Network Codes for Effective Resource Allocation in Cache-Enabled 5G Mobile Networks

by

Mohammed Saif

B.Sc. Faculty of Engineering, Taiz University, 2010
M.Sc., King Fahd University of Petroleum and Minerals, 2017

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE COLLEGE OF GRADUATE STUDIES

(Electrical Engineering)

THE UNIVERSITY OF BRITISH COLUMBIA

(Okanagan)

August 2020

The following individuals certify that they have read, and recommend to the College of Graduate Studies for acceptance, a thesis/dissertation entitled:

Network Codes for Effective Resource Allocation in Cache-Enabled 5G Mobile Networks

submitted by Mohammed Saif in partial fulfilment of the requirements of the degree of Doctor of Philosophy

Dr. Md. Jahangir Hossain, School of Engineering
Supervisor

Dr. Stephen O'Leary, School of Engineering
Supervisory Committee Member

Dr. Anas Chaaban, School of Engineering
Supervisory Committee Member

Dr. Shawn Wang, Irving K. Barber School of Arts and Sciences
University Examiner

Dr. Lin Cai, Electrical and Computer Engineering, University of Victoria (UVIC), Canada
External Examiner

Abstract

Cooperation is a key enabler for delivering popular contents to users in today's centralized wireless communication networks where the cloud-computing base station (CBS) is connected with low power base-stations, also named as remote radio heads (RRHs), via fronthaul links. The performance of today's wireless communication networks is constrained by the scarcity of radio resources and limited-capacity of fronhaul links. Edge caching (EC) is a powerful networking technique, where popular contents are cached in the edge nodes, that reduces the burden on fronthaul links. State-of-the-art literature on content delivery focuses on allocating a single user to each radio resource-block (RRB). However, this needs a large numbers of RRBs to satisfy the tremendous increase in the number of users. Consequently, network coding (NC) is employed to mix users' contents and simultaneously allocate different users to the same RRB. This dissertation designs innovative NC schemes for effective resource allocation in cache-enabled 5G networks, where the goal is to optimize different metrics, such as throughput, CBS offloading, and completion time. We first study the cross-layer throughput maximization problem in cloud-radio access network (C-RAN). For this problem, we develop novel cross-layer NC (CLNC) frameworks that propose to optimize RRHs' transmit powers and users' rates in making the coding decisions. Subsequently, we study CBS offloading and users' quality of service (QoS)-guarantee trade-off for fog-RAN system subject to the cached contents, the required minimum rate, power control, and NC constraints. Next, we focus on developing NC schemes for device-to-device (D2D)-aided F-RAN system. The developed NC schemes exploit the aforementioned F-RAN's optimization factors and potential D2D communications to minimize the completion time of users. Finally, to circumvent

the necessity for CBSs in the above-mentioned centralized systems, we develop a distributed and decentralized game-theoretical NC framework for partially connected D2D networks such that the number of required D2D transmissions until all users receive all their requested contents is minimized. Simulation results are provided to attest the effectiveness of the proposed frameworks against baseline schemes for each of the above systems.

Lay Summary

Imagine that a group of users are located in a hotspot, e.g., playground, and request popular-contents that represent popular-videos. Streaming such users' requests with minimum possible delay while reducing the burden on the core-network is a key requirement for today's cellular networks. Proactive-caching is inevitable for offloading the core-network, where the popular contents, that are frequently requested by users, are cached in edge-servers and streamed to users with low intervention from the core-network. To efficiently utilize the radio-resources and stream popular contents to a large set of users with a minimum delay, network coding is a communication paradigm that was proposed to ensure instantaneous streaming of contents and simultaneous transmissions. We aim to stream popular contents with the goal of optimizing throughput and delay. Towards this objective, we develop network coding schemes for several cache-enabled 5G-systems. Through extensive simulations, we demonstrate the efficiency of the proposed schemes against baseline methods.

Preface

This thesis is based on the research conducted under the supervision of Prof. Md. Jahangir Hossain. For all the chapters, I conducted the required literature review and developed the research ideas. I carried out the required mathematical analysis and prepared simulations. Moreover, I prepared the related manuscripts. My supervisor provided feedback on system models, problems' solutions and their corresponding computational complexities, and improved the presentation of all articles for scholarly publication. Some collaborations on part of this thesis is conducted with Dr. Sameh Sorour and Dr. Ahmed Douik. Both helped me to discuss system models and developed ideas, and improve the presentation of some manuscripts.

Following is the list of the journal and conference publications related to the chapters of the thesis. Both published and submitted journal articles are provided in this list.

Published and Submitted Journal Papers

J1. M.-S. Al-Abiad, A. Douik, S. Sorour, and Md. J. Hossain, "Throughput maximization in cloud-radio access networks using cross-layer network coding," *IEEE Trans. on Mobile Computing,* Early Access, Aug. 2020 (Chapter 2).

J2. M.-S. Al-Abiad, Md. J. Hossain, and S. Sorour "Cross-layer cloud offloading with quality of service guarantees in Fog-RANs," *IEEE Trans. on Commun.,* vol. 67, no. 12, pp. 8435-8449, Jun. 2019 (Chapter 3).

J3. M.-S. Al-Abiad and Md. J. Hossain, "Completion time minimization

in Fog-RANs using D2D communications and rate-aware network coding," submitted for publication at *IEEE Trans. Wireless Commun.* (Chapter 4).

J4. M.-S. Al-Abiad, A. Douik, and Md. J. Hossain, "Coalition formation game for cooperative content delivery in network coding assisted D2D communications," *IEEE Access,* Early Access, Sep. 2020 (Chapter 5).

Referred Conference Publications

C1. M.-S. Al-Abiad, A. Douik, S. Sorour, and Md. J. Hossain, "Throughput maximization in cloud radio access networks using network codes," *EEE International Conference on Communications (ICC' 2018)* Kansas, USA, May, 2018, pp. 1-6 (part of Chapter 2).

C2. M.-S. Al-Abiad, S. Sorour, and Md. J. Hossain, "Cloud offloading with QoS provisioning using cross-layer network coding," *IEEE Globecom'18,* Abu Dhabi, UAE, 2018, pp. 1-6 (part of Chapter 3).

The following papers are also prepared during my PhD study; however, the materials of these papers are not included in the thesis.

1. A. Douik, M.-S. Al-Abiad, and Md. J. Hossain, "An improved weight design for unwanted packets in multicast instantly decodable network Coding," *IEEE Commun. Lett.,* vol. 23, no. 11, pp. 2122-2125, Nov. 2019.

2. M.-S. Al-Abiad, Md. Z. Hassan, A. Douik, and Md. J. Hossain, "Low-complexity power allocation for network-coded scheduling in cloud-RANs," revision pending at *IEEE Commun. Lett.*

3- M.-S. Al-Abiad, Md. Z. Hassan, and Md. J. Hossain, "Rate-aware network codes for distributed content sharing in cache-enabled D2D network," to be submitted for publication at *IEEE Internet of Things* Journal.

4- M.-S. Al-Abiad, Md. Z. Hassan, and Md. J. Hossain, "Cross-layer network codes for content delivery in cache-enabled D2D networks," to be submitted for publication at *IEEE Commun. Lett.*

Table of Contents

List of Tables

List of Figures

List of Acronyms

Acronyms	Definitions
5G	Fifth Generation
EC	Edge Caching
RRB	Radio Resource Block
NC	Network Coding
C-RAN	Cloud Radio Access Network
CLNC	Cross Layer Network Coding
RRH	Remote Radio Head
F-RAN	Fog Radio Access Network
D2D	Device-to-Device
IoT	Internet-of-things
BS	Base Station
CBS	Cloud Base Station
EN	Edge Node
eRRH	Enhanced Remote Radio Head
LTE	Long Term Evolution
RNC	Random Network Coding
ONC	Opportunistic Network Coding
GF	Galois Field
RLNC	Random Linear Network Coding
IDNC	Instantly Decodable Network Coding
NP	Nondeterministic Polynomial
RA-IDNC	Rate-Aware IDNC
PMP	Point to Multipoint
RSNC	Rate Selection and Network Coding
UD	User Device

CFG	Coalition Formation Game
PL	Power Level
SUI	Stanford University Interim
SINR	Signal-to-Interference-Plus-Noise-Ratio
AWGN	Additive White Gaussian Noise
bps	Bits-per-second
MHz	Meaga Hertz
LC	Local Condition
CC	Conflict Condition
KKT	Karush-Kuhn-Tucker
Mb	Meaga-byte
RC	Rate Condition
TC	Transmission Condition
IA-IDNC	Interference-Aware IDNC
IS	Independent Set
NTU	Non Transfer Utility
ACK	Acknowledgment
MC	Merge Condition
SC	Split Condition
ONC	One Coalition Formation

List of Symbols

Symbols	Definitions
$\mathcal{B}$	Set of B RRHs
$\mathcal{U}$	Set of U Users
$\mathcal{F}$	Set of F Files
$\mathcal{P}(\mathcal{F})$	Power set of the set $\mathcal{F}$
$\lvert\mathcal{F}\rvert$	Cardinality of the set $\mathcal{F}$
$\mathcal{Z}$	Set of Z RRHs
$h_{b_n z_m}^{u_i}$	Channel gain from z_m-th RRB in b_n-th RRH to u_i-th user
σ^2	Additive white Gaussian noise (AWGN) power
$P_{b_n z_m}$	Power level (PL) of z_m-th RRB in b_n-th RRH
$P_{b_n z_m}^{\max}$	Maximum PL of z_m RRB in b_n-th RRH
$\mathbf{P}$	Matrix containing the PLs of RRBs
N	Size of file
$\mathcal{R}$	Set of all achievable capacities
$\mathcal{W}_{u_i}$	Set of requested files by u_i-th user
$\mathcal{H}_{u_i}$	Set of received files at u_i-th user
$R_{b_n z_m}$	Transmission rate of z_m-th RRB in b_n-th RRH
$\kappa_{b_n z_m}$	The encoded file of z_m-th RRB in b_n-th RRH
$\tau_{b_n z_m}$	Set of scheduled users to z_m-th RRB in b_n-th RRH
s	Association scheme
$\mathcal{S}_{\text{IDNC}}$	Set of all possible IDNC file combinations
$\mathcal{A}$	Set of all possible associations between RRHs, RRBs, files, and rates
$\mathcal{A}_{z_m}$	All associations in $\mathcal{A}$ that are indexed by z_m-th RRB
$\mathbf{S}$	Feasible schedule
$\mathcal{A}_{\text{fs}}$	Set of all feasible schedules
$\mathcal{G}(\mathcal{V}, \mathcal{E})$	Graph and its vertices $\mathcal{V}$ and edges $\mathcal{E}$

$w(v)$	Weight of vertex v
$\mathtt{C}$	Maximum weight clique
O	Complexity operator
D	Cell size
μ	Caching ratio
$\mathcal{H}_{b_n}$	Set of cached files at b_n-th eRRH
P	Transmitted power of the CBS
h_{u_i}	Channel gain from the CBS to the u_i-th user
$\mathcal{I}$	Independent set
R_{th}	Rate threshold
φ	Mapping function
$\mathtt{I}$	Maximum weight independent set
${}^{n}P_{r}$	Permutation operator
$\binom{n}{k}$	Combination operator
$\mathcal{C}_{u_i}$	Coverage zone of u_i-th user
$\mathtt{R}$	Coverage zone radius
$h^{d2d}_{u_k u_i}$	D2D channel gain between u_k-th and u_i-th users
$\mathbf{R}$	Capacity status matrix (CSM)
$\mathcal{U}_{\text{tra}}$	Set of transmitting users
t	Time index
$\kappa^{d2d}_{u_k}$	XOR file to be sent by u_k-th D2D transmitter
$\tau(\kappa^{d2d}_{u_k})$	Scheduled users to u_k-th transmitter
$T^{c}_{b_n}$	Transmission duration of the b_n-th eRRH
$T^{d2d}_{u_k}$	Transmission duration of the u_k-th D2D transmitter
$T_{\max}$	Maximum transmission duration
$\mathbb{D}_{u_l}$	Accumulated time delay of u_l-th user
$\mathtt{T}_{u_l}$	Completion time of u_l-th user
$\mathtt{T}_{o}$	Overall completion time
$\mathcal{P}$	Set of P packets
$\mathtt{p}_{u_k}$	XOR packet to be sent by u_k-th UD
$\mathcal{T}$	Transmission schedule
$\sigma_{u_k u_l}$	Packet erasure probability from u_k-th UD to u_l-th UD
$\underline{\mathtt{n}}$	Binary vector of the transmitting UDs

$\underline{\mathcal{D}}$	Decoding delay experienced by all UDs
$\underline{\mathcal{I}}$	UDs hearing more than one transmission
$\underline{\mathcal{O}}$	Out of transmission range UDs
$\mathbb{E}[\sigma_{u_k}]$	Expected erasure probability of u_k-th UD
$\delta_{u_k u_m}$	Kronecker symbol between u_k and u_m UDs
$\phi(\mathcal{S}_s)$	Payoff of coalition $\mathcal{S}_s$
$\phi_{u_k}(\mathcal{S}_s)$	Payoff of u_k-th UD in a coalition $\mathcal{S}_s$
Ψ	Coalition structure
$\phi(\Psi)$	Payoff of coalition structure Ψ
$\triangleright$	Preference operator
Ψ_{ini}	Initial coalition structure
Ψ_{fin}	Final converged coalition structure
$\mathbb{D}_{\text{hp}}$	Stable coalition structure
C	Connectivity index
ϵ	CBS-to-UD erasure probability

Acknowledgements

I would like to express my sincere and unlimited gratitude to my advisor Dr. Md. Jahangir Hossain for the continuous support of my Ph.D research, and for his constant encouragement and motivation. He spent a lot of his precious time helping me at each step of discussing system models, improving the manuscripts and revising them for submission. Besides his academic help, I am also extremely grateful to Dr. Hossain for advising me in non-academic aspects.

Besides my advisor, I would like to thank Dr. Sameh Sorour and Dr. Ahmed Douik for their insightful comments, valuable feedback, and for their stimulating discussions. Also, I would like to thank my thesis committee members Dr. Stephen O'leary, Dr. Anas Chaaban, Dr. Shawn Wang and, Dr. Lin Cai, for their willingness to serve on my supervisory committee. I am very thankful to Dr. Stephen O'leary and Dr. Anas Chaaban for providing me their sincere feedback, critical comments, and suggestions during my qualification exam.

Throughout my PhD journey, I have met with different people from UBCO. I would like to thank my fellow labmates and friends from UBCO for the stimulating discussions and for all the fun we have had in this journey.

Finally, I would like to thank my family members for keeping their faith in me. I would like to pay my humble respect and the deepest thanks from my heart to my parents, my brothers, and my sisters for their patience, understanding, and encouragement. I would also like to sincerely thank my wife, the beautiful smile in my life, for her unconditional love and encouragement in all these years. She constantly motivated me during my good time and bad time. My achievements would not have been possible without the continuous support of my wife.

To my parents, my wife, and my kids

Chapter 1

Introduction and Background

1.1 An Overview of Cache-Enabled 5G Networks

Communication networks are a predominant component of our current life. Everywhere around us, people and devices exchange information among each other at different scales, speeds, and throughput. The ever-increasing number of devices combined with their explosive demand for bandwidth-hungry data services constrain today's wireless communication networks, specially cellular networks. A recent data-flow report, prepared by *Cisco Inc.*, shows that by 2023 the number of users will increase seven folds compared to the number of users of today's network [1]. In addition, with the evolution of so-called Internet-of-things (IoT), the number of connected-devices will dramatically increase and reach around up to 30 billions by 2021 [2]. This high growth of data-flow demand and network size has driven the research and industrial communities to explore a new wireless network architecture, commonly referred as fifth generation (5G) wireless communication network. One of the most key potentials of 5G technology is network densification through planning a massive deployment of small-cells and implementing their corresponding low-power base-stations (BSs) [3]. However, deploying a massive number of small cells and managing the BSs are challenging [3]. Cloud radio access network (C-RAN), proposed by *China Mobile Inc.*, is an innovative and practical platform for 5G that efficiently accommodates network densification. In C-RAN, low-power BSs, also named as remote radio heads (RRHs), are connected to a cloud-computing base sta-

tion (CBS) through fronthaul links. Thus, RRH in C-RAN is designed to only perform radio functionality and all other baseband signal processing functionalities and computations are centrally performed at the CBS. Consequently, compared to the traditional BSs, the power consumption and complexity of the RRHs are substantially reduced. Essentially, decreasing the prices of RRHs and increasing the density of RRHs in network is financially affordable. Due to its centralized coordination, C-RAN offers a practical technology that provides effective resource optimization through cloud computing, reliable interference mitigation mechanism through collaborative signal processing, and enhanced energy efficiency [4]. Besides its centralized coordination, heterogeneity through the deployment of various sizes of RRHs and their different transmission powers and ranges is an another important feature key of C-RAN. In C-RAN, the CBS coordinates the different transmit frames of the connected RRHs and shares users' contents among all (or a subset of) these RRHs (depends on the level of coordination, i.e., signal-level or scheduling-level). In signal-level coordination [5], [6], [7], users' contents are shared among various RRHs, which requires high-capacity, low-latency fronthaul links. On the other hand, in scheduling-level coordination [8], [9], the CBS allocates users to the radio resources of the different RRHs, under the system constraint that each user can be connected to at most a single RRH. This system constraint not only overcomes sharing users' contents among multiple RRHs, but also manages interference in C-RAN. As a result, scheduling-level coordination in C-RAN offers simplified, yet efficient, resource coordination frameworks.

The unprecedented increase of wireless systems radically reshaped our current daily lives and led to the proliferation of mobile devices and the popularity of mobile multimedia applications among them. Yet, despite this growth in demand for wireless data, it is observed that users tend to request multiple copies of a particular widespread audio, photo, or video within a short period of time [10]. The plethora of frequent contents streamed from the C-RAN core network (e.g., the CBS) to the RRHs introduce an unnecessary load on the fronthaul links. However, the C-RAN architecture is, in fact, restricted to the practical limitations of fronthaul links, especially

in ultra-dense wireless networks. In order to satisfy the fronthaul demand of C-RAN in such ultra-dense networking scenario, a powerful wireless networking technology, known as edge-caching (EC), has emerged [11]. EC is being widely used and motivated due to the fact that a large portion of mobile multimedia traffic is generated by many duplicate downloads of particular popular contents [10], [11]. Apart from its essential goal of reducing the traffic fronthaul congestion of 5G networks, EC can also provide opportunities for cooperative transmission if there are common contents across the caches of multiple edge-nodes. Further, EC is shown to be an effective way of reducing the delivery time [12], [13]. By pre-fetching popular contents to the local storage of RRHs at off-peak times, RRHs can efficiently stream such contents to requesting users with low or no intervention from the CBS. As such, the traffic can be efficiently delivered while offloading the C-RAN's fronthaul links. These high-capability RRHs are henceforth referred to as enhanced RRHs (eRRHs). With EC, the wireless networks design experiences a paradigm shift from focusing solely on eRRH-user association to the consideration of the content dissemination at the eRRHs. This shift introduces a novel wireless architecture referred as Fog-RAN (F-RAN) [14], [15], [16]. Therefore, F-RAN can exploit the advantages of both centralized processing of C-RAN and fast access of EC [17]. Harvesting these advantages needs careful design of the content distribution among eRRHs and content delivery. The former needs to consider many factors, such as the cache placement (where is it best to put caches?), request-to-cache routing (how are cached content copies to be found?), and the cache capacity (how many contents to cache in each node?). The latter focuses on improving network resource utilization and efficiently delivering contents to their intended users [18]. In this thesis, we focus only on contents delivery problem which draws its importance from the required users' quality of service (QoS).

