User Association for Load Balancing in Cellular Network with Hybrid Cognitive Radio Relays

Hongfu Guo, Fanqin Zhou, Lei Feng, Peng Yu, and Wenjing Li
State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications, China
Beijing, 100876, P. R. China
Email: wjli@bupt.edu.cn

***Abstract*—Hybrid cognitive radio (CR) relays serve cellular users in a two-hop fashion, which jointly utilize both licensed and unlicensed radio spectrums to significantly increase the system capacity. User equipments (UEs) need to be actively associated with the macro-cell BS or CR relays having a more lightly loaded spectrum if the quality of services (QoS) can be guaranteed. To this end, this paper investigates optimal user association for load balancing problem in cellular network with hybrid cognitive radio relays. Firstly, we propose a multi-objective user association optimization model to balance the loads among different tiers while reducing the total resource occupancy. Then, this multi-objective problem is converted into a single one by the linear weighing-sum method and a genetic algorithm is introduced to solve it. The numerical simulation results show that our proposed scheme can obtain more balanced resources occupation, better throughput performance, and lower blocking rate compared with the heuristic and max-power strategies.**

Keywords*—Cognitive radio, relay, load balancing, user association

I. Introduction

With the rapid development of wireless communication, there is an increasing demand for spectrum resources. And spectrum scarcity has become a serious problem, which severely restricts the development of the radio industry. In addition, the licensed spectrum is usually preassigned and can not be flexibly configured according to the actual business situation, which leads to inefficient use of spectrum resources. In order to take full advantage of the very limited licensed spectrum, it is necessary, on the one hand to devise new access technology with higher spectral efficiency, and on the other hand to densify the network by introducing low power nodes (LPNs), like small cells and relays resulting in the heterogeneous networks (HetNets). HetNets is capable of achieving more spectrum-efficient communications by deploying LPNs underlaid on the macro cells [1]. LPNs vary in types, while they share some common properties: low power consumption and operation cost, where the relay node (RN) is considered as a typical low power and low cost device. The relay can be placed at the edge of the cellular network to expand the network coverage, or it can be placed in the area of the high user's business to ensure the user's rate requirements.

In the meanwhile, there are a large number of unlicensed spectrum resources available shared among non-3GPP wireless technologies, and it is difficult to fully cope with the challenges of rapid growth in business needs only by increasing the spectral efficiency and deploying more LPNs. Therefore, using unlicensed spectrum communication becomes a direction of future mobile communications. Cognitive radio (CR) technology [2] which can dynamically detect and access the idle spectrum such that the spectrum efficiency is improved will play a significant role in the use of available spectrum.

With spectrum-sensing ability CR relays are used to detect used and unused parts of the considered spectrum and therefore to provide efficient and fair usage of the available spectrum. The base station (BS) communicates with the CR-enable relays in the cellular spectrum, while the cognitive relays communicate with the user equipment (UE) in unlicensed CR spectrum. Such a network, called hybrid cognitive radio relay network [3]–[5], is promising because it has the potential to improve cellular capacity without expensive upgrades on the legacy network. In order to optimize the use of radio resources, an effective radio resource management program becomes particularly important.

HetNets will play a major role in the face of the growing needs mobile business of the 5G system. However, in order to better reach its potential, there are still some problems to be solved. One of the main challenges of HetNets is the load imbalance between the base station and relay nodes. The user association scheme in the cellular network is generally based on the principle of maximizing signal receiving power. Because users tend to select the high power base station causing the base station overload, while the wireless relay and other LPNs are less associated with the low utilization rate. For HetNets systems, load balancing is particularly important and has been extensively studied [6]–[9]. However, most of schemes are based on that the network spectrum resources are static. In this paper, we study the hybrid CR networks scenario, the system spectrum resources are dynamically changed with the cognitive spectrum resources sensed by the relay nodes, the system load balancing problem becomes more complicated and can not be solved by the existing schemes. In addition to that, most of works consider the number of associated UEs as the load of the BS or RNs, and few take account of QoS requirements that reflect the amount of consumed resources.

This paper investigates the load balancing problem in the cellular network with hybrid CR relays, specifically focused on optimal user association among the BS and the relays

978-3-901882-98-2

and system resources consumption. The BS and RNs can be seen as composing of a group of equal-sized resource blocks (RBs), and each user will occupy a certain number of RBs according to their own desired rate requirements. Based on this, we propose a multi-objective optimization model to improve the system load balance degree while reducing the total resource occupancy. A genetic algorithm is used to solve this problem. Simulation results show that our proposed scheme can significantly improve system performance in terms of the network load balancing, amount of business occupied resources, system blocking rate and throughput.

The contributions are as follows. 1) A user association scheme is proposed to address the problem of load balancing for the hybrid CR network scenario, which effectively implements the system load balancing and improves the system performance. 2) We make full use of sensing ability of hybrid CR relays, which on the one hand allows more users to access cellular network by providing cognitive spectrum resources, and on the other hand detects the spectrum overlap ratio between CR relays so as to reduce the impact of interference between them.

