Achieving High Throughput with Predictive Resource Allocation

Chuting Yao, Jia Guo and Chenyang Yang

Beihang University, Beijing, China, Email: {ctyao, guojia, cyyang}@buaa.edu.cn

***Abstract*—Big data analytics makes predicting human behavior possible, but it is unclear how to exploit the predictable information for improving performance of wireless networks. In this paper, we investigate the potential of predictive resource allocation in supporting high throughput by exploiting excess resources. To this end, we assume that the requests and trajectories of mobile users and the average resource usage status of base stations can be predicted within a window. To fully use resources within the prediction window and reserve resources for the unpredictable traffic arrived after the window, we optimize a resource allocation plan to minimize the maximal transmission completion time. To assist the base stations for user scheduling, we introduce a method to make a transmission plan. These two plans determine where, when and what to transmit to the users with how much resources. Simulation results show that the predictive resource allocation can provide substantial gain over non-predictive strategy in terms of both network throughput and user experience.**

I. Introduction

To support the explosively growing traffic demands, the main trend techniques for the fifth generation (5G) cellular networks are to update network architecture, improve spectral efficiency (SE) by network densification, and explore more spectrum [1]. While increasing SE is always beneficial, in reality the resources are often under-utilized in many base stations (BSs), owing to the time-varying traffic pattern.

More or less inspired by a recent report that human behavior is highly predictable [2], optimizing wireless networks by leveraging the prediction ability endowed by big data is drawing research attention. With big data analytics, the map of traffic load and the mobility pattern [3]–[5] can be predicted, at least within a prediction window. From the traffic map, the average network resource usage status can be estimated. From the trajectory of a mobile station (MS), the average channel gains can be obtained with the help of a radio map [6]. Moreover, the content popularity and even the preferred content of an individual user, are possible to be known before the user(s) truly initiates the request, say by using collaborative filtering that has long been studied for recommendation problems [7], which is a canonical application of big data. Undoubtedly, predicting the behavior related information is challenging. This naturally raises the following question: what performance of cellular networks can be improved by exploiting such valuable information, with how much gain, and how?

With predicted content popularity, caching at the wireless edge can reduce the backhaul cost, offload the traffic in core and access networks, improve user experience and energy efficiency [8]–[10]. Yet how other information able to be predicted could impact the wireless resource management is largely unexplored. When a central processor (CP) connected to BSs can predict or obtain the network resource usage state, user trajectory, the content to be requested and the request arrival time, then how these information can be exploited for improving the performance of wireless networks? With these information, called network level, user level, and application level context information in [11], a long-term resource allocation plan can be made for each user before transmission, including which BSs along the trajectory of a MS should pre-download files to the user, in which duration, and with how many resources. Such *predictive* or *proactive* resource allocation has been proposed for reducing the outage probability [12], saving the energy consumed at the BSs [6], [13] or improve throughput [14]. In most existing works along this line, future average or even instantaneous achievable rate are assumed known [6], [12], [14], which implies that at least the user and network level information are available.

In this paper, we attempt to show the potential of predictive resource allocation in supporting high throughput by exploiting excess resources. To this end, we assume that the three levels of context information is known within a prediction window, although the prediction is never perfect. To fully use the excess resources within the prediction window, we optimize the resource allocation planning for pre-downloading the files to be requested to users with the help of these information to minimize the maximal transmission completion time. Such an optimization is very different from those in [14], which were oriented for full-buffer traffic and less efficient when the resource is under-utilized. To help the BSs schedule users with the pre-assigned resources according to the plan, we proceed to make a transmission plan. Simulation results show that the proposed predictive resource allocation can support higher traffic load for given user satisfaction rate and reduce average waiting time of the users than the traditional non-predictive resource allocation, both dramatically.

II. System Model

Consider a N_b-cell network, where each BS is equipped with N_t antennas. The BSs serve two classes of traffic: real-time (RT) service (e.g, phone call) and content delivery (e.g., file downloading), both requests arrive randomly. Due to the high priority of RT service, the contents can only be delivered by using the residual resources (i.e., bandwidth and transmit power) at each BS, which are stochastic processes in practice.

In this work, we are concerned with predictive resource allocation for the mobile users demanding content delivery, each requesting one file with size of B bits. We refer to these users as the MSs in the rest of the paper.

