Joint User Association and Resource Allocation for Interference Mitigation and Load Balancing in HCNs

Haiqiang Liu, Zhenxiang Gao, Xulong Shao, Xu Shan, and Weihua Zhou
Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
Email: {liuhaiqiang, gaozhenxiang, shaoxulong, shanxu, zhouweihua}@iie.ac.cn

Abstract —The deployment of Heterogeneous Cellular Networks (HCNs) can acquire high system capacity and spectrum efficiency through spatial reuse, compared with the conventional cellular networks. Nevertheless, the co-channel deployment of HCNs and ultra-high connection density bring severe downlink and uplink interference. In addition, the traditional user association schemes, which mainly take downlink power into account, may result in an imbalanced distribution of load because of the massive disparities in cell sizes in HCNs. To solve the aforementioned challenges, we present a QoS-aware user association scheme with resource allocation for load balancing and inference mitigation, which includes both uplink and downlink channels, and formulate it as a network-wide utility maximization problem. A low-complexity distributed algorithm is also proposed via the dual decomposition method to obtain a near-optimal solution. Numerical results demonstrate that compared with the traditional association schemes, the proposed association scheme has some significant advantages in the mass.

Index Terms—Heterogeneous cellular networks, user association, resource allocation, load balancing, interference mitigation, dual decomposition

I. Introduction

Heterogeneous Cellular Network (HCN) is considered as one promising solution to deal with the explosive demands for mobile data traffic mostly driven by the vigorous growth in the population of smart phones and the continuously enriched types of service [1]. In an HCN, the Macro Base Station (MBS) is overlaid with various types of Small Base Stations (SBS), Pico Base Stations (PBS) for example. A higher system capacity and spectrum efficiency can be obtained by the deployment of HCNs [2].

However, there are still many remaining challenges to be addressed, which may lead to a significant decrease in system performance gain. Interference management and user association are two fundamental issues in HCNs. Because of the co-channel deployment of HCNs and ultra-high connection density, there exist severe downlink and uplink interference in HCNs. Besides, owing to the diversity of different tiers of BSs in HCNs, a number of user association issues need to be settled, such as load balancing and Quality of Service (QoS) requirement.

Different metrics have been adopted for traditional user association schemes in cellular networks, such as energy efficiency, QoS, fairness, coverage probability [3]. Because of the co-channel deployment, there is another critical issue, severe interference, in HCNs, so intelligent user association, resource allocation, and interference management schemes are needed in HCNs. Some literature adopts eICIC [4] in user association as an effective interference management method. eICIC is a pretty good way to mitigate the downlink interference. However, according to Shannon's law, uplink interference is also an important factor which can affect the user association scheme through influencing the capacity of the BSs.

In this paper, we concentrate on the joint user association and resource allocation problem, and aim at mitigating interference and achieving a relatively balanced distribution of cell load. The strategy for interference mitigation comprises two parts, uplink and downlink. eICIC, as a resource allocation scheme, is applied to mitigate the downlink interference. In the uplink, uplink power is considered in the user association problem to mitigate the interference. Load balancing is also taken into account in user association problem. We formulate it as a sum-utility maximization problem. Then a low-complexity distributed algorithm is proposed to solve the NP-hard problem. The performance of the proposed association scheme is validated by our numerical results.

The remaining of this paper is organized as follows: Section II describes related work on user association and interference management in HCNs. Section III details the system model and formulates the problem as a utility maximizing problem. In section IV, the distributed algorithm is presented. The simulation results are showed in section V. Finally, section VI draws the conclusion.

II. Related Work

In this part, we will first introduce the user association schemes. Then some resource allocation methods will be presented. Finally, the literature combining the user association and resource allocation will be described.

User association has been studied extensively in HCNs. Range Expansion Association (REA) is proposed to offload users from the MBS to the SBS by setting

Manuscript received May 11, 2016; revised November 21, 2016.
This work was supported by the National High-Tech Research and Development Program under Grant No.2015AA01A706
Corresponding author email: zhouweihua@iie.ac.cn.
doi:10.12720/jcm.11.11.977-983

different bias values for different tiers of BSs when comparing the received signal strength on the user side [5]. Nevertheless, the optimal bias value is very difficult to obtain and sensitive to the spatio-temporal distribution of users [6]. Therefore, some other load-aware schemes were put forward to avoid finding the optimal bias. Based on traffic transfer, Q. Ye [7] considered user association and resource allocation jointly and proposed a distributed algorithm via dual decomposition for downlink HCNs.