The rest of this paper is organized as follows. We discuss the related work in Section II. Section III introduces our system model and problem formulation. The proposed solving algorithm is described in detail in Section IV. Section V presents the performance evaluation results and Section VI concludes the paper.

II. Related work

In the existing literature, the conventional load balancing schemes in HetNets have been widely investigated. The most common bias method [10] is simple and effective, but almost none of the existing studies give an optimal solution for a closed form expression. So in recent years, some other related schemes for load balancing have been proposed [11], especially in the case of high deployment density, with more options to balance the load. However, the above scheme is based on that the network spectrum resources are static and it is difficult to adapt to dynamic hybrid CR network.

There are two key concerns when the UE is associated with a BS or RN: 1) selecting the BS or RN that can provide the optimal QoS to reach the required data rate; 2) controlling the BS or RN load to avoid congestion. The former is user-centric and the latter is network-centric. These two aspects are generally contradictory.

Several approaches have then addressed the above two contradictory problems through two aspects by joint consideration to achieve load balancing. Studies have shown that this is a convex optimization problem [12]–[15]. Authors in [12] use the Hungarian matching algorithm to optimize the user association and iteratively discover the available resources to minimize the system blocking rate. The literature [13] assumes that the user pre-association, [14] relaxes the constraint on number of serving station for each UE. [15] proposes a distributed relaxed association scheme that can utilize idle and non-idle resources to serve UEs. The above-mentioned schemes consider the number of associated UEs as the load standard and do not take into account the UE's specific QoS requirements, which characterizes the resources required by the UE.

In this paper, we study the hybrid CR networks with CR relays. The user association schemes commonly in CR networks have been extensively investigated, and each scheme is based on a specific criterion, such as the maximum received power [16], minimum path loss [17] or maximum link SINR [18]. However, there are relatively few studies on user association based load balancing in hybrid CR networks. The link performance and resource allocation in the cellular networks with hybrid CR relays are studied in [4], where only one relay is considered. [5] studies the system capacity and outage probability with multiple relays. But they adopt the simplest association scheme: each user selects the associated BS or RN based on the shortest distance. However, such association scheme is not optimal because it does not take into account the link quality and the load balancing. In addition to that, the available CR spectrum resources are constantly fluctuating and it may result in system performance decline. [19] implements user association based on the criteria of the maximum access rate, and then through the admission control to achieve load balancing. This will inevitably lead to a decline in system capacity.

To our best knowledge, the user association based load balancing between the base station and relay nodes in hybrid CR networks from the optimization perspective has never been investigated before. Therefore, this paper proposes an optimal model and a corresponding solving algorithm for this kind of problem. And some insights are found by the performance evaluations.

III. System model and problem formulation

A. System Model

Without loss of generality, we consider a cellular network with one BS, M hybrid CR RNs, and N UEs. The BS is placed at the center of the cell, while the RNs are placed at the fixed positions, which are uniformly distributed near the cell edge. Finally, UEs are randomly distributed in the cell area and each user has a desired data rate.

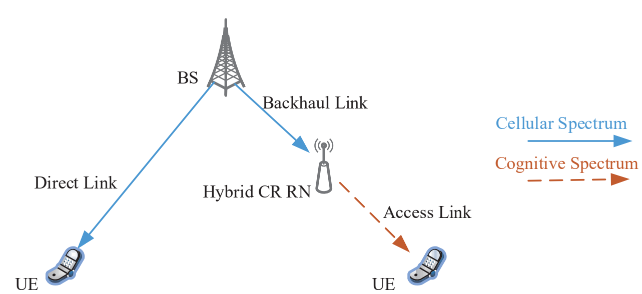

Fig. 1. Communication links for downlink transmission

Here we only consider the downlink transmission, and the results can be extended to the uplink case easily. Each user needs to associate with the BS or one of the RNs. As shown in Fig.1, the system includes two transmission links, 1) direct transmission: the user directly communicates with the base station using cellular spectrum, this link is called the direct link; 2) relay-based transmission: the communication over the first hop from the BS to the RN operates in cellular spectrum called backhaul link, and the second hop from the RN to the UE operates in unlicensed CR spectrum called access link. Cellular and CR spectrum are represented by $W_{cellular}$ and W_{CR}, respectively. The cellular and CR spectrum are composed of a group of equal-sized resource blocks of bandwidth B. The impulse response of each RB is assumed to be independent of each other. In order to simplify the analysis, we consider the overlay spectrum sharing model for the CR relay, but our solution can be implemented in the underlay case by considering more interference. In the CR network, the primary user is a user who is authorized to use some bands, it has the absolute priority to use the spectrum. The CR RN can sense the dynamic changes of the available spectrum resources of the primary user. In the case where the transmission power is limited in the available bandwidth, we assume that the CR RN will not cause harmful interference to the primary user, however, the interference among CR RNs needs to be considered. In addition to this, in order to reduce the interference between the relays, when a relay allocates available resources to a user, it always preferentially allocates resources that are not used by other relays which can be achieved through the cognitive ability of the RNs.

