IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2018.DOI

Interference Coordination and Resource Allocation Planning with Predicted Average Channel Gains for HetNets

KAIYANG GUO[1], TINGTING LIU[1], CHENYANG YANG[1], and ZIXIANG XIONG.[2]

[1]School of Electronics and Information Engineering, Beihang University, Beijing 100191, China (e-mail: {kyguo, ttliu, cyyang}@buaa.edu.cn).
[2]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, USA (e-mail: zx@ece.tamu.edu).

Corresponding author: Tingting Liu (e-mail: ttliu@buaa.edu.cn).

This work was supported by National Natural Science Foundation of China (NSFC) under Grants 61429101, 61731002 and 61671036.

ABSTRACT Future average channel gains are recently reported predictable within a minute-level horizon for a mobile user. Predictive resource allocation for non-realtime service with future average channel gains in a time window of one or more minutes long has been demonstrated effective in improving user experience and network throughput as well as reducing energy consumption. Nonetheless, existing studies for predictive resource allocation never consider inter-cell interference (ICI), which severely limits the user experience and network performance in densely-deployed cellular networks. This paper investigates predictive resource allocation in heterogeneous networks, where some base stations generate strong ICIs to some mobile users requesting non-realtime service. Optimizing the resource allocation in a predictive manner for interference networks is challenging, since how to allocate future resources depends on the future signal to interference-plus-noise ratio, which in turn relies on the assigned resources. To deal with this difficulty, we introduce a predictive interference coordination scheme to divide all BSs and users into groups, where the BS-user pairs in each group can communicate simultaneously in a second-level frame. Then, we optimize a common resource allocation plan for all users in each group. The plan essentially determines which BSs in the group should be muted and the users should be associated with which active BS in each frame of the prediction window. By resorting to graph theory, we obtain the optimal solution and derive a low-complexity algorithm. Simulation results show that the proposed scheme outperforms existing relevant methods in terms of user satisfactory rate in heterogeneous networks with heavy traffic load.

INDEX TERMS Predictive resource allocation, interference coordination, average channel gains, heterogeneous networks.

I. INTRODUCTION

THE prevalent and future cellular networks are heterogeneous, where low-power base stations (BSs) are deployed within the coverage of high-power BSs to offload traffic and improve user's quality of service (QoS) [1]. However, the reuse of spectrum resources and dense deployment of cells in heterogeneous networks (HetNets) cause severe inter-cell interference (ICI), which is a limiting factor to boost network throughput and improve user experience.

Many inter-cell interference coordination (ICIC) techniques have been proposed for HetNets. With time domain ICIC, the macro BSs are muted (i.e., do not transmit signals) in the almost blanking subframe [2]. With frequency domain ICIC, the macro BSs and pico BSs are assigned with different spectrum [3]–[6]. Noticing that user association is able to control the interference level by adjusting the traffic load [7], joint user association and spectrum partition problems were formulated for HetNets with different objectives and the optimal solutions were obtained in [4]–[6]. Considering that user association and resource allocation are coupled, the joint user association and power control was investigated in [8]–[10], and the joint user association and orthogonal resource allocation was studied in [11]–[14]. All existing methods along this line of research mitigate ICI in the time scale of millisecond (when small scale channels vary) or at most in the time scale of seconds (when user locations change), without considering the feature of services.

Wireless industry has recently witnessed an unprecedented growth of data traffic, which is expected to account for over 78% of global traffic by 2021. This type of non-realtime traffic such as video-on-demand or file downloading is delay-tolerant. The QoS of a user requesting such service will be satisfied if its requested video segment or file can be transmitted completely within an expected duration. On the other hand, the total amount of data that should be transmitted cannot exceed the number of bits in a requested segment or file. While radio resource allocation has been investigated extensively, such as the ICIC techniques for HetNets [2]–[14], and various methods proposed for relay systems, small cell networks, vehicular communication networks, and device-to-device communications [15]–[19], these traffic features are rarely taken into consideration.

Meanwhile, big data analysts reveals that human mobility is highly predictable [20]. In particular, trajectories of mobile users can be predicted with various machine learning techniques [21], [22]. By further combining with a radio map [23], which stores or learns the pathloss and shadowing in each location, predicting the average channel gains within a minute-level time window becomes possible [24]. With the predicted channel information, both the system performance such as throughput [25] and energy consumption [26]–[28] and the user experience for non-realtime service [29] can be improved dramatically by allocating radio resources in a predictive manner. The basic idea is to transmit to a mobile user when the user moves to the vicinity of a BS [26]. Such an idea can be realized by optimizing a predictive resource allocation plan for a user to convey the requested file before the expected deadline. The plan determines which BSs alongside the trajectory of a mobile user serve the user in which time to satisfy the user's QoS according to the predicted channel condition. In [26], the potential of predictive resource allocation is demonstrated by optimizing future time resource allocation with channel state information (CSI) (i.e., instantaneous channel gain) perfectly known over a minute-level time horizon. In practice, however, CSI is hard to predict beyond channel coherence time (in millisecond-level) if not impossible. Based on the average channel gains in a prediction window, resource allocation among future frames (in second-level, to be defined later) was developed in [27], [29], either to improve the QoS of each user or to minimize the energy consumption of the system. In [30], a predictive proportional fair scheduling was designed by exploiting the average channel prediction. In [28], a robust optimization framework was proposed to cope with the errors on the predicted data rate.

Different from the non-predictive resource allocation [2]–[19] that is optimized according to current instantaneous or average channel gain, predictive resource allocation is designed by assuming that future average channel gains or data rates within a time window can be obtained by prediction. As a consequence, predictive resource allocation can ensure the QoS of non-realtime user (i.e., transmit the requested file or video segment before the expected deadline) directly by optimizing a resource allocation plan at the start of the prediction window, before the real-time transmission at the start of each time slot with known CSI.

All existing works along this line of research do not take into account ICI (or treating ICI as noise), and consider simplified network topology (e.g., homogeneous network or even a single cell) [25]–[30]. When applied to HetNets where traffic load varies drastic among cells, some ICIs are strong that cannot be treated as noise, and hence the performance achieved by these methods will inevitably degrade. Intuitively, ICI can be controlled more flexibly by leveraging future average channel gains to improve QoS of mobile users with delay tolerant service, since the interference can be coordinated in a much larger time-space range. In the interference networks, however, transmitting to a user with good channel condition does not necessarily lead to high data rate. The resource allocation plan should be optimized based on the future information of signal to interference-plus-noise ratio (SINR) instead of the average channel gain of each user. On the other hand, the SINR depends on how the ICI is to be coordinated and the resource is to be allocated. Owing to such a "chicken-and-egg" problem, predictive resource allocation in interference networks is more complex than interference-free networks. So far, how to optimize the resource allocation plan in interference networks is an open problem.

In this paper, we investigate predictive resource allocation for mobile users requesting non-realtime service in HetNets with ICI. To coordinate interference, we allow a BS muting when the BS generating strong ICI. To deal with the "chicken-and-egg" challenge, we design the resource allocation plan in two steps at the start of a prediction window, when future average channel gains are available. We first optimize a predictive interference coordination scheme to find which BSs can transmit to which users concurrently that ensures a given average SINR. Then, we optimize a common resource allocation plan for this group of BS-user pairs that maximizes a network utility aiming to improve user satisfactory rate. Since the common plan is made for a group of users that request files with different sizes and are with different average SINRs, it is non-trivial to satisfy the QoS for all these users meanwhile not to waste radio resources. By resorting to graph theory, we obtain the optimal interference coordination scheme and the optimal resource allocation plan, and further derive a low-complexity algorithm.

The major contributions are summarized as follows:

- We exploit predicted average channel gains to optimize predictive resource allocation planning in heterogeneous interference networks. Existing works studying predictive resource allocation either do not consider interference at all [26], [28]–[30] or simply treat ICI as noise [25], [27]. In addition, existing works either consider time-division or frequency-division access to avoid multi-user interference (MUI). To our best knowledge, this is the first work to optimize predictive resource allocation with interference coordination, or with spatial-division multiple access.

- We establish a framework for optimizing predictive resource allocation plan in interference networks by first optimizing a predictive interference coordination scheme and then optimizing the resource allocation plan to maximize the user satisfactory rate. To cope with the difficulty in predictive interference coordination with spatial-division multiple access, we divide each BS into multiple virtual sub-BSs. To deal with the difficulty in making a single plan for multiple users, we introduce a logistic function to characterize the network utility that can make more users satisfied without wasting radio resources. Simulation results show that the proposed low-complexity algorithm is near-optimal, and provides much higher user satisfactory rate than existing predictive resource allocation when the traffic load is high and outperforms non-predictive resource allocation remarkably when the prediction window is long.

The remainder of the paper is organized as follows. Section II describes the system model. Section III introduces the predictive interference coordination and Section IV optimizes the resource allocation plan. Section V provides simulation results. Finally, we conclude this paper in Section VI.

Notations: Transposition and expectation are represented by $(\cdot)^T$ and $\mathbb{E}\{\cdot\}$, respectively. $|\mathcal{X}|$ denotes the cardinality of a set $\mathcal{X}$, $\mathcal{X} \setminus \mathcal{Y}$ denotes the set of elements in $\mathcal{X}$ but not in $\mathcal{Y}$, $\varnothing$ denotes an empty set, and $\|\mathbf{x}\|$ denotes the norm of a vector $\mathbf{x}$.

II. SYSTEM MODEL

Consider a downlink HetNet as shown in Figure 1, which consists of multiple-tier BSs (e.g., macro and pico BSs) serving mobile users over a bandwidth of W. A macro BS is equipped with more antennas and higher transmit power than a pico BS, and hence has a larger coverage. To simplify the notations, we do not differentiate the BSs in different tiers. Denote the index set of G BSs as $\mathcal{G} = \{1, \cdots, G\}$, and the index set of K single-antenna users as $\mathcal{K} = \{1, \cdots, K\}$. The gth BS is equipped with M_g antennas and with maximum transmit power P_g, $\forall g \in \mathcal{G}$. Denote the gth BS and the kth user as BS_g and UE_k, respectively.

Assume that the trajectory of each mobile user can be predicted in a time window. At the start of the prediction window, UE_k requests a file with size of B_k bits that needs to be transmitted within T, $k = 1, \cdots, K$. To obtain the performance gain from predictive resource allocation, it is no need to set the duration of the prediction window larger than T. On the other hand, the duration of the prediction window should be large enough such that the average channel gains including the pathloss and shadowing of a mobile user change significantly. If the predictable horizon is less than T, then multiple prediction windows are required to complete the file transmission. For simplicity, we assume that the duration of prediction window equals to T.

The prediction window is divided into N_f frames each with the interval of $T_f = T/N_f$, and each frame is divided into N_s time slots each with the interval of $T_s = T_f/N_s$, as shown in Figure 2(a). The average channel gain is assumed staying constant in each frame but may vary among different frames. The instantaneous channel gain (i.e., the CSI) changes independently among time slots.