A. Context Information and Predictive Resource Allocation

Assume that all the BSs are connected to a CP. Endowed by big data analytics, the CP can predict or can obtain the prediction of three levels of information within a prediction window as follows. (i) The request arrival time and the file to be requested of every MS, i.e., the *application level context information*. (ii) The trajectory of every MS, i.e., the *user level context information*. (iii) The average residual resources remained at each BS after serving the RT traffic, i.e., the *network level context information*.

With the predicted context information, the CP can make a plan of resource allocation and transmission for conveying the files that the MSs will request, which essentially determines where, when, what, and with how much resources to transmit how many bits. After informed by the CP, the BSs along the trajectory of each MS can pre-download the file to the MS before it initiates the request, and continue to transmit the remaining file (if the file has not been conveyed completely) after the MS's request arrives, according to the plan. Such a predictive resource allocation is sharply different from the traditional transmission mechanism, where the MSs are served with best effort after their requests arrive.

B. Channel Model and Achievable Rate

Time is discretized into time slots. The length of the prediction window is T_f frames, where each frame includes T_s time slots. To reflect the variation of the path-loss and shadowing due to user mobility, we assume that the large scale channel gains remain constant within each frame and may vary among frames. The small scale channel gains remain constant within each time slot and vary among time slots independently.

Assume that only the closest BS to a MS may pre-download (or transmit) the file to be requested (or having been requested) to the MS. According to a plan made by the CP, a BS may need to pre-download files to multiple MSs. To avoid multi-user interference, the BS transmits to the MSs in different time slots, then maximum ratio transmission (MRT) is optimal, and the achievable rate of the kth MS (denoted as MS_k) in the tth time slot of the jth frame is

$$R_{j,k}^t = W_{j,k}^t \log_2(1 + g_{j,k}^t p_{\max,j,k}^t), \tag{1}$$

where $W_{j,k}^t$ and $p_{\max,j,k}^t$ are respectively the residual bandwidth and transmit power available for the MS at the closest BS after serving RT traffic, $g_{j,k}^t \triangleq \alpha_k^j \|\mathbf{h}_{j,k}^t\|^2/\sigma^2$ is the equivalent channel gain, $\mathbf{h}_{j,k}^t \in \mathbb{C}^{N_t \times 1}$ is the channel vector, α_k^j is the large scale channel gain, σ^2 is the noise variance, and $\|\cdot\|$ denotes Euclidean norm.

III. Optimizing Predictive Resource Allocation

In order to exploit the information in the prediction window, we first formulate a resource allocation planning problem to minimize the maximal file delivery time, and find the optimal solution. Then, we make a transmission plan, and provide a transmission strategy according to the plan.

A. Making the Resource Allocation Plan

Assume that the prediction window starts at the 1st time slot of the 1st frame, and ends at the T_sth time slot of the T_fth frame, as shown in Fig. 1, where K MSs will initiate requests within the window, and MS_k will send its request at the $t_{k,\mathrm{a}}$th time slot of the $T_{k,\mathrm{a}}$th frame. Denote the request arrival time of MS_k as $J_{k,\mathrm{a}} \triangleq ((T_{k,\mathrm{a}}-1)T_s + t_{k,\mathrm{a}})/T_s$.

By transmission according to the plan, the file is completely delivered to MS_k in the t_kth time slots of the T_kth frame, called *completion time*, denoted as $J_k \triangleq ((T_k-1)T_s + t_k)/T_s$.

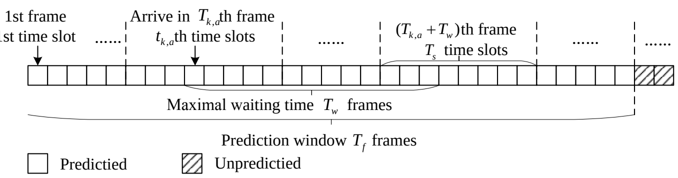

Fig. 1. Illustration of time model: the CP makes the plan for the MSs at the 1st time slot of the 1st frame with context information in the window.