Long range communications through wireless links between eRRHs and users are subject to fading, shadowing, and so on. This results in link failures to deliver users' content(s). Motivated by this fact and evolution of IoT, device-to-device (D2D) communication technology is proposed as a key enabler of 5G networks [19], [20], [21], [22]. Due to its short range commu-

nication, cooperation, and data exchange between connected devices, D2D communication technology offloads the traffic from the eRRHs and CBSs and can thus free up some spectrum for them to boost data rates and serve other users [3, 23]. Thus, the performance of F-RANs can be further improved by implementing D2D communications, where caching at the eRRHs and users' cooperation via D2D communications are exploited. This integrated system is referred as D2D-aided F-RAN [24]. The performance of EC at the eRRHs can be further improved by pushing some popular contents to the devices near to the eRRHs. In the over congested network of beyond-5G era, the size of the popular contents will be significantly large and it will not be possible to cache the overall set of the popular contents in the eRRHs owing to their equipment costs and sizes issues. To circumvent this problem, caching at devices are employed [25], [26]. With distributed caching at eRRHs/devices and D2D-aided F-RAN system, not only the eRRHs transmit their cached contents to requested devices, but devices can also deliver the contents they cache to other devices via D2D communications. Besides its integration with F-RANs for contents delivery, D2D communication is also a core component in many mobile cloud/fog computing systems that distribute heavy processing tasks, e.g., distributed optimization and computation, over a computationally limited CBS.

Several advancements aimed to develop schemes that efficiently utilize radio resources to speed up contents delivery in the aforementioned cache-enabled 5G systems, i.e., C-RAN, F-RAN, D2D communications, and D2D-aided F-RAN. One trend is to assign a single user to each radio resource block (RRB), and thus transmit only one content from that RRB. However, with the tremendous increase of number of users nowdays, it is highly challenging to serve them by assigning a single user to each RRB. As previously explained, it has been observed that users tend to have a common interest in downloading popular contents, especially videos, within a small interval of time. The prior downloads of these popular contents by different users and the erasures of different contents create a pool of side information in the network, which is widely found in the current wireless networks. By combining this side information (i.e., downloaded/cached and erased contents)

using network coding, multiple users can be served simultaneously on the same RRB. Thus, significant improvements can be achieved in transmission efficiency, contents recovery, throughput and delay, and CBS offloading, in many sorts of wireless networks including the aforementioned cache-enabled 5G mobile systems. In accordance with the aforementioned overview, the main theme of this dissertation is to investigate effective resource allocation schemes using network coding for delivering users' contents and ensuring the required QoS in cache-enabled 5G networks.

1.2 Network Coding

1.2.1 Introduction to Network Coding

From its inception in 2000 to now, network coding (NC) has built a strong reputation in improving content[1] recovery process and achieving maximum information flow in both wired and wireless communication systems [27], [28], [29], [30]. Imagine a group of geographically close individuals are using their smartphones to stream popular videos, exchange information on social media, while others are meeting online through teleconferencing applications, and so on. To ensure this simultaneous demand, multicast and broadcast services have become key enablers in the design of wireless systems and networking protocols, e.g., long term evolution (LTE) [31]. These applications, especially video streaming, not only consumes huge amount of network resources, but also they are delay-sensitive. This means packets that are not instantly useful at their reception at users may cause interruption to the stream. Few years ago people would tolerate some delays for a video to be streamed before watching. This is almost unacceptable for any individual nowadays with the advancements of wireless communication and diversity of social media. Therefore, the design of wireless communication protocols should efficiently consider high-rate demand, maximum information flow, and immediate packet decoding to meet the delay requirements and ensure good streaming quality. To this end, NC can achieve maximum

[1] In the context of this thesis, popular content represents a frame of files/packets.

information flow by combining packets at the source and intermediate nodes in the network.

NC, initiated by Ahlswede et al. [27], performs on the network layer and relies on a simple idea; it enables intermediate nodes in the network to send a combination of the source packets they obtained instead of the simple receive-and-forward technique (i.e., uncoded). To attest this fact, let us consider a simple network that contains 1 BS, 2 users, and suppose that there are 2 packets need to be transmitted to the users over erasure-free channels. Assume that user 1 already received packet 2 and user 2 received packet 1. Traditionally, the BS needs two transmissions to complete the reception of all packets at both users. However, NC reduces the number to a single transmission by combining both packets at the BS. Generally, for a system of various users and source packets, the key potential of NC is to ensure a partial (or hopefully full) reception of these source packets at the multiple users by sending network coded packets. With such leverage of NC, different network metrics can be optimized, e.g., maximizing the throughput, minimizing the delay, or the number of transmissions. In the literature, two NC design methods are categorized into random (or full) network coding (RNC) and opportunistic network coding (ONC). In the former trend, the transmitter first combines all source packets using random and independent coefficients from a given Galois Field (GF), and then broadcasts this combined packet to all users. In the latter trend, the sender exploits the diversity of requested and received packets at each user in the selection process of the combined packets. As such, a particular optimization aim is achieved. It is worth mentioning that in wireless networks, channel impairments, such as fading and shadowing result in packet loss. Such a packet loss is seen, at the application layer, as an erasure [32], [33] that affects the delivery of information flow and the ability of users to decode the combined packets [34]. Consequently, a higher information flow in the network achieved by NC does not always mean a lower delay at the users [35], [36]. This is because the combined packets have to be retrieved (if they successfully received at users) first before using them at the application layer. Thus, optimizing one metric (e.g., throughput) degrades the other (e.g., delay), i.e., indeed there

is a throughput-delay trade-off.

The most remarkable example that shows the above-mentioned interplay of throughput and delay is Random Linear NC (RLNC) [37], [38], [39], [40], [41] in broadcast scenarios. To show this interplay, consider that a frame of M_{total} packets need to be transmitted to all users. Using RLNC, the transmitter combines these M_{total} packets and performs fountain property such that each single user successfully receives M_{total} of these independent combinations [42], [43]. From throughput perspective, RLNC achieves optimal throughput, and from delay perspective, each user is able to start decoding only after it has successfully received M_{total} combinations. Clearly, RLNC provides high information flow in the network and optimally in minimizing the number of packet transmissions in broadcast scenarios. Further, it does not require feedback from users for recovery. However, due to its high delay, RLNC is only suitable for delay tolerant applications as it does not provide instant packet decoding. From a complexity point of view, the computational complexity for decoding the packets at users increases cubically with the number of packets [42]. Such high complexity is not affordable on the battery-constrained devices. In addition, RLNC is not sufficient in multicast transmissions where different groups of users are interested in different subsets of packets.

In contrast, ONC is a potential candidate to overcome the aforementioned problems of RLNC. However, ONC has high encoding computational complexity and scalability issues in optimally solving many ONC problems in real time. This is because the optimal solution of online optimization (i.e., real time optimization) requires knowing all future channel realizations and the number of possibilities of users' side information. This is clearly intractable [44]. Fortunately, significant advancements in the analysis and algorithm designs for different optimization metrics have been achieved for a particular ONC, namely instantly decodable NC (IDNC). Lately, Sorour et al. [44], [46], [47], [48], [49], [50] studied deeply the instantaneously decodable codes and introduced the term IDNC. The key potential of IDNC is that it allows that the received coded packets are either immediately decoded at their delivery time or discarded. As a result, IDNC facilitates fast

and simple coding/decoding processes which are essential for future wireless networks. These fast and progressive packet reception features of IDNC balances between throughput-delay interplay with negligible decoding complexity, which ensures a better QoS for users. These simple and powerful features of IDNC attracted much attention in the last few years thanks to the following benefits:

1. The immediate decodability feature of received packets makes them useful at the reception time, which meets the streaming applications requirements [51].

2. Encoding at the transmitter is performed using a simple binary XOR operation, which reduces the coefficient reporting overhead as compared to RLNC. Likewise, decoding at the users is performed using a simple binary XOR operation, which overcomes the high computational complexity of RLNC.

3. Any received packets that are not instantly decodable will be discarded, thus IDNC overcomes the need for buffers to store these non-instantly decodable packets for possible future decoding. This feature not only overcomes the buffer requirements, but it also allows the design of cost and energy efficient devices.

Owing to its aforementioned benefits, IDNC has received a considerable attention in the research community, e.g., relay-aided networks [52], [53], [54], video-on-demand and multimedia streaming [55], [56], [57], [58], [59], [60], and D2D-enabled systems [61], [62], [63], [64], [65], [66], [67], [69], [70]. A survey of different techniques and applications of IDNC in centralized and decentralized networks can be found in [68]. The reason, besides its aforementioned features, of such tremendous investigations in the analysis and the development of simple algorithms for IDNC is its representation as a simple graph model as given below.

1.2.2 IDNC Graph

In order to represent all packet combinations that are instantly decodable by a subset (possibly all) users, the authors in [47] proposed a graph model, known as the IDNC graph. This graph model is first designed in [71], [72] as a heuristic algorithm to solve the index coding problem.
Before constructing the IDNC graph, we first introduce the two following sets for each user u_k in the network:

- The *Has* set: Denoted by $\mathcal{H}_{u_k}$ and defined as the set of packets successfully received at u_k-th user.
- The *Wants* set: Denoted by $\mathcal{W}_{u_k}$ and defined as the set of requested packets by u_k-th user. Intuitively, in the broadcast context, these two sets are complementary.

These two sets are referred as the side information of the users and summarized in a binary $U \times M$ *state matrix* $\mathbf{S} = [s_{u_k,p_l}]$ wherein the entry $s_{u_k,p_l} = 0$ represents that the p_l-th packet is successfully received at u_k-th user and 1 otherwise, U is the number of users, and M is the number of packets. Therefore, the transmitter constructs the IDNC graph solely based on the side information of users.

The key idea of constructing the IDNC graph is that two packets are requested by two users are combinable (coded) if these two users can both benefit from the combination. In other words, two packets can be combined into one combination if these packets are instantly decodable to the corresponding associated users. This scenario occurs if and only if (i) the two users are requesting the same packet, or (ii) the packet requested by each user is already received by the other user. These two constraints for the encoding process at the transmitter can be represented by edges (lines) in the IDNC graph. Likewise, the requested packets and their corresponding requesting users can be represented by vertices (circles) in the IDNC graph. The IDNC graph is an undirected graph, in which its edges have no orientation. Consequently, to construct the IDNC graph, denoted by $\mathcal{G}(\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ represents set of vertices and $\mathcal{E}$ represents set of edges, a vertex

$v_{u_k p_l} \in \mathcal{V}$ is generated for each requested packet $p_l \in \mathcal{W}_{u_k}$ by every user u_k. Two distinct vertices $v_{u_k p_l}$ and $v_{u_{k'} p_{l'}}$ in $\mathcal{V}$ are connected with an edge in $\mathcal{E}$ if one of the two following conditions is satisfied:

- $p_l = p_{l'}$: This represents that the same packet p_l is requested by both users u_k and $u_{k'}$.
- $p_l \in \mathcal{H}_{u_{k'}}$ and $p_{l'} \in \mathcal{H}_{u_k}$: This represents that the XOR combination $p_l \oplus p_{l'}$ is decodable by both users u_k and $u_{k'}$.

The aforementioned conditions emphasize that a feasible packet combination is a combination that helps at least a single user receiving one of its requested packets. Finding the optimal packet combination that provides the highest objective function to a given network's metric requires an exhaustive search over all possible file combinations, which is clearly non-feasible for large network sizes [68].

From the IDNC graph construction, each clique[2] in the graph represents a feasible packet combination that can be decoded by all users having vertices in that clique. Cliques usually show the tangible properties of graphs. For instance, one can consider that vertices in the graph represent people and edges represent friend's relationships. So, a clique in this case is a set of individuals who are all friends with each other. In fact, the term "clique", borrowed from French and used in a 1949's study of the above social network model [73], refers to a group of individuals with a common interest. A maximal clique in a graph is a clique that cannot be expanded to include more adjacent vertices without violating its connectivity conditions, i.e., it is not a part of a larger clique. Likewise, a maximal clique with the highest number of vertices in a graph is called the maximum clique. Unlike clique that provides local properties of the graph, the maximum clique reveals global trends of the graph.

As previously explained, each feasible packet combination can be represented by a clique in the IDNC graph. Therefore, finding the combination that provides the highest objective function is equivalent to finding the

[2]Clique is a set of vertices that are pairwise connected to each other.

maximum clique [68]. Without loss of generality, solving the IDNC related problems (e.g., completion time, decoding delay, so on) is equivalent to solving the maximum clique problem [68], where the maximum clique problem is a problem that finds the largest set of vertices in a graph. Although looks easy to solve, maximum clique problems are NP-hard problems [74], [75], [76]. A detail analysis of the developed algorithms, complexity, and applications of maximum clique problems can be found in [77]. The maximum weight clique problem is similar to the maximum clique problem, but it is applied on weighted graphs. In weighted graphs, every vertex has a particular weight that reflects the contribution of that vertex to optimize a given metric in a network. Throughout this thesis, we consider a weighted IDNC graph, where the weight of a vertex $v_{u_k p_l} \in \mathcal{V}$ is denoted by $w(v_{u_k p_l})$. Consequently, we consider maximum weight clique problems.

To illustrate the construction of the IDNC graph and its corresponding maximum clique, let us consider a simple example in Figure 1.1. We consider an arbitrary users' side information as shown in the feedback matrix on the left side of Figure 1.2. Assume error free transmissions. Based on users' side information and RRH's cached files, the corresponding IDNC graph is shown on the right side of Figure 1.2. For example, packet p_3 that is requested by user u_1 is represented by a vertex $(1, 3)$ in the graph. In order to maximize the number of targeted users for this example, many possible solutions can be found. One possible solution is that RRH combines packets p_1, p_3 and transmits the combination $p_1 \oplus p_3$ to the set of targeted users u_1, u_2, u_3, and u_4. The corresponding vertices of this combination are drawn by yellow color in the figure. Under this scenario, the decoding process at the users are as follows.

- The u_1-th user already has p_1, so it can XOR the combination $(p_1 \oplus p_3)$ with p_1 (i.e., $(p_1 \oplus p_3) \oplus p_1$) to retrieve p_3. In this case, the transmission is instantly decodable for u_1.

- The u_2-th user already has p_3, so it can XOR the combination $(p_1 \oplus p_3)$ with p_3 (i.e., $(p_1 \oplus p_3) \oplus p_3$) to retrieve p_1. In this case, the transmission is instantly decodable for u_2.

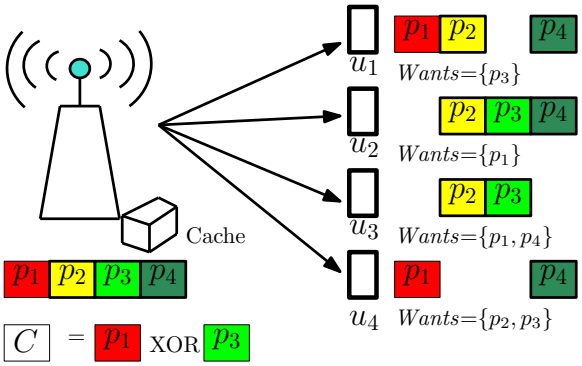


Figure 1.1: A simple model containing 4 users, 4 packets, and 1 RRH. The RRH cached the whole set of packets. The side information of users is also shown, where each user has received some packets and wants some other packets. For example, user u_1 already received packets p_1, p_2, and p_4 and requests p_3.

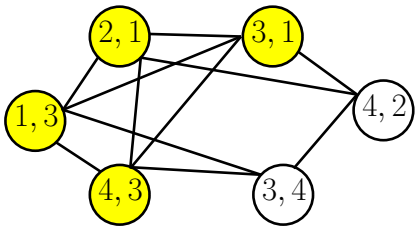
- The u_3-th user already has p_3, so it can XOR the combination $(p_1 \oplus p_3)$ with p_3 (i.e., $(p_1 \oplus p_3) \oplus p_3$) to retrieve p_1. In this case, the transmission is instantly decodable for u_3.
- The u_4-th user already has p_1, so it can XOR the combination $(p_1 \oplus p_3)$ with p_1 (i.e., $(p_1 \oplus p_3) \oplus p_1$) to retrieve p_3. In this case, the transmission is instantly decodable for u_4.

1.2.3 IDNC and Rate-Aware IDNC

The previously explained IDNC technique is solely performed at the network layer and depends on users' side information in the network. Thus, the coding decision at the transmitter solely depends on the requested and cached packets at the users. Essentially, IDNC aims at finding a packet combination that hopefully satisfies all users in a network as shown in the previous example and in Figure 1.2. To satisfy all users, the transmission rate of the transmitter is adopted to the minimum achievable capacity of the assigned users. This results in serving a large number of users but also leads to physical layer throughput degradation. Therefore, IDNC maximizes

	p_1	p_2	p_3	p_4
u_1	0	0	1	0
u_2	1	0	0	0
u_3	1	0	0	1
u_4	0	1	1	0

Feedback matrix giving the requested and received packets for all users. "0" means that a packet is received and "1" means that it is requested



IDNC graph and its corresponding maxmium clique

Figure 1.2: Illustration of the IDNC graph. Each vertex represents a requested packet by one user, i.e., vertex $(2, 1)$ represents the indices of user u_2 and packet p_1. Each clique in the IDNC graph represents a feasible IDNC combination. For example, one possible clique is: $\{(1, 3), (2, 1), (3, 1), (4, 3)\}$ that corresponds to transmitting the combination $p_1 \oplus p_3$ to the targeted users u_1, u_2, u_3, and u_4.

the information flow in the network layer, but it degrades the physical-layer performance. In fact, wireless channels are dynamic with heterogeneous capacities, which requires the transmitters to judiciously adopt the transmission rates. For instance, a packet transmission with a high rate will take a shorter time but will be only successfully received by the users with high channel capacities. In contrast, a packet transmission with a lower rate will be received by more users but will take a longer time. In order to improve the physical layer performance and users' QoS, the channel capacity should be included as another factor in the network coding decisions. As a result, the coding decision depends on both the requested and cached packets at the users and their channel capacities. This is called rate-aware IDNC (RA-IDNC) [78]. Thus, RA-IDNC scheme has the potential to serve significant set of users with relatively fair transmission rate. Thanks to this potential feature, RA-IDNC scheme has been recommended in a number of works, for example [70], [78], [79], [80], [81], [82].

1.2.4 Game Theory

In the centralized scenarios (e.g., C-RAN), the CBS performs the whole process of constructing the IDNC graph, planning IDNC packet combinations, and coordinating transmissions. However, the aim of the distributed systems, i.e., D2D communications, is to offload the CBS. In this regard, game theory is an amazing tool to model distributed systems without a central CBS. In game theory, users, known as game players, interact with other game players to achieve the desired global objective [83]. Two types of game theory are usually used, namely non-cooperative games and coalition games, also known as cooperative games. In non-cooperative games, each player makes its decisions individually and selfishly. In coalition games, players interact with each other to make cooperative and altruistic decisions to achieve further the delay for delivering the contents [83]. In order to improve the distributed solutions using game theory, studying and optimizing the different type of equilibrium, such that the Nash equilibrium and the Pareto optimal [84], is crucial. Game theory has been well investigated to solve NC problems in different D2D models, see the survey paper [68]. In general, the considered games in those works were complete and perfect. The complete information assumption means that the utility function and the strategy that each player can take are known by all other players involved in the game. Likely, the perfect information assumption means that all players know the history of the game perfectly. In this thesis, we are interested in incomplete information game that draws its importance from the fact that limited-coverage zone players are distributed in an area, where players know the information about their neighbors, not all players in the game course. An elegant subclass of coalition games to model NC problems is the coalition formation game (CFG). The game is called CFG because it involves the formation of independent and disjoint coalitions. The key idea of forming coalitions is to study the cooperative behavior of players in coalitions. Indeed, UDs are formed coalitions so that they maximize/minimize the utility (e.g., minimize completion time, minimize delay, maximize throughput, and so on) compared to when they do not form coalitions.

1.3 Literature Review

1.3.1 Throughput Maximization in C-RAN System

C-RANs are promising solutions to satisfy the growing traffic demand in 5G mobile networks. In the C-RANs literature, two common strategies have been proposed, namely *data-sharing* strategy and *compression-based* strategy, to efficiently use the fronthaul links [85]. In the data-sharing strategy, the CBS splits and shares user's message with only a subset of RRHs, which encode these messages and send them over wireless channels to the users [86], [87]. The RRHs in the compression-based strategy transmit only compressed versions of the encoded messages to the users [7]. Different coding schemes [88], [89], [90] have been studied in the literature as a means to maximize the achievable rates. For example, in [89], the authors proposed two lattice-based coding schemes that maximize the achievable sum-rate in uplink multi-cell networks. In [91], the authors built on the code proposed in [90] to further improve the sum rate. Cross-layer design that uses NC has also been proposed to optimize different parameters in different network settings [92], [96]. In [92], the authors used NC to maximize the sum-rate in C-RANs. The authors of [93], [94] proposed a cross-layer design for network planning in Ad-Hoc wireless networks. In [95], [96], the authors considered a cross-layer optimization and NC to optimize the throughput in wireless queuing tandem and wireless multi-hop networks, respectively. However, the aforementioned existing works require optimization over different factors, such as RRHs beamforming vectors, user-RRH association, and message traffic over the fronthaul links. This makes the C-RAN's design complex. Therefore, we consider in this section the downlink of C-RANs and its existing works from a different perspective. Particularly, works considered the joint target of users-RRHs scheduling policy, users' side information, and power control at the CBS, and relay them to the RRHs through low-rate fronthaul links. In general, the existing works addressed the throughput maximization problem and its related factors individually. Particularly, works discussed in [7], [97], [98], [99], [100], [101], [102], [103], [104], [105], [106], [107], [108], [109], [110] considered only the physical-layer

factors, such as resource scheduling and power adaptation. This results in assigning a single user to each RRB. On the contrary, works presented in [44], [48], [68], [111], [112], [113], [114], [115], [116] considered only upper-layer factors, such as users' side information. This allocates many users to each RRB. However, in order to serve all assigned users, each RRB in each RRH adopts its transmission rate to the minimum rate of all its assigned users.