From the predicted trajectories and with the help of a radio map, a center point (CP) can predict the average channel gains between all BSs and all users in each frame within the prediction window. To demonstrate the potential of predictive resource allocation in interference networks, we assume that the future average channel gains in the window are perfectly known. However, we only assume that CSI is perfect in current time slot but do not assume that future CSI is predictable. As a result, the minimal time unit for making a resource allocation plan is frame duration, as illustrated in Figure 2 (b).

A. RESOURCE ALLOCATION PLANNING

At the start of the prediction window, the CP first finds an interference coordination scheme and then makes a resource allocation plan for each user, both with the predicted average channel gains. Then, the CP informs the corresponding BSs that will serve the users about the plans. The plan determines which active BSs serve a user in which frames to meet the request of the user. From another perspective, we can say that the plan determines which BSs should be muted and the users are associated with which BSs in each frame.

To indicate which BSs are able to serve a particular user in each frame, we introduce a candidate BS set. In the lth frame, the candidate BS set of UE_k is denoted by $\mathcal{B}_k^l$, which contains several adjacent BSs who can provide higher average received signal powers to the user in the frame.

When multiple users are associated with the same BS in a frame (say the lth frame) according to the plans, these users constitute a set called candidate user set, denoted by $\mathcal{A}_g^l$. From the candidate BS set of UE_k in the lth frame (i.e., $\mathcal{B}_k^l$), we can obtain the candidate user set of BS_g in the frame as

$$\mathcal{A}_g^l = \left\{k \mid g \in \mathcal{B}_k^l,\ k \in \mathcal{K}\right\},\ \forall k \in \mathcal{K},\ l = 1, \cdots, N_f. \quad (1)$$

We use a binary variable to indicate which BS serves a user in a frame, called resource allocation planning variable, which is

$$x_{k,g}^l \in \{0, 1\},\ \forall g \in \mathcal{B}_k^l,\ k \in \mathcal{K},\ l = 1, \cdots, N_f. \quad (2)$$

When BS_g serves UE_k in the lth frame, $x_{k,g}^l = 1$, otherwise, $x_{k,g}^l = 0$.

We do not consider cooperative transmission among the BSs, so that no more than one BS can serve a user at the same time. Hence, the resource allocation planning variable satisfies

$$\sum_{g \in \mathcal{B}_k^l} x_{k,g}^l \leq 1,\ \forall k \in \mathcal{K},\ l = 1, \cdots, N_f. \quad (3)$$

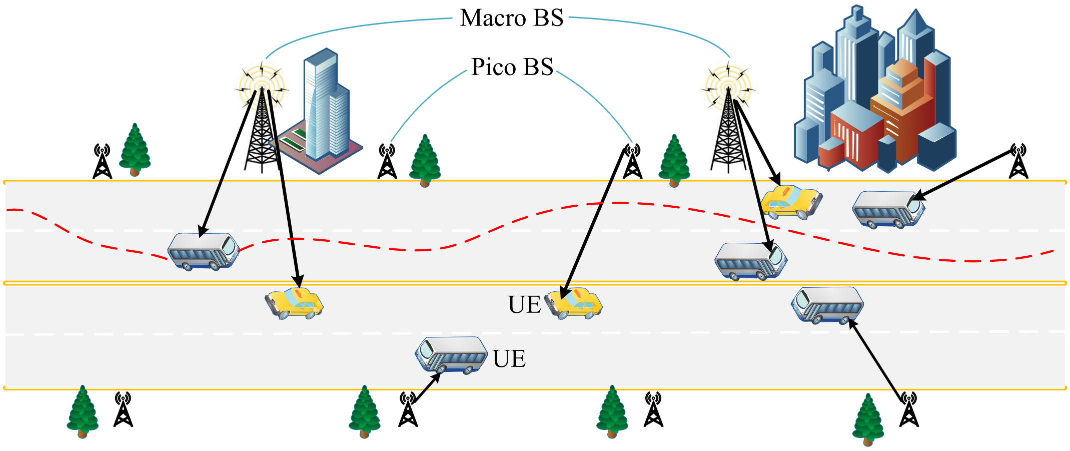
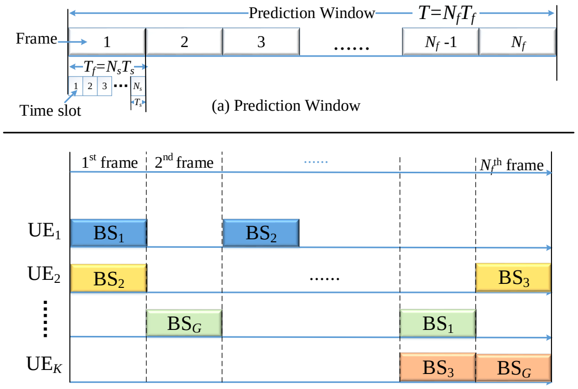


FIGURE 1. An example of downlink HetNet and user trajectory.



(a) Prediction Window

(b) Resource Allocation Plan

FIGURE 2. Prediction window and resource allocation plan.

B. TRANSMISSION SCHEME

At the start of each time slot, the instantaneous channel gains of the users accessed to a BS can be available at the BS by channel estimation. With the CSI, the BS transmits to the users according to the resource allocation plan using the following transmission scheme.

To exploit the antennas at each BS to serve multiple users, we consider multi-user multi-input-multi-output (MIMO) precoder to remove MUI. To simplify the analysis, we employ the zero-forcing (ZF) precoder. Since CSI is assumed perfect, if the number of selected users in each time slot of a frame (say the lth frame) does not exceed the number of antennas at the BS (say the gth BS), i.e.,

$$K_g^l = \sum_{k\in\mathcal{A}_g^l} x_{k,g}^l \le M_g,\ \forall g\in\mathcal{G},\ l=1,\cdots,N_f, \tag{4}$$

then MUI can be eliminated thoroughly.

We consider Rayleigh fading channel. Let $\boldsymbol{w}_{k,g}^{l,t}\in\mathbb{C}^{M_g\times 1}$ denote the ZF precoder of BS$_g$ for UE$_k$, which satisfies

$$\begin{cases} \|\boldsymbol{w}_{k,g}^{l,t}\|^2 = 1, \\ (\boldsymbol{h}_{k,g}^{l,t})^T\boldsymbol{w}_{j,g}^{l,t} = 0,\ \forall x_{j,g}=1,\ j\neq k \end{cases}, \tag{5}$$

where $\boldsymbol{h}_{k,g}^{l,t}\in\mathbb{C}^{M_g\times 1}$ is an M_g-length instantaneous channel vector from BS$_g$ to UE$_k$, which is the zero-mean Gaussian vector satisfying $\|\boldsymbol{h}_{k,g}^{l,t}\|^2 = M_g\alpha_{k,g}^l$, and $\alpha_{k,g}^l$ is the average channel gain in the lth frame.

To simplify the analysis, we consider equal power allocation among users in each frame. Then, the transmit power allocated to UE$_k$ by BS$_g$ in the lth frame is

$$P_{k,g}^l = \begin{cases} \frac{P_g}{K_g^l}, & \forall x_{k,g}^l = 1, \\ 0, & \text{otherwise}. \end{cases} \tag{6}$$

When UE$_k$ is served by BS$_g$ in the tth time slot of the lth frame, its received signal can be expressed as

$$\begin{aligned} y_k^{l,t} =& \sqrt{P_{k,g}^l}(\boldsymbol{h}_{k,g}^{l,t})^T\boldsymbol{w}_{k,g}^{l,t}s_k^{l,t} \\ &+ \sum_{i=1,i\neq g}^{G}\sum_{j\in\mathcal{A}_i^l} x_{j,i}^l\sqrt{P_{j,i}^l}(\boldsymbol{h}_{k,i}^{l,t})^T\boldsymbol{w}_{j,i}^{l,t}s_j^{l,t} + n_k^{l,t}, \end{aligned} \tag{7}$$

where $s_k^{l,t}$ is the symbol transmitted to UE$_k$, satisfying $\mathbb{E}\{s_k^{l,t}\} = 0$ and $\mathbb{E}\{|s_k^{l,t}|^2\} = 1$, $n_k^{l,t}$ is the Gaussian noise satisfying $\mathbb{E}\{n_k^{l,t}\} = 0$ and $\mathbb{E}\{|n_k^{l,t}|^2\} = \sigma_n^2$, and σ_n^2 is the noise power. The second term in (7) is ICI.

C. PERFORMANCE METRIC

The instantaneous SINR of UE$_k$ served by BS$_g$ in the tth time slot of the lth frame is

$$\gamma_{k,g}^{l,t} = \frac{P_g}{K_g^l}\frac{\left|(\boldsymbol{h}_{k,g}^{l,t})^T\boldsymbol{w}_{k,g}^{l,t}\right|^2}{I_{k,g}^{l,t}+\sigma_n^2}, \tag{8}$$

where the ICI power is

$$I_{k,g}^{l,t} = \sum_{i=1, i\neq g}^{G} \sum_{j\in\mathcal{A}_i^l} \frac{x_{j,i}^l P_i}{K_i^l} \left| (\boldsymbol{h}_{k,i}^{l,t})^T \boldsymbol{w}_{j,i}^{l,t} \right|^2 . \tag{9}$$

From (8), the instantaneous data rate in the tth time slot is

$$R_{k,g}^{l,t} = W \log_2 \left(1 + \frac{P_g}{K_g^l} \frac{\left| (\boldsymbol{h}_{k,g}^{l,t})^T \boldsymbol{w}_{k,g}^{l,t} \right|^2}{I_{k,g}^{l,t} + \sigma_n^2} \right). \tag{10}$$

The transmitted data of UE$_k$ in the lth frame is

$$D_{k,g}^l = T_s \sum_{t=1}^{N_s} R_{k,g}^{l,t} = \frac{T_f}{N_s} \sum_{t=1}^{N_s} R_{k,g}^{l,t}. \tag{11}$$

Then, the overall amount of data transmitted to UE$_k$ in the prediction window is $D_k = \sum_{l=1}^{N_f} \sum_{g=1}^{G} x_{k,g}^l D_{k,g}^l$.

To reflect the service quality for UE$_k$, we consider a completion ratio defined as

$$J_k = \frac{D_k}{B_k} = \sum_{l=1}^{N_f} \sum_{g=1}^{G} \frac{x_{k,g}^l D_{k,g}^l}{B_k}. \tag{12}$$

When $J_k = 1$, the demand of UE$_k$ is satisfied, i.e., the required B_k bits are transmitted to the user within duration T. Let $\rho_k = \mathbf{1}(J_k = 1)$ indicate whether or not UE$_k$ is satisfied, where $\mathbf{1}(x)$ is the indicator function. Then, the user satisfactory rate of the network is

$$\rho = \frac{1}{K} \sum_{k=1}^{K} \rho_k = \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}(J_k = 1), \tag{13}$$

which reflects the percentage of users, whose requested files can be completely conveyed before the expected transmission deadline, among all non-realtime users.

