Joint User Association and Resource Allocation in the Downlink of Heterogeneous Networks

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

***Abstract*—In this paper, we consider the intercell interference coordination (ICIC) problem in heterogeneous cellular networks with randomly deployed small-cell base stations (BSs). Current research on ICIC mainly focuses on optimizing the spectrum and power allocations at BSs, whereas the user–BS association is treated as a separate issue. Nevertheless, the user–BS association problem is an important issue in ICIC and should be jointly optimized with resource allocations to achieve global optimality. In this paper, with the objective of maximizing the system sum rate in a distributed manner, we propose a novel belief propagation (BP) algorithm to jointly optimize user association, subchannel assignment, and power allocation. We first develop a factor graph model to decompose the network-wide objective and constraints into multiple local utilities. Then, we transform the maximization of local utilities into the estimations of marginal distributions and propose a distributed BP algorithm to solve the estimations. Simulations show that our distributed BP algorithm dramatically improves the performance compared with the benchmark scheme.**

***Index Terms*—Belief propagation (BP), heterogeneous networks, resource allocation, user association.**

I. Introduction

To cope with the dramatic growth of data traffic in wireless cellular networks, mobile network operators have been deploying increasingly more small-cell base stations (SBSs) to cooperate with traditional macrocell base stations (MBSs) in recent years. The resulting network with mixed SBSs and MBSs is called the heterogeneous cellular network (HCN) [1]. Compared with traditional homogeneous networks, intercell interference is more severe and difficult to manage in HCNs.

Long-Term Evolution (LTE) specifications provide several approaches for intercell interference coordination (ICIC) [2]. One straightforward method is orthogonal transmission among interfering cells. That is, the transmission time or the available bandwidth is divided into several parts and individually assigned to neighboring base stations (BSs), for instance, the almost blank subframe and fractional frequency reuse [3]–[5]. In [5], an optimal scheme on joint fractional frequency reuse and power control is proposed to maximize the long-term log-scale throughput of the system, where the flexibility of user association is considered. Although the orthogonal transmission effectively avoids the intercell interference, the disadvantage is that each BS can only use one part of the resource, which limits the system performance.

As the spectrum available for a cellular system becomes rare and expensive, the cochannel deployments, where MBSs and SBSs share the same spectrum resource, are highly desirable [6]. However, the cochannel transmissions will lead to sever intercell interference. One solution is to optimize the resource allocation among multiple BSs to maximize the system performance. In [7], Lopez-Perez *et al.* proposed a dynamic algorithm to jointly allocate frequency and power to mitigate intercell interference. Apart from resource allocation, user association is considered as another efficient factor in dealing with intercell interference. Qian *et al.* in [8] proposed an algorithm via the classic Benders' decomposition to solve the optimization problem of joint user association and power control. Li *et al.* in [9] proposed an asymptotically optimal solution for the resource allocation problem in HCNs with cooperative relay nodes. However, joint optimization on user association, spectrum allocation, and power allocation is not considered in the two [8] and [9] and, thus, unable to achieve global optimality. In [10] and [11], the joint optimization on user association and resource allocation is formulated, and the performance of different user association rules and resource allocation schemes is investigated. However, the exact solutions are not obtained due to its nonconvexity and NP-hard properties.

In this paper, by considering user association as a flexible parameter, we jointly optimize user association and resource allocation with the aim of maximizing the system sum rate distributively. We propose a distributed belief propagation (BP) algorithm to solve the maximization problem. Specifically, a factor graph model is first developed to decompose this network-wide optimization problem into multiple maximization problems performed at each BS. Then, we propose a novel method to satisfy the inter-BS resource constraints via introducing additional checks in the factor graph. Meanwhile, we transform the maximization problem into estimating the marginal distribution of the utility function. Next, we propose a distributed BP algorithm to solve the estimation problem on the developed factor graph. Complexity analysis suggests that the complexity of the proposed BP algorithm is low and practical for implementation. Simulation results show a great improvement in system throughput compared with the benchmark scheme.

In the remainder of this paper, we describe the system model in Section II. Section III introduces the factor graph model and the BP framework for this problem. The BP algorithm is addressed in Section IV. Section V gives the numerical results highlighting the benefits of this BP method compared with several other schemes. Finally, concluding remarks are drawn in Section VI.

II. System Model and Problem Formulation

We consider the downlink of an HCN with LTE specifications, where the transmission time is divided into multiple transmission time intervals (TTIs), and the transmission bandwidth is divided into multiple subchannels. Each TTI is composed of two time slots of

Manuscript received December 18, 2014; revised April 17, 2015; accepted June 11, 2015. Date of publication July 6, 2015; date of current version July 14, 2016. This work was supported in part by Australian Research Council Programs *DP*120100405; by the National 973 Project 2012*CB*316106; by the National Natural Science Foundation of China under Grant 61328101, Grant 61271230, and Grant 61472190; by the STCSM Science and Technology Innovation Program 13510711200; and by the SEU National Key Lab on Mobile Communications 2013*D*11 and 2013*D*02. The review of this paper was coordinated by Prof. Y.-B. Lin.

Y. Chen, Z. Lin, and B. Vucetic are with the School of Electrical and Information Engineering, University of Sydney, Sydney, NSW 2006, Australia (e-mail: youjia.chen@sydney.edu.au; zihuai.lin@sydney.edu.au; branka.vucetic@sydney.edu.au).

J. Li is with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: jleesr80@gmail.com).

W. Chen is with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: wenchen@sjtu.edu.cn).

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

Digital Object Identifier 10.1109/TVT.2015.2452953

0.5 ms. Each subchannel consists of 12 consecutive subcarriers, where a subcarrier has the equal bandwidth of 15 kHz. A radio resource spanning over one time slot and one subchannel is called a resource block (RB), which can be assigned to a mobile user (MU) for data transmission [12]. In the following, we have three basic assumptions in our system model.