The throughput maximization problem in C-RAN, also known as joint user scheduling and power allocation problem, is shown to be challenging [97], [98], [99]. Indeed, besides its joint combinatorial nature, the mathematical formulation reveals that the power adaptation itself in this setting can be non-convex. A large body of literature, see for examples, [97], [98], [99], [103], [104] tackled the problem by solving it iteratively. The authors of [97], [99] solved the problem via an integer relaxation method in single and multi-cell configurations, respectively, to maximize throughput. In [7], [105], the authors considered inter-cell level coordination in which only low-rate fronthaul link is used. The studies are extended in [106], [107], [108] to optimize the power level of each RRB through a power optimization step. To this end, the authors of [109], [110] considered a scheduling-level coordinated C-RAN to solve the joint user scheduling and power optimization problem using graph theory techniques. As previously stated, all studies mentioned above viewed the network solely from the physical-layer perspective without taking into consideration upper-layer facts. This results in assigning a single user to each RRB. However, it has been observed that users tend to have a common interest in downloading popular files, especially videos, within a small interval of time. This creates a pool of side information in the network. This side information is widely found in the current wireless networks. For example, the channel imperfections, e.g., interference, fading, etc. always cause file's erasures. The diversity in prior downloads of the popular files by different users generates side information for their new requests by others. Also, the rate-based service optimization protocols with users' achievable capacities play a significant role in generating side information. For clarification, let us consider a scenario of two groups of users where both groups

are interested in popular files. The achievable capacities of the first group are much lower than the second one. The modern LTE broadcast will target the first group of users with very low rates. If we use NC, users in the first group would not be served in the current transmission's frame and left to be served on a later transmission's frame when their rates become higher. This results in generating side information, which can be clearly exploited using NC to maximize throughput. In NC, such side information is employed to mix users' requests and aid the coding decision at the CBS. Hence, it leads to serving these users simultaneously on the same RRB.

A large body of literature, see for example, [44], [48], [68], [111], [112], [113], [114], [115], [116] optimized throughput, i.e., maximized the information flow in the network, by taking advantage of users' side information. For example, the authors of [44], [48] reduced the total completion time and the decoding delay in point-to multipoint system (PMP), respectively. In [115], a delay-based analysis is provided to reduce the completion time. However, they ignored the physical-layer factors in the coding decisions. If we incorporate some physical-layer factors in the NC decisions, we can optimize the cross-layer throughput. In the recent literature, RA-IDNC scheme has been recommended in a number of works, see for example [78], [79], [80], [81], [82]. In the RA-IDNC scheme, both the side information and rate heterogeneity of different users to different RRHs are used for making NC decisions. Thus, RA-IDNC scheme has the potential to maximize the system's throughput. The authors of [81] proposed a novel joint rate selection and network coding (RSNC) scheme for increasing the system's benefit in PMP. In [80], [82], the authors proposed RA-IDNC scheme to minimize the actual completion time in PMP and C-RAN systems, respectively. However, in [82], the authors assumed that each RRH maintains at a fixed power. Moreover, the transmission rate of all RRHs is the same and will be limited by the weakest RRH, i.e., the RRH who can support the lowest transmission rate. This may violate the QoS rate guarantee.

From the above discussions, we observe it is crucial to consider both the NC at the network layer and the aforementioned physical layer factors, to improve throughput. However, NC and C-RAN's resources can be

contradicting as their combined effect, if not jointly optimized, can result in a performance degradation. Indeed, by combining users' side information and scheduling multiple users to the same RRB, the total number of targeted users in each RRB increases. For successful transmission for the scheduled users, the transmission rate in each RRB will be dictated by the worst user's rate. Consequently, if such user-RRB associations and rates are not carefully chosen when deciding on file combinations, the throughput of each RRB may decrease compared to serving a single user in it. This results in worse throughput performance. Further, power level employed in each RRB plays a significant role in this trade-off due to the interference on the users served by the same RRB in different RRHs. Therefore, it is important to investigate these trade-offs between users' side information, their associations to RRBs, power adaptation, and rate heterogeneity in C-RAN resource setting. In this context, in Chapter 2, we introduce a fully cross-layer optimization framework for capturing the above-mentioned interplays. We investigate the following two scenarios. In the first scenario, we solve the throughput maximization problem by jointly considering the network-coded user scheduling and power optimization. In the second scenario, we overcome the high implementation complexity of the joint solution by solving the problem iteratively.

1.3.2 CBS Offloading in F-RAN System

CBS offloading is an important performance metric that has to be optimized to compensate the limited capacity of C-RAN's fronthaul links. In order to optimize the offloading of CBS's resources, i.e., time and frequency, RRHs need to pre-store and transmit users' side information at the network edge (i.e., close to the end users). By pre-fetching most users' side information to their local storages at off-peak times, eRRHs can efficiently deliver such files to requesting users with low or no intervention from the CBS. As such, CBS's resources are significantly offloaded. Therefore, F-RAN can exploit the advantages of both centralized processing of C-RAN and fast access of edge caching [17]. Harvesting these advantages needs careful op-

timization of the files' distribution among eRRHs as well as the immediate and fast delivery of requested files to users to satisfy their QoS. Thus, addressing the scheduling of immediate file delivery problem is important to meet the required QoS. If all popular files are cached in eRRHs, the immediate file delivery to users may necessitate the involvement of the CBS in serving the users that cannot be immediately served by the eRRHs. Thus, an efficient scheduling of file delivery from the eRRHs and (possibly) the CBS is required not only to maximize the CBS offloading (i.e., minimize the amount of CBS resources consumed in serving unserved users by the eRRHs), but also to guarantee the required QoS. The joint CBS offloading and QoS guarantee problem in F-RANs involves many factors, such as the different subsets of the files cached in each eRRHs, the users' requested files, their rates, need to be jointly considered. Existing works considered the above factors and their corresponding problems separately. In particular, works presented in [7], [109], [119], [120] considered optimizing user-eRRH associations and power in F-RANs, without any upper-layer considerations. Moreover, these works did not attempt to reduce CBS involvement in delivering popular files. Other works [121], [122], [123], [124], [125] addressed a simplified version of the problem, in which the CBS offloading problem is solved irrespective of users' QoS in F-RAN.

In the recent literature, a fronthaul-upper layer offloading design with one-to-one and one-to-many user-cache connections has been investigated in a number of works, e.g., [121], [122], [123], for small network settings. The focus of these works was mostly on cached files optimization and user service delay, without any immediate request of such service nor exploitation the potentials of the IDNC to maximize fronthaul offloading. IDNC [124] suits most of the critical-time applications that need an immediate delivery of users' requested files. The authors of [125] developed a framework for IDNC download time optimization from multiple CBS storage servers. One step further, the authors of [126] built on that model to employ NC in CBS offloading given the immediate file delivery constraint. The main limitation of the work in [126] is its network-layer view of the problem. It employed NC on both the eRRHs (therein called femto-caches) and CBS (therein called

macro base-station) to minimize the number of CBS channels required to serve the users that are not instantaneously served by the eRRHs. However, this work did not consider the heterogeneity in the rates of these users left to be served by the CBS. The study in [126] is extended one step further by using RA-IDNC to achieve better physical CBS offloading, i.e., reduces the consumed frequency and time resources of the CBS [127]. To the best of our knowledge, [127] was the first attempt to achieve CBS offloading maximization while considering some physical-layer aspects using RA-IDNC. However, the work in [127] did not consider the standard F-RAN physical-layer resource structures that can be possibly allocated to different sets of users. More importantly, it did not consider to optimize rates through power selection in each RRB. Furthermore, no attention was given to the users' achieved QoS. Unlike the RA-IDNC scheme that uses the rates as passive inputs, the proposed cross-layer NC controls these rate-aware codes by deciding on the set of power levels of the RRBs and their corresponding transmission rates in F-RAN setting. Motivated by the above limitations of the existing works, it is crucial to propose an efficient cross-layer resource scheduling of file delivery from the eRRHs and (possibly) the CBS. As such, the CBS offloading is maximized while guaranteeing the required QoS. Clearly, maintaining the aforementioned two objectives in F-RAN settings can be contradicting. For example, by coding more (or judiciously selected) users' requests to each of the RRBs, the number of users left to be served by the CBS (if any) can be reduced. As mentioned before, the transmission rate in each RRB will be limited by the weakest user. This may violate the QoS rate guarantee. On the contrary, pre-setting a minimum rate in each RRB usually results in assigning less users to it. Thus, it increases the burden on the CBS to serve the remaining users. Consequently, in Chapter 3, we study such high trade-off between throughput maximization and CBS offloading.

1.3.3 Cooperative F-RAN and D2D Communications

As mentioned before, F-RAN has been introduced to exploit both EC and C-RAN for carrying out file delivery effectively. In order to further

improve the performance of F-RANs, implementing D2D communications in F-RAN is shown to be a potential technology in 5G and beyond. This integrated system is referred as D2D-aided F-RAN [24]. D2D-aided F-RAN system draws a remarkable benefit for reducing both users' files delivery time and burden on fronthaul links. Thanks to the EC at the eRRHs and users' cooperation via D2D communications, we focus on files delivery problem in D2D-aided F-RAN system. The files delivery problem of interest is motivated by immediate delivery of common popular files for real-time applications, i.e., live video streaming. In particular, we study the scheduling of files delivery problem from both the eRRHs and potential transmitting users in D2D-aided F-RAN system using NC. The problem of delivering a frame of delay-sensitive files to a set of users with minimum possible delay has been a topic of research for a quite some time. This problem is referred as completion time minimization problem. Based on layer functionalities, existing NC solutions for this problem can be classified into upper-layer NC [44], [48], [60], [67], [71] and RA-IDNC [78], [79], [80], [81], [82] methods. As their names indicate, upper layer NC algorithms focused only on NC at the network layer to minimize the number of transmissions. RA-IDNC approaches incorporate both upper and physical layers to minimize the completion time (in second) required to deliver requested files to all requesting users. The latter is more practically relevant as it involves the dynamic nature of wireless channels in the completion time optimization.

The completion time minimization problem in IDNC-based networks was considered in different network settings, e.g., PMP [44], [48], D2D networks [61], D2D F-RAN [67]. In particular, the authors of [44], [48] proposed schemes to deliver the requested files by users with a minimum possible number of transmissions. Recently, in [67], a centralized D2D F-RAN scheme was proposed for completion time reduction. However, the aforementioned works considered IDNC from the perspective of network-layer. The main drawback is that the transmission rate of each RRB is selected based on the user with the weakest channel quality. This results in prolonged file reception time and thus, consumes the time resources of network. Therefore, considering both network layer coding and physical layer factors for the RA-IDNC

is crucial [80]. With RA-IDNC, the completion time minimization problem needs a careful optimization of selecting the IDNC file and transmission rate of each RRB, see for example [78], [79], [80], [81], [82]. The authors of [82] used RA-IDNC in C-RANs for completion time reduction. However, the authors assumed that all RRHs maintain a fixed transmit power level. Moreover, for synchronization purposes, the same transmission rate (i.e., the lowest transmission rate of all RRHs) is selected. This may violate the QoS rate guarantee and lead to a longer time for file transmission. Importantly, the proposed solution did not exploit the high capabilities of D2D communications. Recently, a cross-layer IDNC scheme was proposed for throughput maximization in C-RAN and CBS offloading in F-RAN [117], [118], [128], [129] respectively. Inspired by our preliminary works [118], [129], in Chapter 4, we address the completion time minimization problem in D2D-aided F-RAN system using RA-IDNC and D2D communications. The considered completion time minimization problem in D2D-aided F-RAN involves many factors, such as cached files at the eRRHs, users' limited coverage zones, their requested and previously received files, and their heterogeneous physical-layer capacities. To the best of the our knowledge, the literature does not solve the completion time minimization problem while considering all above factors. Therefore, a balance among the conflicting effects of the aforementioned factors, such as IDNC codes, scheduled users, and transmission rates of eRRHs and users in D2D-aided F-RAN system is crucial to minimize the total frame delivery time. In Chapter 4, we investigate two scenarios as follows. In the first scenario, we solve the completion time problem by first minimizing the transmission time from the eRRHs. Given the obtained transmission time from the eRRHs, we maximize the number of unscheduled users to eRRHs by implementing D2D communications. In the second scenario, we first schedule users via D2D communications and obtain the possible transmission time. Then, the unscheduled users via D2D communications are scheduled to eRRHs over cellular channels during the obtained transmission time from the D2D communications.

1.3.4 Content Delivery in D2D Networks

The use of smartphones and data-hungry applications in radio access networks are increasing dramatically worldwide. This growth impacts the ability of traditional wireless networks to meet the required QoS for its user devices (UDs). As explained earlier, D2D communication, as one of the candidate technology for 5G and beyond [3], supports a massive number of connected UDs and possibly improves the data-rate without fixed infrastructure for content delivery [19], [20], [21], [22]. The decentralized nature of D2D networks allows UDs to communicate with other nearby UDs via D2D links, which is suitable for numerous applications in mobile networks. For example, in wireless cellular networks, a D2D system enables cellular offloading by UD cooperations for content downloading and sharing. Based on [1], most wireless data contents are mainly distributed in hotspot areas, such as a playground, public transport, a conference hall. In such hotspots, there is a tremendous high UD density that requests high volumes of data traffic, which causes network congestion and interruption at UDs. Actually, it is noticed that some UDs tend to have a common interest in downloading the same content, known as popular content. This frequently happens where UDs download a popular application, an electronic map, or streaming multimedia contents, such as videos. This consumes a huge amount of resources of cellular systems [11]. Therefore, D2D communication technique can be exploited to distribute the popular content as such to offload the CBS and reduce network congestion in the cellular networks. For instance, consider that a content consists of 3 packets p_1, p_2, and p_3 is requested by 3 UDs u_1, u_2, and u_3. Suppose that the CBS transmitted the requested contents to the UDs and due to channel impairments UD u_i did not receive packet p_i for $1 \leq i \leq 3$. Traditionally, the missing packets are retransmitted repeatedly from the CBS to each UD. As a result, the CBS needs at least 3 uncoded transmissions for packets delivery, which degrades system performance [130]. However, UDs can be either packet holders providing their received packets to other UDs or packet requesters receiving the requested packets from other UDs. For example, u_1-th UD holds p_2, p_3, and accord-

ingly, transmits binary XOR combination $p_2 \oplus p_3$ to the u_2-th and u_3-th UDs. Then, u_2-th UD holds p_1 and can provide it to u_1-th UD. As a result, 2 transmissions are required for packets delivery. Therefore, D2D techniques can be used with binary XOR combination to combine contents and transmit them to interested UDs via D2D links. Thus, minimizing the number of transmission slots and offloading the CBS's resources. The completion time minimization problem was solved centrally in different network settings for various applications. For example, Raptor codes [131], and RLNC [39] addressed the problem and achieved maximum network throughput. However, they are not attractive techniques for delivering delay-sensitive contents for real-time applications, such as online gaming, and teleconferencing. On the other hand, IDNC improves throughput while allowing progressive decoding of the received contents, e.g., [68] and its references. It has been applied in several real-time broadcast applications wherein received contents need to be used at the application layer immediately to maintain a high QoS, e.g., relay-aided networks [52], [53], [54].

The aforementioned IDNC works that are listed in the survey paper [68], for both fixed infrastructures and D2D networks, are centralized in the sense that they require a central CBS to plan packet combinations and coordinate transmissions. For example, the authors of [67] considered the completion time minimization problem in a partially connected D2D fog network. The problem was solved by optimization under the assumption that the fog is within the transmission range of all UDs and has perfect knowledge of the network topology. Accordingly, the fog selects transmitting UDs and their packet combinations and conveys the information to the UDs for execution. The above mentioned centralized approaches provide good performance for the decentralized system, but they bring large challenges to the system. First, they require high computation cost at the fog units and high power consumption at each UD. Indeed, UDs need to send the status of all D2D channels to the fog after each transmission slot. Second, the fog requires to know the downloading history of UDs for content delivery. Recently, in [132], [133], a non-cooperative game was used to reduce the communication time for content reception. Still, the authors just considered a *fully* connected

scenario that only one player transmits the contents in each transmission slot. The fully connected model is ideal in the sense that all players are connected, and each player makes its decisions individually and selfishly. Thus, it ignored the cooperative and altruistic decisions of players to minimize further the delay for delivering the contents [83]. The cooperative and altruistic decisions of players have shown the potential of coalition games in optimizing different parameters in different models [134], [135], [136], [137], [138], [139]. For example, the tutorial in [134] classified the coalition games and demonstrated the applications of coalition games in communication networks. The authors of [135] proposed a distributed game-theoretic scheme for players' cooperation in wireless networks to maximize the sum-rate. In [139], the authors recommend that UDs can cooperatively obtain the same content through multi-hop D2D communications. However, the aforementioned works are agnostic to the available side information at the network layer, i.e., requested and previously received contents by different UDs. As a result, each UD sends uncoded (i.e., without NC) content that serves a single UD. Such side information can be exploited to efficiently select a combination of contents that can benefit a subset of interested UDs.

In addition to their investigations in uncoded existing works, coalition games for network coding-enabled networks have been extensively explored from various perspectives [140], [141], [142]. Prior network coding works with coalition games aimed to maximize sum rate, e.g., [140], [141]. In particular, the authors in [141] showed the potential of network coding in maximizing the sum rate in network coding-aided D2D communication. Moreover, some prior works analyzed content dissemination when network coding is enabled, e.g., [143], [144]. The authors in [143] presented efficient non-cooperative game theoretic scheme for content dissemination using RLNC. However the above related works [143], [144] used RLNC for content dissemination. As previously mentioned, RLNC is not attractive technique for delivering delay-sensitive contents for real-time applications of interest in this work. Recently, the authors of [145] proposed a coalition game framework for packet recovery reduction in a partially connected D2D network using IDNC. However, the authors assumed that the cluster head selection

is predetermined. Moreover, all clusters are formed once at the beginning of recovery phase. This does not give flexibility to UDs to choose their own groups in the next transfer. Importantly, if no available packets in any cluster, UDs in that cluster have to stop their intra-cluster recovery phase and wait for their cluster head to perform the inter-cluster recovery phase. This results in prolonged packets delivery time. To this end, we consider a more general setting in which a decentralized and distributed framework is developed to: i) derive the rules for optimizing the selection of the transmitting UD in each formed coalition, and ii) give flexibility to UDs to choose their own alliances at each transmission. As such, all packets are delivered to all UDs in only intra-coalition delivery phase with minimum completion time.

Motivated by the above limitations of the existing works, in Chapter 5, we consider a partially connected D2D networks comprising several single-antenna UDs distributed in a hotspot, and each UD is partially connected to other UDs. The completion time minimization problem is motivated by real-time applications that tolerate only low delays. In such applications, the contents, represented by a frame of packets, need to be delivered to UDs via D2D links with few numbers of D2D transmissions. Our proposed model appears in different applications. For example, UDs at the edge of the service area or in dense urban areas often experience packet losses from CBSs due to channel impairments. Our proposed distributed scheme would serve these UDs via D2D links. Moreover, in cell centers with low erasures, our proposed scheme would offload the CBS's resources, e.g., time, bandwidth, and the ability to serve more UDs. To this end, our main goal in Chapter 5 is to develop a distributed framework that models the collaborations among the UDs of an IDNC-based D2D network. In order to do that, we use game theory because it involves a set of players that interact with each other to form a coalition without any coordination from the CBS. Further, as the requested packets are transmitted simultaneously from multiple UDs, we can formulate the content delivery problem as a coalition game, and each group of UDs constitutes a coalition. Thus, each UD acts as a game player to be either packets holder providing its acquired packets to other players or packets requester. Meanwhile, each game player has its own decision to

join or leave the coalition based on its preference as well as other players' preferences in that coalition. Therefore, the interactions among players and their cooperative decisions motivate us to model and solve the completion time problem in a distributed manner using game theory.