Resource allocation is always a hot topic in the wireless communication system, and it also applies to HCNs. Many resource allocation schemes have been proposed to manage the interference in HCNs. eICIC which leaves certain macro-cell subframes blank is proposed to mitigate the downlink inference [4]. eICIC is a time-domain multiplexing technique for improving the performance of co-channel HCNs. eICIC can offer many benefits, including a more equitable distribution of users among all the BSs, consequently leading to better SBS utilizations. Y. Peng [8] explored the interference of macro-pico scenario and introduced corresponding interference avoidance and mitigation solutions, including time, frequency, and power domains. Besides, the possible solutions with performance analysis were also looked into.

In recent years, an increasing number of studies that joint user association and resource allocation can be found. Y. Chen [9] proposed a novel belief propagation (BP) algorithm to jointly optimize user association, subchannel assignment, and power allocation. Q. Kuang [10] proposed a unified framework to analyze and compare a wide range of user association and resource allocation strategies for HCNs, and provided an optimal benchmark for network performance. Y. Jia [11] formulated the user association and ABS ratio issues jointly as a network-wide max-min fairness optimization problem, and solved the problem by decoupling the problem and resolving the two subproblems one by one.

III. System Model and Problem Formulation

Without loss of generality, a two-tier HCN is considered, which consists of one MBS and several PBSs. Let $M=\{1\}$ denote the only MBS, which is overlaid with several PBSs that are denoted by $P=\{1,2,...,P\}$. BSs in the HCN can be represented by $B=M\cup P$. Users are represented by $U=\{1,2,\ldots,U\}$. Aforementioned HCN is illustrated in Fig. 1.

The interference in HCN is categorized in two types, downlink (transmission from the BS to the user) and uplink (transmission from the user to the BS) interference. To mitigate the interference in the downlink channel, eICIC, as a resource allocation scheme, is introduced into the paper. For the interference in the uplink channel, uplink power is taken into consideration in the user association scheme.

A. Resource Allocation for Downlink Interference Mitigation

Many resource allocation schemes are introduced to deal with the strong interference in HCNs. In this paper, eICIC which allocates time domain resources is employed. ABS is introduced to HCNs, which only contains reference signals, no data signals. Once ABS is configured by the MBS, the users served by PBSs in the subframe are only interfered by other PBSs, no MBSs. Fig. 2 demonstrates the ABS.

Let $S=\{S_a,S_n\}$ denote all of the subframes, where S_a denotes ABS subframes, and S_n denotes normal subframes. Let γ denote the fraction of ABS, i.e. $|S_a|=\gamma|S|$, where $|S_a|$ and $|S|$ denote the number of ABSs and whole subframes respectively. We assume no intra-cell interference exist, which is easy to be achieved by using multiple access techniques or other techniques. So if user i is associated with BS j in subframe s, the SINR of the user's side can be derived as follows:

$$SINR_{ijs}=\begin{cases}\dfrac{P_{js}G_{ijs}}{\sum_{n\in P,n\neq j}P_{ns}G_{ins}+\sigma_{js}^{\ 2}} & s=S_a,\ j\in P\\ \dfrac{P_{js}G_{ijs}}{\sum_{n\in B,n\neq j}P_{ns}G_{ins}+\sigma_{js}^{\ 2}} & s=S_a,\ j\in M\\ \dfrac{P_{js}G_{ijs}}{\sum_{n\in B,n\neq j}P_{ns}G_{ins}+\sigma_{js}^{\ 2}} & s=S_n,\ j\in P\\ \dfrac{P_{js}G_{ijs}}{\sum_{n\in B,n\neq j}P_{ns}G_{ins}+\sigma_{js}^{\ 2}} & s=S_n,\ j\in M\end{cases} \quad (1)$$