RN has the global knowledge about which RB is occupied by the primary users and which is idle. Specifically, since the primary users have occupied a part of spectrum resources, defined as C_p, then the available cognitive spectrum bandwidth for RN $j(j \in \{1,2,\cdots,M\})$ in access link is W_j. For RNs and BS in the cell, we define the load factor to represent the RBs utilization of each RN and BS.

$$\eta = \frac{W_{occupied}}{W_{total}}, \tag{1}$$

where $W_{occupied}$ is the already occupied bandwidth and W_{total} is the total bandwidth of each RN and BS.

In the paper, we consider the frequency division duplexing (FDD) scheme, so where η is equivalent to the ratio of the number of RBs occupied by the UEs and the total number of RBs of the BS or RNs.

The CR user provides the desired data rate to the BS or RN through a dedicated feedback channel, and then the BS or RN performs resource allocation. The resources available in the CR network are dynamically changed, so resource allocation may not guarantee all users' requirements.

For the above two different links, we can separately find the transmission rate:

a) Direct transmission. According to the Shannon capacity formula, when the UE i and the BS are associated, the data rate can be calculated as follows

$$R_1^{direct}(i) = N_{i,BS} B * \log_2(1+\gamma_{BS,i}), \tag{2}$$

where $N_{i,BS}$ is the number of RBs occupied by UE i, $\gamma_{BS,i}$ is the received SINR over the link between the BS and UE i and it can be expressed as

$$\gamma_{BS,i} = \frac{|h_{BS,i}|^2 P_{BS}}{\sum_{k=1}^{N_{BS}} |h_{BS_k,i}|^2 P_{BS} + N_{i,BS} N_0 B}, \tag{3}$$

where P_{BS} is the transmission power of BS, $h_{BS,i}$ is the channel gain between BS and UE i, and N_0 denotes the one-side noise power spectrum density. Considering the interference from the surrounding N_{BS} BSs using the cellular spectrum, and $h_{BS_k,i}$ is the channel gain from the nearby BS k.

b) Relay-based transmission. When the UE selects this transmission path, the two links are analyzed:

i) Backhaul link. The data rate of UE i in this link can be given as

$$R_{2b}^{realy}(i,j) = N_{RN_j,BS} B * \log_2(1+\gamma_{RN_j,BS}), \tag{4}$$

where $N_{RN_j,BS}$ is the number of RBs occupied by communication between the BS and RN j, $\gamma_{RN_j,BS}$ is the received SINR of RN j from the BS and the calculation method is as (3).

ii) Access link. The data rate in access link between RN j and UE i is as below.

$$R_{2a}^{realy}(i,j) = N_{i,RN_j} B * \log_2(1+\gamma_{RN_j,i}), \tag{5}$$

where N_{i,RN_j} is the number of RBs occupied by communication between the RN j and UE i, $\gamma_{RN_j,i}$ is the received SINR of UE i from the RN j.

Note that only the interference from other CR RNs is considered in CR spectrum. Considering that the available spectrum is part of the total bandwidth, only signals that use the same part of the spectrum can cause interference. The definition $\alpha_{j,j'} \in [0,1]$ is used to denote the overlap ratio between the available RBs sensed by RN j and the available RBs sensed by RN j'. The relay can obtain the spectrum utilization of each station. E.g., $\alpha_{j,j'} = 1$ if the all available resources overlap; $\alpha_{j,j'} = 0.5$ if half of the available RBs number of RN j overlap with the RBs of RN j'. Therefore the UEs received SINR from RN j is

$$\gamma_{RS_j,i} = \frac{\left|h_{RN_j,i}\right|^2 P_{RN}}{\sum\limits_{j'=1,j'\neq j}^{M} \left|h_{RN_{j'},i}\right|^2 P_{RN} \alpha_{j,j'} + N_{i,RN_j} N_0 B}, \tag{6}$$

where P_{RN} is the transmission power of RN, $h_{RN_j,i}$ is the channel gain between RN j and UE i.

The backhaul link and access link operate in different spectrum resources, so the RNs work in a full-duplex mode while transmitting and receiving. So the data rate when UE i is associated with RN j can be expressed as

$$R_2^{realy}(i,j) = \min\left(R_{2b}^{realy}(i,j), R_{2a}^{realy}(i,j)\right). \tag{7}$$

Consequently, when UE i has the desired rate R_i, we can convert it to a request for the minimum number of RBs on each link:

a) Direct transmission

$$N_{i,BS} = \arg \min_{N_{i,BS}} \left\{ R_1^{direct}(i) \geq R_i \right\}. \tag{8}$$

b) Relay-based transmission. Similarly can be obtained, the number of RBs required for access link and backhaul link are separately N_{i,RN_j} and $N_{RN_j,BS}$ when UE i is associated with RN j. Then we define (9) to denote the sum of the number of RBs required for these two sub-links.

$$N_{i,j}^{relay} = N_{i,RN_j} + N_{RN_j,BS}. \tag{9}$$

Because the backhaul link is usually optimized(line-of-sight transmission), the resources used are relatively lower compared to the resources consumed on the direct and access links.