III. PREDICTIVE INTERFERENCE COORDINATION

In interference networks, the resource allocation planning variables should be optimized based on the future information of the SINR. From (8) and (9) we know that the SINR (and hence the amount of data transmitted to UE$_k$ in the lth frame, $D_{k,g}^l$, as well as J_k) and the resource allocation planning variables are coupled. As a consequence, when we make the resource allocation plan to maximize the user satisfactory rate of interference networks, the optimization is challenging.

To circumvent this difficulty, we introduce a predictive interference coordination scheme. The basic idea of the scheme is to avoid strong interference in each frame by BS muting and allow the BS-user pairs with weak interference among each other to transmit simultaneously in a frame with a satisfactory average QoS. After we find such a scheme, i.e., the BS-user pairs with weak interference, we can optimize the predictive resource allocation plan to these BS-user pairs to meet the user requirements.

To improve the performance of all users, we coordinate the interference to ensure that the SINR exceeds a threshold. Considering that only the average channel gains are available at the start of the prediction window, we use the average SINR in each frame $\bar{\gamma}_{k,g}^l \triangleq \mathbb{E}\{\gamma_{k,g}^{l,t}\}$ to reflect the QoS. The interference coordination scheme is designed to guarantee that the average SINR satisfies

$$\bar{\gamma}_{k,g}^l \geq \gamma_{\mathrm{T}}, \tag{14}$$

when BS$_g$ serves UE$_k$ in the lth frame.

Figure 3 illustrates the idea of predictive interference coordination scheme, where only strong ICIs are coordinated to meet the QoS constraint in (14). To make full use of resources, a BS can share the resource with the BSs who generate weak ICIs to the users served by the BS.

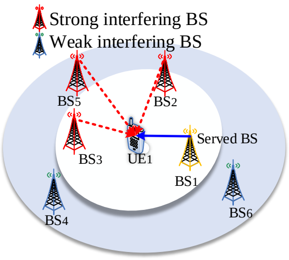


FIGURE 3. Predictive interference coordination

In this section, we design the prediction interference coordination scheme. We first study which interfering BSs should be muted for a user to meet the QoS requirement of the user. Then, we divide the BS-user pairs in the network into different groups, and find the BS-user pairs with weak interference, resorting to a graph-based method.

A. BS SET GENERATING STRONG ICI TO A UE

To coordinate strong ICIs to meet (14), we use an interfering BS set to indicate which interfering BSs should be muted when BS$_g$ serves UE$_k$ in the lth frame, denoted by $\mathcal{I}_{k,g}^l$.

To avoid the ICIs from the BSs in $\mathcal{I}_{k,g}^l$, the resource allocation planning variables need to satisfy

$$x_{k,g}^l + \sum_{j\in\mathcal{A}_i^l} x_{j,i}^l \leq 1, \ \forall\, i \in \mathcal{I}_{k,g}^l. \tag{15}$$

It indicates that when BS$_g$ serves UE$_k$ in the lth frame, the BSs in $\mathcal{I}_{k,g}^l$ should be muted in the same frame.

To derive the interfering BS set for a user, in what follows we derive the average SINR in each frame. After the strong ICIs are coordinated, the residual ICIs are weak, which can be approximated as Gaussian noise based on the Gaussian approximation in [31]. Then, we can approximate the residual ICI power as its average power in one frame. From (9), the average power of the residual ICIs generated to UE$_k$ in the lth frame can be expressed as

$$I_{k,g}^l = \mathbb{E}\{I_{k,g}^{l,t}\} = \sum_{i=1, i\neq g}^{G} \rho_i^l P_i \mathbb{E}\left\{ \left| (\boldsymbol{h}_{k,i}^{l,t})^T \boldsymbol{w}_{j,i}^{l,t} \right|^2 \right\}, \tag{16}$$

where $\rho_i^l = \mathbf{1}(\sum_{j\in\mathcal{A}_i^l} x_{j,i}^l > 0)$ is a binary variable denoting whether or not BS$_i$ is active on the lth frame.

Recall that the elements of $\boldsymbol{h}_{k,i}^{l,t}$ are Gaussian random variables with mean 0 and variance $\alpha_{k,i}^l$. With ZF precoder, the elements of $\boldsymbol{w}_{j,i}^{l,t}$ are independent of those of $\boldsymbol{h}_{k,i}^{l,t}$ and satisfy $\|\boldsymbol{w}_{j,i}^{l,t}\|^2 = 1$. Hence, $(\boldsymbol{h}_{k,i}^{l,t})^T\boldsymbol{w}_{j,i}^{l,t}$ is a Gaussian random variable with mean 0 and variance $\alpha_{k,i}^l$, i.e., $\mathbb{E}\{|(\boldsymbol{h}_{k,i}^{l,t})^T\boldsymbol{w}_{j,i}^{l,t}|^2\} = \alpha_{k,i}^l$. Upon substituting into (16), we have

$$I_{k,g}^l = \sum_{i=1,i\neq g}^{G} \rho_i^l P_i \alpha_{k,i}^l. \tag{17}$$

Since the strong ICIs are coordinated, i.e., the BSs in $\mathcal{I}_{k,g}^l$ mute, we have $\rho_i^l = 0,\ \forall i \in \mathcal{I}_{k,g}^l$. Besides, the other BSs only generate weak ICIs, whether they are active or not has little impact on $I_{k,g}^l$. Therefore, the average residual ICI power at UE$_k$ in the lth frame can be approximated as

$$I_{k,g}^l \approx \sum_{i=1,i\neq g}^{G} P_i\alpha_{k,i}^l - \sum_{i\in\mathcal{I}_{k,g}^l}^{G} P_i\alpha_{k,i}^l. \tag{18}$$

When computing the SINR with (8), the number of served users $K_g^l = \sum_{k\in\mathcal{A}_g^l} x_{k,g}^l$ is unknown. Here, we consider a conservative estimation, i.e., approximating K_g^l as its upper-bound $\hat{K}_g^l = \min\{M_g, |\mathcal{A}_g^l|\}$. Then, the SINR in the tth time slot of the lth frame is approximated as

$$\gamma_{k,g}^{l,t} \approx \frac{P_g}{\hat{K}_g^l}\frac{\left|(\boldsymbol{h}_{k,g}^{l,t})^T\boldsymbol{w}_{k,g}^{l,t}\right|^2}{I_{k,g}^l + \sigma_n^2}. \tag{19}$$

When the number of time slots in each frame is large, the elements of $\boldsymbol{h}_{k,g}^{l,t}$ over different time slots are ergodic. Since the residual interference is approximated as the Gaussian noise independent of $\boldsymbol{h}_{k,g}^{l,t}$, the instantaneous SINR is ergodic over these time slots. From the analysis in [32], the average SINR in the lth frame can be obtained as

$$\bar{\gamma}_{k,g}^l = \frac{P_g}{\hat{K}_g^l}\frac{\alpha_{k,g}^l(M_g - \hat{K}_g^l + 1)}{I_{k,g}^l + \sigma_n^2}. \tag{20}$$

By substituting (20) into (14), the power of residual ICI after interference coordination should meet

$$I_{k,g}^l \leq \frac{P_g\alpha_{k,g}^l(M_g - \hat{K}_g^l + 1)}{\gamma_{\mathrm{T}}\hat{K}_g^l} - \sigma_n^2. \tag{21}$$

Upon substituting into (18), the power of coordinated strong ICI should satisfy

$$\sum_{i\in\mathcal{I}_{k,g}^l}^{G} P_i\alpha_{k,i}^l \geq \sum_{i=1,i\neq g}^{G} P_i\alpha_{k,i}^l + \sigma_n^2 - \frac{P_g\alpha_{k,g}^l(M_g - \hat{K}_g^l + 1)}{\gamma_{\mathrm{T}}\hat{K}_g^l}. \tag{22}$$

The interfering BS set $\mathcal{I}_{k,g}^l$ satisfying (22) contains several strongest interfering BSs, which can be constructed in a recursive way. In particular, we first set $\mathcal{I}_{k,g}^l = \varnothing$ and $\mathcal{J}_{k,g}^l = \mathcal{G} \setminus \{g\}$ to initialize, and then find the BS with the maximal interference power in $\mathcal{J}_{k,g}^l$ and put it into $\mathcal{I}_{k,g}^l$, i.e.,

$$i = \arg\max_{j\in\mathcal{J}_{k,g}^l} \left\{P_j\alpha_{k,j}^l\right\}, \tag{23}$$
$$\mathcal{I}_{k,g}^l \leftarrow \mathcal{I}_{k,g}^l \cup \{i\},\ \mathcal{J}_{k,g}^l \leftarrow \mathcal{J}_{k,g}^l \setminus \{i\}.$$

By repeating (23) until (22) is satisfied, $\mathcal{I}_{k,g}^l$ is finally obtained.

B. BS-UE PAIRS WITH WEAK ICI

From an interfering BS set for UE$_k$ when the user is associated with BS$_g$ in the lth frame (i.e., $\mathcal{I}_{k,g}^l$), we know which BSs should be muted. According to the interfering BS sets of all users in the network, we can find strong ICIs existing among which BS-user pairs, i.e., conflicts exist among the pairs. Then, we can obtain the BS-user pairs with weak ICI in each frame, which are the group of BSs and users able to communicate concurrently in a frame to satisfy the QoS constraint in (14). Since a graph is easy to describe the conflict relationships among the BS-user pairs, we resort to graph theory [33] to find such group, i.e., the interference coordination scheme.

Since (3) and (4) are satisfied when we derive (19), if the constraints in (3), (4), and (15) with the interfering BS set obtained from (23) are satisfied, the constraint in (14) will be satisfied. We construct a graph to denote these constraints, which is called conflict graph.

1) Conflict Graph

Definition 1 (Conflict Graph): A conflict graph is an undirected graph, where the vertexes denote all possible BS-user pairs and the edges denote the conflicts among these BS-user pairs.

Let $\mathcal{C}^l = \left(\mathcal{V}^l, \mathcal{E}^l\right)$ denote the conflict graph in the lth frame, where $\mathcal{V}^l = \left\{(k,g)\ \middle|\ g\in\mathcal{B}_k^l,\ k\in\mathcal{K}\right\}$ is the vertex set, and $\mathcal{E}^l = \left\{((k,g),(j,i))\ \middle|\forall x_{k,g}^l + x_{j,i}^l \leq 1\right\}$ is the edge set. The vertex set can be obtained from the candidate BS sets of all users directly. In the following, we construct the edge set to represent the constraints in (3), (4), and (15).

The constraint in (3) is equivalent to

$$x_{k,g}^l + x_{k,i}^l \leq 1,\ \forall i \neq g,\ i\in\mathcal{B}_k^l, \tag{24}$$

which mean that if UE$_k$ is served by BS$_g$, the user cannot be served by other BS in the frame. Then, an edge connecting (k,g) and (k,i), $\forall i \neq g$ can denote the constraint, which is called BS conflict edge.