1.4 Thesis Objective

In accordance with the aforementioned introduction and background, the main theme of this dissertation is to investigate NC schemes for throughput maximization and completion time minimization in cache-enabled 5G systems. In particular, we consider instantaneous NC schemes for the following cache-enabled 5G networks: (i) C-RAN downlink communication system; (ii) F-RAN downlink communication system; (iii) D2D-aided F-RAN system; and (iv) partially connected D2D networks. To this end, this thesis is built on the following four research objectives.

1. Develop cross-layer NC schemes for C-RAN downlink communication system to efficiently stream a set of popular files to users with the maximum throughput.

2. Investigate joint cross-layer CBS offloading and users' QoS guarantee optimization in F-RAN system by considering cross-layer NC, power optimization, and distributed caching strategy at the eRRHs.

3. Develop joint and iterative rate-aware NC schemes for completion time minimization (in second) in F-RAN architecture with D2D communications.

4. Develop a distributed and decentralized game-theoretical NC framework for partially connected D2D networks to minimize the completion time in terms of number of D2D transmissions.

1.5 Thesis Contributions

The key contribution of this thesis is to design innovative NC schemes for effective resource allocation in cache-enabled 5G networks, where the goal is to optimize different metrics, such as throughput, CBS offloading, and completion time. Specifically, in this thesis, we make the following four-fold contributions.

- *Development of cross-layer NC schemes for C-RAN system:* The main novelty of this contribution is the development of cross-layer NC schemes for C-RAN resource setting. The developed joint and iterative cross-layer NC schemes maximize the cross-layer throughput of C-RAN system by performing user scheduling, file encoding, and power allocation. The main contributions of this part as follows. First, we introduce a new graph, known as CRAN-CLNC graph, to formulate the cross-layer throughput maximization problem as a maximum weight clique problem. Second, using this graph, we develop a joint solution that greedily selects the maximum weight clique. For the high implementation complexity of the joint scheme, we develop an efficient iterative scheme that has relatively low-complexity. Numerical results reveal that the proposed joint and iterative schemes offer improved throughput performances as compared to the existing algorithms in the literature. Compared to our proposed joint scheme, the iterative scheme has a certain degradation, roughly in the range of 9%-14%. This small degradation in the throughput performance of the iterative scheme comes at the achieved low computational complexity as compared to the complexity of the joint scheme.

- *CBS offloading, QoS trade-off in F-RAN system:* The main novelty of this contribution is to propose smart scheduling of file delivery schemes for maximizing the CBS offloading while ensuring the required QoS in F-RAN system. In particular, for F-RAN system, we develop framework(s) where eRRHs and CBS judiciously collaborate to immediately deliver a set of requested popular files to users. As such, CBS offload-

ing maximization and required QoS guarantee trade-off is investigated. The main contributions of this part are as follows. First, we use the ONC graph to formulate the joint cross-layer CBS offloading and QoS guarantee optimization problem, and we show its NP-hardness. To tackle the problem, we design a graph, known as FRAN-CLNC graph, that considers both aforementioned C-RAN's optimization factors, the different subsets of the files cached in each eRRHs, the required QoS. In this FRAN-CLNC graph, cross-layer vertex weight is designed, where one vertex weight corresponds to CBS offloading while the other corresponds to QoS guarantee (i.e., rate maximization). The joint problem is, then, formulated on the FRAN-CLNC graph and shown to be equivalent to maximum-weight independent set and maximum weight coloring problems. Afterward, a greedy vertex search and vertex coloring schemes are developed to efficiently solve the joint problem. Second, we proposed a simpler and efficient solution that mainly performs on solving the joint cross-layer CBS offloading and QoS guarantee optimization problem iteratively. Presented numerical results reveal that both proposed schemes achieve significant gains in users' throughput compared to the existing solutions. Compared to the QoS un-aware algorithm, our proposed schemes have a certain degradation in CBS offloading.

– *Development of RA-IDNC schemes for completion time minimization in D2D-aided F-RAN system:* The main novelty is to introduce novel optimization framework(s) taking network coding, rate adaptation, potential D2D communications, and users' limited coverage zones into account. As such, we minimize the completion time for sending all requested files to all users in D2D-aided F-RAN system. Given the intractability of solving the completion time minimization problem over all possible future NC decisions, we reformulate the problem at each transmission with the constraints on user scheduling, their limited coverage zones, transmission rates, and QoS rate guarantee. Then, through transmission time analysis, we decompose the problem into

two subproblems, namely transmission time minimization subproblem and number of served users maximization subproblem. The first part in this contribution aims at proposing a joint approach that develops network-coded transmissions from both eRRHs and potential users for the aforementioned subproblems. The second part aims at proposing an efficient and effective coordinated scheduling solution from both D2D transmitters and eRRHs. Simulation results have shown that our proposed schemes can effectively minimize the frame delivery time as compared to conventional schemes.
The main observations of the above three-fold contributions are as follows.

1. It is advantageous to serve many (possibly all) users with NC files as in the classical IDNC algorithm, but selecting the minimum transmission rate of all scheduled users degrades its performance, e.g., throughput, CBS offloading, completion time, and so on. Thus, this scheme is impractical from physical-layer perspective.
2. It is advantageous to allocate users experiencing high rates to the RRBs as in the uncoded (i.e., without NC) scheme. However, with the increase of number of users, it is highly challenging to serve them by assigning a single user to each RRB. For a system consists of Z_{total} RRBs, the maximum number of scheduled users is Z_{total}, which limits the overall gain in the average throughput of the uncoded scheme. Further, at each transmission, the number of transmitted uncoded-files from all RRBs is limited to only Z_{total} files. As a result, a large number of transmissions is required for transmitting a frame of files. Consequently, employing this scheme in practical is highly challenging as it consumes network resources.
3. Our developed solutions strike a balance between the aforementioned aspects by judiciously scheduling a significant set of user given their requested files are combined using NC, adopting the transmission rates of each RRB and optimizing the transmission

power of each RRB. Such novel solutions offer improved performance in terms of throughput, CBS offloading and QoS improvement, and completion time in different resource settings, such as C-RAN, F-RAN, and D2D-aided F-RAN.

– *Game-theoretical development for contents delivery in D2D networks:* Our main novelty in this part is the development of a distributed and decentralized game-theoretical framework for contents delivery minimization in partially connected D2D networks. As such, we overcome the need for the CBSs while minimizing the completion time. The main contributions of this part are given as follows. First, we develop a decentralized IDNC-assisted D2D game theoretical framework to minimize the completion time while offloading the CBSs. In particular, the completion time minimization problem is formulated and modeled as a coalition game. Given the difficulty of expressing the problem as a coalition game with non-transfer utility (NTU), we relax it to a coalition formation game (CFG). Second, in the relaxed CFG, the completion time metric is written as the utility function, which is transferred to each player's payoff in each coalition. We derive the rules for associating players, selecting the transmitting player, and finding its packet combination that is beneficial for a set of interested UDs in each disjoint altruistic coalition. Afterward, we develop a coalition formation distributed algorithm in each transmission slot based on merge-and-split rules. Finally, the proposed coalition formation algorithm is proved to converge to a Nash-stable equilibrium, and its complexity and communication overhead are analyzed theoretically. We validate our theoretical finding using comprehensive numerical simulations, which reveal that our distributed scheme can significantly outperform existing centralized and fully distributed methods. For the presented network setups, our proposed decentralized scheme offers almost the same performance as the centralized F-RAN D2D scheme.

1.6 Thesis Organization

This thesis is organized into six chapters. In what follows, we summarize the material of each chapter.

Chapter 1 is divided into two main parts, introduction and background and literature review. In the first part, we presented a brief background on the development of cache-enabled 5G networks along with the competence of EC technology for 5G fronthaul network offloading. We, then, provided a comprehensive introduction to network coding and its sub-classes. We further showed how to construct the IDNC graph and find its corresponding maximum clique. A brief description of both RA-IDNC and game theory is also provided. In the second part, we comprehensively discussed and presented existing related works on throughput, CBS offloading, and completion time optimizations. The research objectives of this thesis are also presented. Finally, the contributions of the overall thesis are summarized.

In Chapter 2, we present the C-RAN system model and provide the required RA-IDNC materials for formulating the cross-layer throughput maximization problem. Then, we introduce a new graph, called the CRAN-CLNC graph that transforms the cross-layer throughput maximization problem into a maximum weigh clique problem. Given this designed graph, we develop a joint solution and analyze its computational complexity. Afterward, we develop an iterative scheme that solves cross-layer throughput maximization problem efficiently and overcomes the high complexity of the joint scheme. Numerical results demonstrate that our proposed joint and iterative schemes significantly outperform the conventional coded and uncoded schemes. Compared to the proposed joint solution, our proposed iterative scheme has a certain degradation in the range of 9%-14%. This small degradation in the performance of the iterative scheme comes at the achieved low computational complexity as compared to the high complexity of the joint solution. This chapter has been included in published conference and submitted journal articles [117], [118] respectively.

In Chapter 3, we investigate the trade-off between throughput maximization and CBS offloading in F-RAN setting. First, we overview the

F-RAN model and its assumptions. Then, we formulate the joint CBS offloading and QoS guarantee optimization problem in the downlink F-RAN, given its cross-layer architecture (i.e., physical and network layers) and the well-known ONC graph. Then, we propose a joint approach to solve the formulated optimization problem of joint cross-layer CBS offloading and QoS guarantee. This graph-based approach designs cross-layer weight for each vertex in the FRAN-CLNC graph that reflects both maximizing CBS offloading and ensuring the required QoS for users. In addition, we propose a low complexity iterative approach. Extensive simulations are finally performed to illustrate the potential CBS offloading and user throughput trade-offs of the proposed approaches. Presented numerical results revealed that both proposed schemes achieve significant gains in users' throughput compared to the existing solutions. Compared to the QoS un-aware algorithm, our proposed schemes have a certain degradation in CBS consumed time. This chapter has been included in two published conference and journal articles [128], [129], respectively.

In Chapter 4, we study the completion time minimization (in seconds) problem in D2D-aided F-RAN system. First, we develop a framework where eRRHs and users collaborate to minimize the completion time. In particular, given the intractability of solving the completion time minimization problem over all possible future NC decisions, we reformulate the problem at each transmission with the constraints on user scheduling, their limited coverage zones, transmission rates, and QoS rate guarantee. By analyzing the problem, we decompose it into two subproblems. Given the decomposed subproblems, we develop two graph-based approaches that efficiently minimize the completion time in D2D-aided F-RAN system. Using simulation results, we compare our proposed schemes with existing coded and uncoded (without NC) schemes. Selected numerical results demonstrate that the proposed schemes can effectively improve completion time performance. The materials presented in this chapter have been included in a submitted journal article [146].

In Chapter 5, we develop a distributed game-theoretical framework for a partially connected D2D network using coalition game and IDNC opti-

mization. Our proposed model is formulated as a coalition formation game with non-transferable utility, and a fully distributed coalition formation algorithm is proposed. We show that the proposed distributed algorithm is converged to a Nash-stable coalition structure using split-and-merge rules while accounting for the altruistic players' preferences. We comprehensively evaluate the completion time and game performances of the proposed distributed solution, which demonstrate that the proposed distributed solution offers almost the same completion time performance similar to the centralized F-RAN D2D network. Several advantageous of our proposed scheme compared to the one in F-RAN D2D are given as follows. First, our distributed and decentralized scheme has roughly same performance of the F-RAN D2D centralized model while offloading the CBS. Such offloading policy will allow the CBS to serve more UDs. Second, our scheme alleviates the required high power at UDs to deliver their channels' status to the CBS by using reliable and short D2D links. Third, if some UDs move from one CBS to another, their lost/received content status have to be transferred to the new CBS. This requires implementing handover procedure at the CBSs which is not an issue in our distributed scheme as each UD receives its requested packets from other nearby UDs. The materials presented in this chapter have been included in a submitted journal article [147].

Finally, Chapter 6 summarizes concluding remarks on the accomplished works of the entire thesis. In addition, potential future research topics related to this thesis are also discussed in this chapter.

Remark 1: In all chapters, we use the term "user", but in Chapter 5 that considers D2D networks, we use the term "user devices (UDs)". This is well-known notation for D2D communications. Furthermore, in the context of game theory in Chapter 5, each UD acts as a game player, and thus, UDs and players are used interchangeably.

Remark 2: In Chapter 2, Chapter 3, and Chapter 4, users' side information refers to the cached and requested files by the users. In these chapters, we consider network's and physical's layers factors. In Chapter 5, since the D2D network is viewed only from the network layer, we consider a frame of packets where the aim is to deliver these packets to UDs within minimum

possible delay. This is well-known in NC at network-layer for packet recovery and packet delivery problems [68]. Therefore, in Chapter 5, UDs' side information refers to the cached and requested packets by the UDs.

Remark 3: In Chapter 2, we consider C-RAN setting and thus low-power BSs are represented by RRHs. This is standard in C-RAN system model. However, in Chapter 3 and Chapter 4, we consider F-RAN and D2D-aided F-RAN models, respectively, and thus these RRHs are implemented by caching capabilities and called eRRHs.

Chapter 2

Throughput Maximization in C-RAN using Cross-Layer Network Coding

In this chapter, we study throughput maximization for C-RAN setting, in which a set of users already have asymmetric side information due to prior downloads of some popular files (constituting a finite frame of files). This frame of files is already cached at the different RRHs during off-peak times. Therefore, when each of the users requests one of these files (that it does not have) to be streamed to it at a given time instant, these streaming requests are time-critical (as is usually the case of caching), and must be done with the maximum possible throughput (unlike pre-loads that can be done at much lower rates or at off-peak times). Such work's setup is practically relevant in current and future 5G wireless networks. This chapter's organization is given as follows. The accomplished works and research contributions are summarized in Section 2.1. The C-RAN and cross-layer NC models are discussed in Section 2.2. In this section, we also formulate the throughput maximization problem. In Section 2.3 and Section 2.4, we develop joint and iterative scheduling and power adaptation approaches. In Section 2.5, we present some selected numerical results, and in Section 2.6, we provide some concluding remarks.

2.1 Accomplished Works and Research Contributions

The contributions of this chapter are summarized as follows. We address the cross-layer throughput maximization problem in C-RAN downlink systems by jointly considering the problem's related factors, such as resource scheduling and power adaptation at the physical layer and users' side information at the network layer. Two different cross-layer NC schemes are developed, namely, the joint cross-layer NC scheme and iterative cross-layer NC scheme. In the joint scheme, we jointly consider network-coded user scheduling and rate/power adaptations of each RRB in each RRH's frame. On the other hand, in the iterative scheme, network-coded scheduling and power allocation are considered individually. New graphs are designed for the above two techniques, namely, CRAN-CLNC and coordinated scheduling graphs such as we transform the throughput maximization problem into maximum weight clique problems. Selected numerical results reveal the effectiveness of the proposed schemes in improving throughput performances as compared to the existing schemes.

2.2 System Model and Problem Formulation

2.2.1 C-RAN Model

We consider the downlink communication in C-RAN in which a CBS is connected to B RRHs, denoted by the set $\mathcal{B} = \{b_1, b_2, \cdots, b_B\}$. These RRHs are distributed in different geographic locations within a cell and cooperated in serving U users, denoted by the set $\mathcal{U} = \{u_1, u_2, \cdots, u_U\}$. The B RRHs are under the CBS coordination and communicate with the CBS through low-rate fronthaul links. Similar to [109], each user and each RRH is assumed to be equipped with one single-antenna. Figure 2.1 shows a C-RAN with 3 RRHs cooperating to serve 6 users simultaneously. The transmit frame of each RRH consists of Z orthogonal time/frequency RRBs that are denoted by the set $\mathcal{Z} = \{z_1, z_2, \cdots, z_Z\}$, as shown in Figure 2.2.

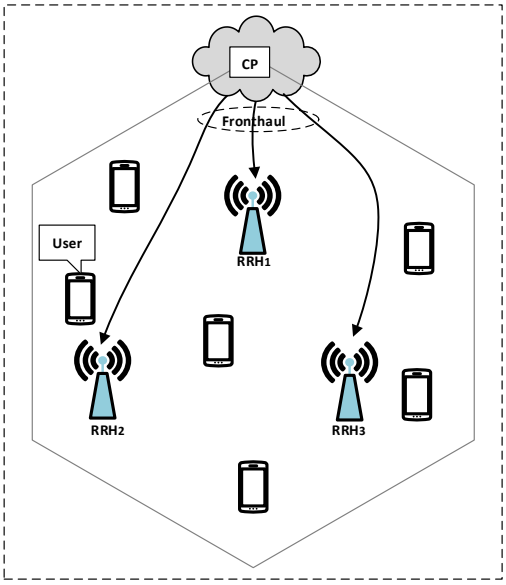


Figure 2.1: A cloud radio access network composed of a central CBS, 6 users, and 3 RRHs. The CBS communicates with the RRHs through low-rate fronthaul links.

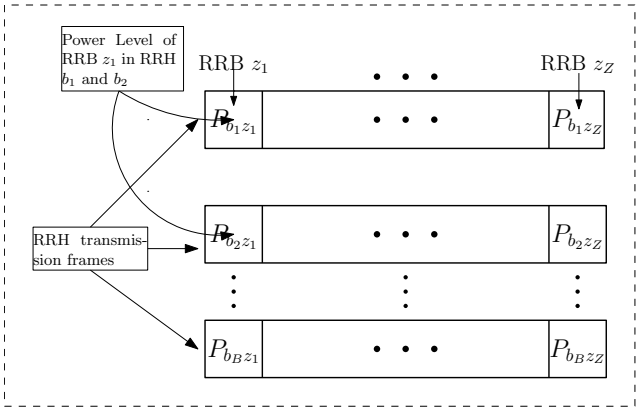


Figure 2.2: Frame structure of the RRHs. Each RRH possesses Z orthogonal RRBs synchronized with the RRBs of the remaining RRHs.

Therefore, the total number of RRBs in the system is $Z_{tot} = Z$. Let $P_{b_n z_m}$ be the power allocation level (PL) of the z_m-th RRB in the b_n-th RRH. Let $\mathbf{P} = [P_{b_n z_m}]$ be a $B \times Z$ matrix containing the PLs of the RRBs/RRHs in the considered system. From practical constraints, the power level of each RRB is bounded by $P_{b_n z_m} \leqslant P_{\max}$, where $P^{\max}_{b_n z_m}$ is the maximum power level of the RRBs in the system. Let $h^{u_i}_{b_n z_m}$ be the complex channel gain from the z_m-th RRB in the b_n-th RRH to the u_i-th user. The channel is assumed to be constant during the transmission time of a single uncoded/coded file and to change from one transmission to another. In our simulation setup, we used the SUI model in which the channel information $\{h^{u_i}_{b_n z_m}\}_{b_n \in \mathcal{B}, z_m \in \mathcal{Z}, u_i \in \mathcal{U}}$ is affected by multiple factors, e.g., fading, shadowing, but the location of the user within the service area is the dominant factor. Such channel model leads to heterogeneous physical-layer rates from different RBBs/RRHs to different users. The achievable rate of the u_i-th user assigned to the z_m-th RRB in the b_n-th RRH can be expressed as

$$R^{u_i}_{b_n z_m} = \log_2(1 + \mathrm{SINR}^{u_i}_{b_n z_l}(\mathbf{P})), \tag{2.1}$$

where $\mathrm{SINR}^{u_i}_{b_n z_m}(\mathbf{P})$ is the corresponding signal-to-interference plus noise-ratio experienced by the u_i-th user when it is assigned with the z_m-th RRB of the b_n-th RRH, and can be expressed as

$$\mathrm{SINR}^{u_i}_{b_n z_m}(\mathbf{P}) = \frac{P_{b_n z_m} |h^{u_i}_{b_n z_m}|^2}{\sigma^2 + \sum_{b_{n'} \in \mathcal{B}, b_{n'} \neq b_n,} P_{b_{n'} z_m} |h^{u_i}_{b_{n'} z_m}|^2}, \tag{2.2}$$

where σ^2 is the additive white Gaussian noise (AWGN) power. As stated earlier, the transmit frame of each RRH consists of Z orthogonal RRBs. Therefore, interference at the z_m-th RRB is seen only from the same z_m-th RRB in the other RRHs. In other words, $\mathrm{SINR}^{u_i}_{b_n z_m}(\mathbf{P})$ depends solely on the scheduled users in z_m-th RRB across the remaining $b_{n'} \neq b_n$ RRHs and the corresponding power level $P_{b_{n'} z_m}$. We consider that the reception of an uncoded/encoded file sent in the z_m-th RRB of the b_n-th RRH is successful at the u_i-th user if $R_{b_n z_m} \leqslant R^{u_i}_{b_n z_m}$, where $R_{b_n z_m}$ is the transmission rate of that RRB. The set of achievable rates of all users in all RRBs across all

RRHs can be represented by the set

$$\mathcal{R} = \bigotimes_{(b_n, z_m, u_i) \in \ \mathcal{B} \times \mathcal{Z} \times \mathcal{U}} R_{b_n z_m}^{u_i}, \tag{2.3}$$

where the symbol $\bigotimes$ represents the product of the set of the achievable capacities.