B. Problem Formulation

In this system, the user can choose the BS to transmit directly or relay-based transmission. Our goal is to achieve user association and load balancing while minimizing system resource consumption as much as possible. In order to measure the load balancing of the whole network, we introduce the variance of the load factor as follows. The size of the value is inversely proportional to the degree of network load balancing. So the goal of our load balancing is to achieve f minimization.

$$\min f = \min \frac{1}{1+M} \sum_{j=1}^{1+M} \left(\eta(j) - \frac{1}{1+M} \sum_{j=1}^{1+M} \eta(j) \right)^2, \tag{10}$$

where $\eta(j)$ is as defined in (1), corresponding to the load factor of the BS or RN. For further measurement the status of load balance of the entire work, we use Jain's fairness index as follows

$$\xi = \frac{\left(\sum_{j=1}^{M+1} \eta_j \right)^2}{(M+1) \sum_{j=1}^{M+1} {\eta_j}^2}, \tag{11}$$

where the load balance index takes the value in the interval $\left[\frac{1}{M+1}, 1 \right]$. The value of ξ closer to 1, the more balanced of the network load. When the value is $\frac{1}{M+1}$, the network load is the most imbalanced.

Another optimization objective is to ensure that the system resource consumption h is minimized and the objective function is given by

$$\min h = \min \sum_{i=1}^{N} y_i \cdot N_{i,BS} + \sum_{j=1}^{M} \sum_{i=1}^{N} x_{i,j} \cdot N_{i,j}^{relay}, \tag{12}$$

where $x_{i,j}$ and y_i are the binary decision variables, $x_{i,j} = 1$ if UE i is associated with RN j, $x_{i,j} = 0$ otherwise; and $y_i = 1$ if UE i is associated with the BS, $y_i = 0$ otherwise. The above two objectives as the multi-objective optimization model, coupled with the constraints, the final optimization model can now be formulated as

$$\min f = \min \frac{1}{1+M} \sum_{j=1}^{1+M} \left(\eta(j) - \frac{1}{1+M} \sum_{j=1}^{1+M} \eta(j) \right)^2, \tag{13a}$$

$$\min h = \min \sum_{i=1}^{N} y_i \cdot N_{i,BS} + \sum_{j=1}^{M} \sum_{i=1}^{N} x_{i,j} \cdot N_{i,j}^{relay}, \tag{13b}$$

$$\text{s.t.} \sum_{j=1}^{M} x_{i,j} + y_i \leq 1, \forall i \in \{1, 2, ...N\}, \tag{13c}$$

$$\text{s.t.} \sum_{i=1}^{N} y_i \cdot N_{i,BS} + \sum_{j=1}^{M} \sum_{i=1}^{N} x_{i,j} \cdot N_{RN_j,BS} \leq W_{cellular}, \tag{13d}$$

$$\text{s.t.} \sum_{i=1}^{N} x_{i,j} N_{i,RN_j} + C_p \leq W_{CR}, \forall j \in \{1, 2, ...M\}. \tag{13e}$$

The objective function presented in (13a) aims at the minimization of the network load rate variance. The objective function (13b) represents the minimization of system resources consumption by minimizing the sum of the radio resources consumed at BS and RNs. Constraints (13c) denotes that each UE is associated with either a BS or a combination of a BS and an RN. Constraints (13d) and (13e) are capacity constraints and imposed that the resources consumed at BS and each RN do not exceed their capacity.

In order to solve the above-mentioned multi-objective optimization problem, we utilize the linear weighing-sum method to convert it to the single objective optimization problem. Considering that there is a large difference in the magnitude of each sub-objective function, we need to make a slight deformation. In this case, the min-max normalization method is used to linearly transform the original data and mapping the resulting values to [0-1], then subjected to linear weighting. So our final joint optimization goal is

$$ObjV = \min \left((1-\lambda) \frac{f - f_{\min}}{f_{\max} - f_{\min}} + \lambda \frac{h - h_{\min}}{h_{\max} - h_{\min}} \right). \tag{14}$$

By only considering the network load balancing as a single objective corresponding to $\lambda = 0$, we can obtain the ideal value of the load factor variance f_{min} and the max value of the total occupied resources h_{max}. Similarly, we can obtain the ideal value of the total occupied resources h_{min} and the max value of load factor variance f_{max} by only taking the lowest total occupied resources as a single objective corresponding to $\lambda = 1$. λ is the corresponding weighting factor that can be used to characterize the relative importance of the two optimization goals.

IV. Genetic algorithm solving for user association problem

In this paper, genetic algorithm (GA) is used to solve the nonlinear optimization problem of (14). GA is a kind of stochastic search algorithm which is based on natural selection mechanism and genetic mechanism. Its search is independent of the gradient information and it is suitable for solving the nonlinear problem which is difficult to solve by the traditional method, and GA has good robustness. For the model of this paper, the specific steps of GA are as follows.