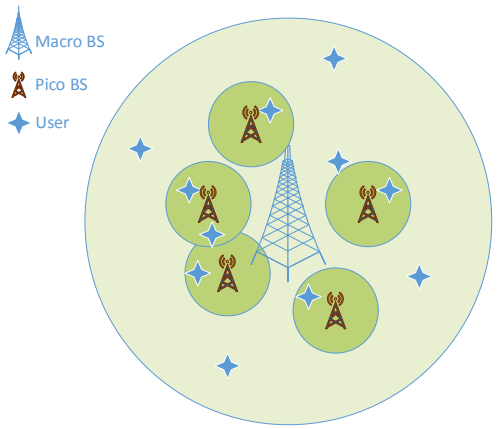

Fig. 1. Illustration of a two-tier HCN

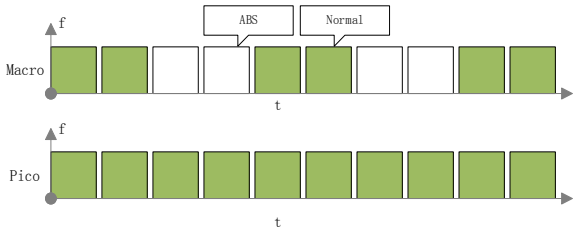

Fig. 2. eICIC technology: ABS

where P_{js} denotes the power BS j uses to transmit data in subframe s. G_{ijs} denotes the channel gain from BS j to user i in subframe s, which includes path loss, antenna gain and shadowing gain. σ^2 is the noise power. As applying some methods, time-averaging [12] for example, the impact of the Rayleigh fading can be eliminated. So in this paper, Rayleigh fading is out of our consideration

According to Shannon's law, it is easy to get the achievable instantaneous rate:

$$R_{ijs} = \begin{cases} \alpha & s = S_a,\ j \in M \\ \log\left(1 + SINR_{ijs}\right) & otherwise \end{cases} \tag{2}$$

when user i is associated with the MBS in subframe S_a, the rate of the user should be 0, but for distributed algorithm proposed below, we just set the user's rate as α, which is a very small constant number.

We assume QoS refers to the minimum required rate. r_i is used to denote the rate demand of user i. Then the resource required can be derived by

$$c_{ijs} = \begin{cases} 0 & s = S_a,\ j \in M \\ r_i / R_{ijs} & otherwise \end{cases} \tag{3}$$

as MB do not transmit data signals in ABS, the amount of consumed resource of ABS of MB is set to be 0.

Considering the load, we use some definitions:

Definition 1: the load of BS j in subframes s is the resource consumed by users associated with BS j in subframe s, i.e. $L_{js} = \sum_i a_{ijs} c_{ijs}$, a_{ijs} is an indicator that implies whether user i is allocated subframe s of BS j, i.e. $a_{ijs} = 1$ indicates that subframe s of BS j is allocated to user i, 0 versus.

Definition 2: the efficient rate of user i is $e_{ijs} = R_{ijs} / L_{js}$ when it is allocated subframe s of BS j, whose load is L_{js}.

B. Uplink Inference-Aware User Association

As is generally known, the interference is proportional to the transmission power. So taking uplink power into consideration in user association scheme is an intuitive way to mitigate the interference in the uplink channel. As we can see from [13], the uplink power can be calculated by

$$UP_{ij} = \min\left\{ 10^{\frac{P_i^{\max}}{10}}, \tau * \sigma^2 / 10^{\frac{-PL_{ij}}{10}} \right\} \tag{4}$$

$P_i^{\max}$ denotes the max power user i can transmit which is set to 23dBm, τ represents the threshold signal to noise ratio which is set to 10dB, and PL_{ij} denotes the path loss between user i and BS j.

Taking uplink power into account, our problem can be formulated as below:

$$\begin{aligned} \max_{\boldsymbol{a}} \quad & \sum_i \sum_j \sum_s a_{ijs} c_{ijs} U_{ijs}\left(\frac{e_{ijs}}{UP_{ij}}\right) \\ s.t. \quad & a_{ijs} \in \{0,1\}\ \forall i \in U, \forall j \in B, s \in S \\ & \sum_i a_{ijs} c_{ijs} \le M_{js} \quad \forall j \in B, s \in S \\ & \sum_j \sum_s a_{ijs} = 1 \quad \forall i \in U \end{aligned} \tag{5}$$

$\boldsymbol{a} = \{a_{ijs}, i \in U, j \in B, s \in S\}$ is the indicator set. $U_{ijs}\left(e_{ijs} / UP_{ij}\right)$ denotes the utility user i provides to the whole network when user i is allocated subframe s of BS j. when the utility function U_{ijs} is a strictly monotone increasing function, the utility has a positive correlation with e_{ijs} which reflects the load balancing effect, and a negative correlation with UP_{ij} which reflects the interference. The utility function simply reveals our goal which is achieving a good load balancing effect and low interference in the meanwhile. The objective function can be regarded as a maximizing weighted sum of user's utility in the network, where the weight is the corresponding amount of consumed resources. The first constraint denotes the association indicator. The second constraint demonstrates that the special kind of subframe s of BS j has a resource limit M_{js}. The last implies that each user should be served.