The constraint in (15) can be rewritten as

$$x_{k,g}^l + x_{j,i}^l \leq 1,\ \forall j\in\mathcal{A}_i^l,\ i\in\mathcal{I}_{k,g}^l, \tag{25}$$

which means that if BS$_g$ serves UE$_k$, BS$_i$ that is serving UE$_j$ should be muted to avoid generating strong ICI. An edge connecting (k,g) and (j,i), $\forall j \neq k,\ i\neq g,\ i\in\mathcal{I}_{k,g}^l$ can represent this constraint, which is called ICI coordination conflict edge.

However, the constraint in (4) can not be directly reflected in conflict graph except in two extreme systems. One is massive MIMO system and the other is single-antenna system, where each BS is equipped with a single antenna. In a massive MIMO system, since the number of candidate users is far less than the number of antennas, the BS is able to serve all users and the constraint (4) is unnecessary. While in a single-antenna system, since (4) reduces to $x^l_{k,g} + x^l_{j,g} \leq 1,\ \forall j \neq k$, an edge connecting (k, g) and (j, g) can denote the constraint.

To use the conflict edges to denote (4) for the systems with general antenna configurations, we divide a BS equipped with multiple antennas into multiple virtual sub-BSs each with a single antenna. Then, BS_g becomes M_g sub-BSs whose indexes are denoted as a set $\mathcal{G}_g = \{g_1, \cdots, g_{M_g}\}$. When the different sub-BSs from one BS serve different users, the interference among different sub-BS and user pairs is MUI rather than ICI. Since the MUI can be eliminated by the ZF precoder, there is no need to introduce any conflict edge between these sub-BS and user pairs. From the resource allocation planning variables of the sub-BSs, we can obtain the planning variable of the original BS as $x^l_{k,g} = \sum_{s\in\mathcal{G}_g} x^l_{k,s}$.

By dividing each BS into multiple virtual sub-BSs, the edges connecting (k, s) and (j, s), $\forall s \in \mathcal{G}_g$ can represent the constraint in (4). However, when dividing all BSs, the number of vertexes grows, which increases the complexity of using the conflict graph to obtain feasible solutions. Therefore, we should judge whether or not a BS needs to be divided. When $M_g \geq |\mathcal{A}^l_g|$, since (4) is always satisfied, it is unnecessary to divide these BSs. Then, we only need to divide the BSs in $\tilde{\mathcal{G}}^l = \{g \mid M_g < |\mathcal{A}^l_g|\}$. Therefore, (4) is equivalent to

$$x^l_{k,s} + x^l_{j,s} \leq 1,\ \forall j \neq k, s \in \mathcal{G}_g,\ g \in \tilde{\mathcal{G}}^l. \tag{26}$$

Then, the constraint is denoted by an edge connecting (k, s) and (j, s), $\forall j \neq k$, which is called user conflict edge.

For BS_g, $\forall g \in \tilde{\mathcal{G}}^l$, the BS index in the candidate BS set $\mathcal{B}^l_k$ and interfering BS set $\mathcal{I}^l_{k,g}$ should be replaced by the corresponding virtual sub-BS indexes, which are denoted by $\tilde{\mathcal{B}}^l_k$ and $\tilde{\mathcal{I}}^l_{k,g}$, respectively.

Finally, the conflict graph in the lth frame is constructed as $\mathcal{C}^l = \left(\mathcal{V}^l, \mathcal{E}^l\right)$, where $\mathcal{V}^l$ is

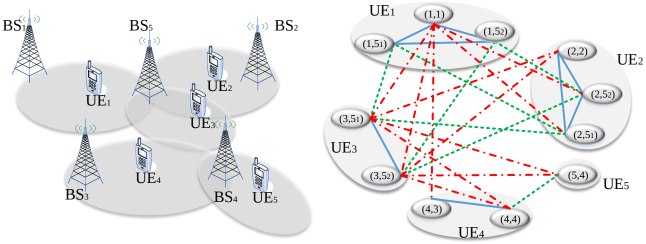
$$\mathcal{V}^l = \left\{(k, g) | g \in \tilde{\mathcal{B}}^l_k,\ k \in \mathcal{K}\right\}, \tag{27}$$

and $\mathcal{E}^l$ is obtained from (24), (25), and (26), which is

$$\mathcal{E}^l = \left\{ ((k,g),(j,i)) \,\middle|\, \begin{array}{l} j = k,\ i \neq g,\ g \in \tilde{\mathcal{B}}^l_k, \text{ or} \\ j \neq k,\ i = g,\ i \in \mathcal{G}_s, s \in \tilde{\mathcal{G}}^l, \text{ or} \\ j \neq k,\ i \neq g,\ i \in \tilde{\mathcal{I}}^l_{k,g} \end{array} \right\}. \tag{28}$$

2) An Example of Conflict Graph

In the following, we illustrate how to construct a conflict graph in each frame. We consider an example in Figure 4, where five BSs serve five users. $\text{BS}_1 \sim \text{BS}_4$ have one antenna and BS_5 has two antennas.



(a) Interference Network

(b) Conflict Graph

FIGURE 4. Interference network and conflict graph for one frame.

In Figure 4(a), from each ellipse region we can obtain the candidate BS sets. According to the BSs and users' locations, we can obtain the interfering BS sets. Both sets are listed in Table 1, where $\mathcal{I}_{3,5}$ indicates the interfering BS set when BS_5 serves UE_3 in the lth frame (we remove the superscript that indicating the frame index l for simplicity).

TABLE 1. Original candidate and interfering BS sets of different UEs

UE_k	Candidate BS set	Interfering BS set
$k = 1$	$\mathcal{B}_1 = \{1, 5\}$	$\mathcal{I}_{1,1} = \{3, 5\}$, $\mathcal{I}_{1,5} = \{1\}$
$k = 2$	$\mathcal{B}_2 = \{2, 5\}$	$\mathcal{I}_{2,2} = \{5\}$, $\mathcal{I}_{2,5} = \{2\}$
$k = 3$	$\mathcal{B}_3 = \{5\}$	$\mathcal{I}_{3,5} = \{2, 4\}$
$k = 4$	$\mathcal{B}_4 = \{3, 4\}$	$\mathcal{I}_{4,3} = \varnothing$, $\mathcal{I}_{4,4} = \{3\}$
$k = 5$	$\mathcal{B}_5 = \{4\}$	$\mathcal{I}_{5,4} = \varnothing$

From the candidate BS sets of all users in a frame, we obtain the candidate user sets of different BSs in the frame and list them in Table 2 (again we remove the superscript l).

TABLE 2. Candidate UE sets of different BSs

BS_g	Num. of antennas	Candidate user set	Num. of users
$g = 1$	$M_1 = 1$	$\mathcal{A}_1 = \{1\}$	$\lvert\mathcal{A}_1\rvert = 1$
$g = 2$	$M_2 = 1$	$\mathcal{A}_2 = \{2\}$	$\lvert\mathcal{A}_2\rvert = 1$
$g = 3$	$M_3 = 1$	$\mathcal{A}_3 = \{4\}$	$\lvert\mathcal{A}_3\rvert = 1$
$g = 4$	$M_4 = 1$	$\mathcal{A}_4 = \{4, 5\}$	$\lvert\mathcal{A}_4\rvert = 2$
$g = 5$	$M_5 = 2$	$\mathcal{A}_5 = \{1, 2, 3\}$	$\lvert\mathcal{A}_5\rvert = 3$

Since $M_5 > 1$ and $M_5 < |\mathcal{A}_5|$, it is necessary to divide BS_5 into $M_5 = 2$ virtual sub-BSs, denoted by BS_{5_1} and BS_{5_2}. Then, the final candidate BS and interfering BS sets are obtained as shown in Table 3, where $\mathcal{I}_{3,5_1}$ indicates the interfering BS set when BS_{5_1} (i.e., the first antenna of BS_5) serves UE_3.

TABLE 3. Final candidate and interfering BS sets of different UEs

UE_k	Candidate BS set	Interfering BS set
$k = 1$	$\tilde{\mathcal{B}}_1 = \{1, 5_1, 5_2\}$	$\tilde{\mathcal{I}}_{1,1} = \{3, 5_1, 5_2\}$, $\tilde{\mathcal{I}}_{1,5_1} = \tilde{\mathcal{I}}_{1,5_2} = \{1\}$
$k = 2$	$\tilde{\mathcal{B}}_2 = \{2, 5_1, 5_2\}$	$\tilde{\mathcal{I}}_{2,2} = \{5_1, 5_2\}$, $\tilde{\mathcal{I}}_{2,5_1} = \tilde{\mathcal{I}}_{2,5_2} = \{2\}$
$k = 3$	$\tilde{\mathcal{B}}_3 = \{5_1, 5_2\}$	$\tilde{\mathcal{I}}_{3,5_1} = \tilde{\mathcal{I}}_{3,5_2} = \{2, 4\}$
$k = 4$	$\tilde{\mathcal{B}}_4 = \{3, 4\}$	$\tilde{\mathcal{I}}_{4,3} = \varnothing$, $\tilde{\mathcal{I}}_{4,4} = \{3\}$
$k = 5$	$\tilde{\mathcal{B}}_5 = \{4\}$	$\tilde{\mathcal{I}}_{5,4} = \varnothing$

According to (27), we obtain all vertexes of the conflict graph, which are $(1,1)$, $(1,5_1)$, $(1,5_2)$, $(2,2)$, $(2,5_1)$, $(2,5_2)$, $(3,5_1)$, $(3,5_2)$, $(4,3)$, $(4,4)$, and $(5,4)$.

From Table 3, we can see that BS$_3$ or BS$_4$ can serve UE$_4$, but only one BS can serve the user in the frame. Therefore, we plot a solid line between $(4,3)$ and $(4,4)$ to denote a BS conflict edge. BS$_4$ can serve UE$_4$ or UE$_5$, but only serves one user. Hence, we plot a dashed line between $(4,4)$ and $(5,4)$ to denote a user conflict edge. Moreover, when BS$_3$ and BS$_1$ serve UE$_4$ and UE$_1$, respectively, UE$_1$ suffers from the strong ICI from BS$_3$. To denote the ICI, we plot a dot-dash line between $(4,3)$ and $(1,1)$. Following the same way, we obtain all BS, user, and ICI coordination conflict edges. The resulting conflict graph is shown in Figure 4(b), where each small ellipse indicates a vertex.

3) Independent Set

After constructing the conflict graph, we can obtain the possible interference coordination results, which are the independent sets of the conflict graph.

Definition 2 (Independent Set): Given a graph $\mathcal{C} = (\mathcal{V}, \mathcal{E})$, an independent set is a subset of vertexes $\mathcal{D} \in \mathcal{V}$, such that no two vertexes in $\mathcal{D}$ are connected, denoted by $\mathcal{D} \in \mathrm{IS}(\mathcal{C})$. A maximal independent set (MIS) is an independent set that is not a subset of any other independent set [33].

