k=1}^{L} \frac{2\pi\lambda_{k}}{\alpha} \left(\frac{\delta P_{k}}{P_{l}}\right)^{\frac{2}{\alpha}} B\left(\frac{2}{\alpha}, 1-\frac{2}{\alpha}\right) r^{2}\right) r dr$$

$$= \frac{1}{2} \int_{0}^{\infty} \exp\left(-\sum_{k=1}^{L} \lambda_{k} \left(\frac{P_{k}}{P_{i}}\right)^{\frac{2}{\alpha}} \frac{2\pi}{\alpha} \delta^{\frac{2}{\alpha}} B\left(\frac{2}{\alpha}, 1-\frac{2}{\alpha}\right) t\right) dt$$

$$= \frac{1}{2\sum_{k=1}^{L} \lambda_{k} \left(\frac{P_{k}}{P_{l}}\right)^{\frac{2}{\alpha}} \frac{2\pi}{\alpha} \delta^{\frac{2}{\alpha}} B\left(\frac{2}{\alpha}, 1-\frac{2}{\alpha}\right)}{4\pi B\left(\frac{2}{\alpha}, 1-\frac{2}{\alpha}\right) \delta^{\frac{2}{\alpha}}} \frac{P_{l}^{\frac{2}{\alpha}}}{\sum_{k=1}^{L} \lambda_{k} \left(P_{k}\right)^{\frac{2}{\alpha}}}.$$
(42)

By substituting the above Z into (23) and (24), we can obtain (27) and (28) respectively. This completes the proof.

APPENDIX F Proof of Corollary 2

The density (sparsity) of an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is defined as:

$$S = \frac{2|\boldsymbol{\mathcal{E}}|}{|\boldsymbol{\mathcal{V}}|(|\boldsymbol{\mathcal{V}}|-1)},\tag{43}$$

where $\frac{1}{2}|\mathcal{V}|(|\mathcal{V}|-1)$ represents the maximum edges in this graph. In our factor graph model, the edges only exist between variable nodes and factor nodes. Thus, the density of our factor graph can be rewritten as

$$S = \frac{|\mathcal{E}|}{|\mathcal{I}||\mathcal{J}|}.$$
(44)

From the perspective of variable nodes \mathcal{I} , it is easy to know that $|\mathcal{E}| = \sum_{i \in \mathcal{I}} D_i$, where D_i denotes the degree of node *i*. Due to the HPPP distribution of the BSs and MUs, each MU

averagely has the same degree, i.e., $D_i = D$. We can obtain that $S = \frac{D|\mathcal{I}|}{|\mathcal{I}||\mathcal{J}|} = \frac{D}{|\mathcal{J}|}$.

We consider a typical MU located at the origin, and an area dx around the position x. The total number of BSs from all the tiers in this area can be given by $\sum_{l=1}^{L} \lambda_l dx$. The number of possible serving BSs (or the number of edges in the factor graph) for the typical MU in this area from all tiers can be formulated as $\sum_{l=1}^{L} \lambda_l dx \Pr(\rho(x, P_l) > \delta)$, where $\rho(x, P_l)$ represents the received SINR at the typical MU from the *l*-th tier BS located at x. Then sparsity in dx can be calculated as

$$S(x) = \frac{\sum_{l=1}^{L} \lambda_l \Pr\left(\rho(x, P_l) > \delta\right) \mathrm{d}x}{\sum_{l=1}^{L} \lambda_l \mathrm{d}x}.$$
 (45)

Then we can obtain the average of the sparsity when x ranges in the whole area R^2 , i.e.,

$$\mathbb{E}(S) = \int_{\mathbb{R}^2} S(x)p(x)dx$$

$$= \int_{\mathbb{R}^2} \frac{\sum_{l=1}^L \lambda_l \Pr\left(\rho(x, P_l) > \delta\right)}{\sum_{l=1}^L \lambda_l} p(x)dx$$

$$= \frac{1}{\sum_{l=1}^L \lambda_l} \sum_{l=1}^L \lambda_l \int_0^R \int_0^R \Pr\left(\rho(\delta, r, P_l) > \delta\right) dx$$

$$= \frac{D_u}{\sum_{l=1}^L \lambda_l R^2}.$$
(46)

This completes the proof.