To model the user association problem, sigmoidal, exponential and logarithmic function are in the candidate list to present the utility. In this paper, we prefer logarithmic function $\log x$, as the logarithmic function is a concave function and has a diminishing return, which means an extra resource is preferred to allocate to users of low rate rather than users of high rate. Moreover, it has a good nature that it can easily convert complicated multiplication and division to simple addition and subtracting, i.e.

$\log(xy) = \log x + \log y$ and $\log x / y = \log x - \log y$.

As logarithmic utility function is applied to this paper, the problem can be formulated as follows:

$$\begin{aligned} \max_{\boldsymbol{a}} \quad & \sum_i \sum_j \sum_s a_{ijs} c_{ijs} \log\left(\frac{e_{ijs}}{UP_{ij}}\right) \\ s.t. \quad & a_{ijs} \in \{0,1\}\ \forall i \in U, \forall j \in B, s \in S \\ & \sum_i a_{ijs} c_{ijs} \le M_{js} \quad \forall j \in B, s \in S \\ & \sum_j \sum_s a_{ijs} = 1 \quad \forall i \in U \end{aligned} \tag{6}$$

By introducing the nature of logarithmic function and relaxing a_{ijs} from {0, 1} to (0, 1), the problem can be transformed as below:

$$\begin{aligned} \max_{\boldsymbol{a}} \quad & \sum_i \sum_j \sum_s a_{ijs} c_{ijs} d_{ijs} - \sum_j \sum_s L_{js} \log L_{js} \\ s.t. \quad & a_{ijs} \in (0,1) && \forall i \in U, \forall j \in B, s \in S \\ & L_{js} = \sum_i a_{ijs} c_{ijs} && \forall j \in B, s \in S \\ & L_{js} \le M_{js} && \forall j \in B, s \in S \\ & \sum_j \sum_s a_{ijs} = 1 && \forall i \in U \end{aligned} \quad (7)$$

where $d_{ijs} = \log R_{ijs} / UP_{ij}$ is a constant when user *i*, BS *j* and subframe *s* are all fixed. L_{js} is the load of BS *j* in subframe *s*. Above problem is a convex optimization problem [7].

IV. Distributed Algorithm

To solve the formulated problem proposed in III, a distributed algorithm is presented in this section. The objective function (7) is decomposed into two sections, one contains indicators a_{ijs}, the other one contains load L_{js}. In view of the special form of the objective function, we are inspired to use the dual decomposition method to eliminate the correlation between a_{ijs} and L_{js} in (7), Lagrange multiplier $\boldsymbol{\lambda}$ is introduced into the problem. Then the dual problem can be formulated as

$$\min_{\boldsymbol{\lambda}} \; D(\boldsymbol{\lambda}) = f(\boldsymbol{\lambda}) + g(\boldsymbol{\lambda}) \quad (8)$$

where

$$f(\boldsymbol{\lambda}) = \begin{cases} \max\limits_{\boldsymbol{a}} & \sum_i \sum_j \sum_s a_{ijs} c_{ijs} \left(d_{ijs} - \lambda_{js} \right) \\ s.t. & a_{ijs} \in (0,1) \; \forall i \in U, \forall j \in B, s \in S \\ & \sum_j \sum_s a_{ijs} = 1 \qquad \forall i \in U \end{cases} \quad (9)$$

and

$$g(\boldsymbol{\lambda}) = \max_{L_{js} \le M_{js}} \sum_j \sum_s L_{js} \left(\lambda_{js} - \log L_{js} \right) \quad (10)$$

As the primal problem is convex, so solving the dual problem is equivalent to solving the primal problem. $f(\boldsymbol{\lambda})$ and $g(\boldsymbol{\lambda})$ can be recognized as the user's side utility and the BS's side utility respectively. In subproblem $f(\boldsymbol{\lambda})$ and $g(\boldsymbol{\lambda})$, $\boldsymbol{\lambda}$ is seen as a constant. Once $\boldsymbol{\lambda}$ is fixed, the subproblem can be easily solved. To find the optimal $\boldsymbol{\lambda}$, we introduce three-step algorithm [14], which is an improved algorithm of gradient project algorithm. The basic idea of the algorithm is that the updating direction of *t*+2 iteration is opposed to the average of the subgradients of the t iteration and the t+1 iteration.