1) Intracell interference is successfully avoided, since the scheduling of a BS will ensure that one subchannel will not be simultaneously assigned to more than one MU.
2) There is a maximum transmission power for each BS. This maximum transmission power in different BSs can vary, whereas MBSs usually transmit at stronger power than SBSs.
3) In each TTI, a BS serves (associates with) several MUs simultaneously among those MUs who can be potentially served by this BS (i.e., can receive strong enough signals from this BS). On the other hand, although an MU may have several potentially serving BSs, it can only be served by one BS in each TTI.

Suppose that the HCN consists of L BSs, including MBSs and SBSs, and M MUs covered by these BSs. Moreover, there are N subchannels available in the network. Let $\boldsymbol{\mathcal{B}} = \{\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_L\}$ denote the set of BSs, where $\mathcal{B}_l$, $l \in \boldsymbol{\mathcal{L}} \triangleq \{1, 2, \ldots, L\}$ represents the lth BS. Denote by $\boldsymbol{\mathcal{U}} = \{\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_M\}$ the set of MUs, where $\mathcal{U}_m$, $m \in \boldsymbol{\mathcal{M}} \triangleq \{1, 2, \ldots, M\}$ represents the mth MU. Let $\boldsymbol{\mathcal{S}} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_N\}$ denote the set of subchannels, where $\mathcal{S}_n$, $n \in \boldsymbol{\mathcal{N}} \triangleq \{1, 2, \ldots, N\}$ represents the nth subchannel.

A. User Association and Resource Allocation

For $\mathcal{B}_l$, we denote by $\boldsymbol{\mathcal{U}}(l)$ the set of MUs in its coverage, by $\boldsymbol{\mathcal{S}}(l)$ the set of its owned subchannels, and by P_l its maximum transmission power. Different BSs may have different sets of MUs and available subchannels. Furthermore, the maximum transmission power of different BSs can also be different. The maximum transmission power of MBSs is much larger than that of SBSs. During one scheduling period, the BS shall carefully choose the MUs to associate with and properly assign its resources, i.e., subchannels and power, to these selected MUs. In the following, we will mathematically formulate the user association and resource allocation in each BS. We unify the size of the resource allocation matrix in all BSs as $M \times N$.

1) User Association: Use the vector $\mathbf{k}^l \triangleq [k_1^l, k_2^l, \ldots, k_M^l]$ to denote the user association at $\mathcal{B}_l$ in one TTI. The entry $k_m^l \in \{0, 1\}$ indicates whether $\mathcal{U}_m$ is served by $\mathcal{B}_l$ or not. We have

$$k_m^l = \begin{cases} 1, & \text{if } \mathcal{U}_m \in \boldsymbol{\mathcal{U}}(l), \text{ and } \mathcal{U}_m \text{ is associated with } \mathcal{B}_l \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Since each MU only can be served by one BS in each TTI, we have $\sum_{l \in \boldsymbol{\mathcal{L}}} k_m^l = \{0, 1\}\, \forall\, m \in \boldsymbol{\mathcal{M}}$.

2) Subchannel Assignment: After the user association, each BS will assign its spectrum resource to its associated MUs. We use the matrix $\mathbf{Y}^l$ of size $M \times N$ to denote the subchannel assignment at $\mathcal{B}_l$. The entry $y_{m,n}^l \in \{0, 1\}$ in $\mathbf{Y}^l$ indicates whether $\mathcal{S}_n$ is assigned to $\mathcal{U}_m$, i.e.,

$$y_{m,n}^l = \begin{cases} 1, & \text{if } \mathcal{S}_n \in \boldsymbol{\mathcal{S}}(l), \text{ and } \mathcal{S}_n \text{ is assigned to } \mathcal{U}_m \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Due to the constraint that each subchannel can only be assigned to one MU at each BS, we have $\sum_{m \in \boldsymbol{\mathcal{M}}} y_{m,n}^l = \{0, 1\}\, \forall\, n \in \boldsymbol{\mathcal{N}}$.

3) Power Allocation: Following the subchannel assignment, each BS needs to decide the level of transmission power for an assigned subchannel. We denote by the vector $\mathbf{p}^l \triangleq [p_1^l, p_2^l, \ldots, p_N^l]$ the allocations of transmission power to different subchannels at $\mathcal{B}_l$, where p_n^l represents the transmission power allocated to $\mathcal{S}_n$. Due to the power constraint at $\mathcal{B}_l$, we have $\sum_{n \in \boldsymbol{\mathcal{N}}} p_n^l \le P_l$.

4) Integration: We use a matrix $\mathbf{X}^l$ of size $M \times N$, which is named the scheduling matrix, to integrate user selection, subchannel assignment, and power allocation at $\mathcal{B}_l$. The entry of $\mathbf{X}^l$, $x_{m,n}^l = k_m^l \times y_{m,n}^l \times p_n^l$, represents the transmission power allocated to $\mathcal{U}_m$ in subchannel $\mathcal{S}_n$.