For the conflict graph $\mathcal{C}$, we can obtain all its independent sets with standard tools of graph theory [33]. Taking the conflict graph in Figure 4 as an example, there are many different independent sets. Due to the lack of space, we only list some independent sets as follows,

$$\begin{aligned}
&\mathcal{D}_1 = \{(1,1),(2,2),(4,4)\}\\
&\mathcal{D}_2 = \{(1,5_1),(2,5_2),(4,3),(5,4)\},\\
&\mathcal{D}_3 = \{(1,5_1),(2,5_2),(4,4)\},\\
&\mathcal{D}_4 = \{(1,5_1),(3,5_2),(4,3)\},\\
&\mathcal{D}_5 = \{(2,5_1),(3,5_2),(4,3),(5,4)\},\\
&\quad\cdots
\end{aligned}$$

From the independent set, we know which BS-user pairs can use the same frame after muting the BSs that generate strong interference. Let $\mathcal{S}^l \in \mathrm{IS}(\mathcal{C}^l)$ denote the chosen independent set in the lth frame.

Then, the resource allocation planning variables for the users capable to be served by the BSs that belong to the independent set in the lth frame can be set as identical, i.e.,

$$x_{k,g}^l = \begin{cases} 1, & \forall (k,g) \in \mathcal{S}^l, \\ 0, & \text{otherwise}. \end{cases} \tag{29}$$

This indicates that the resource (i.e., each frame) can be allocated to a group of users rather than a single user.

IV. PREDICTIVE RESOURCE ALLOCATION PLANNING

In this section, we optimize the resource allocation plan to the BS-user groups obtained from the independent sets, based on the average channel gains in the prediction window. We first formulate the optimization problem to maximize a network utility to improve user satisfactory rate, and then find the globally optimal solution. Finally, we develop a low-complexity algorithm.

A. PROBLEM FORMULATION

The user satisfactory rate defined in (13) can be applied for evaluating the performance of the proposed solution, but is not appropriate for serving as an optimization objective for resource allocation. This is because ρ_k is a step function of the completion ratio J_k, which cannot reflect the contribution of the allocated resources to J_k when a file has not been completely transmitted.

To properly adjust the allocated resources to UE$_k$ with different values of J_k, we take a logistic function with a "S" shape, $\mathrm{U}(J_k)$, as the performance metric of each user, where

$$\mathrm{U}(x) = \frac{1}{1+\exp(-a(x-x_0))}, \tag{30}$$

and $a > 0$ controls the steepness of the curve, and x_0 is the x-value of the sigmoid's midpoint, as shown in Figure 5. When $a = 1$ and $x_0 = 0$, $\mathrm{U}(J_k)$ is the standard sigmoid function. When $a \to \infty$ and $x_0 = 1$, $\mathrm{U}(J_k)$ is equal to ρ_k, which indicates whether or not UE$_k$ is satisfied.

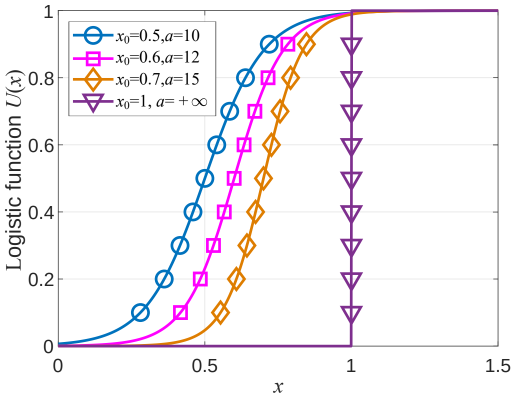


FIGURE 5. Logistic functions with different parameters.

$\mathrm{U}(x)$ is a monotonically increasing function from 0 to 1. The growth is approximately exponential in the initial stage, then slows down in the saturation stage, and stops in the maturity stage. The rationale to select such a function as the weight function is explained as follows.

In predictive resource allocation for the users requesting files, a common way to avoid assigning unnecessary frames to UE$_k$ is to impose a constraint of $D_k \leq B_k$. However, our analysis in Section III-B3 indicates that the resource can be assigned to a group of users at the same time rather than a single user after the interference coordination. When the file transmission of some users in the group are completed but those of others are not, such a constraint may cause that the files of other users cannot be completely transmitted. As shown in Figure 5, the considered logistic function can adjust its growth according to the completion ratio. When J_k increases and approaches to one, the growth slows down and finally stops. With such function, more resources will be allocated to a user whose file transmission will be completed soon (say a cell-central user) and less resources will be allocated to a user who has little hope to satisfy. Besides, instead of using a simple exponential function, the saturation stage of the function allows more users to satisfy, since the weight for an almost-completed user is already very large. By choosing appropriate parameters of the function, we can reduce the waste of resources by not allocating frames to the users who have been satisfied, without imposing the constraint of $D_k \leq B_k$.

Then, the network utility is defined as

$$\begin{aligned} \mathcal{U} &= \sum_{k=1}^{K} \mathrm{U}(J_k) \\ &= \sum_{k=1}^{K} \mathrm{U}\left(\sum_{l=1}^{N_f} \sum_{g=1}^{G} \frac{x_{k,g}^{l} \hat{D}_{k,g}^{l}}{B_k} \right), \end{aligned} \tag{31}$$

where $\hat{D}_{k,g}^{l}$ is the amount of data transmitted to UE$_k$ by BS$_g$ in the lth frame, which can be obtained from (11) as

$$\hat{D}_{k,g}^{l} = T_f \hat{R}_{k,g}^{l}, \tag{32}$$

where $\hat{R}_{k,g}^{l} = 1/N_s \sum_{t=1}^{N_s} \hat{R}_{k,g}^{l,t}$ is the time-average rate in the lth frame after the predictive interference coordination.

To obtain the achieved data rate of UE$_k$ when served by BS$_g$ in the tth time slot of the jth frame after the interference coordination, $\hat{R}_{k,g}^{l,t}$, we need to derive the corresponding SINR. Recall that (21) is derived from (14), which is the target of the ICI coordination. When the equality of (14) holds, (21) is satisfied after the coordination. By substituting the right-hand side of (21) into (19), the SINR achieved after the interference coordination becomes

$$\gamma_{k,g}^{l,t} \approx \frac{\gamma_{\mathrm{T}} \left| (\boldsymbol{h}_{k,g}^{l,t})^T \boldsymbol{w}_{k,g}^{l,t} \right|^2}{\alpha_{k,g}^{l} (M_g - \hat{K}_g^l + 1)}, \tag{33}$$

which follows Gamma distribution with shape $m = M_g - \hat{K}_g^l + 1$ and scale $\theta = \gamma_{\mathrm{T}} / (M_g - \hat{K}_g^l + 1)$ [34].

Then, from (33) we have

$$\hat{R}_{k,g}^{l,t} = W \log_2 \left(1 + \frac{\gamma_{\mathrm{T}} \left| (\boldsymbol{h}_{k,g}^{l,t})^T \boldsymbol{w}_{k,g}^{l,t} \right|^2}{\alpha_{k,g}^{l} (M_g - \hat{K}_g^l + 1)} \right), \tag{34}$$

which depends on the CSI that is unavailable at the start of the prediction window.

To make the resource allocation plan with future average channel gains, we need to derive the average data rate. When the number of time slots in each frame is large and the instantaneous SINR is ergodic over the time slots, we have

$$\begin{aligned} \lim_{N_s \to \infty} \hat{R}_{k,g}^{l} = \bar{R}_{k,g}^{l} &= W \mathbb{E} \left\{ \log_2 \left(1 + \gamma_{k,g}^{l,t} \right) \right\} \\ &= W \int_0^{\infty} \log_2 (1 + x) f_{\gamma}(x) dx, \end{aligned} \tag{35}$$

where $f_{\gamma}(x)$ is the probability distribution function (PDF) of the SINR $\gamma_{k,g}^{l,t}$. The analysis in [27] indicates that when $N_s \geq 100$, $\hat{R}_{k,g}^{l} \approx \bar{R}_{k,g}^{l}$, which is accurate.

From the analysis in [34], for Rayleigh fading channels, the ensemble-average data rate $\bar{R}_{k,g}^{l}$ can be obtained as

$$\bar{R}_{k,g}^{l} = \frac{W}{\ln 2 \Gamma(m) \theta^m} \cdot \mathbb{G}_{2,3}^{3,1} \left(\begin{array}{c} -m, -m+1 \\ -m, -m, 0 \end{array} \middle| \frac{1}{\theta} \right), \tag{36}$$

where $\Gamma(x)$ is the Gamma function and $\mathbb{G}(x)$ is the Meijer G-function. Recalling the expressions of m and θ below (33), $\bar{R}_{k,g}^{l}$ in (36) only depends on the average channel gain.

By replacing $\hat{R}_{k,g}^{l}$ in (32) with $\bar{R}_{k,g}^{l}$, we have $\bar{D}_{k,g}^{l} = T_f \bar{R}_{k,g}^{l}$. After the interference coordination, we can make the resource allocation plan to the BS-user pairs that can communicate at the same time. Such a plan determines which BSs in the network transmit signals to which users in a frame. This amounts to find the independent sets that maximize the network utility in all frames in the window. Then, the optimization problem can be formulated as follows,

$$\begin{aligned} \mathbf{P1}: \ \max_{\mathcal{S}^l} \mathcal{U} &= \sum_{k=1}^{K} \mathrm{U}\left(\sum_{l=1}^{N_f} \sum_{g=1}^{G} \frac{x_{k,g}^{l} \bar{D}_{k,g}^{l}}{B_k} \right) \\ \text{s.t. } & \mathcal{S}^l \in \mathrm{IS}(\mathcal{C}^l), \ (29), \ l = 1, \cdots, N_f. \end{aligned}$$

B. MAXIMAL INDEPENDENT SET AND OPTIMAL SOLUTION

Given two resource allocation solutions $\mathcal{S}_1 = \{\mathcal{S}_1^1, \cdots, \mathcal{S}_1^{N_f}\}$ and $\mathcal{S}_2 = \{\mathcal{S}_2^1, \cdots, \mathcal{S}_2^{N_f}\}$, let $\mathcal{U}(\mathcal{S}_1)$ and $\mathcal{U}(\mathcal{S}_2)$ denote the network utility of $\mathcal{S}_1$ and $\mathcal{S}_1$, respectively. Considering that $\mathrm{U}(x)$ is a monotonically increasing function, it is not hard to show that if $\mathcal{S}_1^l \subseteq \mathcal{S}_2^l$, $l = 1, \cdots, N_f$, then

$$\mathcal{U}(\mathcal{S}_1) \leq \mathcal{U}(\mathcal{S}_2). \tag{37}$$

Thus, we can obtain the optimal solution from the MISs of the conflict graphs, since an MIS is an independent set that is not a subset of any other independent set. This indicates that the globally optimal solution of problem **P1**, denoted as $\mathcal{S} = \{\mathcal{S}^1, \cdots, \mathcal{S}^{N_f}\}$, satisfies

$$\mathcal{S}^l \in \mathrm{MIS}(\mathcal{C}^l), \ l = 1, \cdots, N_f, \tag{38}$$

where $\mathrm{MIS}(\mathcal{C}^l)$ denotes a set consisting of all MISs of $\mathcal{C}^l$.