Eq. (9) can be interpreted as sum of the utility of every user. As the constraints in (9) implies that each user is supposed to associate with one and only one BS and subframe, so for each user, (9) can be simplified to

$$\max_{js} c_{ijs} \left(d_{ijs} - \lambda_{js} \right) \quad (11)$$

which means each user will choose BS *j* and subframe *s* that can bring the max $c_{ijs} \left(d_{ijs} - \lambda_{js} \right)$.

As the problem (10) is differential, we can apply KKT condition [15] on (10), which means taking a derivative with respect to L_{js}. The optimal load of subframe *s* of BS *j* can be derived as follows

$$L_{js} = \min \left(M_{js}, exp\left(\lambda_{js} - 1 \right) \right) \quad (12)$$

After getting the association status a_{ijs} and load status L_{js}, Lagrange multiplier $\boldsymbol{\lambda}$ can be updated by:

$$\lambda_{js}^{\,t+1} = \lambda_{js}^{\,t} - \delta^t \left(\frac{L_{js}^{\,t} - \sum_i a_{ijs}^{\,t} c_{ijs} + L_{js}^{\,t-1} - \sum_i a_{ijs}^{\,t-1} c_{ijs}}{2} \right) \quad (13)$$

where $\delta^t > 0$ is a sufficient small stepsize to update λ_{js}.

So the distributed algorithm is shown in Algorithm 1 (user side) and algorithm 2 (BS side):

Algorithm 1 user side

1 **if** t=0 **then**
2 Evaluate power consumption and rate
3 **else**
Receive λ_{js} broadcast by each BS
Connect to BS j^* according to (11).
If there are more than one optimal BS, just choose any one of them.
4 Feedback association information to BS j^*.
5 **end if**

Algorithm 2 BS side

1 **if** t=0 **then**
2 Set the initial λ_{js} and precise ξ.
3 **else**
4 Receive information a_{ijs}
Apply KKT condition to update L_{js} by (12).
5 According to the outcome of 4, update λ_{js} by (13).
6 Broadcast λ_{js} to all users
7 **end if**

After a few iterations, the algorithm 2 is guaranteed to converge [14].

Following is the complexity analysis for the proposed algorithm. As (13) implies, the update of λ_{js} only depends on the status of subframe *s* of BS *j*, no global information needed. So it is a completely distributed algorithm. At each iteration, at the user's side, each user needs to compare all of λ_{js}, so the complexity is $O(|N||S|)$, where $|N|$ and $|S|$ are the total number of all BSs and subframes respectively. At the BS side, each BS needs to calculate λ_{js} for all of its subframes, so the complexity is $O(|U||S|)$, where $|U|$ represents the number of all users.

V. Simulation Results

We consider a two-tier HCN consisting of one MBS and several PBSs. The location of the MBS is modelled

to be fixed according to traditional cellular network framework, while the PBSs are deployed according to a Poisson distribution based on the fixed location of the MBS. Users are scattered into each macro cell in a random way. The transmission power is 45 and 30 dBm for the MBS and PBS respectively. The path loss model for the MBS is represented as 34+40log(d), and the one for PBSs is represented as 37+30log(d), where d is the distance between the user and BS. The power density of the noise is assumed to be -104dBm/Hz. Besides, a log-normal shadowing with a standard deviation σ_s =8dB is considered. The above parameters are specified by reference to [7].

To highlight the performance of our algorithm, UARA (user association and resource allocation), we introduce two association schemes, i.e. Traditional Association scheme and REA scheme, which are explained as below:

Traditional Association scheme (TA): The utility is defined as the sum rate of users. To maximize the utility of the whole network, user $i \in U$ will be allocated subframe s^* of BS j^* which can bring it the maximal rate, i.e. $\{j^*, s^*\} = \max_{j \in B, s \in S} R_{ijs}$.

REA: user $i \in U$ will be allocated subframe s^* of BS j^* that meets the condition $\{j^*, s^*\} = \max_{j \in B, s \in S} \eta_j R_{ijs}$; η_j is the bias factor of BS j. Different bias factors are applied to different kinds of BSs. η_j =1 if BS j is the MBS, $\eta_j > 1$ if BS j is a PBS.