2.2.2 Cross-Layer Network Coding (CLNC) Model

We assume that all users are interested in streaming one file or more out of a set $\mathcal{F} = \{f_1, f_2, \cdots, f_F\}$ containing a finite window of F files. These files are deemed popular due to their previous multiple downloads by different subsets of the users over a short period of time. Popular files in this work represent frames from video-on-demand streaming. A user can start playing the video after some (short) time for buffering in z_m-th RRB, while download goes on in other $z_m \neq z_{m'}$ RRBs in the same RRH. If users request more than one file simultaneously, they can get them from the same RRH by listening to multiple RRBs, each delivering one of its required files. All files in $\mathcal{F}$ are assumed to be stored in the CBS with the same size of N bits. The size of any XOR encoded file (binary encoding operation) is also N bits. It is assumed that each RRH holds/caches the whole set of files $\mathcal{F}$ that they receive from the CBS and updates the CBS of the indices of the downloaded files by users after each transmission. Furthermore, the CBS, which is a high computational device, keeps a log file to temporarily store all the requested and downloaded popular files by the users under its coverage. Then, it can track the downloaded files by users and use them to do instantly decodable NC (IDNC) combinations. Finally, the log entries of users of the current popular files are transferred as part of this user's handover procedure if this user moves from one CBS to another. RRHs and users use the previously mentioned asymmetric users' downloaded/cached files to perform XOR encoding and decoding operations, respectively, when new files requested by users. In particular, users need the cached files to extract the wanted files immediately. The entire process of receiving and

decoding these requested files takes a small duration of time. Thus, in any arbitrary scheduling epoch, the files of $\mathcal{F}$ can be classified for u_i-th user as follows.

- The *Has* set $\mathcal{H}_{u_i}$ containing files previously cached by the u_i-th user.
- The *Lacks* set $\mathcal{L}_{u_i} = \mathcal{F} \backslash \mathcal{H}_{u_i}$ containing the non-cached files.
- The *Wants* set $\mathcal{W}_{u_i} \subset \mathcal{L}_{u_i}$ containing files requested by the u_i-th user in the current scheduling frame.

The CBS exploits the diversity of side information to transmit encoded files in order to maximize the number of successfully received bits, i.e., throughput, in each scheduling frame. Let $\tau_{b_n z_m}$ denotes the targeted set of users benefiting from the encoded file $\kappa_{b_n z_m}$ that transmitted over the z_m-th RRB of the b_n-th RRH, where $\kappa_{b_n z_m}$ is an element of the power set $\mathcal{P}(\mathcal{F})$ representing a coded combination of a set of files. An instantly decodable combination $\kappa_{b_n z_m}$ is used to retrieve a new wanted file by u_i-th user if and only if

1. $R_{b_n z_m} \leqslant R^{u_i}_{b_n z_m}$: The u_i-th user can properly receive the combination with a rate below its achievable capacity on the z_m-th RRB of the b_n-th RRH.
2. $|\mathcal{W}_{u_i} \cap \kappa_{b_n z_m}| = 1$: The u_i-th user can re-XOR $\kappa_{b_n z_m}$ with its previously cached files to retrieve a new file. In particular, user u_i can re-XOR $\kappa_{b_n z_m}$ with its previously received packets that are in the combination $\kappa_{b_n z_m}$.

Therefore, the set of targeted users when the z_m-th RRB in b_n-th RRH transmits the IDNC combination $\kappa_{b_n z_m}$ is

$$\tau_{b_n z_m} = \left\{ u_i \in \mathcal{U} \;\middle|\; |\mathcal{W}_{u_i} \cap \kappa_{b_n z_m}| = 1 \;\&\; R_{b_n z_m} \leqslant R^{u_i}_{b_n z_m} \right\} \tag{2.4}$$

The above-mentioned concepts can be illustrated in the following example.

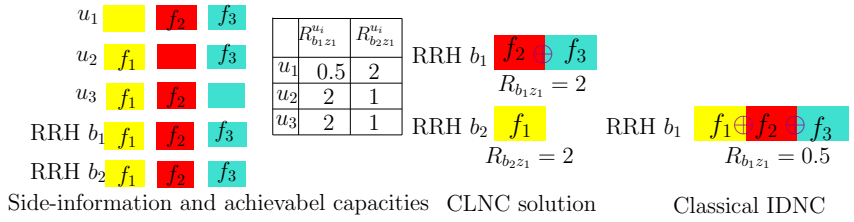


Figure 2.3: A simple C-RAN system composed of 3 users, 3 files, 2 RRHs, and 1 RRB in each RRH's transmit frame.

Example 1: This example considers a simple model in Figure 2.3 that consists of 2 RRHs, 1 RRB in each RRH's transmit frame, 3 users, users' side information and their rates. Each user u_i possesses 2 files and wants 1 file. For example, u_1 possesses f_2, f_3, and wants f_1. The whole set of files is cached by each RRH. Given a certain power level to each of the RRHs, assume that users' rates in bits/s are provided in a table on the middle of Figure 2.3. In order to maximize throughput for this example, many possible solutions can be found as follows.

- **CLNC solution:** The b_1-th RRH XORes f_2 and f_3 into $\kappa_{b_1z_1} = f_2 \oplus f_3$ and transmits $\kappa_{b_1z_1}$ with a rate of 2 bits/s to the set of users $\tau_{b_1z_1} = (u_2, u_3)$. The b_2-th transmits $\kappa_{b_2z_1} = f_1$ with a rate of 2 bits/s to the set $\tau_{b_2z_1} = u_1$ by sending. Thus, $\kappa_{b_1z_1}$ in b_1-th RRH is instantly decodable for all served users $\tau_{b_1z_1}$, and $\kappa_{b_2z_1}$ in b_2-th RRH is instantly decodable only for u_1 in $\tau_{b_2z_1}$. Given this, we have the following decoding process at the users. User u_2 gets f_2 by XORing $\kappa_{b_1z_1}$ with f_3 and user u_3 gets f_3 by XORing $\kappa_{b_1z_1}$ with f_2. Therefore, the best achievable overall throughput in this scenario is 6 bits/s as each served user will simultaneously receive 2 bits/s.

- **Uncoded solution:** The b_1-th and b_2-th RRHs transmits f_2 and f_1 with an equal rate of 2 bits/s, respectively, to u_2 and u_1. This results in 4 bits/s throughput achievable.

- **Classical IDNC solution:** The b_1-th XORes f_1, f_2, and f_3 into $\kappa_{b_1z_z} = f_1 \oplus f_2 \oplus f_3$ and transmits $\kappa_{b_1z_z}$ to to the set of users $\tau_{b_1z_1} =$

$\{u_1, u_2, u_3\}$. This can be done from the network coding viewpoint but would degrade the throughput than the uncoded and CLNC solutions. Indeed, to serve all users given their files are combined using NC, this solution will limit the rate sent to 0.5 bit/s, resulting in total throughput of 1.5 bits/s.

From the completion time viewpoint, CLNC needs a total of N seconds, the uncoded solution consumes $1.5N$ seconds, and classical IDNC requires $2N$ seconds. This clearly shows that CLNC is more efficient in maximizing the throughput and minimizing the completion time. These results can be further improved by optimizing the power allocation for each RRB in each RRH.

2.2.3 Problem Formulation

The joint user scheduling and power optimization problem considered in this paper is equivalent to assigning users to the RRBs of the RRHs and adapt the power levels under the following network connectivity constraints (**CC**).

- **CC1**: Each user can connect to at most one RRH, but possibly to many RRBs in that RRH.
- **CC2**: Each power level PL is bounded by a value.

Let $X^{u_i}_{b_n z_m}$ be a binary variable that is equal to 1 if u_i-th user is assigned to z_m-th RRB of b_n-th RRH, and zero otherwise. Typically, $X^{u_i}_{b_n z_m}$ will be 0 if b_n-th RRH does not store the requested file by u_i-th user. Let $Y^{u_i}_{b_n}$ be a binary variable that is set to 1 if u_i-th user is assigned to b_n-th RRH, and zero otherwise. The joint coordinated scheduling and power allocation

problem can be formulated as follows

$$\max \sum_{b_n \in \mathcal{B}} \sum_{z_m \in \mathcal{Z}} \sum_{u_i \in \mathcal{U}} X^{u_i}_{b_n z_m} \min_{u_{i'} \in \tau_{b_{n'} z_{m'}}} \log_2(1 + \mathrm{SINR}^{u_{i'}}_{b_{n'} z_{m'}}(\mathbf{P})) \tag{2.5a}$$

$$\text{s.t. } Y^{u_i}_{b_n} = \min\left(\sum_{z_m} X^{u_i}_{b_n z_m}, 1\right), (b_n, u_i) \in \mathcal{B} \times \mathcal{U}, \tag{2.5b}$$

$$\sum_{b_n} Y^{u_i}_{b_n} \leqslant 1, u_i \in \mathcal{U}, \tag{2.5c}$$

$$0 \leq P_{b_n z_m} \leq P^{\max}, (b_n, z_m) \in \mathcal{B} \times \mathcal{Z}, \tag{2.5d}$$

$$X^{u_i}_{b_n z_m}, Y^{u_i}_{b_n} \in \{0, 1\}, \kappa_{b_n z_m} \in \mathcal{P}(\mathcal{F}), (u_i, b_n, z_m) \in \mathcal{U} \times \mathcal{B} \times \mathcal{Z}, \tag{2.5e}$$

where the optimization is carried over the variables $X^{u_i}_{b_n z_m}$, $Y^{u_i}_{b_n}$, $\kappa_{b_n z_m}$, $R_{b_n z_m}$, and $P_{b_n z_m}$. The variables $X^{u_i}_{b_n z_m}$ and $Y^{u_i}_{b_n}$ are discrete optimization parameters that represent the user-RRH and user-RRB associations, respectively. On the other hand, the variables $\kappa_{b_n z_m}$, R_{bz} and $P_{b_n z_m}$ account for the file combination, the transmission rate, and the PLs for the z_m-th RRB of the b_n-th RRH, respectively. Constraints (2.5b) and (2.5c) translate **CC1**. Constraint (2.5d) corresponds to constraint **CC2**. The problem in (2.5) is NP-hard. To show its hardness, we consider a simpler case when all users have the same rates to all RRBs/RRHs. This makes the objective function constant. In this case, the problem in (2.5) is reduced to be an NC problem, which is shown to be NP-hard [116]. Therefore, the problem in (2.5) has the worst case exponential solution but can be solved using the branch and bound algorithm whose performance depends on the quality of lower and upper bounds. To reach a tractable solution to (2.5), one can discretize the continuous power $P_{b_n z_m}$ into a set of discrete PLs $\mathbf{P}$ while constructing the CRAN-CLNC graph simultaneously.

2.3 Joint Scheduling and Power Adaptation Solution using CLNC

In this section, we first design the CRAN-CLNC graph to reformulate the throughput maximization problem as a maximum-weight clique problem.

Then, using the deigned graph, we propose an efficient greedy joint approach.

2.3.1 CLNC Graph Design

In this subsection, we first generate all possible IDNC file combinations for configuring the power control subgraph for each z_m-th RRB in the network. Afterward, we design the CRAN-CLNC graph from all power control subgraphs.

Let $\mathcal{S}$ be the set of all possible associations between users and files (pairs of users and their requested files), which is defined by $\mathcal{S} = \mathcal{U} \times \mathcal{F}$, and s is an element in $\mathcal{S}$. For example, the set of all possible associations between users and their requested files in Example 1 is $\mathcal{S} = \{(u_1, f_1), (u_2, f_2), (u_3, f_3)\}$, $s = (u_1, f_1)$ represents the association of u_1-th user and its f_1-th requested file. For simplicity, $u_i(s)$ represents the u_i-th user in the s-th association. The generation of the set of the possible IDNC file combinations, denoted by $\mathcal{S}_{\text{IDNC}}$, relies on the fact that the corresponding users in the combined associations in $\mathcal{S}$ can be instantly served by one IDNC file. Therefore, any two distinct associations $s \in \mathcal{S}$ and $s' \in \mathcal{S}$ are combined if one of the following conditions is satisfied.

- **C1**: $u_i(s) \neq u_i(s')$ and $f_k(s) = f_k(s')$. The two associations s, s' are induced by the same file f_k that is requested by two distinct users.
- **C2**: $u_i(s) \neq u_i(s')$ and $f_k(s) \in \mathcal{H}_{u_i(s')}$ and $f_k(s') \in \mathcal{H}_{u_i(s)}$. This represents that two different files are wanted by two different users. Meanwhile, each wanted file in each association is in the *Has* set of the user that induced the other association.

These conditions ensure that any IDNC file combination in $\mathcal{S}_{\text{IDNC}}$ is always decodable for the users represented by the corresponding combined associations. Therefore, $\mathcal{S}_{\text{IDNC}}$ consists of all possible IDNC file combinations resulted from the associations in $\mathcal{S}$, and $\mathtt{s} = (\kappa, \tau)$ is an element in $\mathcal{S}_{\text{IDNC}}$, where κ and τ are defined in Section 2.2.2. For example, associations $s = (u_1, f_1) \in \mathcal{S}$ and $s' = (u_2, f_2) \in \mathcal{S}$ satisfy **C1** and **C2**, and accordingly, we can combine them into $\mathtt{s} = (\kappa, \tau) = \{(f_1 \oplus f_2), (u_1, u_2)\}$. Let $\mathcal{A}$ be the

set of all possible associations between RRHs, RRBs, IDNC file combinations, and achievable capacities, i.e., $\mathcal{A} = \mathcal{B} \times \mathcal{Z} \times \mathcal{S}_{\text{IDNC}} \times \mathcal{R}$. For example, one possible association in $\mathcal{A}$ is $(b_n, z_m, \mathtt{s}, R)$ which represents the RRH b_n, RRB z_m, IDNC combination $\mathtt{s}$, and rate R. Basically, each element in $\mathcal{A}$ contains combined associations of $s \in \mathcal{S}$ representing an IDNC combination $\mathtt{s}$ that is transmitted from RRH b_n on RRB z_m with its corresponding rate R. Let $\mathcal{A}_{z_m} \subset \mathcal{A}$ represent the associations relative to the z_m-th RRB across all RRHs. In other words, $\mathcal{A}_{z_m}$ represents all associations in $\mathcal{A}$ that are indexed by z_m.

Now let the power control subgraph of the z_m-th RRB in the network be denoted by $\mathcal{G}_{z_m}(\mathcal{V}_{z_m}, \mathcal{E}_{z_m})$ wherein $\mathcal{V}_{z_m}$ and $\mathcal{E}_{z_m}$ refer to the set of vertices and edges of this subgraph, respectively. The set of vertices in the $\mathcal{G}_{z_m}$-th subgraph is generated by merging all possible associations in $\mathcal{A}_{z_m}$ for the different RRHs under the system condition **CC1**. Any possible merged associations in $\mathcal{A}_{z_m}$ is referred to as feasible schedule and denoted by $\mathbf{S}$. In $\mathbf{S}$, the same associated user is not scheduled to different RRHs as stated in the condition **CC1**. The set of all possible feasible schedules $\mathbf{S}$ corresponding to the z_m-th RRB is represented by $\mathcal{S}_{z_m,\text{fs}}$. Therefore, a vertex $v \in \mathcal{V}_{z_m}$ associated with $\mathbf{S} \in \mathcal{S}_{z_m,\text{fs}}$ satisfies $b_n(s) \neq b_n(s')$ and $u_i(s) \neq u_i(s')$ for all $s \neq s' \in \mathbf{S}$. As a result, each vertex $v \in \mathcal{V}_{z_m}$ represents the partial schedule of users to the z_m-th RRB across all RRHs.

As mentioned above, the set of all vertices $\mathcal{E}_{z_m}$ represents the set of all merged associations. Therefore, each $v \in \mathcal{V}_{z_m}$ representing the feasible schedule $\mathbf{S} \in \mathcal{S}_{z_m,\text{fs}}$ satisfies the following local conditions **(LC)**.

- **LC1**: For all $(s, s') \in \mathbf{S}$ such that $b_n(s) = b_n(s')$, we have $R(s) = R(s')$. This condition guarantees that all associations in the same RRB z_m and RRH b_n have the same transmission rate.

- **LC2**: For all $(s, s') \in \mathbf{S}$ such that $b_n(s) \neq b_n(s')$, we have $\tau \cap \tau' = \varnothing$. This condition guarantees that each user is scheduled to at most a single RRH.

Suppose that the power distribution $\mathbf{P}$ will be computed later, then a vertex v representing the schedule $\mathbf{S}$ has a weight that reflects the total contribution

of the vertex to the network. This weight can be expressed as

$$w(v) = \sum_{s \in \mathbf{S}} \min_{u_i(s) \in \tau_{b_n(s) z_m(s)}} \log_2(1 + \mathrm{SINR}^{u_i(s)}_{b_n(s) z_m(s)}(\mathbf{P})). \tag{2.6}$$

We design the CRAN-CLNC graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ by generating the Z power control subgraphs. Without loss of generality, the set of all schedules in the network is the union of schedules of all RRBs, i.e., $\mathcal{S}_{\mathrm{fs}} = \bigcup_{z_m \in \mathcal{Z}} \mathcal{S}_{z_m,\mathrm{fs}}$. Therefore, the vertex set of the CRAN-CLNC graph is simply the union of vertices of all the power control subgraphs, i.e., $\mathcal{V} = \bigcup_{z_m \in \mathcal{Z}} \mathcal{V}_{z_m}$. Since we already described the edges between vertices within each subgraph $\mathcal{G}_{z_m}$ in the previous subsection, the rest of this section describes remaining edges corresponding to different subgraphs. Following similar philosophy as before, two different vertices belonging to two different subgraphs are adjacent if their combination results in a feasible schedule. In particular, two vertices are connected if no user is scheduled to different RRHs. To mathematically formulate this, let vertex $v \in \mathcal{G}_{z_m}$ be corresponding to the schedule $\mathbf{S} \in \mathcal{S}_{z_m,\mathrm{fs}}$ and vertex $v' \in \mathcal{G}_{z_{m'}}$ corresponding to the schedule $\mathbf{S}' \in \mathcal{S}_{z_{m'},\mathrm{fs}}$. Vertices v and v' are adjacent if the associations they represent satisfy the following general condition **(GC)**.

- **GC**: For all $(s, s') \in \mathbf{S} \times \mathbf{S}'$ such that $\tau \cap \tau' \neq \varnothing$, we have $b_n(s) = b_n(s')$. The condition guarantees that the same user can be associated only to a unique RRH.

The design of the CRAN-CLNC graph $\mathcal{G}$ makes any maximal clique satisfies the following criterion.

- All targeted users can retrieve a new file(s) from the transmission schedule of all RRBs and RRHs.
- Each user is scheduled to a single RRH.
- Each RRB identified by the vertices in a maximal clique adopts the transmission rate identified by the vertex. This rate supports the smallest channel capacity of all scheduled users to that RRB.

Given the optimal power allocation $\mathbf{P}$, the following theorem reformulates the problem (2.5).