Consequently, the optimal solution $\mathcal{S}$ can be obtained by first finding $\mathrm{MIS}(\mathcal{C}^l)$ in each frame and then exhaustively searching the combination of MISs in N_f frames to maximize $\mathcal{U}(\mathcal{S})$. By finding the optimal solution from the MISs rather than all independent sets, the computational complexity can be reduced. Although finding MISs is still NP-hard, there exist efficient algorithms to find the MISs, such as [35].

With the solution $\mathcal{S}$, we can find the optimal resource allocation planning variables $x_{k,g}^{l}, k = 1, \cdots, K, g = 1, \cdots, G, l = 1, \cdots, N_f$ from (29), after removing the redundant frames assigned to a user that are unnecessary.

To help understand the optimal resource allocation plan, we again consider the example in Section III-B2 for illustration. We set the number of frames in the prediction window as $N_f = 3$. For easy exposition, we assume that the conflict graphs of all frames in the prediction window are same. Each user requires a file with size of $B_k = 2$ megabytes (MB), and the transmitted data after interference coordination in each frame is $\hat{D}_{k,g} = 1$ MB. Therefore, it is necessary to assign each user two frames. By exhaustively searching the MISs that maximizes $\mathcal{U}$, we obtain the optimal solution $\mathcal{S}$ as

$$\mathcal{S}^1 = \mathcal{D}_2 = \{(1,5_1),(2,5_2),(4,3),(5,4)\},$$
$$\mathcal{S}^2 = \mathcal{D}_4 = \{(1,5_2),(3,5_1),(4,3)\},$$
$$\mathcal{S}^3 = \mathcal{D}_5 = \{(2,5_1),(3,5_2),(4,3),(5,4)\}.$$

It means that the first, second, and third frames are assigned to the BS-user pairs in the independent sets $\mathcal{D}_2$, $\mathcal{D}_4$ and $\mathcal{D}_5$. We can see that UE$_1$, UE$_2$, UE$_3$, and UE$_5$ are assigned with two frames, while UE$_4$ are assigned with three frames. Since two frames are enough for each user, it is not necessary to assign the third frame to UE$_4$. Therefore, the optimal resource allocation planning variables in the three frames are

$$x^1_{1,5} = x^1_{2,5} = x^1_{4,3} = x^1_{5,4} = 1,$$
$$x^2_{1,5} = x^2_{3,5} = x^2_{4,3} = 1,$$
$$x^3_{2,5} = x^3_{3,5} = x^3_{5,4} = 1.$$

C. LOW-COMPLEXITY ALGORITHM

To obtain the globally optimal solution of problem **P1**, we need to first find all MISs of N_f conflict graphs and then find the combinations of the MISs. Letting z^l denote the number of MISs in $\mathcal{C}^l$, the searching space is $\mathcal{O}\left(\prod_{l=1}^{N_f} z^l\right)$. As the number of frames N_f increases, the complexity becomes unacceptable.

In what follows, we develop a low-complexity algorithm, which is inspired by an alternating optimization algorithm in [30]. For each frame, given the assignment results of other frames, we can find the optimal assignment in the frame. By searching iteratively, we can obtain a suboptimal solution.

When assigning the resources in the lth frame, we aim to maximize the network utility

$$\mathcal{U}^l = \sum_{k=1}^{K} \mathrm{U}\left(\frac{x^l_{k,g}\bar{D}^l_{k,g}}{B_k} + \eta^l_k\right), \tag{39}$$

where $\eta^l_k = \sum_{\tau=1,\tau\neq l}^{l-1}\sum_{i=1}^{G} x^\tau_{k,i}\bar{D}^\tau_{k,i}/B_k$ is the total completion ratio in other frames.

Considering that $x^l_{k,g}$ is a binary variable, it is not difficult to rewrite (39) as

$$\mathcal{U}^l = \sum_{k=1}^{K} x^l_{k,g}\Delta U^l_{k,g} + \mathrm{U}\left(\eta^l_k\right), \tag{40}$$

where

$$\Delta U^l_{k,g} = \mathrm{U}\left(\frac{\bar{D}^l_{k,g}}{B_k} + \eta^l_k\right) - \mathrm{U}(\eta^l_k) \tag{41}$$

is the utility increment when $x^l_{k,g} = 1$.

In (40), since $\mathrm{U}(\eta^l_k)$ is independent of the planning variables in the lth frame, we can maximize $\sum_{k=1}^{K}\sum_{g=1}^{G} x^l_{k,g}\Delta U^l_{k,g}$ instead of maximizing $\mathcal{U}^l$. As a result, we can formulate the problem as

$$\mathbf{P2}: \max_{\mathcal{S}^l} \sum_{k=1}^{K}\sum_{g=1}^{G} x^l_{k,g}\Delta U^l_{k,g}$$
$$\text{s.t. } \mathcal{S}^l \in \mathrm{MIS}(\mathcal{C}^l), (29).$$

In graph theory [33], a maximal weighted independent set (MWIS) is an independent set with maximum total weight. By setting the weight of vertex (k,g) in the conflict graph $\mathcal{C}^l$ as $\Delta U^l_{k,g}$, the optimal solution of problem **P2** can be obtained from the MWIS of $\mathcal{C}^l$. According to the near-optimal solution to find MWIS in [36], we obtain a low-complexity algorithm with details shown in Algorithm 1.

Algorithm 1 Low-complexity Algorithm

```
Step 1:Initialization
    Step 1.1: Construct the conflict graph in all frames $\mathcal{C}^l = \left(\mathcal{V}^l, \mathcal{E}^l\right)$, $l = 1, \cdots, N_f$, according to (27) and (28).
    Step 1.2: Choose a MIS of $\mathcal{C}^l$ denoted as $\mathcal{S}^l$ and set the initial value of $x^l_{k,g}$, $l = 1, \cdots, N_f$.
Step 2: Find resource allocation plan
For $l = 1 : N_f$
    Step 2.1: Set the weight of each vertex in $\mathcal{C}^l$ as $\Delta U^l_{k,g}$ in (41).
    Step 2.2: Solve the MWIS of $\mathcal{C}^l$ denoted as $\mathcal{S}^l$.
    Step 2.3: Update the planning variables $x^l_{k,g}$ from $\mathcal{S}^l$.
Return Step 2 until the network utility stops increasing.
```

Since the computational complexity of solving MWIS by the method in [36] is $\mathcal{O}(n^3)$, where n is the number of vertexes in graph, the complexity of Algorithm 1 is

$$\mathcal{O}\left(b\sum_{l=1}^{N_f}|\mathcal{V}^l|^3\right) \leq \mathcal{O}\left(b\sum_{l=1}^{N_f}\left(\sum_{g=1}^{G}|\mathcal{A}^l_g|M_g\right)^3\right). \tag{42}$$

where b is the number of iterations.

V. SIMULATION RESULTS

In this section, we evaluate the performance of the proposed solution by simulation.

Consider a HetNet, where N_m macro BSs and N_p pico BSs respectively equipped with M_m and M_p antennas serve mobile users. Macro BSs are located along one side of a straight road, with inter-site distance of 500 m. Pico BSs are randomly located along both sides of the road, with the minimal inter-site distance of 80 m. The distance from each macro BS to the road is 120 m, and the distance from the pico BSs to the road is uniformly distributed from 40 m to 60 m. The users move along the road with random speeds of 50 km/h $\sim$ 60 km/h. At the beginning of a prediction window, each of K single-antenna users starts to request a file with size of B MBytes in a random location on the road. In each frame, three BSs providing the highest signal powers to a user belong to the candidate BS set of the user. The road is with length of L m. To avoid the edge effect, a user arriving the end of the road will re-enter the road from the other side. Other system parameters are set according to Third Generation Partnership Project (3GPP) specifications in [37], which are summarized in Table 4. All results are obtained by averaging over 500 simulation trails. In each trail, the locations of pico BSs are random, the locations where the users initiate the requests are random, the moving speed and direction of each user are random, the shadowing is random according to log-normal distribution with 50 m coherence distance, and small scale channels among time slots are randomly generated according to Rayleigh distribution.

TABLE 4. Simulation Setup

Parameters	Values	
System bandwidth	$W = 10$ MHz	
Duration of the prediction window	60 s	
Duration of a frame	1 s	
Duration of a time slot	10 ms	
Noise power	-95 dBm	
	Macro BS	Pico BS
Transmission power	46 dBm	30 dBm
Path loss at 1 meter	15.3 dB	30.6 dB
Path loss exponent	37.6	36.7
Standard deviation of shadowing	8 dB	10 dB

A. ACCURACY OF THE APPROXIMATION

In Figure 6, we provide the cumulative distribution function (CDF) of the average data rate in each frame numerically obtained with (36) and from simulation to evaluate the accuracy of the approximated average data rate.

We can see that the approximation is accurate when $\gamma_T = 15$ dB or 25 dB and is less accurate when $\gamma_T = 5$ dB. This is because when deriving the approximated rate, we conservatively estimate K_g^l as its maximal value, so that the approximated data rate is a lower bound of the real data rate. Besides, when γ_T is large, only a few BSs in the network are active. Then, more users are severed by an active BS, and the active BSs need to apply full multiplexing, hence the conservative estimation of K_g^l is accurate in this case.

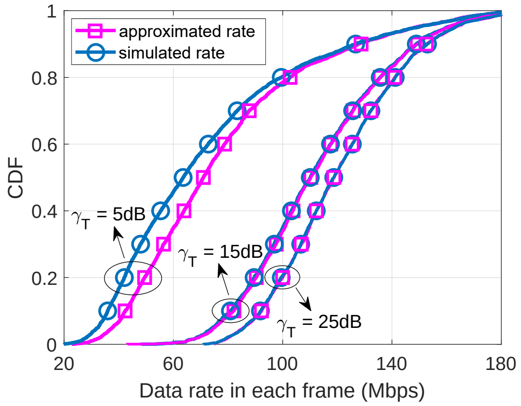


FIGURE 6. Accuracy of approximated average rate in each frame.

B. PERFORMANCE LOSS OF THE LOW-COMPLEXITY ALGORITHM

In Figure 7, we show the performance gap between the low-complexity algorithm and the optimal solution of problem **P1** in a small-scale network, where we employ the user satisfactory rate of the network (i.e., ρ in (13)) to evaluate the performance.

The network parameters are $N_m = 1$, $N_p = 10$, $M_m = 2$, $M_p = 1$, $K = 5$, $B = 30$ MB, $L = 500$ m, $T = 5$ s and $\gamma_T = 25$ dB. The related parameters in the proposed low-complexity algorithm are $x_0 = 0.7$ and $a = 15$, respectively. We can see that the low-complexity algorithm performs closely to the optimal solution.

C. IMPACT OF KEY PARAMETERS

To show the impact of the key parameters and evaluate the performance of the proposed solution by simulating the network with a more practical scale, we only consider the low-complexity algorithm in the rest of this section.