B. System Sum Rate

Let matrix $\mathbf{G}$ of size $M \times N \times L$ denote the channel gains between MUs and BSs in different subchannels. The entry $g_{m,n,l}$ denotes the channel gain from $\mathcal{B}_l$ to $\mathcal{U}_m$ on $\mathcal{S}_n$ including the path loss, shadowing, and the antenna gain. The signal-to-interference-plus-noise ratio (SINR) experienced by $\mathcal{U}_m$ from $\mathcal{B}_l$ on $\mathcal{S}_n$ can be expressed as

$$\gamma_{m,n,l} = \frac{g_{m,n,l} \cdot x_{m,n}^l}{\sum_{v \in \boldsymbol{\mathcal{L}} \setminus l} \left(g_{m,n,v} \sum_{q \in \boldsymbol{\mathcal{M}}} x_{q,n}^v \right) + \sigma^2} \tag{3}$$

where σ^2 is the variance of Gaussian white noise at the receiver of each MU. The sum rate of the HCN can thus be calculated as

$$F = \sum_{l \in \boldsymbol{\mathcal{L}}} \sum_{m \in \boldsymbol{\mathcal{M}}} \sum_{n \in \boldsymbol{\mathcal{N}}} \times \log_2 \left(1 + \frac{g_{m,n,l} \cdot x_{m,n}^l}{\sum_{v \in \boldsymbol{\mathcal{L}} \setminus l} \left(g_{m,n,v} \sum_{q \in \boldsymbol{\mathcal{M}}} x_{q,n}^v \right) + \sigma^2} \right). \tag{4}$$

C. Problem Formulation

By defining $\mathbf{X} \triangleq [\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^L]$, we formulate the sum-rate optimization problem as

$$\max_{\mathbf{X}} \quad F(\mathbf{X}) \tag{5a}$$

$$\text{s.t.} \quad \left(\sum_{n \in \boldsymbol{\mathcal{N}}} x_{m,n}^l \right) \left(\sum_{n \in \boldsymbol{\mathcal{N}}} x_{m,n}^v \right) = 0 \tag{5b}$$

$$\forall\, m \in \boldsymbol{\mathcal{M}} \quad \forall\, l, v \in \boldsymbol{\mathcal{L}} : l \ne v \tag{5c}$$

$$x_{i,n}^l x_{j,n}^l = 0 \quad \forall\, i, j \in \boldsymbol{\mathcal{M}} : i \ne j \tag{5d}$$

$$\sum_{m \in \boldsymbol{\mathcal{M}}} \sum_{n \in \boldsymbol{\mathcal{N}}} x_{m,n}^l \le P_l \quad \forall\, l \in \boldsymbol{\mathcal{L}}. \tag{5e}$$

Constraints (5b), (5d), and (5e) correspond to user association, subchannel assignment, and power allocation, respectively. Specifically, (5b) represents that an MU can only associate with one BS, (5d) means that each subchannel of a BS can only be assigned to one MU, and (5e) ensures the power constraint at each BS.

III. Factor Graph Model

Here, we will develop a factor graph for our BP algorithm to represent the relationship between the resource allocations, user associations, and the system sum rate. Conventionally, a factor graph is composed of variable nodes and factor nodes. Variable nodes represent the variables (or parameters) to be optimized, and factor nodes represent the local utilities as the results of the decomposition of the objective function.

In our factor graph, as the optimization problem in (5) has constraints, to apply BP, we will introduce two kinds of factor nodes, namely, *primary factor nodes* and *auxiliary factor nodes*. The former is defined according to the decomposition of the objective F, and the latter is defined to coordinate the constraints in (5).

A. Variable Nodes

We define $\mathbf{X}^l$, $l = 1, \ldots, L$, in (5) as L variable nodes in the factor graph, representing the variables to be optimized in (5). The variable

$\mathbf{X}^l$ is related to the lth BS $\mathcal{B}_l$, since $\mathbf{X}^l$ needs to be optimized by $\mathcal{B}_l$. Note that the constraints in (5) on $\mathbf{X}^l$ should be satisfied during the BP optimization process. While (5d) and (5e) are intravariable constraints that are satisfied within each individual variable node $\mathbf{X}^l$, (5b) represents intervariable constraints that can only be satisfied with the help of coordination among variable nodes. The function of coordination is executed by the auxiliary factor nodes, which will be explained later.

B. Primary Factor Nodes

To obtain the factor nodes, we decompose the objective F into L local utility functions as $t^1, t^2, \ldots, t^L$, where

$$t^l = \sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}} \times \log_2\left(1+\frac{g_{m,n,l}\cdot x^l_{m,n}}{\sum_{v\in\mathcal{L}\setminus l}\left(g_{m,n,v}\sum_{q\in\mathcal{M}} x^v_{q,n}\right)+\sigma^2}\right) \quad (6)$$

and we have $F=\sum_{l\in\mathcal{L}} t^l$. To maximize F, we need to maximize each individual t^l at the lth BS. Hence, the local utilities t^l, $l=1,\ldots,L$, represent the factor nodes of our factor graph, with each factor node related to a BS. As t^l is related to the objective F, we call the factor nodes t^l primary factor nodes.

C. Auxiliary Factor Nodes

To meet (5b) in BP, we define a kind of auxiliary factor nodes to coordinate the variable nodes for the intervariable constraints. In a distributive manner, the MUs will act as the coordinators that exchange information with the BSs and regulate the optimizations on legitimate $\mathbf{X}^l$ satisfying (5b). We define M auxiliary factor nodes related to the M MUs, respectively, with the utility function at the mth auxiliary factor node as

$$c^m=\begin{cases}0, & \text{if } \left(\sum_{n\in\mathcal{N}} x^l_{m,n}\right)\left(\sum_{n\in\mathcal{N}} x^v_{m,n}\right)=0 \quad \forall\, l,v\in\mathcal{L}: l\neq v\\ -\infty, & \text{otherwise.}\end{cases} \quad (7)$$

The reason why the values of c^m are chosen as in (7) will be explained later in this section. The local utilities c^m, $m=1,\ldots,M$ represent the auxiliary factor nodes and, combined with the primary factor nodes, constitute the set of factor nodes.