REFERENCES

- F. Boccardi, R. Heath, A. Lozano, T. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, Feb. 2014.
- [2] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A survey on 3GPP heterogeneous networks," *IEEE Wireless Communications*, vol. 18, no. 3, pp. 10–21, Jun. 2011.
- [3] Y. Kishiyama, A. Benjebbour, T. Nakamura, and H. Ishii, "Future steps of LTE-A: evolution toward integration of local area and wide area systems," *IEEE Wireless Communications*, vol. 20, no. 1, pp. 12–18, Feb. 2013.
- [4] T. Nakamura, S. Nagata, A. Benjebbour, Y. Kishiyama, T. Hai, S. Xiaodong, Y. Ning, and L. Nan, "Trends in small cell enhancements in LTE advanced," *IEEE Communications Magazine*, vol. 51, no. 2, pp. 98–105, Feb. 2013.
- [5] P. Bahl, M. Hajiaghayi, K. Jain, S. Mirrokni, L. Qiu, and A. Saberi, "Cell breathing in wireless LANs: Algorithms and evaluation," *IEEE Transactions on Mobile Computing*, vol. 6, no. 10, pp. 164–178, Feb. 2007.
- [6] Y. Bejerano and S.-J. Han, "Cell breathing techniques for load balancing in wireless LANs," *IEEE Transactions on Mobile Computing*, vol. 8, no. 6, pp. 735–749, Jun. 2009.
- [7] Huawei, "System simulations for downlink co-channel interference scenario (R1-130507)," in *3GPP TSG RAN WGC Meetig-72*, Jan. 2013.
- [8] A. Khandekar, N. Bhushan, J. Tingfang, and V. Vanghi, "LTE-advanced: Heterogeneous networks," in *European Wireless Conference (EW)*, Apr. 2010, pp. 978–982.
- [9] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A survey on 3GPP heterogeneous networks," *IEEE Wireless Communications*, vol. 18, no. 3, pp. 10–21, Jun. 2011.
- [10] S. Corroy, L. Falconetti, and R. Mathar, "Dynamic cell association for downlink sum rate maximization in multi-cell heterogeneous networks," in *IEEE International Conference on Communications (ICC)*, Jun. 2012, pp. 2457–2461.

- [11] Y. Bejerano, S.-J. Han, and L. Li, "Fairness and load balancing in wireless lans using association control," IEEE/ACM Transactions on Networking, vol. 15, no. 3, pp. 560-573, Jun. 2007.
- [12] L. Wang, W. Chen, and J. Li, "Congestion aware dynamic user association in heterogeneous cellular network: A stochastic decision approach," in IEEE International Conference on Communications (ICC), Jun. 2014.
- [13] R. Han, C. Feng, and H. Xia, "Optimal user association based on topological potential in heterogeneous networks," in IEEE 24th International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), Sept. 2013, pp. 2409-2413.
- [14] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "Distributed αoptimal user association and cell load balancing in wireless networks," IEEE/ACM Transactions on Networking, vol. 20, no. 1, pp. 177-190, Feb. 2012.
- [15] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. Andrews, "User association for load balancing in heterogeneous cellular networks," IEEE Transactions on Wireless Communications, vol. 12, no. 6, pp. 2706-2716. Jun. 2013.
- [16] W. Li, S. Wang, Y. Cui, X. Cheng, R. Xin, M. Al-Rodhaan, and A. Al-Dhelaan, "AP association for proportional fairness in multirate WLANs," IEEE/ACM Transactions on Networking, vol. 22, no. 1, pp. 191-202, Feb 2014
- [17] T. Bu, L. Li, and R. Ramjee, "Generalized proportional fair scheduling in third generation wireless data networks," in IEEE International Conference on Computer Communications (INFOCOM), Apr. 2006, pp. 1 - 12
- [18] W.-H. Kuo and W. Liao, "Utility-based resource allocation in wireless networks," IEEE Transactions on Wireless Communications, vol. 6, no. 10, pp. 3600-3606, Oct. 2007.
- [19] A. Sang, X. Wang, M. Madihian, and R. Gitlin, "Coordinated load balancing, handoff/cell-site selection, and scheduling in multi-cell packet data systems," Wireless Networks, vol. 14, no. 1, pp. 103-120, 2008.
- [20] N. Roseveare and B. Natarajan, "An alternative perspective on utility maximization in energy-harvesting wireless sensor networks," IEEE Transactions on Vehicular Technology, vol. 63, no. 1, pp. 344-356, Jan. 2014
- [21] M. Ding, H. Luo, and W. Chen, "Polyblock algorithm based robust beamforming for downlink multi-user systems with per-antenna power constraints," IEEE Transactions on Wireless Communications, vol. 13, no. 8, pp. 4560-4573, Aug. 2014.
- [22] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," IEEE Transactions on Information Theory, vol. 47, no. 2, pp. 498-519, Feb. 2001.
- [23] F. Kelly, "Charging and rate control for elastic traffic," European Transactions on Telecommunications, vol. 8, no. 1, pp. 33-37, Jan. 1997.
- [24] K. Eriksson, S. Shi, V. Nikola, M. Schubert, and E. G. Larsson, "Globally optimal resource allocation for achieving maximum weighted sum rate," in IEEE Global Telecommunications Conference (GLOBECOM), Dec. 2010, pp. 1-6.
- [25] S. Rangan and R. Madan, "Belief propagation methods for intercell interference coordination in femtocell networks," IEEE Journal on Selected Areas in Communications, vol. 30, no. 3, pp. 631-640, Apr. 2012.
- [26] C. C. Moallemi and B. Van Roy, "Resource allocation via message passing," INFORMS Journal on Computing, vol. 23, no. 2, pp. 205-219, 2011.
- [27] N. Noorshams and M. Wainwright, "Stochastic belief propagation: A low-complexity alternative to the sum-product algorithm," IEEE Transactions on Information Theory, vol. 59, no. 4, pp. 1981-2000, Apr. 2013.
- [28] H. Dhillon, R. Ganti, F. Baccelli, and J. Andrews, "Modeling and analysis of K-tier downlink heterogeneous cellular networks," IEEE Journal on Selected Areas in Communications, vol. 30, no. 3, pp. 550-560, Apr. 2012.
- [29] 3GPP, "Further advancements for E-UTRA physical layer aspects," www.3gpp.org, Tech. Rep. v.9.0.0, Mar. 2010.
- S. Singh, H. Dhillon, and J. Andrews, "Offloading in heterogeneous [30] networks: Modeling, analysis, and design insights," IEEE Transactions on Wireless Communications, vol. 12, no. 5, pp. 2484-2497, May 2013.
- [31] B. Soret, H. Wang, K. Pedersen, and C. Rosa, "Multicell cooperation for lte-advanced heterogeneous network scenarios," *IEEE Wireless Commu*nications, vol. 20, no. 1, pp. 27-34, Feb. 2013.
- [32] S. N. Chiu, D. Stoyan, W. S. Kendall, and J. Mecke, Stochastic geometry and its applications. Hoboken: Wiley, 2013.