We can get a general idea of the features of the above schemes. TA is only rate-aware, while REA is rate-aware and partly load-aware, as users are more likely to associate with PBSs than TA by setting an unequal bias factor for different kinds of BSs. Moreover, the both above schemes are neither QoS-aware nor power-aware, whereas the proposed scheme is QoS-aware, load-aware and power-aware as seen from the objective function.

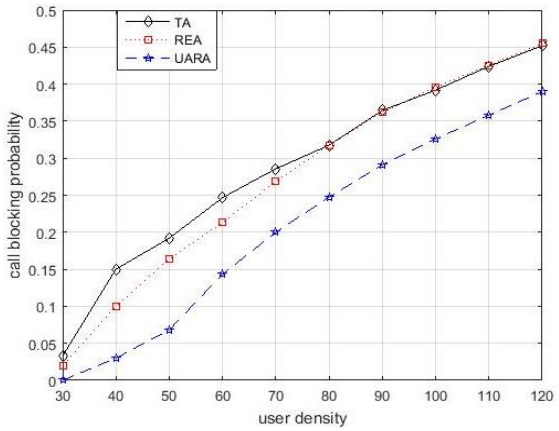

Fig. 3. Call blocking probability

Fig. 3 gives the QoS status of the whole network. To quantify the QoS status, call blocking probability (CBP) is defined, which is expressed as $CBP = 1 - K/|U|$ [16], where K denotes the number of scheduled users, U is the total number of the network, so $K/|U|$ is the serviced ratio in the network. The reason for CBP is that the limited resources cannot support so many users, so there are some users who are unable to associate with BSs. The scheduling scheme adopts max-rate achievable rate first. As depicted in Fig. 3, as user density increases, the total number of users in the range of the network is also increasing, but resources remain unchanged, so there will be more users that no BS can provide service for. Consequently, the CBP is decreasing with the increase of user density in all three schemes. As the MBS attracts most of the users because of the overwhelming transmission power in TA, the MBS is overloaded, whereas the PBSs are underutilized, so it is easy to understand why TA has a highest CBP. As REA has an offloading capacity to some extent, REA has a lower CBP than TA at first, but as user density goes up, CBP of the two schemes tend to be the same. This occurs mainly because the MBS and the PBSs are both overloaded. UARA obtains the lowest CBP compared to the other two schemes, as it takes users' QoS requirement into consideration.

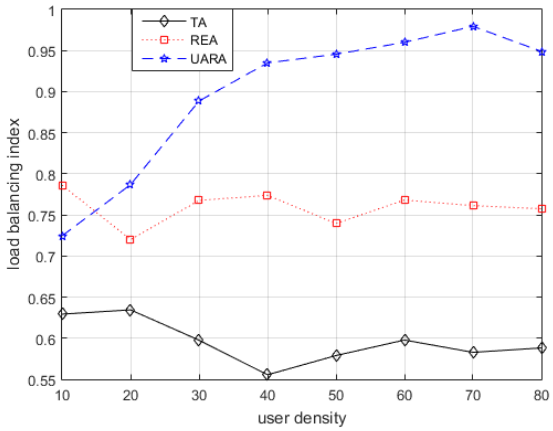

Fig. 4. Load balancing index

Fig. 4 represents the load balancing status by load balancing index (LBI), which reflects the disparity of the consumed resources of different BSs, i.e. $\mathrm{LBI} = (\sum_{j \in B} \sum_{s \in S} L_{js})^2 / \sum_{j \in B} \sum_{s \in S} L_{js}^2$. The distribution of load is more balanced if LBI is bigger. As user density increases, LBI tends to be stable, which is probably because the increase of the user's number reduces the randomness of the topology of the network. As load balancing is out of consideration in TA, the LBI of the scheme is the lowest. REA simply sets an offset to offload users from the MBS to the PBSs, so the index of REA is higher than TA. In UARA, load balancing is always in consideration, so it can achieve the highest LBI among the three schemes.

Fig. 5 demonstrates the CDF of uplink power. We can derive the ratio by the scope in different reference intervals. As Fig. 5 shows, when the reference power is low, which means that the distance between the user and serving BS is short, the ratios in these intervals are the

same among the three schemes. It can be easily explained through the fact that if the distance between the BS and the user is short enough, distance is the dominant factor, and the three schemes tend to associate the user with the nearest BS. With the reference power increasing, we can see that in UARA and REA, uplink power is more distributed in small reference power interval, i.e. [-20, 0] dB, and less distributed in big reference power interval, i.e. [0, 20] dB. The reason is that REA and UARA are both load-aware and tend to associate users with PBSs, which are usually nearer than the MBS.