1) Design Chromosome Encoding: In this paper, the integer coding method is adopted. For the N users of the association problem, each user has $M+1$ choices, so the chromosome is divided into N segments, the value $x_i (i \in \{1, 2, \cdots, N\})$ of the i-th segment corresponds to the number of the associated station of user i. Therefore, the chromosome encoding of this model is as follow.

$$X = (x_1, x_2, x_3, ..., x_N). \tag{15}$$

2) Initial population: After completing the chromosome coding, an initial population must be generated as a starting solution, so it is first necessary to determine the number of initialized populations. The number of initialized populations is generally based on experience. In general, the number of populations is determined by the size of the user, and the value is floating between 50 and 200, where we take a population size of 100.

3) Fitness function: In view of the above-mentioned final optimization model, the individual fitness function can be designed as

$$fitness = \frac{1}{ObjV}. \tag{16}$$

That is, the fitness function is the reciprocal of our final optimization objective value. The optimization goal is to select the chromosome with the fitness function as large as possible. The greater the fitness value is, the better the chromosome is. Considering the constraint problem, here we determine the binding for each chromosome, and for the chromosomes that do not satisfy the constraint, the fitness function value is set to minus infinity.

4) Selection operation: Selection Operation means that selects the individual from the old group to the new group with a certain probability. The probability that an individual is selected is related to the fitness value, and the greater the individual fitness value, the greater probability of being selected.

5) Cross operation: According to the empirical value, the crossover probability is 0.5. Determine the individuals for cross operation, then divide them into some groups and each group comprising two individuals. Each group makes two-point cross operation: two cross locations are randomly arranged in the two individual gene strings that are paired and then exchange two gene string positions between the two intersections set.

6) Mutation operation: Whether a chromosome is mutated is determined by the mutation probability, the mutation probability value is usually selected between 0.001-0.1, so here to take the mutation probability value of 0.05. The strategy of the mutation is to randomly select two locations and change the values corresponding to the two locations. For this paper, generate two random integers N_1 and N_2 in the range $[1, N]$ and exchange the values at their corresponding positions.

7) Evolutionary reversal operation: In order to improve the local search ability of the GA, a continuous evolutionary reversal operation is introduced after selection, cross and mutation. Here, "evolution" refers to the unidirectional nature of the reversal operator, that is, only the fitness value is improved to accept it after the reversal, otherwise the reversal is invalid. For this paper, generate two random integers N_1 and N_2 in the range $[1, N]$ and reverse the values of the interval.

Each individual is cross-mutated and then evaluated into the fitness function. The individuals with greater fitness values are selected for the next generation of crossover and mutation and evolutionary reversal. By the loop operation, determine whether the genetic algebra is satisfied MAXGEN=100, the calculation of the fitness value is jumped if it is not satisfied; otherwise, the genetic operation is ended.

The detailed algorithm is shown as Algorithm 1.

Algorithm 1: The Genetic Algorithm (GA)

input : crossover probability *PC*;
mutation probability *PM*;
maximum genetic algebra *MAXGEN*.
output: the best individual

Generate initial population using integer coding;
for $time = 1$ **to** *MAXGEN* **do**
  Determine the fitness of each individual via (16);
  Select the individuals that satisfy constraints via the roulette selection strategy;
  Optionally apply the two-point cross operation on chromosome i and j with probability *PC*;
  Optionally apply the mutation operation on the new chromosome with probability *PM*;
  Apply the evolutionary reversal operation;
end
return *The optimal solution has obtained.*

The specific steps for solving the model are as follows.

Step 1: Solve the optimization problem of load balancing as a single objective, and obtain the load rate variance $f_{\min}$ and the total number of RBs $h_{\max}$ consumed by all users, corresponding to $\lambda = 0$.

Step 2: Solve the optimization problem of resource consumption as a single objective, and obtain the load rate variance $f_{\max}$ and the total number of RBs $h_{\min}$ consumed by all users, corresponding to $\lambda = 1$.

Step 3: The multi-objective optimization is transformed into single-objective optimization, and the GA is used to solve the problem and obtain the best user association solution (the corresponding chromosome to obtain the optimal fitness value).

V. Performance evaluation

A. Simulation Setting

In order to evaluate the performance of the proposed scheme, we simulate a cellular network with a radius of 500m, which contains a BS located in the center and a number of hybrid CR RNs deployed in the network. 6 nearby BSs are simulated for calculating the interference. The total bandwidth of the cellular spectrum is a fixed value $W_{cellular}$=10MHz, while each CR RN can sense and utilize a portion of the CR spectrum bandwidth W_{CR}=10MHz. The initial load factor η of BS and RNs is randomly selected from the interval $[0, 1]$. The remaining details of the simulation settings are list in Table I.