Jun Li (M'09) received Ph. D degree in Electronic Engineering from Shanghai Jiao Tong University, Shanghai, P. R. China in 2009. Then he worked as a Postdoctoral Fellow and a Research Fellow at the School of Electrical Engineering and Telecommunications, the University of New South Wales and the School of Electrical Engineering, the University of Sydney, respectively. From June 2015 to now, he is a Professor at the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, China. His research interests

include network information theory, channel coding theory, wireless network coding and cooperative communications.



Youjia Chen received the B.S. and M.S degrees in communication engineering from Nanjing University, Nanjing, China, in 2005 and 2008, respectively. She is currently pursuing the Ph.D. degree in wireless engineering at The University of Sydney, Sydney, Australia. Her current research interests include resource management, load balancing, and caching strategy in heterogeneous cellular networks.



Zihuai Lin received the Ph.D. degree in Electrical Engineering from Chalmers University of Technology, Sweden, in 2006. Prior to this he has held positions at Ericsson Research, Stockholm, Sweden. Following Ph.D. graduation, he worked as a Research Associate Professor at Aalborg University, Denmark and currently at the School of Electrical and Information Engineering, the University of Sydney, Australia. His research interests include graph theory, source/channel/network coding, coded modulation, MIMO, OFDMA, SC-FDMA, radio resource

management, cooperative communications, small-cell networks, 5G cellular systems, etc.



Guoqiang Mao (S'98-M'02-SM'08) is a Professor of Wireless Networking, Director of Center for Realtime Information Networks at the University of Technology, Sydney. He has published more than 100 papers in international conferences and journals, which have been cited more than 3000 times



Branka Vucetic (M'83-SM'00-F'03) currently holds the Peter Nicol Russel Chair of Telecommunications Engineering at the University of Sydney. During her career she has held various research and academic positions in Yugoslavia, Australia, UK and China. Her research interests include wireless communications, coding, digital communication theory and machine to machine communications. Prof Vucetic co-authored four books and more than four hundred papers in telecommunications journals and conference proceedings. She has been elected to the grade of IEEE Fellow for contributions to the theory and applications of

channel coding.