In order not to compromise user experience, the file requested by each MS should be conveyed before an expected time instant after it sends the request, called *deadline*. When the completion time of MS_k is before the deadline, the MS is satisfied. To control the user satisfaction rate, we impose an constraint on the completion time: the file requested by each MS should be conveyed before T_w frames after it sends the request, i.e., $J_k \leq J_{k,\mathrm{a}} + T_\mathrm{w}$, where $J_{k,\mathrm{a}} + T_\mathrm{w}$ is called *maximal waiting time*. To fully use the excess network resources within the window, the completion time should also satisfy $J_k \leq T_f$ since the CP can only make the plan for future T_f frames. Therefore, the completion time should satisfy

$$J_k \leq J_{k,\mathrm{mw}} \triangleq \min(J_{k,\mathrm{a}} + T_\mathrm{w}, T_f). \tag{2}$$

In order to reserve resources for the unpredictable requests arrived after the prediction window, we minimize the maximal completion time of all the MSs arrived in the window.

For easy exposition, we first formulate a problem with all instantaneous information known in the future $T_f T_s$ time slots, including the instantaneous residual resources and small scale channel gains. Denote $m_{j,k}^t \in \{1, 0\}$ as an indicator. When $m_{j,k}^t = 1$ or 0, the file to be requested by MS_k *will* or *will not* be pre-downloaded by its closest BS in the tth time slot of the jth frame. The optimization problem is as follows,

$$\min_{\mathbf{M}_1,\ldots,\mathbf{M}_K} \ \max_k \ J_k \tag{3a}$$

$$s.t. \sum_{j=1}^{T_f} \sum_{t=1}^{T_s} m_{j,k}^t R_{j,k}^t \Delta_t = B, \tag{3b}$$

$$J_k \leq J_{k,\mathrm{mw}}, k = 1, \ldots, K, \tag{3c}$$

$$m_{j,k}^t = 0, \forall j > T_k, m_{T_k,k}^t = 0, \forall t > t_k, \tag{3d}$$

$$\sum_{k \in \mathcal{K}_{j,i}} m_{j,k}^t \leq 1, i = 1, \ldots, N_b, \tag{3e}$$

where $\mathbf{M}_k = [\mathbf{m}_{1,k}, \dots, \mathbf{m}_{T_f,k}]$, $\mathbf{m}_{j,k} = [m_{j,k}^1, \dots, m_{j,k}^{T_s}]^H$, (3b) means that B bits should be conveyed to each MS within the prediction window, (3c) is from (2), (3d) indicates that a MS will not be served after the completion time, (3e) indicates that each BS only transmits to a single MS in each time slot, $\mathcal{K}_{j,i}$ is the set of MSs that enter the coverage of the ith BS in the jth frame, and Δ_t is the duration of each time slot.

It is worthy to note that problem (3) is not viable in practice, because the required instantaneous information in the future is hard to predict if not impossible. To make a resource allocation plan only with the statistical channel and network status information, we assume that the small scale channel gains and residual resources are ergodic, and $T_s \to \infty$ (such an assumption has been justified in [15], which is valid when T_s = 100). Then, the left hand side of (3b) approaches

$$\lim_{T_s\to\infty} \sum_{j=1}^{T_f} \sum_{t=1}^{T_s} m_{j,k}^t R_{j,k}^t \Delta_t \overset{(a)}{=} \sum_{j=1}^{T_f} \sum_{t=1}^{T_s} m_{j,k}^t \bar{R}_{j,k} \Delta_t \overset{(b)}{=} \sum_{j=1}^{T_k} s_{j,k} \bar{R}_{j,k} T_s \Delta_t,$$

where $\bar{R}_{j,k} \triangleq \mathbb{E}\{R_{j,k}^t\}$ is the average achievable rate in each time slot in the jth frame, (a) holds because time average equals ensemble average, (b) comes from defining $s_{j,k} \triangleq \sum_{t=1}^{T_s} m_{j,k}^t / T_s \in [0,1]$, which is the fraction of time resource allocated to MS_k in the jth frame, and $\mathbb{E}\{\cdot\}$ represents expectation.

Constraint (3c) indicates that no data is transmitted to the MS in the time slots after the completion time. This can be relaxed to a constraint on the fraction of time resources employed in each frame, which is

$$s_{j,k} = 0, \forall j > \lceil J_{k,\text{mw}} \rceil, s_{\lceil J_{k,\text{mw}} \rceil} \le J_{k,\text{mw}} - \lceil J_{k,\text{mw}} \rceil + 1, \quad (4)$$

where $\lceil \cdot \rceil$ is ceiling function.