D. Edges

The edges in the factor graph connect factor nodes to the related variable nodes. Recall that $\boldsymbol{\mathcal{U}}(l)$ denotes the set of MUs in the coverage of $\mathcal{B}_l$, and $\boldsymbol{\mathcal{S}}(l)$ denotes the set of subchannels owned by $\mathcal{B}_l$. The edges emanated from the factor node c^m connect to the variable node $\mathbf{X}^l$ if $\mathcal{U}_m\in\boldsymbol{\mathcal{U}}(l)$. For the factor node t^l, its edges consist of two parts. The first part includes the edges emanated from t^l to the variable node $\mathbf{X}^l$. This kind of edges exists since the calculation of t^l depends on $\mathbf{X}^l$ as the signal power. The second part includes the edges emanated from t^l to the variable node $\mathbf{X}^v$, $\forall\, v\in\mathcal{L}$ and $\boldsymbol{\mathcal{S}}(l)\bigcap\boldsymbol{\mathcal{S}}(v)\neq\emptyset$. That is, $\mathcal{B}_v$ and $\mathcal{B}_l$ share the same spectrum resource. This part of edges exists since t^l depends on $\mathbf{X}^v$ as the intercell interference. Fig. 1 shows a factor graph based on an HCN with $L=2$ BSs and $M=3$ MUs. Accordingly, the factor graph has two variable nodes, two primary factor nodes, and three auxiliary factor nodes.

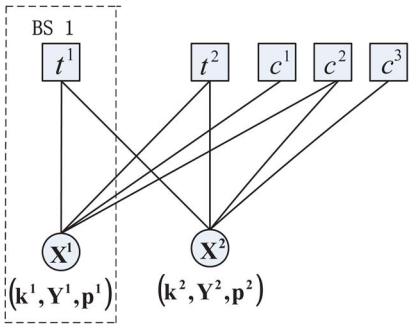

Fig. 1. Factor graph model.

E. Problem Conversion

With the factor graph, the network-wide optimization problem in (5) can be converted into

$$\max_{\mathbf{X}} \sum_{l\in\mathcal{L}} t^l + \sum_{m\in\mathcal{M}} c^m. \quad (8)$$

It is clear that $\sum_{l\in\mathcal{L}} t^l$ is the original objective F in (5), whereas the purpose of $\sum_{m\in\mathcal{M}} c^m$ is to satisfy the intervariable constraint (5b) coordinated by the auxiliary factor nodes, as previously discussed. We now explain why the values of c^m in (7) can guarantee the equivalence between optimization problems (5) and (8). We consider the following two cases: 1) If constraint (5b) is satisfied, F in (5) will be exactly (8); 2) if constraint (5b) is not satisfied, then we obtain a negatively infinite value of (8), which will be ruled out when maximizing (8). Eventually, the maximum of (8) is equivalent to the maximum of F in (5).

IV. Distributed Belief Propagation With Gaussian Approximations

Here, we will introduce a novel BP algorithm to distributively solve the optimization problem with low complexity based on the proposed factor graph model.

A. Transformation of the Local Optimization

We define a probability mass function (PMF) of the variable $\mathbf{X}$ based on the network-wide utility function as [13], i.e., $p(\mathbf{X})\triangleq(1/Z)\exp(\mu F(\mathbf{X}))$, where μ is a positive number, and Z is utilized to normalize this expression. According to [14], when $\mu\to\infty$, $p(\mathbf{X})$ concentrates around the maxima of $F(\mathbf{X})$, i.e., $\lim_{\mu\to\infty}\mathbb{E}(\mathbf{X})=\arg\max_{\mathbf{X}} F(\mathbf{X})$, where $\mathbb{E}(\cdot)$ denotes the expectation. Thus, once we obtain $\mathbb{E}(\mathbf{X})$, we will have a good estimation for the maximization of $F(\mathbf{X})$.

Based on (8), the maximization of F can be decomposed into multiple maximization problems of local utilities t^l and c^m at individual factor nodes. Correspondingly, the estimation on $\mathbf{X}$ is decomposed into estimations on $\mathbf{X}^u$ at t^l $\forall\,\mathbf{X}^u\in\boldsymbol{\mathcal{H}}(t^l)$, and $\mathbf{X}^v$ at c^m $\forall\,\mathbf{X}^v\in\boldsymbol{\mathcal{H}}(c^m)$.

B. Iterative Message Passing

In the classic BP algorithm, the PMF of $\mathbf{X}^l$ is the message updated between the variable node $\mathbf{X}^l$ and its neighboring factor nodes, i.e., $p(\mathbf{X}^l)=\{\Pr(\mathbf{X}^l=\boldsymbol{\chi}^l_i)\,\forall\, i\}$, where $\boldsymbol{\chi}^l_i$ denotes the ith possible value of $\mathbf{X}^l$. To simplify the notations in some occasions, we use a to denote the variable node $\mathbf{X}^l$ and use b to denote one of a's neighboring factor nodes, i.e., the local utility function t^l or c^m. The steps for our distributed BP are given as follows.

1) Initialization: At each variable node a, set an initial PMF for $\mathbf{X}^l$ $\forall\, l\in\mathcal{L}$, which can be a uniform distribution. That is, $p_a(\mathbf{X}^l)$.