Unless otherwise specified, in the sequel we consider the following network setup: $N_m = 4$, $N_p = 30$, $M_m = 4$, $M_p = 2$, $K = 30$, $B = 150$ MB, $L = 1500$ m, $T = 60$ s and $\gamma_T = 25$ dB, the related parameters in the low-complexity algorithm are $x_0 = 0.7$ and $a = 15$, respectively.

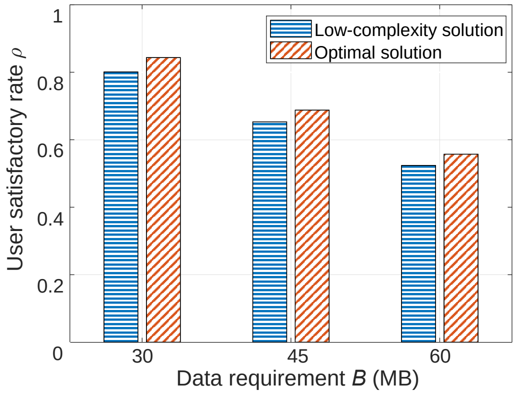


FIGURE 7. User satisfactory rate of optimal and low-complexity solutions.

In Figure 8, we show the impact of the two parameters used in the utility function on the behavior of the resource allocation planning. We provide the complementary cumulative distribution function (CCDF) of completion ratio J_k achieved with different values of a and x_0, recalling that the network utility depends on the parameters of logistic function. The user satisfactory ratio can be obtained from the value of CCDF at the point of the curve where J_k equals to one at the first time. We can see that when a and x_0 increase, user satisfactory rate grows. Besides, for larger values of a and x_0, more users are with $J_k = 0$ or $J_k = 1$. This indicates that the resource allocation planning is more aggressive. That is, if a user will not be satisfied at the end of the prediction window due to not experiencing good channel condition, almost no resource is allocated to the user. On the other hand, if a user is more likely to be satisfied, then more resource is allocated to the user.

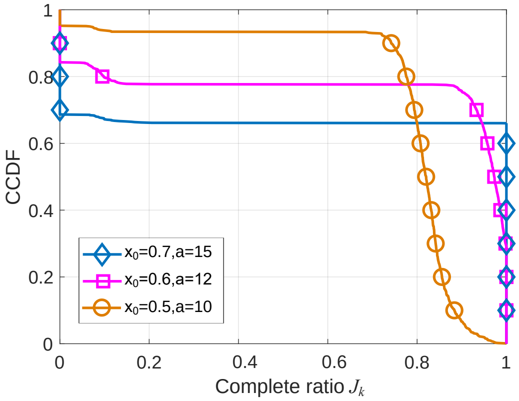


FIGURE 8. Impact of parameters a and x_0 on the completion ratio J_k.

In Figure 9, we show the impact of the SINR threshold γ_T on the performance, ρ. We can see that the value of ρ first increases and then decreases with γ_T, i.e., there exists an optimal SINR threshold to maximize the user satisfactory rate. This is because the SINR threshold has a two-fold impact on user satisfactory rate. For large value of γ_T, most BSs are muted to avoid ICI, so that fewer BS-UE pairs can be transmitted concurrently and hence fewer users in different cells can be served in a frame. On the other hand, because most ICI are avoided through BS muting, the achieved SINR and thus the achievable rate of each user are high. This suggests a tradeoff between the resource reuse and the achieved data rate. Hence, there exists the best value of γ_T leading to the maximal user satisfactory rate.

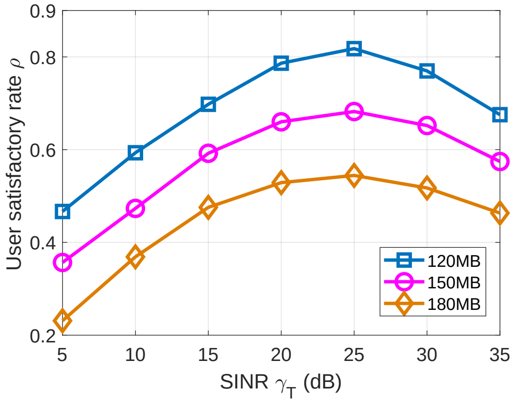


FIGURE 9. Impact of the SINR threshold γ_T.

D. COMPARISON WITH EXISTING SOLUTIONS

Finally, we compare the proposed low complexity algorithm (with legend “Pred RA (w/ ICIC)”) with existing solutions.

Because there are no methods in the literature that have the same objective function, optimization variables, and constraints as ours, we modify two most relevant methods proposed in [14], [25] for a fair comparison. To show the impact of not considering ICI or not harnessing predicted average channel gains, we compare with three reference resource allocation strategies:

- Predictive resource allocation without ICI coordination (with legend “Pred RA (w/o ICIC)”): This strategy can be regarded as a revised version of an existing predictive resource allocation strategy in [25], which also exploits the predicted average channel gains in the prediction window but simply treats ICI as noise and does not employ multi-user MIMO precoder. In each frame, the BS with highest average signal power is the candidate BS for a user. There is no BS muting. The revised strategy is obtained from Algorithm 1 by setting $\gamma_T \to -\infty$ (in particular, we set $\gamma_T = -1000$ dB for this strategy in the simulation).
- Non-predictive resource allocation with ICI coordination (with legend “Non-pred RA (w/ ICIC)”): This strategy can be regarded as a revised version of an existing interference coordination technique in [14], which optimizes user association and resource allocation based on CSI without exploiting future information. The revised strategy is obtained by initializing all the resource allocation planning variables as zero and then applying step 2 in Algorithm 1 once (i.e., set $N_f = 1$ for this strategy).
- Non-predictive resource allocation without ICI coordination (with legend “Non-pred RA (w/o ICIC)”): The only difference from “Pred RA (w/o ICIC)” lies in that this strategy does not employ predicted average channel gains.

In Figure 10, we show the impact of traffic load on the performance ρ by changing the file size of each user B, as well as the impact of the prediction errors. To reflect the prediction errors on the average channel gains in the prediction window, we first add errors on the predicted trajectory of each user, and then simulate the average channels gains in each frame of each user with all BSs by using the pathloss and shadowing setup stated in the start of this section. According to the prediction results for vehicular users in city roads with intersection and traffic lights in [24], the trajectory prediction errors are bounded within 6 m for a 60-seconds prediction window. To provide a conservative evaluation, we set the prediction errors as uniform distribution between 0 m and 10 m. The results are with legend “Pred RA (w/ ICIC) w/error”, which show that the prediction errors of future average channel gains cause marginal performance loss.

In Figure 11, we further show the impact of traffic load by changing the number of users K.

We can see from Figures 10 and 11 that the proposed method outperforms the reference strategies for different values of B and K. No matter if the interference is coordinated, predictive resource allocation is always superior to the non-predictive counterpart. When $B = 180$ MB and $K = 50$, the values of ρ achieved by “Non-pred RA (w/o ICIC)” are almost zero, hence cannot be seen in the figures. When the traffic load is light (i.e., in the case of $B = 60$ MB), both predictive strategies achieve 100% of user satisfactory rate and outperform the non-predictive strategies. When the traffic load is heavy (i.e., in the cases of $B = 120$, 150, 180 MB), "Pred RA (w/o ICIC)" is even inferior to "Non-pred RA (w/ ICIC)". This suggests the necessity of interference coordination for predictive resource allocation in heavy load scenarios. By comparing the proposed method with "Pred RA (w/o ICIC)", we can observe the remarkable gain of the interference coordination.

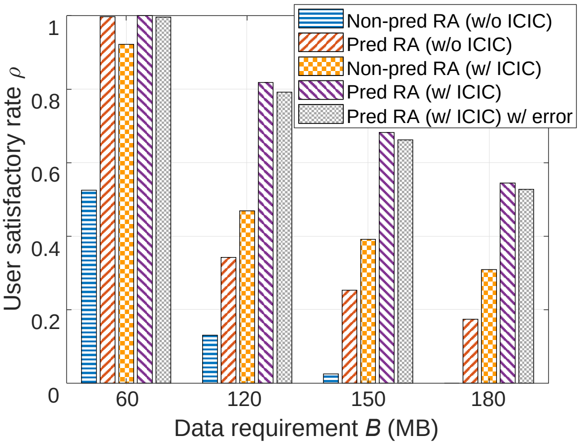


FIGURE 10. User satisfactory rate with different file sizes, $K = 30$, $N_f = 60$.

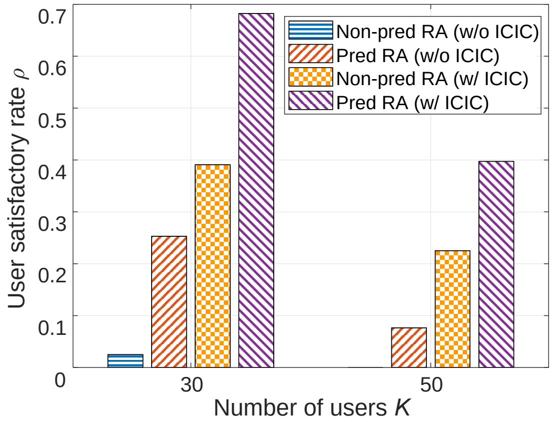


FIGURE 11. User satisfactory rate with different number of users, $N_f = 60$, $B = 150$ MB.

In Figure 12, we show the impact of prediction window on the performance by changing the number of frames in the window, i.e., N_f. We can see that the user satisfactory rates achieved by all the four strategies increase with N_f, since the available resources to transmit the file of each user increase. When the value of N_f increases, the user satisfactory rate of 'Pred RA (w/ ICIC)' grows faster than that of 'Non-pred RA (w/ ICIC)'. This is because the proposed strategy exploits the predicted information to coordinate interference in a proactive manner, where a longer prediction window provides more flexibility for the coordination.

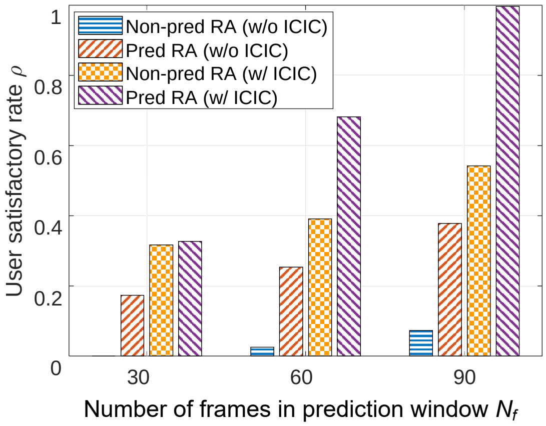


FIGURE 12. User satisfactory rate with different prediction window durations, $K = 30$, $B = 150$ MB.