TABLE I
SIMULATION PARAMETERS

Parameter	Value
Transmit Power(dBm)	BS:46; RN:30;
Path loss model (d is in km)	BS-UE:$138.5 + 32.23\log_{10}(d)$; RN-UE:$127.0 + 30\log_{10}(d)$;
Noise density(dBm/Hz)	-174
Number of RNs	1-10
Location of RNs(m)	350,450
Number of UEs	50,60,70,80,90,100
Inter-site distance(m)	500
Population size	100
Crossover rate	0.5
Mutation rate	0.05
MAXGEN	100

For the purpose of studying the impact of RNs on hybrid CR network performance and the algorithm itself, and to achieve a simple deployment of the RNs, we mainly simulate and analyze the number and location of the RN. Two options regarding distance from the RNs to the BS are considered: 350 and 450 meters. For each option we carried out simulations for 1) 3 RNs placed at the vertices of a triangle with the BS placed at the center, and 2) 6 RNs placed at the vertices of a hexagon with the BS placed at the center. Furthermore, the number of UEs is varied from fifty to one hundred. Through the performance analysis of system blocking probability and throughput, the optimal RN number and deployment location for the scenario of this paper are determined.

Based on the above work, we can get the optimal number and location of RN. Then we compare our proposed scheme against two other schemes from the literature in terms of load balancing, resource consumption and usage, system sum rate, and blocking probability. In particular, the following schemes are considered:

a) Heuristic scheme: According to this scheme in [20], a UE is associated with network node BS or RN so that the overall resource consumption is minimized.

b) Max-Power:According to this scheme, a UE is associated with the BS or RN that provides the stronger received power.

B. Weighting Factor Selection

The determination of the weighting factor often depends on the degree of emphasis of the decision makers about various objectives in the system. In order to reasonably determine the weighting factor, we simulate on the different values separately. The normalized results are shown in the Fig.2. $\lambda = 0$ denotes the result of the single objective with the minimum load factor variance and $\lambda = 1$ denotes the result of the single objective with the minimum total resource occupancy. $0.2 < \lambda < 0.8$, the values of the two single objectives change slowly and close to their ideal values. In the following part, we make $\lambda = 0.4$.

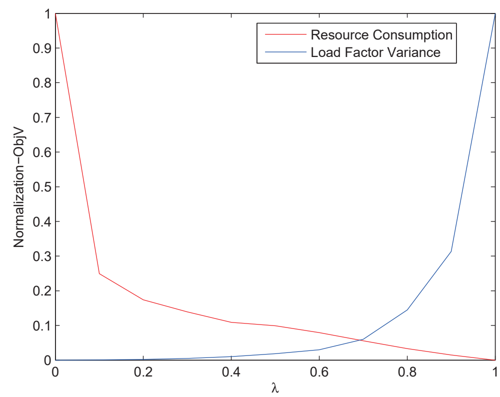

Fig. 2. Weighting factor set

C. Determine the Number and Location of RNs

Fig. 3 shows the results of the blocking probability under the above scenario. It is shown that with the increase of the number of CR relays, the system blocking probability is decreased correspondingly. As the total sensed available resources of the system are increased, and the system capacity and coverage are improved accordingly, resulting in a corresponding decrease in system blocking probability. It can also be seen that the closer the relay is to the BS, the higher the system blocking probability is. This is because that more users will prefer to directly associate the BS, but taking into account the role of load balancing, when the user is associated with the relay, it will consume more resources, resulting in faster system resource consumption and system blocking probability.

Fig. 4 is a comparison of the system sum rate with the change in the number of RNs. Multi-objective GA by jointly considering the resource consumption and load condition, compared to the other two schemes, achieves a much higher system rate. In addition, we can also see that the number of RNs in the system is not the more the better. Although with the increase in the number of RNs, the system available sensing spectrum resources correspondingly increase, the interference between the RNs also increases accordingly. For our simulation scenario, the system performance remains basically stable when the number of RNs reaches 6.

Therefore, the next simulation work we choose that the number of RNs is 6 and the location is at 450m.

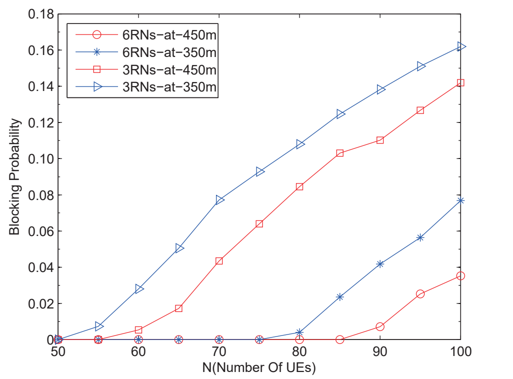

Fig. 3. Blocking probability versus UEs number with different RNs numbers and locations for GA

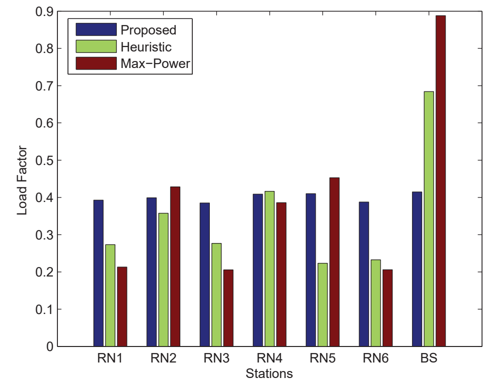

Fig. 5. Comparison of load balancing performance for various user association schemes