2) Variable Node Update: Generally, in the BP algorithm, a message from a variable node a to a factor node b is the product of the messages sent from a's neighboring factor nodes except b [15]. Thus, in the τth iteration, the variable node a updates $p(\mathbf{X}^l)$ for the factor node b based on the messages sent from its neighboring factor nodes other than b, which is denoted by $\boldsymbol{\mathcal{H}}(a) \setminus b$. Then, we have

$$p_{a\to b}^{(\tau)}(\mathbf{X}^l) = \frac{1}{Z_a^{(\tau)}} \prod_{\bar{b}\in\boldsymbol{\mathcal{H}}(a)\setminus b} p_{\bar{b}\to a}^{(\tau-1)}(\mathbf{X}^l) \tag{9}$$

where $Z_a^{(\tau)}$ is for the purpose of normalization, i.e., $\sum_i \Pr_{a\to b}^{(\tau)}(\mathbf{X}^l = \boldsymbol{\chi}_i^l) = 1$.

Since in (9), b represents either a primary factor node t^l or an auxiliary factor node c^m, we denote by $\Phi_{a\to b}^{(\tau)}$ the message sent from a to t^l and by $\Psi_{a\to b}^{(\tau)}$ the message from a to c^m, in the τth iteration.

First, we focus on $\Phi_{a\to b}^{(\tau)}$. Since we adopt Gaussian approximation when calculating the intercell interference [13] to reduce the computational complexity at t^l, a does not need to transmit the PMF $p(\mathbf{X}^l)$ in (9) as in the general BP algorithm. Instead, a only needs to transmit the mean and variance of $\mathbf{X}^l$ based on the PMF $p(\mathbf{X}^l)$, i.e.,

$$\Phi_{a\to b}^{(\tau)}(\mathbf{X}^l) \triangleq \left\{\mathbb{E}_{a\to b}^{(\tau)}(\mathbf{X}^l)\mathbb{D}_{a\to b}^{(\tau)}(\mathbf{X}^l)\right\} \tag{10}$$

where $\mathbb{E}_{a\to b}^{(\tau)}(\mathbf{X}^l) = \sum_i \boldsymbol{\chi}_i^l \Pr_{a\to b}^{(\tau)}(\mathbf{X}^l = \boldsymbol{\chi}_i^l)$, and $\mathbb{D}_{a\to b}^{(\tau)}(\mathbf{X}^l) = \sum_i (\boldsymbol{\chi}_i^l - \mathbb{E}_{a\to b}^{(\tau)}(\mathbf{X}^l))^2$.

Then, we focus on $\Psi_{a\to b}^{(\tau)}$. The message sent by a to an auxiliary factor node c^m is the probability that $\mathcal{U}_m$ is selected by $\mathcal{B}_l$, i.e.,

$$\Psi_{a\to b}^{(\tau)} \triangleq \sum_i \Pr_{a\to b}^{(\tau)}\left(\mathbf{X}^l = \boldsymbol{\chi}_i^l\right) \cdot \mathbf{1}\left(\left(\boldsymbol{\chi}_i^l\right)_m \neq \mathbf{0}\right) \tag{11}$$

where $\left(\boldsymbol{\chi}_i^l\right)_m$ denotes the mth row of $\boldsymbol{\chi}_i^l$, $\mathbf{1}(\cdot)$ is the indicator function, and $\left(\boldsymbol{\chi}_i^l\right)_m \neq \mathbf{0}$ represents the event that $\mathcal{U}_m$ is selected by $\mathcal{B}_l$.

3) Factor Node Update: Generally, in the BP algorithm, a message from a factor node b to a variable node a is the product of the local function at b with the messages sent from b's neighboring variable nodes except a, marginalized over these variables [15]. That is

$$\begin{aligned} p_{b\to a}^{(\tau)}(\mathbf{X}^l) &= \sum_{\sim\mathbf{X}^l} \left(\exp\left(\mu b\left(\boldsymbol{\mathcal{H}}(b)\right)\right) \prod_{\bar{a}\in\boldsymbol{\mathcal{H}}(b)\setminus a} p_{\bar{a}\to b}^{(\tau)}\right) \\ &= \mathbb{E}_{\sim\mathbf{X}^l}\left(\exp\left(\mu b\left(\boldsymbol{\mathcal{H}}(b)\right)\right)\right) \end{aligned} \tag{12}$$

where $\sum_{\sim\mathbf{X}^l}$ denotes the summary over all other variables except $\mathbf{X}^l$. Note that in (12), the first equation calculates the marginal distribution of variable $\mathbf{X}^l$, which is equivalent to the expectation on all the variables other than $\mathbf{X}^l$ in the second equation. Furthermore, $b(\boldsymbol{\mathcal{H}}(b))$ in (12) represents the local function of its neighboring variables $\boldsymbol{\mathcal{H}}(b)$. We have

$$\begin{aligned} p_{b\to a}^{(\tau)}(\mathbf{X}^l) = \Big\{ &\Pr_{b\to a}^{(\tau)}\left(\mathbf{X}^l = \boldsymbol{\chi}_i^l\right) = \mathbb{E} \\ &\times \left(\exp\left(\mu b\left(\boldsymbol{\mathcal{H}}(b)|_{\mathbf{X}^l=\boldsymbol{\chi}_i^l}\right)\right)\right) \quad \forall\, i \Big\}. \end{aligned} \tag{13}$$

For the primary factor node b, e.g., t^l, as mentioned, we adopt Gaussian approximation on the intercell interference to reduce the complexity. Denote by $\mathbf{Z}^l$ the $M \times N$ intercell interference matrix, and we have $\mathbf{Z}^l \sim \mathcal{N}(\mathbb{E}(\mathbf{Z}^l), \mathbb{D}(\mathbf{Z}^l))$, where the mean $\mathbb{E}(\mathbf{Z}^l)$ and variance $\mathbb{D}(\mathbf{Z}^l)$ can be obtained based on $\mathbb{E}_{a\to b}^{(\tau)}(\mathbf{X}^l)$ and $\mathbb{D}_{a\to b}^{(\tau)}(\mathbf{X}^l)$, respectively, sent from the variable nodes. According to [13], the PMF $p_{b\to a}^{(\tau)}(\mathbf{X}^l = \boldsymbol{\chi}_i^l)$ can be updated with the Gaussian distributed interference from (12).