Similarly, constraint (3d) can be relaxed as

$$s_{j,k} = 0, \forall j > \lceil J \rceil, s_{\lceil J \rceil,k} \le J - \lceil J \rceil + 1, \quad (5)$$

where $J \triangleq \max_k J_k$, which is the *maximal completion time* among all the K MSs whose requests arrive in the window.

Then, the resource allocation plan, i.e., how much time resource should be allocated to these MSs, can be made by finding the solution from the following problem that only needs the statistical information,

$$\min_{J, \mathbf{s}_1, \dots, \mathbf{s}_K} J \quad (6a)$$

$$s.t. \sum_{j=1}^{T_f} s_{j,k} \bar{R}_{j,k} T_s \Delta_t = B, \quad (6b)$$

$$\sum_{k \in \mathcal{K}_{j,i}} s_{j,k} \le 1, i = 1, \dots, N_b, \quad (6c)$$

$$(4), (5),$$

where $\mathbf{s}_k = [s_{1,k}, \dots, s_{T_f,k}]^H$ is the time resource allocated to MS_k, (6b) guarantees that B bits can be delivered to each user within the prediction window, and (6c) is relaxed from (3e), which ensures the transmission time of each BS in each frame not exceeding one frame duration.

Since for a given J, problem (6) is a linear programming, the optimal solution of $\mathbf{s}_k, k = 1, \dots, K$ can be obtained if J is feasible. Therefore, problem (6) can be equivalently decoupled into two subproblems. The inner subproblem is a linear programming problem with given J. The outer subproblem is to find the minimal J that makes the inner subproblem feasible, which can be solved with binary search since the number of feasible solutions increases with J.

B. Making the Transmission Plan

With the optimal resource allocation plan $\mathbf{s}_k^*$, the CP decides how much time resources are allocated to MS_k in each frame. To assist each BS for scheduling users in each time slot such that the maximal number of MSs will be satisfied, the CP also needs to decide how many bits should be transmitted to each MS at some critical time instance (called time-stamps) in the prediction window, i.e., to make a transmission plan. Since such a decision should be made without knowing future instantaneous information, it is more reasonable to determine the data amount ought to be accumulatively conveyed at the time-stamps, called *transmission progress*.

Since (4) and (5) are the relaxed constraints of the completion time, to serve more MSs before their maximal completion time, the time-stamps denoted as a set $\mathcal{J} = \{\tilde{J}_1, \dots, \tilde{J}_{T_f+K}\}$ includes the ends of the frames and the maximal waiting time of the K MSs, i.e., $\{1, \dots, T_f, J_{1,\text{mw}}, \dots, J_{K,\text{mw}}\}$, but with an ascending order.

According to the resource allocation plan, the transmission progress for MS_k at the time-stamp $\tilde{J}_n$, which is the overall number of bits ought to be transmitted before $\tilde{J}_n$, should be

$$\Lambda_B(k, \tilde{J}_n) = \sum_{j=1}^{\lceil \tilde{J}_n \rceil - 1} s_{j,k}^* \bar{R}_{j,k} T_s \Delta_t + (\tilde{J}_n - \lceil \tilde{J}_n \rceil + 1) s_{\lceil \tilde{J}_n \rceil,k}^* \bar{R}_{\lceil \tilde{J}_n \rceil,k} T_s \Delta_t. \quad (7)$$

After the resource allocation and transmission plans are made for every MS in the first time slot of the first frame, the CP can inform the plans and trajectories of the K MSs in the prediction window to the BSs who may pre-download or transmit to these users, simply by broadcasting.

C. Transmission Strategy

When a MS enters the coverage of a BS who is planned to pre-download the file to the MS, the BS starts to estimate the instantaneous channel information of the MS, and selects the user for transmitting in each time slot if the BS needs to pre-download files to multiple MSs according to the resource allocation plan. To maximize the number of satisfied MSs (i.e., the users whose files are completely conveyed before their expected deadline), the BS schedules the MSs according to their transmission plans by exploiting all its residual resources.