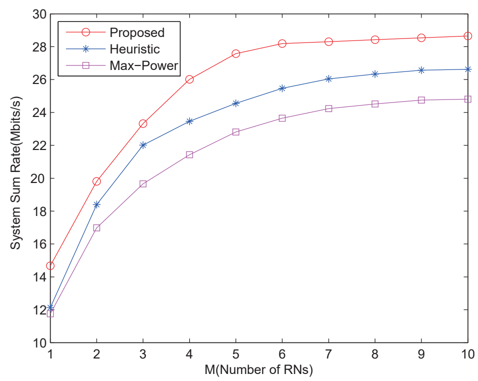

Fig. 4. System sum rate versus RNs number for various user association schemes

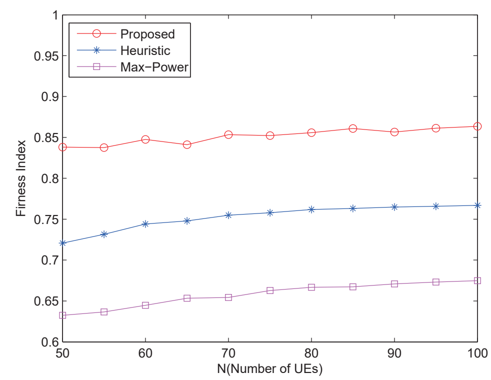

Fig. 6. Fairness index under different UE number

D. Network Performance Improvement

In Fig. 5, we compare our proposed scheme with the other two schemes on the load balancing performance. It is obvious that for our proposed scheme, the load factor of each station is very close, indicating that the GA works well. For the heuristic scheme, the system load rate variance is obviously larger than the GA proposed in this paper because it only considers the single objective of the minimization of system resources consumption without taking into account the system load balancing. For the access scheme based on the maximum received power criterion, the load balancing performance is the worst. As UEs tend to associate the strongest node BS, the direct link of the BS is depleted quickly while the RNs remain underutilized. Fig. 6 indicates the relationship between the fairness index and UE number that is one of most important factor influencing load. It is obvious that our proposed scheme has a larger fairness index than the two other schemes.

In Fig. 7 we depict the resource utilization using the above three schemes. As the multi-objective GA takes the lowest resource consumption as one of the optimization objectives, effectively guarantees the efficient use of limited resources, the total system resources consumption is less than the maximum received power scheme. But the heuristic scheme targets lowest resource consumption as the only optimization goal, the resources consumption is less than the multi-objective optimization algorithm. As the number of UEs increases, the system resource utilization of each scheme increases slowly, because the system begins to block. However, since the heuristic scheme does not consider the load balancing, the system blocking rate is greater than the GA which can also get the same conclusion in Fig. 8 so the gap between the two schemes gradually increases.

Fig. 8 shows the comparing results of the blocking probability. As we can see, our proposed scheme shows better performance in reducing the system blocking rate. This is because the heuristic scheme targets the lowest resource consumption as the criterion for UEs access essentially, and it does not consider the system load. In the scheme based on the maximum received power, more UEs tend to choose the directly associated BS resulting in that cellular spectrum resources are rapidly declining while the access link of the RNs as well as the backhaul link remain underutilized. However, GA determines the user association scheme based on two considerations: load balance and minimum resource consumption, and it considers the system's resource utilization, so as to achieve maximum of the remaining available resources in the hybrid CR network. Effectively increase the number of access UEs, reduce the system blocking rate, and increase the system capacity.

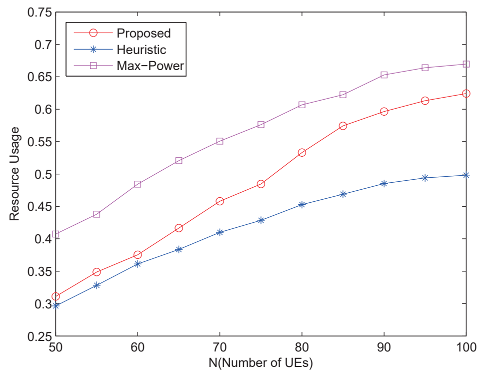

Fig. 7. Resource usage versus UEs number for various user association schemes

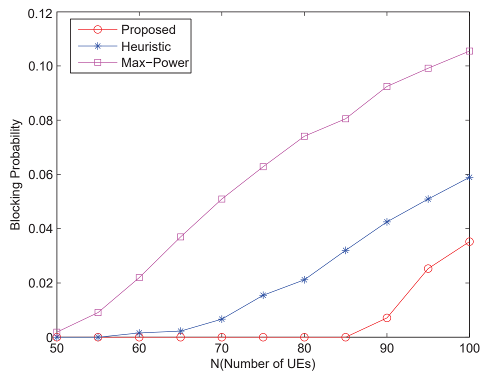

Fig. 8. Blocking Probability versus UEs number for various user association schemes

VI. Conclusion

In this paper, we investigate the user association and load balancing problem in cellular networks with hybrid CR relays. An effective multi-objective optimization model is proposed to formulate this problem, which aims at minimizing load factor variance and the total resource consumption. On this basis, we adopt the genetic algorithm to solve. Simulation results prove that the proposed scheme improves overall system performance on the load balancing, system blocking probability and system throughput.