Theorem 2.1. *The CRAN joint coordinated scheduling and power allocation problem (2.5) is equivalent to a maximum-weight clique problem over the CRAN-CLNC graph, and can be written as*

$$\arg\max_{\mathtt{C}\in\mathcal{C}} \sum_{v\in\mathtt{C}} w(v), \tag{2.7}$$

where $\mathtt{C}$ *is a maximal clique in the CRAN-CLNC graph,* $\mathcal{C}$ *is the set of all possible maximal cliques, and* $w(v)$ *is defined in (2.6). The vertices in* $\mathtt{C}$ *represents the targeted users and the file combinations in each RRB across all RRHs.*

For proof, please refer to Appendix A.

The maximum-weight clique problem in Theorem 2.1 is NP-hard. However, we use one of the existing algorithms in the literature, e.g., [148], [149], [150] to solve problem (2.5) heuristically for one particular transmission. The proposed algorithm selects a number of vertices with the maximum weights, wherein each vertex's weight represents the sum of transmission rates of the z_m-th RRB across all RRHs that suitable for only a subset of associated users.

2.3.2 Joint Scheduling and Power Allocation Solution

This subsection presents the joint CRAN-CLNC solution to (2.5). The idea is to solve the power control problem for each vertex of the power control subgraph. This would allow the design of the CRAN-CLNC graph. Afterward, we heuristically solve the reformulated problem in Theorem 2.1 using greedy graph algorithm.

We discretize a set of power levels to solve (2.5) efficiently. More specifically, we show that by simultaneously computing the optimal power allocations $\mathbf{P}$ while designing the CRAN-CLNC graph, we can achieve a tractable solution to (2.5). Consider the z_m-th power control subgraph

and a vertex $v \in \mathcal{V}_{z_m}$ in that subgraph is associated with the schedule $\mathbf{S} = \{s_1, s_2, \cdots, s_{|\mathbf{S}|}\} \in \mathcal{S}_{z_m,\text{fs}}$, where $|\mathbf{S}|$ is the degree of $\mathbf{S}$, i.e., $|\mathbf{S}| = \sum_{b_n \in \mathcal{B}} |\tau_{b_n z_m}|$. The optimal power levels PLs $(P_{b_1 z_m}, \cdots, P_{b_B z_m})$ that maximize the throughput for that vertex v are the solution to the following problem

$$\max_{P_{b_n z_m}} \sum_{b_n \in \mathcal{B}} |\tau_{b_n z_m}| * \min_{u_i \in \tau_{b_n z_m}} \log_2(1 + \text{SINR}_{b_n z_m}^{u_i}(\mathbf{P}))$$
$$\text{s.t. } 0 \leqslant P_{b_n z_m} \leqslant P_{b_n z_m}^{\max}, \ \forall b_n \in \mathcal{B}, \tag{2.8}$$

where the optimization is over the power levels $P_{b_n z_m}, \forall b_n \in \mathcal{B}$. The power optimization problem (3.5) is a well-known non-convex problem [151] and can be solved optimally [152]. However, such a solution requires a high computation complexity and converges slowly. To find a reasonable solution to (3.5) with a low computational load, we use one of the power optimization algorithms proposed in [151].

The joint coordinated scheduling and power allocation algorithm can be implemented at the CBS as follows.

First, we design the CRAN-CLNC graph as follows. For each RRB $z_m \in \mathcal{Z}$, we configure the power control subgraph $\mathcal{G}_{z_m}$ by generating each vertex $v \in \mathcal{G}_{z_m}$ corresponding to the feasible schedule $\mathbf{S} \in \mathcal{S}_{z_m,\text{fs}}$ using conditions **LC1** and **LC2**. Afterward, we calculate the optimal PLs for each vertex by solving the optimization problem (3.5). Then, we append the computed PLs of each vertex and the corresponding rates to that vertex. We repeat steps above for all RRBs $z_m \in \mathcal{Z}$. We merge all subgraphs of all RRBs and connect them using condition **GC**. Second, in each iteration, we find the maximum-weight clique among all other maximal cliques in CRAN-CLNC graph $\mathcal{G}$, which can be implemented as follows. We compute the vertices' weights using (2.6). Then, we select the maximum-weight vertex v^* and add it to `C`, i.e., `C` is initially empty. Then, we update the graph $\mathcal{G}$ by removing v^* and all the vertices that are not adjacent to it so as the next chosen vertex is not in feasible transmission conflict with the already chosen ones in `C`. We execute this process until no more vertices exist in $\mathcal{G}$. Clearly, the

Algorithm 1: Joint coordinated scheduling and power allocation algorithm

Data: $\mathcal{U}, \mathcal{F}, \mathcal{B}, \mathcal{Z}, \mathcal{H}_{u_i}, \mathcal{W}_{u_i}$ and $h^{u_i}_{b_n z_m}, (u_i, b_n, z_m) \in \mathcal{U} \times \mathcal{B} \times \mathcal{Z}$;
Initialization: maximum-weight clique $\mathtt{C} = \phi$.
for $z_m \in \mathcal{Z}$ **do**
  Initialization: $\mathcal{G}_{z_m} = \phi$.
  for *each* $\mathbf{S} = \{s_1, s_2, \cdots, s_{|\mathbf{S}|}\} \in \mathcal{S}_{z_m,fs}$ **do**
    Solve (3.5) to calculate the optimal power levels $\mathbf{P} = \{(P^*_{b_1 z_m}, P^*_{b_2 z_m}, \cdots, P^*_{b_B z_m})\}$.
    Create $v = \{\{(s_1, R^*_{b_1 z_m}, P^*_{b_1 z_m}), \cdots, (s_{|\tau_{b_1 z_m}|}, R^*_{b_1 z_m}, P^*_{b_1 z_m})\}, \cdots, \{(s_B, R^*_{b_B z_m}, P^*_{b_B z_m}), \cdots, (s_{|\tau_{b_B z_m}|}, R^*_{b_B z_m}, P^*_{b_B z_m})\}\}$.
    Compute $w(v)$ using (2.6).
    Set $\mathcal{G}_{z_m} = \mathcal{G}_{z_m} \cup \{v\}$.
  end
end
Set $\mathcal{G} = \bigcup_{z_m \in \mathcal{Z}} \mathcal{G}_{z_m}$.
Connect vertices of $\mathcal{G}$ using **GC**.
Solve maximum-weight clique problem over $\mathcal{G}$.
Output $\mathtt{C}$.

number of vertices in the selected maximum-weight clique is at most Z. The detailed procedures of the solution are provided in Algorithm 1.

For understanding the design of the CLNC-CRAN graph in Figure 2.2 and Algorithm 1, please refer to the illustration in Appendix B.

2.3.3 Complexity Analysis of Joint Solution

The computational complexity of the aforementioned joint solution is divided into two phases: i) the complexity of constructing the CRAN-CLNC graph, and ii) the complexity of finding the maximum weight clique. These phases are explained as follows.

Phase I: CRAN-CLNC construction complexity: The computational complexity of constructing the CRAN-CLNC graph is divided into the following steps: i) generating a single subgraph for each RRB, ii) computing the power levels of the RRHs, and iii) connecting all the subgraphs in the

system to construct the CRAN-CLNC graph. These steps are explained as follows.

First step: The complexity of this step corresponds to the computational complexity of generating all vertices in any subgraph. Consider computing the complexity of subgraph $\mathcal{G}_{z_m}$ of the z_m-th RRB. As explained in Section 2.3.1, the total number of vertices in each subgraph is the total number of possible feasible schedules $|\mathcal{S}_{z_m,\text{fs}}|$. Each schedule $\mathbf{S} \in \mathcal{S}_{z_m,\text{fs}}$ contains many possible associations that represent IDNC file combinations to be transmitted to a set of users from a set of RRHs with their corresponding transmission rates/powers. $|\mathcal{S}_{z_m,\text{fs}}|$ depends on all possible IDNC file combinations $\mathcal{S}_{\text{IDNC}}$ in the system. Any possible IDNC combination contains combined files correspond to associated requesting users. We consider the worst-case complexity of generating the IDNC combinations as follows. First, note that each associated user with any IDNC file combination can receive at most one file. In other words, each user can instantly decode at most one file from its associated IDNC combination. Therefore, the maximum number of associations in any IDNC combination is $\min(U, F)$. In our setting, the number of files is greater than the number of users. Thus, the upper bound of the IDNC combinations is F choose U, which can be computed as $\frac{F!}{(F-U)!U!}$. For simplicity, it can be denoted by $C = \frac{F!}{(F-U)!U!}$. Second, note that $\mathcal{G}_{z_m}$ corresponds to the z_m-th RRB across all RRHs. Since all RRHs cache the whole set of files, each RRH can generate the same set of IDNC combinations that can be generated by the other RRHs. Further, the same IDNC combination and its corresponding users can not be associated to different RRHs in the same schedule $\mathbf{S}$ as this violates the condition **LC2** that explained in Section 2.3.1. Hence, the possible IDNC combinations for all RRHs in the system can be permuted between all RRHs, and each permutation represents schedule $\mathbf{S}$. As previously mentioned, schedule $\mathbf{S}$ is represented by vertex v in $\mathcal{G}_{z_m}$. Therefore, the total number of vertices in $\mathcal{G}_{z_m}$ is the number of permutations of these IDNC combinations among B RRHs, which is $\frac{C!}{(C-B)!}$.

Second step: To compute the power levels, we solve (3.5) for each vertex v in each subgraph. Solving (3.5) for each v depends on the number

of RRHs, number of users in $\mathbf{S}$, and interference pricing messages from the RRHs. The number of users in each schedule $\mathbf{S}$ is upper bounded by U, i.e., the case when all users' files can be combined using NC. For the worst-case complexity of computing (3.5), Algorithm 1 needs for each scheduled user in RRH b_n to compute the interference of all other RRHs, i.e., $b_{n'} \neq b_n$ and the pricing messages from each RRH $b_{\tilde{n}} \neq b_{n'}$. Thus, the computational complexity of calculating the power allocation is upper bounded by $O\left(U(B(B-1)+B(B-1))\right) = O(2U(B^2-B)) = O\left(2UB^2\right)$.

Third step: The number of subgraphs is the number of RRBs in the system. Thus, Z subgraphs constitutes the CRAN-CLNC graph, and it has $O\left(\frac{C!}{(C-B)!}Z\right)$ vertices. Therefore, the computational complexity of constructing the CRAN-CLNC graph and computing the power allocations is $O\left(\frac{C!}{(C-B)!}Z\left(2UB^2\right)\right)$.

Phase II: Maximum-weight search complexity: First, we build the adjacency matrix of the CRAN-CLNC graph. Particularly, we need to check the **General Condition (GC)** that explained in Section 2.3.1 of each pair of $\frac{C!}{(C-B)!}Z$ vertices to determine whether they should be connected by an edge. This adjacent connection between vertices needs a complexity of $O\left(\left(\frac{C!}{(C-B)!}Z\right)^2\right)$. The computational complexity of the maximum-weight search algorithm can be decomposed in sum weights and vertex search computations as follows. Each iteration requires a complexity of $O\left(\frac{C!}{(C-B)!}Z\right)$ for weight calculations of its maximum-weight clique. Note that each maximum-weight clique has at most Z vertices as each subgraph can contribute at most with one vertex per transmission. Therefore, building the adjacency matrix and executing the maximum-weight search have a complexity of $O\left(\left(\frac{C!}{(C-B)!}Z\right)^2 + \frac{C!}{(C-B)!}Z + Z\right) = O\left(\left(\frac{C!}{(C-B)!}Z\right)^2\right)$. Given the aforementioned computational complexities of the different algorithm components, the total per-iteration computational complexity of Algorithm 1 in terms of main system's parameters is upper bounded by $O\left(\frac{\left(\frac{F!}{(F-U)!U!}\right)!}{\left(\frac{F!}{(F-U)!U!}-B\right)!}Z\left(2UB^2\right) + \left(\frac{\left(\frac{F!}{(F-U)!U!}\right)!}{\left(\frac{F!}{(F-U)!U!}-B\right)!}Z\right)^2\right)$. Since the second term dominates the overall complexity, we can write the upper bound overall

complexity as $O\left(\left(\frac{\left(\frac{F!}{(F-U)!U!}\right)!}{\left(\frac{F!}{(F-U)!U!}-B\right)!}Z\right)^2\right)$. This complexity can be approximated using Stirling's approximation as $O\left(\left(\frac{\sqrt{2\pi C}(\frac{C}{e})^C}{\sqrt{2\pi K}(\frac{K}{e})^K}Z\right)^2\right)$, where $K = (C-B)!$. In general, the computational complexity of the joint solution is less than that as not all files can be combined due to the instant decodability conditions that explained in Section 2.2.2. In the uncoded case [109], the computational complexity is $O\left(\frac{U!}{(U-B)!}Z\left(2B^3\right)+\left(\frac{U!}{(U-B)!}Z\right)^2\right)$. For the high implementation complexity of the joint solution in large networks, we develop in the next section an efficient and alternative coordinated scheduling solution that has low implementation complexity.

Remark on the complexity of the joint solution: If we increase the number of files without changing the number of users and RRHs, the complexity of the whole system will decrease. This is because for a large number of files, any set of randomly picked uncoded files by the RRHs has a high chance to satisfy all users. Thus, no need to perform network coding in this scenario.

2.4 Iterative Optimization for Coordinated Scheduling and Power Adaptation

We present in this section a low-complexity iterative solution to the optimization problem (2.5). In particular, we solve the coordinated NC schedule and power control separately and iteratively upon convergence. Towards that target, we fix the power levels PLs in (2.5) and solve the following simplified problem to refine NC and user-RRB/RRH schedule

$$\max \sum_{b_n\in\mathcal{B}}\sum_{z_m\in\mathcal{Z}}\sum_{u_i\in\mathcal{U}} X_{b_n z_m}^{u_i} \min_{u_{i'}\in\tau_{b_{n'}z_{m'}}} R_{b_{n'}z_{m'}}^{u_{i'}} \tag{2.9a}$$

$$\text{s.t. } (2.5b), (2.5c), (2.5e). \tag{2.9b}$$

The optimization is carried over the variables $X_{b_n z_m}^{u_i}$, $Y_{b_n}^{u_i}$, and $\kappa_{b_n z_m}$. Then, we use the resulting NC and user-RRB/RRH schedule in (2.5) and solve the

following problem per RRBs basis

$$\max_{P_{b_n z_m}} \sum_{b_n \in \mathcal{B}} |\tau_{b_n z_m}| * \min_{u_i \in \tau_{b_n z_m}} \log_2(1 + \mathrm{SINR}_{b_n z_m}^{u_i}(\mathbf{P}))$$
$$\text{s.t. } 0 \leqslant P_{b_n z_m} \leqslant P_{b_n z_m}^{\max}, \ \forall b_n \in \mathcal{B}, \tag{2.10}$$

where $P_{b_n z_m}, \forall b \in \mathcal{B}$ is the set of optimization parameters, $\mathcal{B}$ is the set of RRHs, and u_i is the user that belongs to the set of targeted users $\tau_{b_n z_m}$ that have associations in the resulting schedule. In what follows, we show how to solve problems (2.9) and (2.10), respectively.

2.4.1 Coordinated Scheduling: Solution to Problem (2.9)

Let $\mathcal{S}$ be the set of all possible associations (all pairs) between RRHs, RRBs, users, files and the achievable capacity, i.e., $\mathcal{S} = \mathcal{B} \times \mathcal{Z} \times \mathcal{U} \times \mathcal{F} \times \mathcal{R}$, and s is an element in $\mathcal{S}$. For example, $s = (b_n, z_m, u_i, f_k, R) \in \mathcal{S}$ represents the RRH b_n, RRB z_m, user u_i, file f_k, and rate R. Let the coordinated scheduling graph be denoted by $\mathcal{G}(\mathcal{V}, \mathcal{E})$ wherein $\mathcal{V}$ and $\mathcal{E}$ refer to the set of vertices and edges of this graph, respectively. This graph is constructed by generating a vertex v for each possible association $s \in \mathcal{S}$. In the same z_m-th RRB and b_n-th RRH, two vertices $v \in \mathcal{V}$ associated with $s \in \mathcal{S}$ and $v' \in \mathcal{V}$ associated with $s' \in \mathcal{S}$ are connected by an edge if one of IDNC conditions (**C1** or **C2**) in Section 2.3.1 and $R(s) = R(s')$ are satisfied. This satisfaction ensures that all users represented by the associations have the same transmission rate, and receive always decodable transmission. Similarly, two different vertices belonging to two different RRHs/RRBs are then set adjacent by an edge if their combination results in a feasible schedule. Let vertex $v \in \mathcal{G}$ be corresponding to the association $s \in \mathcal{S}$ and vertex $v' \in \mathcal{G}$ corresponding to the association $s' \in \mathcal{S}$. The vertices v, v' are adjacent if one of the general conditions **(GC)** is satisfied.

- **GC1**: $u_i(s) = u_i(s')$ and $b_n(s) = b_n(s')$, $\forall (s, s') \in \mathcal{S}$. This condition represents that the same user can be served from multiple RRBs within the same RRH.

- **GC2**: $(b_n(s) = b_n(s')$ and $f_k(s) = f_k(s'))$ OR $(b_n(s) = b_n(s')$ and $f_k(s) \in \mathcal{H}_{u_i(s')}$ and$f_k(s') \in \mathcal{H}_{u_i(s)})$. This condition guarantees that the encoded combinations of the same users can be served by multiple RRBs within the same RRH.

- **GC3**: $u_i(s) \neq u_i(s')$ and $(b_n(s), z_m(s)) \neq (b_n(s'), z_m(s'))$. This condition completes the adjacencies in the graph for any two vertices not opposing the **CC1** constraint for any two different users.

Example 2: We show in this example the coordinated scheduling graph and its corresponding cliques of the network presented in Example 1. In Figure 2.4, each vertex v is labeled by the subscripts $nmikR$ of the association $s = (b_n, z_m, u_i, f_k, R)$, where n, m, i, k and R represent the indices of RRHs, RRBs, users, files, and rates, respectively. The dashed lines in Figure 2.3 represent either NC conditions (**C1**, **C2**) or rate condition $(R(s) = R(s'))$, and the solid lines represent **GC1**. There are many possible maximal cliques in $\mathcal{G}$ that are represented by connected vertices. Each clique represents the potential network-coded scheduling of the RRHs/RRBs that maximizes throughput. For example, the potential cliques in this example represented in Figure 2.4 by solid lines are: $\{11111, 21221\}$,$\{21332, 11111\}$, $\{21221, 11332\}$, $\{21111, 21221, 21331\}$, $\{11111, 11221, 11331\}$, $\{21112, 11222, 11332\}$, and $\{11223, 21112, 21332\}$. The achieving total throughputs of these cliques are: 2, 3, 3, 3, 3, 6, and 7 bits/s, respectively. Clearly, the vertices (circles) connected by red solid lines represent the last maximal clique, and should be the one selected as it offers the maximum throughput. It can be noted that any clique in Figure 2.4 satisfies the instant decodablity transmission to all users have vertices in it.

The following theorem characterizes the solution to the coordinated scheduling problem (2.9).

Theorem 2.2. *The coordinated scheduling problem in (2.9) is equivalent to a maximum-weight clique problem over the coordinated scheduling graph, wherein the weight of a vertex $v \in \mathcal{V}$ corresponding to the association $s =$*

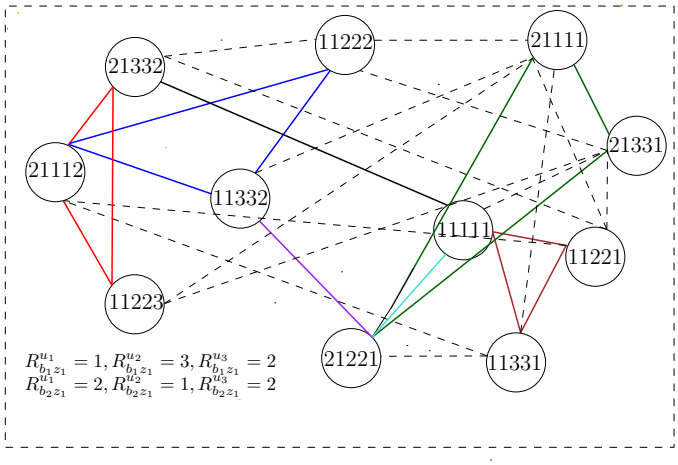


Figure 2.4: The coordinated scheduling graph of the network presented in Figure 2.3 using Algorithm 2.