In the tth time slot in the jth frame when $\tilde{J}_{n-1} < ((j-1)T_s + t)/T_s < \tilde{J}_n$, the set of MSs whose files are planned to be downloaded by the ith BS but have not caught up the transmission progress can be expressed as $\tilde{\mathcal{K}}_{j,i} \triangleq \{k \in \mathcal{K}_{j,i} | \Lambda_B(k, \tilde{J}_n) - (\sum_{l=1}^{j-1} \sum_{\tau=1}^{T_s} R_{l,k}^\tau + \sum_{\tau=1}^{t-1} R_{j,k}^\tau) \Delta_t > 0\}$.

Then, the ith BS selects the MS with the maximal instantaneous achievable rate from this MS set, i.e., according to the following rule

$$k^* = \arg\max_k \{R_{j,k}^t | s_{j,k}^* > 0 \text{ and } k \in \tilde{\mathcal{K}}_{j,i}\}. \quad (8)$$

Then, the ith BS transmits the file to the k^*th MS with MRT using the instantaneous residual transmit power and residual bandwidth $W_{j,k}^t$ and $p_{\max,j,k}^t$.

IV. Simulation Results

In this section, we illustrate the performance of the proposed predictive resource allocation by simulations.

Consider a N_b-cell system with cell radius $D = 250$ m, where $N_b = 13$ BSs each equipped with $N_t = 6$ antennas are located along a straight line. The MSs move along three roads of straight lines with minimum distance from the BSs as 50 m, 100 m and 150 m, respectively. Each MS requests a file with $B = 30$ Mbytes. The prediction window contains $T_f = 300$ frames. Each frame is with duration of one second, and each time slot is with duration $\Delta_t = 10$ ms, i.e., each frame contains $T_s = 100$ time slots. To reflect the fluctuation of traffic load, the content delivery requests of the MSs randomly arrive only between the 101th frame and 200th frame in the prediction window. To characterize the different resource usage status of the BSs by serving the RT traffic in a underutilized network, we consider two types of BSs: busy BS with average residual bandwidth $\overline{W} = 1$ MHz and idle BS with $\overline{W} = 10$ MHz, which are alternately located along the line as idle, idle, busy, busy, idle, idle, and so on. The maximal transmit power of each BS is 40 W and cell-edge SNR is set as 5 dB, where the intercell interference is implicitly reflected. The path loss model is $36.8 + 36.7 \log_{10}(d)$, where d is the distance between the BS and user in meter. The results are obtained from 500 Monte Carlo trails. In each trail, the trajectories change randomly with speed uniformly distributed in $(2.5, 12.5)$ m/s, the requests of the MSs arrive after the 100th frame according to Poisson process with given average arrival rate, the small-scale channel in each time slot changes independently according to Rayleigh fading, and the residual bandwidth at each BS in each time slot uniformly varies with mean value of $\overline{W}$.

Since context information is not predicted for free, a nature question is whether only one level of information can obtain most of the performance gain. Since user and network level context information should be employed jointly to predict average data rate, the following three stratgies are simulated.

- "All Context": The proposed predictive resource allocation with three levels of context information.
- "A Context": The CP only knows the application level information at the start time of the prediction window. The CP informs the BSs who are closest to the MSs to download the files to the MSs no matter if their requests actually arrive with best efforts, i.e., each BS employs all its instantaneous residual bandwidth and transmit power to transmit to the MS who can achieve highest instantaneous data rate in each time slot.
- "No Context": This is the traditional non-predictive resource allocation, where the transmission begins right after the requests truly arrive, again with best efforts.

To show the performance in terms of throughput and reflect the impact of the pre-determined deadline, in Fig. 2(a) we evaluate the maximal carrying traffic load, defined as the maximal request arrival rate given the user satisfaction rate (which is the number of satisfied MSs divided by the overall number of MSs in the prediction window). We consider two cases where 95% or 99% MSs's files can be completely conveyed before the deadline.

To reflect the performance in terms of user experience under different average request arrival rates, in Fig. 2(b) we evaluate the average waiting time, which is the average duration from the MSs initiating requests to the actual completion time.