For the auxiliary factor nodes b, e.g., c^m, the message sent back to the variable node a is the PMF of $\mathbf{X}^l$. Given $\mathbf{X}^l = \boldsymbol{\chi}_i^l$, we have its probability based on (12) as

$$\begin{aligned} &p_{b\to a}^{(\tau)}\left(\mathbf{X}^l = \boldsymbol{\chi}_i^l\right) \\ &= \begin{cases} \prod_{\bar{a}\in\boldsymbol{\mathcal{H}}(b)\setminus a}\left(1-\Psi_{\bar{a}\to b}^{(\tau)}\right), & \text{if } \left(\boldsymbol{\chi}_i^l\right)_m \neq \mathbf{0} \\ \sum_{\bar{a}\in\boldsymbol{\mathcal{H}}(b)\setminus a} \Psi_{\bar{a}\to b}^{(\tau)} \prod_{\check{a}\in\boldsymbol{\mathcal{H}}(b)\setminus\{a,\bar{a}\}}\left(1-\Psi_{\check{a}\to b}^{(\tau)}\right), & \text{otherwise.} \end{cases} \end{aligned} \tag{14}$$

4) Final Decision: Suppose there are T iterations in the distributed BP algorithm. After T iterations, the PMF in variable node a can be calculated as

$$\Pr\left(\mathbf{X}^l = \boldsymbol{\chi}_i^l\right) = \frac{1}{Z^{(T)}} \prod_{b\in\boldsymbol{\mathcal{H}}(a)} p_{b\to a}^{(T)}\left(\mathbf{X}^l = \boldsymbol{\chi}_i^l\right) \tag{15}$$

which is the probability that $\mathcal{B}_l$ chooses the scheduling matrix $\boldsymbol{\chi}_i^l$. Based on (15), the scheduling decision can be made by $\boldsymbol{\mathcal{B}}_l$ by choosing the scheduling option with the maximum posterior probability. That is, $\boldsymbol{\chi}_{\hat{i}}^l$ is selected, where $\hat{i} = \arg\max_i \Pr(\mathbf{X}^l = \boldsymbol{\chi}_i^l)$.

C. Communication Complexity

Note that the L variable nodes are related to L BSs, the L primary factor nodes are related to the BSs, and the M auxiliary factor nodes are related to the M MUs. The belief messages between the variables and primary factor nodes are exchanged among the BSs or inside each BS. Specifically, in each iteration, a variable node $\mathbf{X}^l$ related to $\mathcal{B}_l$ needs to transmit two real numbers, e.g., the mean and variance of $\mathbf{X}^l$ according to (10), to each of its neighboring factor nodes. On the other hand, a primary factor node t^l related to $\mathcal{B}_l$ needs to transmit the whole PMF $p(\mathbf{X}^l)$ to each of its neighboring variable nodes, which includes multiple probability values and may incur high communication complexity. Since the BSs can communicate with each other via backhaul channels, the communication complexity among BSs will be tolerable.

At the same time, the messages between the variable nodes and auxiliary factor nodes are exchanged between the BSs and MUs. Specifically, in each iteration, a variable node $\mathbf{X}^l$ related to $\mathcal{B}_l$ needs to transmit one real number to each of its neighboring factor nodes, according to (11). On the other hand, an auxiliary factor node c^m related to $\mathcal{U}_m$ needs to transmit one real number to each of its neighboring variable nodes, according to (14). Therefore, the communication complexity between the BSs and MUs is low and reasonable in practice.

D. Computational Complexity

We denote by H_v the average number of neighboring factor nodes of a variable node and by H_f the average number of a factor node's neighboring variable nodes. Moreover, we denote by W the average number of the scheduling options in variable nodes.

1) Computational Complexity in Variable Nodes: In each iteration, a variable node should first calculate its PMF in (9), then use the PMF to calculate its expectation and variance in (10), or only the expectation in (11). Thus, in one iteration, the computational complexity in a variable node is equal to $O(WH_v^2)$.

2) Computational Complexity in Primary Factor Nodes: In (12), a primary factor node first calculates the expectation and variance of the interference, whose computational complexity is $O(H_f)$. Combining with W signal options, the computational complexity for each message is $O(WH_f)$. Thus, the total computational complexity in one iteration is $O(WH_f^2)$.

3) Computational Complexity in Auxiliary Factor Nodes: According to (14), the total computational complexity in one iteration is $O(WH_f^2)$.

TABLE I
PARAMETER TABLE FROM 3GPP SPECIFICATION [16]

Parameter	Macro Cell	Pico-Cell
Maximum Transmission Power	43dBm	33dBm
Sub-channel Bandwidth	180kHz	
Carrier Frequency	2GHz	
Noise Power	-110dBm	
Path Loss (d is distance in meters)	128.1+37.6lg(d/1000)	140.7+36.7lg(d/1000)
Shadowing Standard Deviation	8dB	10dB
SINR Threshold (δ)	-3dB	

E. Convergence and Optimality

The results of the BP algorithm operating in a cycle-free factor graph are proved to converge to the exact marginal. Unfortunately, the precise conditions under which, in the cycle case, the BP algorithm will converge are still not well understood. In most simulation instances, we observe that the proposed BP algorithm converges to a fixed point. Regarding optimality, as we are solving an NP-hard problem, optimality cannot be guaranteed. However, the performance of the BP algorithm is empirically shown to approach optimality with a very small gap and with relatively low complexity.