Acknowledgment

This work was supported by the China Postdoctoral Science Foundation funded project 2017M610827, National Natural Science Foundation of China (NSFC) (Grant No. 61672108) and Tibet Natural Science Foundation(Grant No. 2016ZR-15-63).

References

[1] J. G. Andrews, S. Singh, Q. Ye, X. Lin, and H. Dhillon, "An overview of load balancing in hetnets: Old myths and open problems," *IEEE Wireless Communications*, vol. 21, no. 2, pp. 18–25, 2014.

[2] N. Miloevi, Z. Nikoli, F. Jelenkovi, V. Nejkovi, M. Toi, and I. ekar, "Spectrum sensing for the unlicensed band cognitive radio," in *Telecommunications Forum Telfor*, 2016, pp. 250–252.

[3] N. H. Mahmood, F. Yilmaz, G. E. Oien, and M. S. Alouini, "On hybrid cooperation in underlay cognitive radio networks," in *Signals, Systems and Computers*, 2013, pp. 308–312.

[4] X. Hong, C. Zheng, J. Wang, J. Shi, and C. X. Wang, "Optimal resource allocation and ee-se trade-off in hybrid cognitive gaussian relay channels," *IEEE Transactions on Wireless Communications*, vol. 14, no. 8, pp. 4170–4181, 2015.

[5] C. Liu, Q. Yang, H. Wang, X. Hong, J. Shi, and B. Tang, "System capacity analysis of hybrid cellular networks with cognitive radio relay," in *International Conference on Wireless Communications and Signal Processing*, 2015, pp. 1–5.

[6] H. Dahrouj, A. Douik, and O. Dhifallah, "Resource allocation in heterogeneous cloud radio access networks: advances and challenges," *Wireless Communications IEEE*, vol. 22, no. 3, pp. 66–73, 2015.

[7] 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, 2012.

[8] N. Prasad, M. Arslan, and S. Rangarajan, "Exploiting cell dormancy and load balancing in lte hetnets: Optimizing the proportional fairness utility," *IEEE Transactions on Wireless Communications*, vol. 62, no. 10, pp. 3706–3722, 2014.

[9] D. Bethanabhotla, O. Bursalioglu, H. Papadopoulos, and G. Caire, "Optimal user-cell association for massive mimo wireless networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 3, pp. 1835–1850, 2016.

[10] H. S. Jo, Y. J. Sang, P. Xia, and J. G. Andrews, "Heterogeneous cellular networks with flexible cell association: A comprehensive downlink sinr analysis," *IEEE Transactions on Wireless Communications*, vol. 11, no. 10, pp. 3484–3495, 2011.

[11] J. Andrews, S. Singh, Q. Ye, X. Lin, and H. Dhillon, "An overview of load balancing in hetnets: Old myths and open problems," *IEEE Wireless Communications*, vol. 21, no. 2, pp. 18–25, 2013.

[12] S. Mishra, S. Rangineni, and C. S. R. Murthy, "Exploiting an optimal user association strategy for interference management in hetnets," *Communications Letters IEEE*, vol. 18, no. 10, pp. 1799–1802, 2014.

[13] J. Yinghao and Q. Ling, "Joint user association and interference coordination in heterogeneous cellular networks," *IEEE Communications Letters*, vol. 17, no. 12, pp. 2296–2299, 2013.

[14] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 6, pp. 2706–2716, 2013.

[15] Q. Ye, M. Al-Shalashy, C. Caramanis, and J. G. Andrews, "On/off macrocells and load balancing in heterogeneous cellular networks," pp. 3814–3819, 2013.

[16] 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, 2011.

[17] S. Mishra and A. Trivedi, "Relay selection with channel allocation for cognitive radio relay channels in crn," in *Eleventh International Conference on Wireless and Optical Communications Networks*, 2014, pp. 1–4.

[18] M. Mezzavilla, K. Somasundaram, and M. Zorzi, "Joint user association and resource allocation in ue-relay assisted heterogeneous networks," in *IEEE International Conference on Communications Workshops*, 2014, pp. 628–634.

[19] G. Li, Y. Zhao, and K. Bian, "Efficient user association in cellular networks with hybrid cognitive radio relays," *IEEE Communications Letters*, vol. 20, no. 7, pp. 1413–1416, 2016.

[20] O. Siviloglou, A. Rouskas, and G. Karetsos, "Association of mobile stations in cellular networks with relay nodes," in *IEEE International Conference on Telecommunications and Signal Processing*, 2015, pp. 1–5.