$(b_n, z_m, u_i, f_k, R) \in \mathcal{S}$ *is given by*

$$w(v) = R. \tag{2.11}$$

Problem in Theorem 2.2 is NP-hard. Therefor, we make use one of the exiting algorithms, e.g., [148], [149], [150], to solve the problem in Theorem 2.2 heuristically in linear time with respect to its size as shown in Algorithm 2.

2.4.2 Power Allocation: Solution to Problem (2.10)

Given the resulting NC and user-RRB/RRH schedule in problem (2.10), finding the power level PL of each RRB is a non-convex problem. Thus, we focus on a numerical solution to achieve at least a local optimum solution using the Karush-Kuhn-Tucker (KKT) iteration approach. In particular, the corresponding power levels for the resulting NC and user-RRB/RRH schedule must satisfy the first derivative, i.e., KKT condition. Therefore, the objective function of the problem (2.10) which is optimized over the set

Algorithm 2: Coordinated NC scheduling algorithm

Data: $\mathcal{U}, \mathcal{F}, \mathcal{B}, \mathcal{H}_{u_i}, \mathcal{W}_{u_i}, P_{b_n z_m}, \mathcal{R}$, and $h^{u_i}_{b_n z_m}$,
$(b_n, z_m, u_i, f_k) \in \mathcal{B} \times \mathcal{Z} \times \mathcal{U} \times \mathcal{F}$;
Initialization: maximum-weight clique $\mathbf{C} = \phi$.
Construct $\mathcal{G}$ using Section 2.4.1.
Solve maximum-weight clique problem over $\mathcal{G}$ as follows.
while $\mathcal{G} \neq \phi$ **do**
  Compute $w(v)$ using (2.11), $\forall v \in \mathcal{G}$.
  Select $v^* = \max_{v \in \mathcal{G}}\{w(v)\}$.
  Set $\mathtt{C} = \mathtt{C} \cup v^*$ and set $\mathcal{G} = \mathcal{G}(v^*)$.
  Continue only with the vertices adjacent to v^*.
end
Output C.

of powers on an RRB-by-RRB basis, can be expressed as

$$R(P_{b_1 z_m}, ..., P_{b_B z_m}) = \sum_{b_n \in \mathcal{B}} \sum_{u_i \in \tau_{b_n z_m}} \log_2 \left(1 + \frac{P_{b_n z_m} |h^{u_i}_{b_n z_m}|^2}{\sigma^2 + \sum\limits_{b_{n'} \neq b_n} P_{b_{n'} z_m} |h^{u_i}_{b_{n'} z_m}|^2}\right)$$

$$\text{s.t. } 0 \leqslant P_{b_n z_m} \leqslant P^{\max}_{b_n z_m}, \ \forall b_n \in \mathcal{B}. \tag{2.12}$$

The steps of solving the power optimization problem in (2.10) is given as follows.

First, we take the first derivative of the objective function (2.12) with respect to $P_{b_n z_m}$ as in (2.14) and (2.15) at the top of the next page. Second, we let the gradient in (2.14) equal to zero. Then, by manipulating the optimality condition, we obtain this manipulation for optimizing the power

$$\begin{aligned} &\frac{1}{P_{b_n z_m}} \sum_{u_i \in \tau_{b_n z_m}} \left(\frac{\text{SINR}^{u_i}_{b_n z_m}}{1 + \text{SINR}^{u_i}_{b_n z_m}}\right) \\ &= \sum_{b_{n'} \neq b_n} \sum_{u_{i'} \in \tau_{b_{n'} z_m}} \frac{|h^{u_{i'}}_{b_n z_m}|^2}{P_{b_{n'} z_m} |h^{u_{i'}}_{b_{n'} z_m}|^2} \left(\frac{(\text{SINR}^{u_{i'}}_{b_{n'} z_m})^2}{1 + \text{SINR}^{u_{i'}}_{b_{n'} z_m}}\right) \end{aligned} \tag{2.13}$$

$$\frac{\partial R}{\partial P_{b_n z_m}} = \frac{\partial}{\partial P_{b_n z_m}} \sum_{u_i \in \tau_{b_n z_m}} \log_2 \left(1 + \frac{P_{b_n z_m} |h^{u_i}_{b_n z_m}|^2}{\sigma^2 + \sum\limits_{b_{n'} \neq b_n} P_{b_{n'} z_m} |h^{u_i}_{b_{n'} z_m}|^2}\right)$$
$$+ \frac{\partial}{\partial P_{b_n z_m}} \sum_{b_{n'} \neq b_n} \sum_{u_{i'} \in \tau_{b_{n'} z_m}} \log_2 \left(1 + \frac{P_{b_{n'} z_m} |h^{u_{i'}}_{b_{n'} z_m}|^2}{\sigma^2 + \sum\limits_{b_{\tilde{n}} \neq b_{n'}} P_{b_{\tilde{n}} z_m} |h^{u_{i'}}_{b_{\tilde{n}} z_m}|^2}\right) = \frac{1}{P_{b_n z_m}} \sum_{u_i \in \tau_{b_n z_m}} \left(\frac{\mathrm{SINR}^{u_i}_{b_n z_m}}{1 + \mathrm{SINR}^{u_i}_{b_n z_m}}\right)$$
$$- \sum_{b_{n'} \neq b_n} \sum_{u_{i'} \in \tau_{b_{n'} z_m}} \frac{|h^{u_{i'}}_{b_n z_m}|^2}{P_{b_{n'} z_m} |h^{u_{i'}}_{b_{n'} z_m}|^2} \left(\frac{(\mathrm{SINR}^{u_{i'}}_{b_{n'} z_m})^2}{1 + \mathrm{SINR}^{u_{i'}}_{b_{n'} z_m}}\right), \tag{2.14}$$

where,

$$\mathrm{SINR}^{u_{i'}}_{b_{n'} z_m} = \left(1 + \frac{P_{b_{b'} z_m} |h^{u_{i'}}_{b_{n'} z_m}|^2}{\sigma^2 + \sum\limits_{b_{\tilde{n}} \neq b_{n'}} P_{b_{\tilde{n}} z_m} |h^{u_{i'}}_{b_{\tilde{n}} z_m}|^2}\right) \tag{2.15}$$

and $u_i \in \tau_{b_n z_m}$ and $u_{i'} \in \tau_{b_{n'} z_m}$ are the scheduled users of the b_n-th RRH, and the $b_{n'}$-th RRH at the z_mth RRB for $\forall b_n, b_{n'} \in \mathcal{B}$, respectively.

Therefore,

$$P_{b_n z_m} = \frac{\sum\limits_{u_i \in \tau_{b_n z_m}} \left(\frac{\mathrm{SINR}_{b_n z_m}^{u_i}}{1+\mathrm{SINR}_{b_n z_m}^{u_i}} \right)}{\sum\limits_{b_{n'} \neq b_n} \sum\limits_{u_{i'} \in \tau_{b_{n'} z_m}} \frac{|h_{b_n z_m}^{u_{i'}}|^2}{P_{b_{n'} z_m} |h_{b_{n'} z_m}^{u_{i'}}|^2} \left(\frac{(\mathrm{SINR}_{b_{n'} z_m}^{u_{i'}})^2}{1+\mathrm{SINR}_{b_{n'} z_m}^{u_{i'}}} \right)} \tag{2.16}$$

Finally, we compute the optimal $[P_{b_n z_m}]_{n=1}^{B}$ by interpreting the power condition iteratively. More precisely, given the power allocation at the first iteration, the right-hand side of (2.16) can be computed using the current power, then the new results are used to update the new power allocations.

Furthermore, to account for the RRB power constraint, an additional projection on $[.]_0^{P^{\max}}$ can be taken such that in each iteration, the power level $P_{b_n z_m}$ is updated according to the following

$$P_{b_n z_m, \mathrm{new}} = \left[\frac{\sum\limits_{u_i \in \tau_{b_n z_m}} \left(\frac{\mathrm{SINR}_{b_n z_m}^{u_i}}{1+\mathrm{SINR}_{b_n z_m}^{u_i}} \right)}{\sum\limits_{b_{n'} \neq b_n} t_{b_{n'} z_m}} \right]_0^{P^{\max}}, \tag{2.17}$$

where,

$$t_{b_{n'} z_m} = \sum_{u_{i'} \in \tau_{b_{n'} z_m}} \frac{|h_{b_n z_m}^{u_{i'}}|^2}{P_{b_{n'} z_m} |h_{b_{n'} z_m}^{u_{i'}}|^2} \left(\frac{(\mathrm{SINR}_{b_{n'} z_m}^{u_{i'}})^2}{1+\mathrm{SINR}_{b_{n'} z_m}^{u_{i'}}} \right) \tag{2.18}$$

and the notation $[a]_y^x$ represents a upper bounded above by x and lower bounded below by y. Note that the iterative process of (2.17) must converge to a point either lie in the interior of the constrained set or on the boundary, so as to satisfy the "KKT condition" of the optimization problem (for details see [151], [153]). It is important to note that, the nominator on the right-hand side of (2.17) represents the effect power of z_m-th RRB in b_n-th RRH on all corresponding RRHs in the schedule, i.e., it is the derivative of the $b_{n'}$-th RRHs terms with respect to the z_m-th RRB power in the bth RRH. In other words, it summarizes the interfering effect of $P_{b_n z_m}$ on the $b_{n'}$th

Algorithm 3: Overall iterative approach for solving problem (2.5)

Initialization: Set the initial values for $P_{b_n z_m}$ $\forall (b_n, z_m) \in \mathcal{B} \times \mathcal{Z}$ and set $t = 1$.

Repeat:

- Solve problem (2.9) based on Algorithm 1 and obtain the NC and user-RRB/RRH association according to Theorem 2.2.
- Solve problem (2.10) as in Section 2.4.2 and obtain the power levels of the corresponding RRBs.
- $t = t + 1$

Until convergence

RRH. Moreover, it depends on the transmit power, SINR, and the ratio of the direct and the interfering channel gains. The dominator on the right-hand side of (2.17) shows the effect of the combined noise and interference in z_m-th RRB of b_n-th RRH.

The overall two-step algorithm to problem (2.5) is summarized in Algorithm 3.

Complexity Analysis

In this part, we analyze the complexity of constructing the coordinated scheduling graph, i.e., generating the vertices and connecting them, and executing the maximum-weight search.

First, we construct the scheduling graph by generating $O(UF|\mathtt{R}|)$ vertices, where $|\mathtt{R}| = UZ$. These vertices represent the requested popular files by users, which can be targeted by one RRB with one of the achievable rates. However, vertices representing different rates cannot be part of one clique in each RRB in the same RRH. Clearly, avoiding such vertices reduces the complexity to $O(UF)$. Generalizing this to the whole system has a complexity of $O(UFZB)$. Then, we iteratively search over these vertices for each rate as will be explained later. The adjacency matrix that connects vertices by edges needs $\mathcal{O}\left(U^2F^2ZB\right)$ for coding conditions, and $O(U^2F^2Z^2B^2)$ for

general conditions. Thus, the complexity of building the adjacency matrix is $O(U^2F^2ZB + U^2F^2Z^2B^2) = O(U^2F^2Z^2B^2)$.

Second, the maximum-weight search algorithm iteratively search over all vertices for each rate of each RRB in each RRH. This needs at most $|\mathtt{R}|$ maximal cliques. Since each user cannot be scheduled to more than one RRH, each maximal clique has at most U vertices. Further, each user can be targeted by at most one file (i.e., one vertex) from one RRB across all RRHs. Thus, the calculation of the weights in each maximal clique and its number of vertices (considering the lowest rate that involves all users) has a complexity of $O(B|\mathtt{R}|[(UF + U)])$. In the whole system, the complexity is $O(BZ|\mathtt{R}|[(UF + U)])$.

Finally, combining all the complexities of the different algorithm elements yields the total complexity of the algorithm per iteration of $O(UFZB + U^2F^2Z^2B^2 + BZ|\mathtt{R}|[(UF + U)])$. Overall, roughly, the second term dominates the total complexity as $|\mathtt{R}| \leqslant UZ$, which abstracts the complexity of Algorithm 2 to $O(U^2F^2Z^2B^2)$.

Communication and Storage Overheads

It is important to recall that the whole process in this work is done by the CBS. Furthermore, the CBS knows users' side information and stores them in a log ledger. Therefore, compared to the standard uncoded solution, coded scheme needs nothing. However, the coding scheme may need overhead for the headers of each encoded transmission (to notify users of the XORed files in each transmission), but this will not exceed a few bytes. It is also worth noting that the encoded file has the same size as the source files, which simplifies both the coding overhead at the RRBs/RRHs and the users' decoding complexity [44]. On the other hand, for storage overheads, users need to cache some files to be used later as side information. Such cached file is refreshed over a short period of time so as it does not increase with time, i.e., newer popular files are replacing older non-used files. Furthermore, increasing the capabilities of devices as well as reducing their memory costs enable that easily. Thus, as long as the number of cached files is bounded, i.e., window size in our case, the cached files would not consider

Table 2.1: Numerical parameters

Bandwidth	10 MHz
Cellular layout	Hexagonal
Channel estimation	Perfect
Channel model	SUI-Terrain type B
Path loss model	$128.1 + 37.6 \log_{10}(\text{dis.[km]})$
Background noise power	-168.60 dBm/Hz
Shadowing variance	0 dB
Distribution of users	Uniform

as overheads (for more details about caching, see for example [122]).

2.5 Numerical Results

In this section, we validate our developed joint CRAN-CLNC and iterative schemes using numerical results. We consider a downlink C-RAN system that described in Section 2.2 and plotted in Figure 2.1. The system setting in this section follows the setup studied in [82] and [109]. The number of RRHs is fixed to 3 unless stated. The additional numerical parameters are summarized in Table 2.1. Each presented value in each plot for the joint and the iterative schemes is obtained by averaging the throughput expressions in Theorem 2.1 and Theorem 2.2 over a large number of iterations, respectively. At each run, users' side information is randomly generated. To assess the performances of our proposed schemes, we simulate various scenarios: number of users U, number of RRBs per each RRH's frame, maximum power $P^{\max}$, and cell size D. For numerical comparison, the proposed solutions are compared with the following two baseline schemes.

– **Classical IDNC**: This scheme optimizes the selection of XOR file combinations with the absence in physical layer aspects. For full file's reception, the RRB's transmission rate is limited by the weakest user who experiences the lowest transmission capacity.

– **RLNC**: This scheme associates users to a single RRB to which it has the maximum capacity. RLNC is employed for each RRB that has

multiple associated users. The encoding is done by mixing all files with different random coefficients, and the selected transmission rate in each RRB is limited by the minimum achievable capacity of the assigned users.

For completeness of our work, we also compare our proposed schemes with the recent RA-IDNC and uncoded schemes that were studied in [82], [109], respectively.

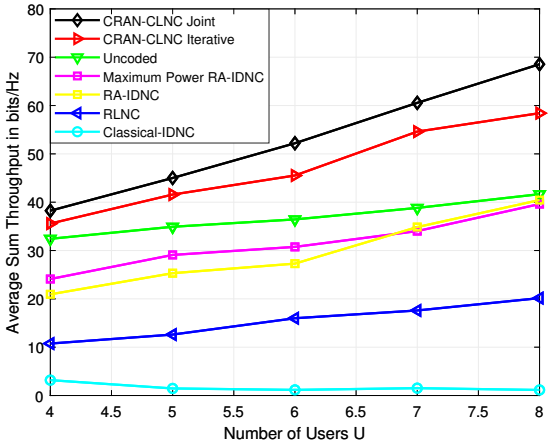


Figure 2.5: Average sum throughput in bits/Hz. vs the number of users U.

In Figure 2.5, we plot the average sum throughput versus the number of users U for a C-RAN composed of 2 RRBs per RRH's frame, $F = 10$ files, a file size of 1Mb, a maximum transmit power $P^{\max} = -42.60$ dBm, and a cell size $D = 500$m. From the figure, we see that the performance gain of the proposed joint and iterative schemes outperform the performances of the other schemes in terms of throughput maximization. In particular, when U is sufficiently small ($U = 4, 5$), the uncoded scheme works closely to the proposed schemes, i.e., joint CRAN-CLNC scheme offers 28% improvement over uncoded for $U = 5$. This is due to the fact that the proposed CRAN-CLNC schemes suffer from a small number of encoded scheduling opportunities in each RRB/RRH. In contrast, when U is moderate ($U = 7, 8$), the proposed

joint scheme achieves an appreciable performance improvement (53% for $U = 7$ and 63% for $U = 8$) over the uncoded scheme. This is because the encoded opportunities of mixing users' flow in each RRB/RRH increases as the number of users U increases.

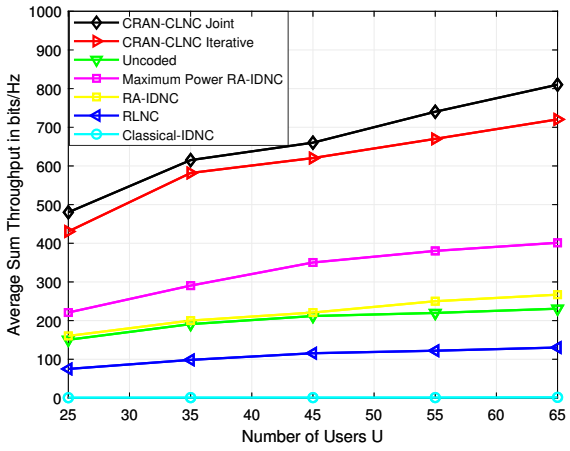


Figure 2.6: Average sum throughput in bits/Hz. vs a large number of users U.

In order to evaluate the performance of the proposed schemes in a large C-RAN, Figure 2.6 illustrates average sum throughput versus a large number of users U for a C-RAN composed of 6 RRHs, 5 RRBs per RRH's frame, and $F = 68$ files, each with a size of 1Mb. The figure suggests that, for the uncoded strategy, the average throughput increases as the number of users U increases, but uncoded scheme serves only one user per RRB (serves in total ZB users), which limits the overall gain in the average throughput. Conversely, if the number of users is large enough compared to ZB, joint CRAN-CLNC scheme achieves significant gain over the uncoded scheme (the average throughput is twice for $U = 30$ and almost three times higher for $U = 60$). We can also see that, with the limitation of rate equality for all RRHs, RA-IDNC scheme offers improved performance as compared to the uncoded scheme at sufficiently large U. Finally, the maximum power RA-IDNC scheme, which exploits the NC abilities, serves a good number of

users in each transmission. Thus, it works better than the uncoded one in a large C-RAN system.

In Figure 2.7, we investigate the impact of the number of RRBs Z on the C-RAN's average sum throughput. We set the parameters $U = 7$, $F = 10$, $N = 1$Mb, $P^{\max} = -42.60$ dBm, and $D = 500$m. From this figure, it can be seen that the average throughput of all schemes increases as the number of RRBs Z increases. This is because all schemes agree on serving the same set of users in different RRBs in the same RRH. However, the performances of the proposed joint scheme over the uncoded and the RA-IDNC scheme is pronounced. In fact, the proposed joint scheme emphasizes the different relative behavior of the uncoded and RA-IDNC schemes. In particular, the uncoded scheme only focuses on the high achievable rates at the expense of transmitting at most one file to a single user from each RRB in all RRHs. The RA-IDNC scheme leverages the potential of rate-aware codes in maximizing the throughput, but it limits the transmission rate of the same RRB in all RRHs to the minimum, which affects the overall gain. Moreover, the proposed joint scheme offers improved performance as compared to the iterative one for larger Z as the search space becomes larger for larger Z. This becomes more advantageous for the proposed CRAN-CLNC joint algorithm.

In Figure 2.8, we plot the average sum throughput versus the maximum power $P^{\max}$, for a C-RAN setting composed of 2 RRBs in each RRH's transmit frame, 7 users, 9 files, a file size $N = 1$Mb, and a cell size $D = 500$m. We can see from the figure that the achieved throughput of the CRAN-CLNC joint scheme with a larger $P^{\max}$ is significantly outperformed the iterative strategy. This is because, in the regime of large $P^{\max}$, i.e., high inter-RRHs interference, the iterative scheme does not consider the whole search space in the system and rather partial consideration. In contrast, the joint scheme as a method of power optimization considers the overall interference's effect in the system.