V. SIMULATIONS

We consider an HCN with a coverage area of 700 $\times$ 700 square meters. In the HCN, an MBS is located in the center, several SBSs are uniformly deployed in the cell edges for enhancing the signals of the weak-coverage area, and 20 MUs are uniformly distributed in the coverage area. Here, we assume that the SBSs are picocell BSs (PBSs). The detailed parameters are listed in Table I according to Third-Generation Partnership Project (3GPP) specifications [16]. The number of subchannels is set as two. Moreover, we have three power levels as $(1/3)P_l$, $(2/3)P_l$, and P_l for power allocations in a specific subchannel. All the results are obtained by 200 independent Monte Carlo simulations. Empirically, the average number of iterations needed in our simulation scenario is around three. Thus, we chose $T = 3$ in our simulation. Furthermore, $T = 5$ is adopted for comparison.

We denote by the "BP" scheme our distributed BP algorithm in all the figures, where we set the maximum number of iterations to five. For comparison, we set up a benchmark scheme for joint user association and resource allocations, using the traditional max-SINR criterion as the user association strategy and opportunistically allocating resources (power and subchannels) for the MUs. Explicitly, each MU chooses the BS that provides the strongest received SINR as its serving BS, and each BS serves the MU who has the best channel quality in each subchannel. We name our benchmark "MaxSINR+Oppo" in the figures.

Fig. 2 shows cumulative distribution function (cdf) curves of the system throughput (sum rate) with different schemes, where there are four PBSs. We can see that the BP algorithm triples the system throughput obtained by "MaxSINR+Oppo." At the same time, we can see that the results of the BP algorithm with two iterations approach the results with five iterations. This means that our BP algorithm almost converges after only two iterations. In simulations, the average iteration number needed to converge is less than three.

Moreover, in the figure, we can see the performance gap between the proposed distributed BP algorithm and the global optimality achieved by the centralized exhaustive search algorithm. This performance loss is caused by the suboptimal distributed BP algorithm and the Gaussian approximation of the intercell interference. However, the computational complexity of exhaustive search is W^L, where W denotes the average number of scheduling options in each BS, and L denotes the number of BSs in the network. Thus, compared with the optimal search, the proposed distributed BP algorithm is more practical.

Then, we consider three scenarios where the numbers of PBSs deployed in the area are set to be 2, 4, and 6. Fig. 3 plots the system

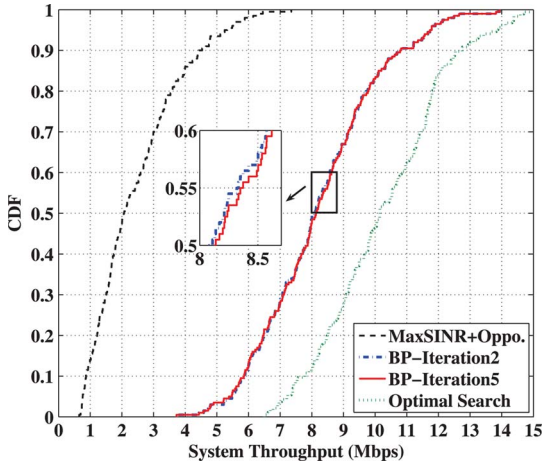

Fig. 2. CDF curves of the system throughput for the four schemes.

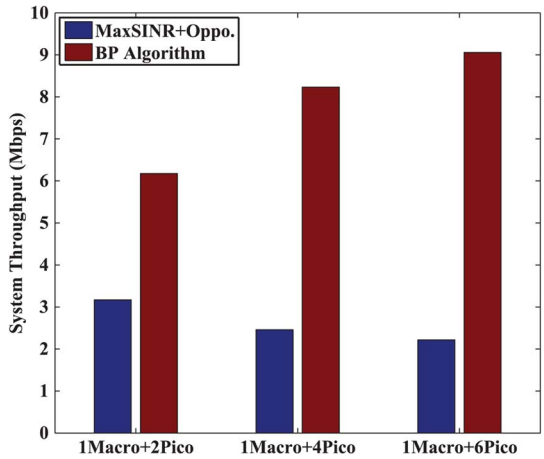

Fig. 3. Average system throughput of different schemes with different numbers of SBSs.

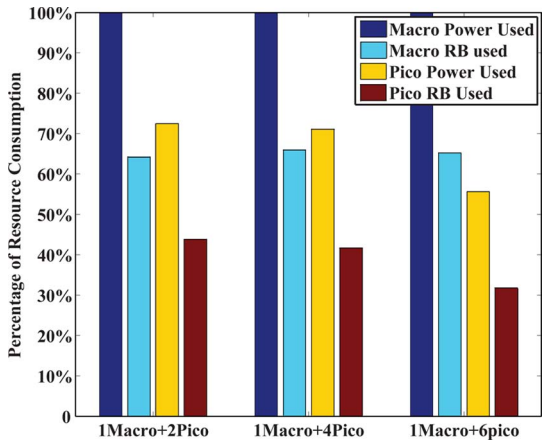

Fig. 4. Percentage of resource usage under different schemes with different numbers of SBSs.