Finally, we briefly address the complexity issue. Recall that the complexity of the proposed low-complexity algorithm is less than $\mathcal{O}\big(b\sum_{l=1}^{N_f}(\sum_{g=1}^{G}|\mathcal{A}_g^l|M_g)^3\big)$, where the algorithm is applied only at the start of each prediction window. In [25], a series of linear programming problems need to solved to find the optimal solution, each with KN_f variables and $(K+GN_f)$ constraints. The complexity to find the optimal solution is $\mathcal{O}(m[N_f(K+G)]^{3.5})$, where m is the number of the linear programming problems. Again, the solution needs to be found at the start of each prediction window. For the considered scale of the network, where $K = 30$, $N_f = 60$, $G = 34$, $M_m = 4$ and $M_p = 2$, the complexity of the proposed algorithm is much lower than the algorithm proposed in [25]. In [14], the proposed algorithm is non-predictive, which needs to be implemented at the start of each time slot, and hence is not comparable with our algorithm operated in a much larger time scale.

VI. CONCLUSIONS

In this paper, we investigated predictive resource allocation for non-realtime service in downlink interference HetNets. To address the challenge of optimizing predictive resource allocation planning in interference networks, we first optimized an interference coordination scheme before making the resource allocation plan, both harnessing the predicted average channel gains in a prediction window. The coordination scheme determines the BS-user pairs that can be communicated simultaneously in each of the future frames to ensure an average SINR. The resource allocation plan determines which BSs are muted and the users are associated to which active BSs in each frame to maximize a network utility aiming at improving user satisfactory rate. By resorting to the maximal independent set and maximal weighted independent set, we obtained the globally optimal solution and a low-complexity algorithm for finding the plan. Simulation results demonstrated that the proposed method provides much higher user satisfactory rate than existing predictive resource allocation when traffic load is heavy and the non-predictive resource allocation when the prediction window is long (say over one minute). The performance gain is larger when the prediction window is longer.

To obtain the promising performance gain of the proposed strategy, future average channel gains need to be predicted. The information can be obtained by first predicting user trajectory [38] with machine learning techniques [22] and then integrating with the radio map. It is also possible to be predicted directly from the measured average channels, considering that establishing radio map via drive test is expensive. Since the requests from minority of the users account for majority of the traffic load according to real data analysis [39], only the average channel gains of a small fraction of mobile users need to be predicted, which may justify the computational complexity introduced by the prediction. Besides, a center point with strong computing ability should be deployed in the network to gather information, record historical data, and make the prediction. Considering the predicability of the average channel gains in a required prediction horizon, the framework could also be applicable for drone BSs [40].

REFERENCES

[1] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka et al., “Scenarios for 5G mobile and wireless communications: the vision of the METIS project,” IEEE Commun. Mag., vol. 52, no. 5, pp. 26–35, May 2014.

[2] 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 Commun., vol. 21, no. 2, pp. 18–25, April 2014.

[3] S. Singh and J. G. Andrews, “Joint resource partitioning and offloading in heterogeneous cellular networks,” IEEE Trans. Wireless Commun., vol. 13, no. 2, pp. 888–901, Feb. 2014.

[4] W. Bao and B. Liang, “Rate maximization through structured spectrum allocation and user association in heterogeneous cellular networks,” IEEE Trans. Commun., vol. 63, no. 11, pp. 4510–4524, Nov. 2017.

[5] Y. Lin, W. Bao, W. Yu, and B. Liang, “Optimizing user association and spectrum allocation in HetNets: A utility perspective,” IEEE J. Sel. Areas Commun., vol. 33, no. 6, pp. 1025–1039, Jun. 2015.

[6] S. Sadr and R. S. Adve, “Tier association probability and spectrum partitioning for maximum rate coverage in multi-tier heterogeneous networks,” IEEE Commun. Lett., vol. 18, no. 10, pp. 1791–1794, Oct. 2014.

[7] D. Liu, L. Wang, Y. Chen, M. Elkashlan, K.-K. Wong, R. Schober, and L. Hanzo, “User association in 5G networks: A survey and an outlook,” IEEE Commun. Surveys Tuts., vol. 18, no. 2, pp. 1018–1044, 2nd Quart 2016.

[8] K. Shen and W. Yu, “Distributed pricing-based user association for downlink heterogeneous cellular networks,” IEEE J. Sel. Areas Commun., vol. 32, no. 6, pp. 1100–1113, Jun. 2014.

[9] R. Madan, J. Borran, A. Sampath, N. Bhushan, A. Khandekar, and T. Ji, “Cell association and interference coordination in heterogeneous LTE-A cellular networks,” IEEE J. Sel. Areas Commun., vol. 28, no. 9, pp. 1479–1489, Dec. 2010.

[10] M. Hong and Z.-Q. Luo, “Distributed linear precoder optimization and base station selection for an uplink heterogeneous network,” IEEE Trans. Signal Process., vol. 61, no. 12, pp. 3214–3228, Jan. 2013.

[11] 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 Trans. Wireless Commun., vol. 12, no. 6, pp. 2706–2716, Jun. 2013.

[12] S. Deb, P. Monogioudis, J. Miernik, and J. P. Seymour, “Algorithms for enhanced inter-cell interference coordination (eICIC) in LTE HetNets,” IEEE/ACM Trans. Netw., vol. 22, no. 1, pp. 137–150, Feb. 2014.

[13] H. Zhou, S. Mao, and P. Agrawal, “Approximation algorithms for cell association and scheduling in femtocell networks,” IEEE Trans. Emerging Topics Comput., vol. 3, no. 3, pp. 432–443, Sep. 2015.

[14] Y. Chen, J. Li, W. Chen, Z. Lin, and B. Vucetic, “Joint user association and resource allocation in the downlink of heterogeneous networks,” IEEE Trans. Veh. Technol., vol. 65, no. 7, pp. 5701–5706, July 2016.

[15] L. Xu, J. Wang, H. Zhang, and T. A. Gulliver, “Performance analysis of iaf relaying mobile D2D cooperative networks,” Elsevier Journal of the Franklin Institute, vol. 354, no. 2, pp. 902–916, 2017.

[16] D. Wu, Q. Wu, Y. Xu, and Y.-C. Liang, “QoE and energy aware resource allocation in small cell networks with power selection, load management and channel allocation,” IEEE Trans. Veh. Technol., vol. 66, no. 8, pp. 7461–7473, 2017.

[17] Q. Han, B. Yang, G. Miao, C. Chen, X. Wang, and X. Guan, “Backhaul-aware user association and resource allocation for energy-constrained HetNets,” IEEE Trans. Veh. Technol., vol. 66, no. 1, pp. 580–593, 2017.

[18] J. Mei, K. Zheng, L. Zhao, Y. Teng, and X. Wang, “A latency and reliability guaranteed resource allocation scheme for LTE V2V communication systems,” IEEE Trans. Wireless Commun., vol. 17, no. 6, pp. 3850–3860, 2018.

[19] L. Wang, H. Tang, H. Wu, and G. L. Stüber, “Resource allocation for D2D communications underlay in Rayleigh fading channels,” IEEE Trans. Veh. Technol., vol. 66, no. 2, pp. 1159–1170, 2017.

[20] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, “Limits of predictability in human mobility,” Science, vol. 327, no. 5968, pp. 1018–1021, 2010.

[21] A. Nadembega, A. Hafid, and T. Taleb, “A destination and mobility path prediction scheme for mobile networks,” IEEE Trans. Veh. Technol., vol. 64, no. 6, pp. 2577–2590, Jun. 2015.

[22] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, “Machine learning for wireless networks with artificial intelligence: A tutorial on neural networks,” arXiv preprint arXiv:1710.02913, 2017.

[23] J. Chen, U. Yatnalli, and D. Gesbert, “Learning radio maps for UAV-aided wireless networks: A segmented regression approach,” in IEEE ICC, 2017.

[24] W. Zhang, Y. Liu, T. Liu, and C. Yang, “Trajectory prediction with recurrent neural networks for predictive resource allocation,” in IEEE ICSP, 2018.

[25] C. Yao, J. Guo, and C. Yang, “Achieving high throughput with predictive resource allocation,” in IEEE GlobalSIP, 2016.

[26] H. Abou-Zeid and H. Hassanein, “Toward green media delivery: Location-aware opportunities and approaches,” IEEE Wireless Commun., vol. 21, no. 4, pp. 38–46, Aug. 2014.

[27] C. Yao, C. Yang, and Z. Xiong, “Energy-saving predictive resource planning and allocation,” IEEE Trans. Commun., vol. 64, no. 12, pp. 5078–5095, Dec. 2016.

[28] R. Atawia, H. S. Hassanein, H. Abou-Zeid, and A. Noureldin, “Robust content delivery and uncertainty tracking in predictive wireless networks,” IEEE Trans. Wireless Commun., vol. 16, no. 4, pp. 2327–2339, 2017.

[29] Z. Lu and G. De Veciana, “Optimizing stored video delivery for mobile networks: The value of knowing the future,” in IEEE INFOCOM, 2013.

[30] R. Margolies, A. Sridharan, V. Aggarwal, R. Jana, N. Shankaranarayanan, V. A. Vaishampayan, and G. Zussman, “Exploiting mobility in proportional fair cellular scheduling: Measurements and algorithms,” IEEE Trans. Networking, vol. 24, no. 1, pp. 355–367, Feb. 2016.

[31] S. Verdu et al., Multiuser detection. Cambridge university press, 1998.

[32] D. Liu, L. Wang, Y. Chen, T. Zhang, K. K. Chai, and M. Elkashlan, “Distributed energy efficient fair user association in massive MIMO enabled HetNets,” IEEE Commun. Lett., vol. 19, no. 10, pp. 1770–1773, Oct. 2015.

[33] J. A. Bondy and U. S. R. Murty, Graph theory with applications. Macmillan, 1976, vol. 290.

[34] R. W. Heath Jr, T. Wu, Y. H. Kwon, and A. C. Soong, “Multiuser MIMO in distributed antenna systems with out-of-cell interference,” IEEE Trans. Signal Process., vol. 59, no. 10, pp. 4885–4899, Oct. 2011.

[35] E. Tomita, Y. Sutani, T. Higashi, and M. Wakatsuki, “A simple and faster branch-and-bound algorithm for finding a maximum clique with computational experiments,” IEICE Trans. Inf. Syst., vol. 96, no. 6, pp. 1286–1298, 2013.

[36] S. Busygin, “A new trust region technique for the maximum weight clique problem,” Elsevier Discrete Applied Mathematics, vol. 154, no. 15, pp. 2080–2096, 2006.

[37] 3GPP Long Term Evolution, “Further advancements for E-UTRA physical layer aspects,” TR 36.814 v9.0.0, 2010.

[38] M. Chen, M. Mozaffari, W. Saad, C. Yin, M. Debbah, and C. S. Hong, “Caching in the sky: Proactive deployment of cache-enabled unmanned

aerial vehicles for optimized quality-of-experience," IEEE J. Sel. Areas Commun., vol. 35, no. 5, pp. 1046–1061, 2017.

[39] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding traffic dynamics in cellular data networks," in Proc. IEEE INFOCOM, 2011.

[40] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Unmanned aerial vehicle with underlaid device-to-device communications: Performance and tradeoffs," IEEE Trans. Wireless Commun., vol. 15, no. 6, pp. 3949–3963, 2016.