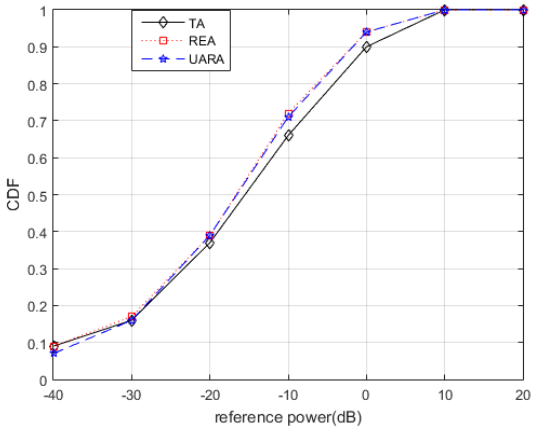

Fig. 5. CDF of uplink power

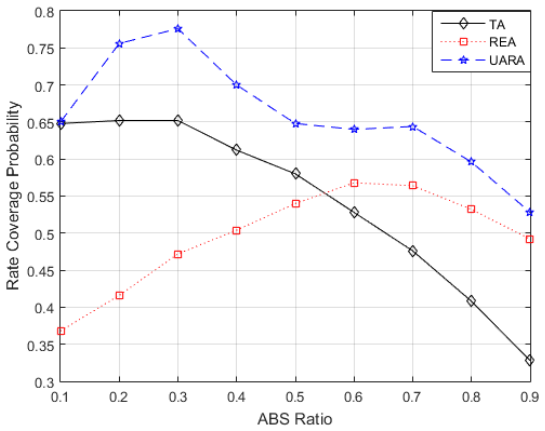

Fig. 6. Resource allocation

Fig. 6 illustrates the rate coverage probability (RCP), which means the proportion of scheduled users whose effective rate is greater than a given target rate. As is illustrated in Fig. 6, RCP of the three schemes are increasing at first, then decreasing with the increasing γ. The main reason is that at first γ is small, there are still sufficient resources for the MBS to serve for its user, and there are more ABSs for PBSs to serve for users with less interference, so an increasing RCP is obtained, but as γ keeps on increasing, the users served by the MBS is decreasing because of the insufficient resources. As shown in Fig. 6, REA has a lower RCP than TA at first, as the users of PBSs are in favor of transmitting data in ABS for a better rate, but at first there not enough resources for the PBS to serve so many users, so many users remain unconnected, resulting a low RCP. As UARA considers resource consumption, it has a highest performance in the three schemes whatever value γ is.

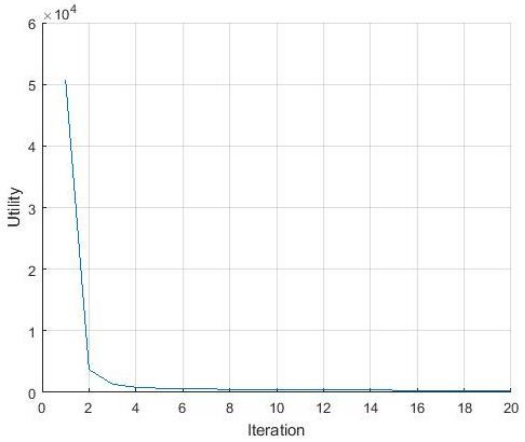

Fig. 7. Convergence

Fig. 7 depicts the convergence of the UARA. As we can see from (12), the utility decreases exponentially, so UARA has a very fast convergence rate. As shown in Fig. 7, the number of iteration is no more than 10, even 1, if the initial parameters are set properly.

VI. Conclusion

In this paper, user association and resource allocation are jointly considered to mitigate the severe interference and obtain a balanced distribution of cell load in a two-tier HCN. We firstly formulate the problem as a nonlinear combinatorial problem which also takes QoS into consideration. To mitigate the severe interference in HCNs, eICIC, as a resource allocation method, is introduced to this paper. Besides, uplink power and load balancing are considered in the user association scheme. In view of the special form of the problem, we are inspired to divide the problem into two sub problems by dual decomposition method. Then a low-complexity distributed algorithm is proposed. The algorithm has a very fast convergence rate. Finally simulation results show that the proposed algorithm achieves a great performance gain in terms of CBP and LBI when compared with TA and REA. Less interference can also be obtained by reducing the uplink power and adoption of eICIC. In future work, dynamic ABS ratio and backhaul capacity can be taken into consideration.