The results demonstrate remarkable gain of "All Context" in supporting high traffic load and in improving the user experience by exploiting the predicted information over "No Context". The maximal carrying traffic loads supported by both "All Context" and "No Context" increase almost linearly with the expected waiting time, i.e., the deadline subtracting the request arrival time. When the user satisfaction rate is high (say 99%) and the expected waiting time is 120 s, the gain in terms of maximal carrying traffic load of "All Context" is 120% over "A Context" and 728% over the traditional transmission. The results in other system settings are similar, which are not shown due to space-limitation.

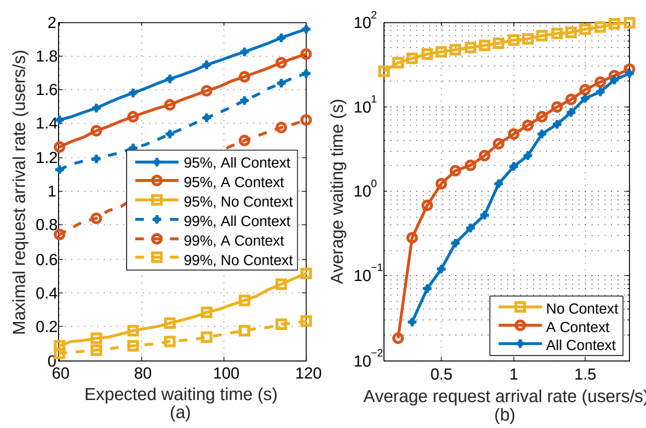

Fig. 2. Maximal carrying traffic load and average waiting time

V. Conclusions

In this paper, we investigated the predictive resource allocation by using three level context information within a prediction window, with which the excess resources in the window can be exploited to boose throughput. We formulated and solved a resource allocation planning problem to minimize the maximal transmission completion time, and introduced a method for making the transmission plan to help user scheduling. Simulation results demonstrated that the gain of predictive resource allocation in supporting high arrival rate and reducing average waiting time is dramatic over non-predictive resource allocation, and is evident over that only exploiting application level information when the user satisfaction rate is high.

References

[1] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. Sukhavasi, C. Patel, and S. Geirhofer, "Network densification: the dominant theme for wireless evolution into 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 82–89, Feb. 2014.

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

[3] M. Mardani and G. B. Giannakis, "Estimating traffic and anomaly maps via network tomography," *IEEE Trans. Netw.*, vol. Early access, 2016.

[4] J. Froehlich and J. Krumm, "Route prediction from trip observations," Soc. Automotive Eng. World Congress, Tech. Rep., 2008.

[5] 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, 2015.

[6] H. Abou-Zeid and H. S. Hassanein, "Toward green media delivery: location-aware opportunities and approaches," *IEEE Wireless Commun.*, vol. 21, no. 4, pp. 38–46, Aug. 2014.

[7] Y. Shi, M. Larson, and A. Hanjalic, "Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges," *ACM Comput. Surveys*, vol. 47, no. 1, pp. 1–45, May 2014.

[8] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Commun. Mag.*, vol. 51, no. 4, pp. 142–149, Apr. 2013.

[9] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5g wireless networks," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82–89, Aug. 2014.

[10] D. Liu and C. Yang, "Energy efficiency of downlink networks with caching at base stations," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 907–922, Apr. 2016.

[11] C. Park, Y. Seo, K. Park, and Y. Lee, "The concept and realization of context-based content delivery of NGSON," *IEEE Commun. Mag.*, vol. 50, no. 1, pp. 74–81, Jan. 2012.

[12] J. Tadrous, A. Eryilmaz, and H. Gamal, "Proactive resource allocation: Harnessing the diversity and multicast gains," *IEEE Trans. Info. Theory*, vol. 59, no. 8, pp. 4833–4854, Aug. 2013.

[13] C. Yao, C. Yang, and Z. Xiong, "Power-saving resource allocation by exploiting the context information," in *IEEE PIMRC*, 2015.

[14] H. Abou-zeid, H. Hassanein, and S. Valentin, "Optimal predictive resource allocation: Exploiting mobility patterns and radio maps," in *IEEE GLOBECOM*, 2013.

[15] C. Yao and C. Yang, "Role of large scale channel information on predictive resource allocation," in *IEEE WCNC*, 2016.