TABLE II
AVERAGE THROUGHPUT OF BSS IN DIFFERENT SCENARIOS

Cells	1 Macro with 2 Picos		1 Macro with 4 Picos		1 Macro with 6 Picos	
Schemes	MaxSINR+Oppo.	BP Algorithm	MaxSINR+Oppo.	BP Algorithm	MaxSINR+Oppo.	BP Algorithm
Macro Cell	2.9683 Mbps	3.6135 Mbps	2.1550 Mbps	3.3288 Mbps	1.5816 Mbps	3.1401 Mbps
Pico Cells	0.2021 Mbps	2.5632 Mbps	0.3023 Mbps	4.9030 Mbps	0.6347 Mbps	5.9163 Mbps

throughput in these three different scenarios. We can see that with the increase in the number of PBSs, the intercell interference in the heterogeneous network is more severe. Hence, without proper ICIC, the throughput of "MaxSINR+Oppo" decreases due to the impact of severe intercell interference. In contrast, the throughput of our BP algorithm dramatically increases, since it controls the interference via excellent user association and resource allocations and makes better use of the resources from the PBSs.

We now investigate the resource usage in the three given scenarios. Fig. 4 shows the usage percentage of transmission power and subchannels (denoted by "RB" in the figure) in the three scenarios. We can see that the resource used by the MBS remains the same despite the number increase of the PBSs. Among all the BSs, the MBS contributes the most to the system throughput. Hence, the MBS makes the best use of the resources although it causes interference to other PBSs. Furthermore, when the number of PBSs increases, our BP will reduce PBSs' transmission power and used spectrum resources to mitigate the interference to other BSs.

Table II presents the detailed average throughput of MBS and PBSs in the three scenarios. We can see that the throughput provided by the MBS dramatically decreases when the number of interfering PBSs increases in "MaxSINR+Oppo," while it is almost kept constant in our "BP." Furthermore, the throughput of PBSs rapidly increases when there are more PBSs. This is because our BP algorithm has good management of intercell interference by jointly optimizing subchannel assignment and power allocation.

VI. CONCLUSION

In this paper, we have considered user association as a flexible parameter and involved the user association step into the resource allocation scheme. With the aim of maximizing the system throughput, an optimization problem of joint user association, subchannel assignment, and power allocation is formulated. First, a factor graph model is developed to decompose this network-wide maximization problem into multiple local maximization problems performed in each BS and translate the constraints to maximization problems performed in each MU. Then, based on this model, the distributed BP algorithm is proposed to solve this optimization problem by transforming it into a marginal distribution estimation problem. From the numerical results, we can see that the distributed BP algorithm jointly optimizing three factors performs much better than the scheme, which considers user association and resource allocation as two independent steps. Moreover, compared with the benchmark scheme, the distributed BP algorithm shows excellent ICIC performance, particularly when there is severe intercell interference.

REFERENCES

[1] F. Boccardi, R. Heath, A. Lozano, T. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.
[2] B. Soret, H. Wang, K. Pedersen, and C. Rosa, "Multicell cooperation for LTE-advanced heterogeneous network scenarios," *IEEE Wireless Commun.*, vol. 20, no. 1, pp. 27–34, Feb. 2013.
[3] T. D. Novlan, R. K. Ganti, A. Ghosh, and J. G. Andrews, "Analytical evaluation of fractional frequency reuse for OFDMA cellular networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 12, pp. 4294–4305, Dec. 2011.
[4] N. Saquib, E. Hossain, and D. I. Kim, "Fractional frequency reuse for interference management in LTE-advanced HetNets," *IEEE Wireless Commun.*, vol. 20, no. 2, pp. 113–122, Apr. 2013.
[5] Q. Li, R. Hu, Y. Xu, and Y. Qian, "Optimal fractional frequency reuse and power control in the heterogeneous wireless networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2658–2668, Jun. 2013.
[6] P. Bhat *et al.*, "LTE-advanced: An operator perspective," *IEEE Commun. Mag.*, vol. 50, no. 2, pp. 104–114, Feb. 2012.
[7] D. Lopez-Perez, X. Chu, and J. Zhang, "Dynamic downlink frequency and power allocation in OFDMA cellular networks," *IEEE Trans. Commun.*, vol. 60, no. 10, pp. 2904–2914, Oct. 2012.
[8] L. P. Qian, Y. J. Zhang, Y. Wu, and J. Chen, "Joint base station association and power control via Benders' Decomposition," *IEEE Trans. Wireless Commun.*, vol. 12, no. 4, pp. 1651–1665, Apr. 2013.
[9] Q. Li, R. Hu, Y. Qian, and G. Wu, "Intracell cooperation and resource allocation in a heterogeneous network with relays," *IEEE Trans. Veh. Technol.*, vol. 62, no. 4, pp. 1770–1784, May 2013.
[10] J. Ghimire and C. Rosenberg, "Resource allocation, transmission coordination and user association in heterogeneous networks: A flow-based unified approach," *IEEE Trans. Wireless Commun.*, vol. 12, no. 3, pp. 1340–1351, Mar. 2013.
[11] D. Fooladivanda and C. Rosenberg, "Joint resource allocation and user association for heterogeneous wireless cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 248–257, Jan. 2013.
[12] "Physical layer aspect for evolved Universal Terrestrial Radio Access (UTRA)," Third-Generation Partnership Project, Sophia Antipolis Cedex, France, Tech. Rep. v.7.0.0, Jun. 2006.
[13] S. Rangan and R. Madan, "Belief propagation methods for intercell interference coordination in femtocell networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 631–640, Apr. 2012.
[14] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. New York, NY, USA: Springer-Verlag, 1998.
[15] F. Kschischang, B. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 498–519, Feb. 2001.
[16] "Further advancements for E-UTRA physical layer aspects," Third-Generation Partnership Project, Sophia Antipolis Cedex, France, Tech. Rep. v.9.0.0, Mar. 2010.