References

[1] I. S. Han, *et al*., "New paradigm of 5G wireless internet," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 3, pp. 474-482, Mar. 2016.

[2] M. Liu, Y. Teng, and M. Song, "Tradeoff analysis between spectrum efficiency and energy efficiency in heterogeneous networks (HetNets) using bias factor," *Journal of Communications*, vol. 10, no. 10, pp. 784-789, Oct. 2015.

[3] D. Liu, *et al*., "User association in 5G networks: A survey and an outlook," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1018-1044, 2016.

[4] ETSI, "Evolved universal terrestrial radio access (E-UTRA) and evolved universal terrestrial radio access network (E-UTRAN)," Overall description; Stage 2 (Release 10), 3GPP TS 36.300 V10.8.0, Jun. 2012.
[5] B. Xie, Z. Zhang, R. Q. Hu and Y. Qian, "Spectral efficiency analysis in wireless heterogeneous networks," in *Proc. IEEE International Conference on Communications*, Kuala Lumpur, 2016, pp. 1-6.
[6] J. Andrews, S. Singh, Q. Ye, X. Lin, and H. S. Dhillon, "An overview of load balancing in hetnets: old myths and open problems," *IEEE Wireless Communications*, vol. 21, no. 2, pp. 18-25, Apr. 2014.
[7] Q. Ye, B. Rong, *et al*., "User association for load balancing in heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 6, pp. 2706-2716, Jun. 2013.
[8] Y. Peng and F. Qin, "Exploring het-net in LTE-Advanced system: Interference mitigation and performance improvement in Macro-Pico scenario," in *Proc. IEEE International Conference on Communications Workshops*, Kyoto, 2011, pp. 1-5.
[9] Y. Chen, J. Li, W. Chen, Z. Lin, and B. Vucetic, "Joint user association and resource allocation in the downlink of heterogeneous networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 7, pp. 5701-5706, July 2016.
[10] Q. Kuang, W. Utschick, and A. Dotzler, "Optimal joint user association and multi-pattern resource allocation in heterogeneous networks," *IEEE Transactions on Signal Processing*, vol. 64, no. 13, pp. 3388-3401, July 2016.
[11] Y. Jia, M. Zhao, and W. Zhou, "Joint user association and eICIC for max–min fairness in HetNets," *IEEE Communications Letters*, vol. 20, no. 3, pp. 546-549, March 2016.
[12] J. S. Kim, J. K. Kim, and J. H. Kim, "Advanced handover scheme considering downlink and uplink service traffic in asymmetric channel," *Computer Networks*, vol. 89, pp. 1-13, Oct. 2015.
[13] X. Chen and R. Hu, "Joint uplink and downlink optimal mobile association in a wireless heterogeneous network," in *Proc. IEEE Global Communications Conference*, Anaheim, 2012, pp. 4131-4137.
[14] X. Zhen, *Practical Numerical Analysis*, Tsinghua University Press, 2006 (in Chinese).
[15] B. Stephen and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.
[16] T. Zhou, Y. Huang, L. Fan, and L. Yang, "Load-aware user association with quality of service support in heterogeneous cellular networks," *IET Communications*, vol. 9, no. 4, pp. 494-500, 2015.

Haiqiang Liu was born in Sichuan Province, China, in 1992. He received the B.S degree in communication engineering from Nankai University in 2014. He is now a postgraduate in Institute of Information Engineering, Chinese Academy of Sciences. His research interests include next generation network and optimization theory.

Zhenxiang Gao received his Ph.D. degree from Beijing University of Posts and Telecommunications (BUPT) in 2015. He is currently a research staff of IIE, CAS. His current research interests include mobile communication technology, mobile social networks and future networks.

Xulong Shao was born in Zhejiang Province, China, in 1992. He received the B.S degree from Nankai University in 2014. He is now a postgraduate in IIE, CAS. His research interests include mobile management and CoMP in the dense heterogeneous networks.

Xu Shan was born in Shandong province, China, in 1989. He received the master degree from Beijing Jiaotong University (BJTU). He is currently a research staff of IIE, CAS. His research interest is secure communication.

WeiHua Zhou received the B.S degree from Xi'dian University, China, in 1999, and received the PhD degree from BUPT in 2004. He is currently a senior engineer with IIE, CAS. His current research interests include 5G networks, mobile virtual networks and secure communication.