In Figure 2.9, we show the average sum throughput versus the cell size D, for a C-RAN setting composed of 7 users, 9 files, 2 RRBs in each RRH's transmit frame, each RRB has a maximum transmit power $P^{\max} = -26.98$

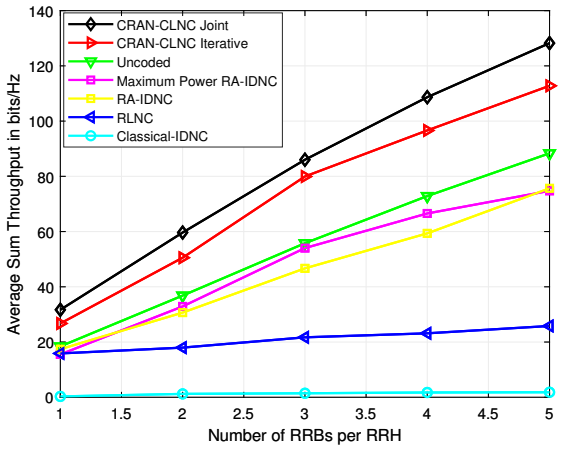


Figure 2.7: Average sum throughput in bits/Hz. vs. the number of radio resource blocks Z.

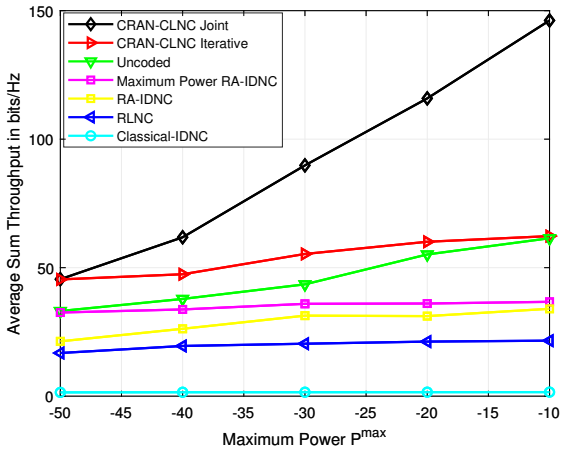


Figure 2.8: Average sum throughput in bits/Hz. vs maximum power $P^{\max}$.

dBm, and a file size $N = 1$Mb. The figure confirms the result in Figure 2.8 that if the cell size is sufficiently small, i.e., high interference level, the proposed joint scheme can provide high gain in average throughput over the iterative one. As the cell size increases, i.e., low interference level, the

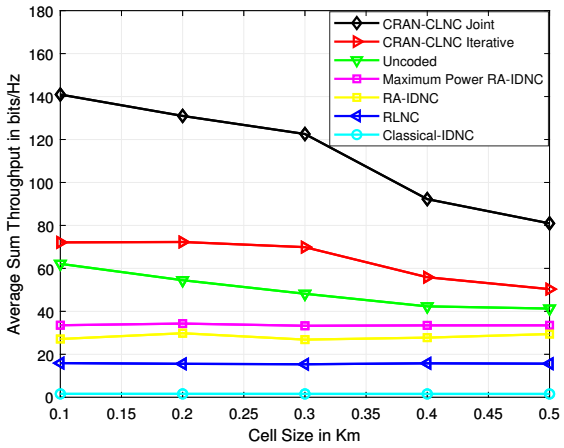


Figure 2.9: Average sum throughput in bits/Hz. vs cell size D in K_m.

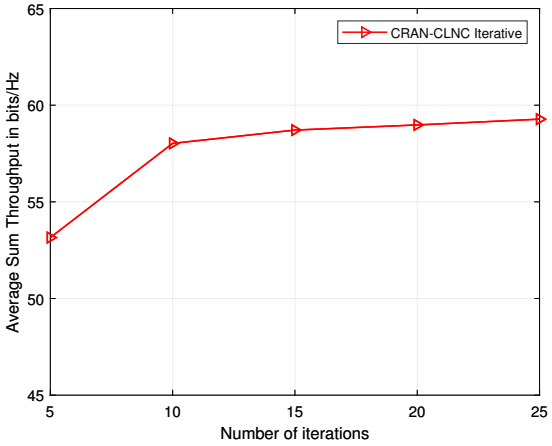


Figure 2.10: Average sum throughput in bits/Hz. vs the number of iterations.

performance of the proposed joint solution over the iterative one decreases and gains only 10% improvement.

We further study the convergence behavior of the iterative approach that explained in Algorithm 3. We set the parameters $U = 8$, $M = 10$, $Z = 2$,

$P^{\max} = -26.98$ dBm, $N = 1$Mb, and cell size $D = 500$m. We plot the average sum throughput versus the number of iterations, as shown in Figure 2.10. As we can see from the figure, solving the coordinated scheduling and the power allocation problems converges iteratively. This confirms the fact that both problems are increasing in the objective function.

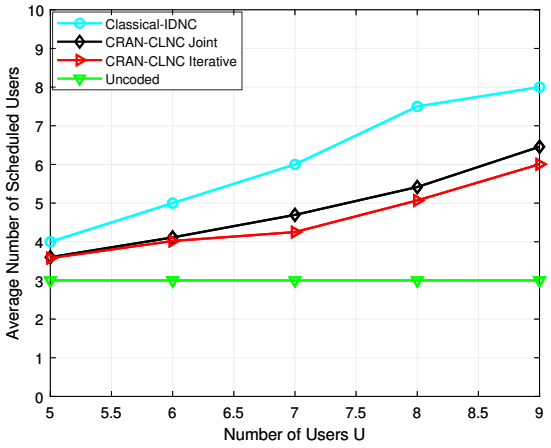


Figure 2.11: Average number of scheduled users vs the total number of users U.

Finally, we study the number of scheduled users by the RRBs/RRHs versus the number of total number of users in C-RAN system. We set the parameters $M = 9$, $Z = 1$, $P^{\max} = -26.98$ dBm, $N = 1$Mb, and cell size $D = 500$m. It can be seen from Figure 2.11 that the number of scheduled users in the classical IDNC scheme is larger than the number of scheduled users in all other schemes. This confirms the fact that classical IDNC scheme works well from the network layer perspective. However, selecting the minimum rate of the scheduled users to the RRBs/RRHs degrades its throughput performance as explained in the aforementioned figures. The uncoded scheme always severs BZ users as shown in the figure. As expected, the proposed schemes strike a balance between the classical IDNC and uncoded schemes in terms of the number of scheduled users.

It is worth remarking that classical IDNC scheme offers completion time

reduction of a frame of files in a number of NC works. However, as mentioned before that serving many users in each RRB while limiting the transmission rate of the RRB to the minimum exhibits a poor performance from a cross-layer perspective. Therefore, by means of extensive numerical results, we showed in this work that CRAN-CLNC joint and iterative schemes provide a more effective way to use NC, power adaptation, and simultaneous transmissions in C-RAN setting than all other schemes.

2.6 Chapter Summary

In this chapter, the cross-layer throughput maximization problem that considers network-coded scheduling and transmit power adaptation in C-RAN downlink communication has been investigated. Two schemes have been developed using graph theory technique, namely joint and iterative CRAN-CLNC schemes. Our numerical results have showed that the proposed schemes achieve improved performances (53% for small C-RAN size and almost three times higher for large C-RAN) over the uncoded solution.

Chapter 3

Cross-Layer Cloud Offloading with Quality of Service Guarantees in F-RANs

In Chapter 2, we introduced a cross-layer NC framework that maximizes the throughput in C-RANs setting. However, the developed framework did not consider the caching strategy at the RRHs and required minimum rate for users' QoS. To this end, in this chapter, we study the joint design of CBS and edge processing for F-RAN setting in which the eRRHs are equipped not only with the functionalities of standard RRHs in C-RAN, but also with local cache. Therefore, the introduced framework in Chapter 2 for C-RAN setting is developed to solve the immediate file delivery in F-RAN with the goal of maximizing the CBS offloading while guaranteeing users QoS, subject to the eRRH's cached files and minimum required rate. The organization of this chapter is given as follows. We summarize the accomplished works and research contributions in Section 3.1. The system model and general assumptions are discussed in Section 3.2. The joint cross-layer CBS offloading and QoS guarantee problem is modeled and formulated in Section 3.3. In Section 3.4 and Section 3.5, we develop joint and iterative cross-layer CBS offloading and QoS guarantee approaches. In Section 3.6, we present some selected numerical results, and in Section 3.7, we summarize the chapter.

3.1 Accomplished Works and Research Contributions

Our contributions in this chapter are summarized as follows.

1. In the first part of this chapter, we formulate the joint cross-layer CBS offloading and QoS guarantee problem in an F-RAN setting using over an ONC graph. This formulated problem is divided into two subproblems, namely the *inner throughput maximization* problem and the *outer CBS consumed time reduction* problem.

2. In the second part of this chapter, we show that the inner problem is equivalent to a maximum-weight independent set problem over a new designed graph, herein called the FRAN-CLNC graph. The designed graph jointly considers the main optimization parameters of the inner problem that are mentioned as follows: 1) **User Association Mechanism:** Each vertex represents a group of users assigned to one of the RRBs of the eRRHs, given that their requests can be combined using NC and that they can be served in this RRB with the required minimum rate; and 2) **Power Level:** Each vertex represents a power control solution of a specific RRB that achieves a given rate for its targeted users, given the interference caused by the same RRB in different eRRHs. Afterward, given the developed joint solution of the inner problem, we propose a greedy coloring solution to the outer problem. In addition, we explain the required computational complexity of our proposed joint scheme.

3. In the third part of this chapter, we propose a low-complexity iterative approach to mitigate the high complexity of the above joint approach. Extensive simulations are finally performed to illustrate the potential CBS offloading and user throughput trade-offs of the proposed approaches. Our simulation results reveal the following three insights: (i) the cross-Layer QoS-unaware technique performs fairly well in terms of offloading the CBS's resources, but it exhibits a poor

performance from users' QoS perspective and ii) our developed solutions strike a balance between the CBS offloading and QoS guarantee by targeting users that have weak channel capacities from the CBS by the eRRHs with the required QoS, and then scheduling the unscheduled users from eRRHs to CBS's channel with minimum possible CBS consumed time.

3.2 System Model and Assumptions

As illustrated in Figure 3.1, we consider the downlink communication in F-RAN, where B eRRHs, denoted by the set $\mathcal{B} = \{b_1, b_2, \cdots, b_B\}$, in cooperation with the CBS serve U users, denoted by the set $\mathcal{U} = \{u_1, u_2, \cdots, u_U\}$. It is assumed that the B eRRHs communicate with the CBS using low-rate fronthaul links. The library F of popular files, denoted by the set $\mathcal{F} = \{f_1, f_2, \cdots, f_F\}$, constitutes the set of most possible files to be requested by the users for within a given time frame. The considered model generalizes the C-RAN set-up studied in [82], [109]. In addition to the conventional RRHs' functions in C-RAN, each eRRH b_n in F-RAN is equipped with a cache, which can pre-fetch a subset of files $\mathcal{H}_{b_n}$ that represents the eRRH *Has* set. We assume that the number of cached files in any eRRH is equal, i.e., $|\mathcal{H}_{b_n}| = \mu F, \forall\ b_n \in \mathcal{B}$, where μ is the caching ratio, which is defined as the ratio of the number of files in the Has set of each eRRH to the total number of files in the library F. It is also assumed that $\mathcal{B}$ eRRHs collectively cache all files in the library, i.e., $\bigcup_{n=1}^{B} \mathcal{H}_{b_n} = \mathcal{F}$, the distribution of files among eRRHs is fixed, and the same file can be stored in different eRRH's caches. Each eRRH's transmit frame consists of Z orthogonal time/frequency RRBs. Let $\mathcal{Z} = \{z_1, z_2, \cdots, z_Z\}$ be the set of RRBs of the frame of the eRRH, in which each RRB is maintained at a fixed duration T, as shown in Figure 3.2(a). Therefore, the total number of available RRBs in the system is $Z_{\text{tot}} = Z$. The CBS has multiple C orthogonal channels, denoted by the set $\mathcal{C} = \{c_1, c_2, \cdots, c_C\}$, as shown in Figure 3.2(b).

On the user side, each user u_i has some stored files from prior downloads in its *Has* set $\mathcal{H}_{u_i}$, and is interested to download only *one* file from a library

of F popular files, denoted as the *Wants* set $\mathcal{W}_{u_i}$ of u_i. All files are assumed to be equally sized of N bytes, and are delivered by the eRRHs across a number of RRBs (or possibly by the CBS across a number of channels). It is important to note that if the transmission rate of any RRB is not sufficient to guarantee the successful download of the requested file, user u_i will download that file in parts from the same eRRH by listening to various RRBs that denoted by $\mathbf{Z}_{u_i}$ until completion. If user requests multiple files simultaneously, he/she will get them from the same eRRH by listening to multiple RRBs, each RRB would deliver each file if its transmission rate is capable of doing so. Since the proposed schemes send the requested popular files with the maximum possible throughput (the least number of RRBs), it would provide more available RRBs for serving other users. If some users still require more files, they can receive them either from the CBS (each user will get one file) or from the eRRHs in the next transmissions.

The eRRHs are designed to be cost-efficient devices, so their role is confined to sending all the relevant information to the CBS through low rate-wireless fronthaul, such as the indices of the requested/downloaded files by each user. On the other hand, the CBS, which has sufficient processing capabilities and memory, stores all the information received from the eRRHs in a log file so as to determine which files should be XORed and transmitted to the users, i.e., users-eRRH/CBS resources associations. It is further responsible for synchronizing the transmission of the different frames across all eRRHs and determining the power control policies of the network. Therefore, the whole process in this work is carried at the CBS in a fully centralized fashion. Furthermore, since the CBS is responsible to do the whole process, it is the one that needs to track the download history of users, which is known in the NC literature by the side information. Thus, the CBS can temporarily store this side information of users and use them later to combine files using ONC XOR operations. Even if user moves from one CBS to another, its log entries of the current popular files can be transferred as part of this user's handover procedure.

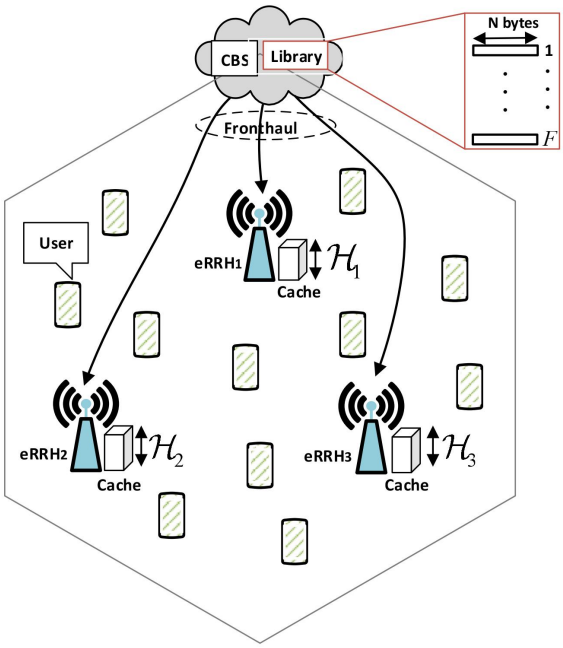


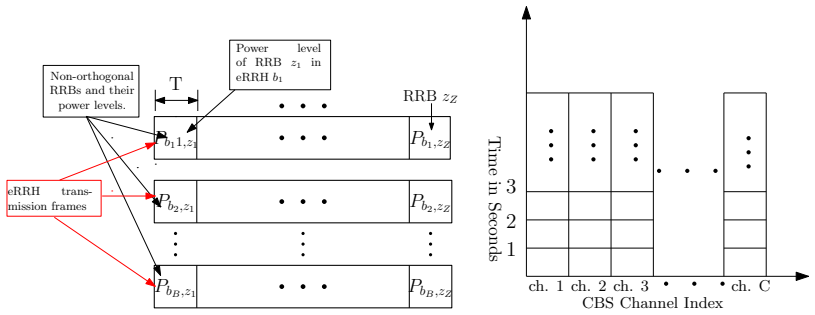
Figure 3.1: A fog radio access network composed of CBS, 11 users, and 3 eRRHs. Only low-rate fronthaul links are required to connect the CBS with the eRRHs.

3.3 Modeling and Problem Formulation

This section formulates the joint CBS offloading and QoS guarantee optimization problem in the downlink F-RAN, given its cross-layer architecture (i.e., physical and network layers) and the well-known ONC graph [22].

3.3.1 Physical-Layer

For eRRH-user transmission strategy, files are transmitted over different RRBs via a wireless link. The achievable rate of the u_i-th user assigned to the z_m-th RRB in the b_n-th eRRH at the t-th transmission can be expressed $R^{u_i}_{b_n z_m}(t) = \log_2(1+\text{SINR}^{u_i}_{b_n z_m}(\mathbf{P}))$, where $\text{SINR}^{u_i}_{b_n z_m}$ denotes the corresponding signal-to-interference plus noise-ratio experienced by the u_i-th user when



(a) eRRHs transmission frames. (b) CBS's resources.

Figure 3.2: Illustration of the transmission frames of the eRRHs and CBS's resources, i.e., time in seconds of the CBS's orthogonal channels.

it is assigned to the RRB z_m in the b_n-th eRRH. This $\text{SINR}^{u_i}_{b_n z_m}$ is given by $\text{SINR}^{u_i}_{b_n z_m}(\mathbf{P}) = \dfrac{P_{b_n z_m}|h^{u_i}_{b_n z_m}|^2}{\sum\limits_{b_{b'} \neq b_n} P_{b_{b'} z_m}|h^{u_i}_{b_{b'} z_m}|^2 + \sigma^2}$, where $h^{u_i}_{b_n z_m}$ denotes the complex channel gain from the z_m-th RRB in the b_n-th eRRH, $P_{b_n z_m}$ is the power allocation level (PL) of the z_m-th RRB in the b_n-th eRRH, $\mathbf{P} = [P_{b_n z_m}]$ is the $B \times Z$ matrix containing the PLs of the considered network, and σ^2 denotes the AWGN variance. For brevity, the time argument t is often omitted when it is clear from the context. It is assumed that the channel gains $\{h^{u_i}_{b_n z_m}\}_{b_n \in \mathcal{B}, z_m \in \mathcal{Z}, u_i \in \mathcal{U}}$ are mostly affected by the location of the users within the service area and can be affected by other channel impairments, e.g., short term fading and shadowing. Thus, each user-RRB/eRRH pair has a different channel condition, i.e., $\{h^{u_i}_{b_n z_m}\}_{b_n \in \mathcal{B}, z_m \in \mathcal{Z}, u_i \in \mathcal{U}}$. Such channel model leads to heterogeneous physical-layer rates from different RRBs/eRRHs to different users. From practical constraints, the power level of RRB z_n in eRRH b_n is subject to the maximum power $P^{\max}_{b_n z_m}$ i.e., $P_{b_n z_m} \leqslant P^{\max}_{b_n z_m}$. Since the transmit frame of each eRRH consists of Z orthogonal RRBs, $\text{SINR}^{u_i}_{b_n z_m}(\mathbf{P})$ depends solely on the scheduled users in z-th RRB across the remaining $b_{n'} \neq b_n$ eRRHs and the corresponding power level $P_{b_{n'} z_m}$.

At the CBS, users' requests (if any) are transmitted via multiple orthog-

onal channels. So, the achievable rate of the u_i-th user from the CBS is expressed as

$$r_{u_i}^{\mathrm{CBS}}(t) = \log_2\left(1 + \frac{P|h_{u_i}|^2}{\sigma^2}\right), \tag{3.1}$$

where h_{u_i} denotes the channel gain from the CBS to the u_i-th user, and P is the transmitted power of the CBS, and assumed to be fixed during the whole transmission process. We consider the standard assumption that both $\{h^{u_i}_{b_n z_m}\}_{b_n \in \mathcal{B}, z_m \in \mathcal{Z}, u_i \in \mathcal{U}}$ and $\{h_{u_i}\}_{u_i \in \mathcal{U}}$ remain constant during each transmission interval and change from one transmission interval to another transmission interval. The z_m-th RRB of b_n-th eRRH and the c_k-th channel of the CBS can transmit at a rate which is at most equal to the minimum achievable capacity of their assigned users, i.e., $R_{b_n z_m} \leqslant R^u_{b_n z_m}$ and $r^{\mathrm{CBS}}_{c_k} \leqslant r^{\mathrm{CBS}}_{u_i}$, where $c_k \in \mathcal{C}$. The set of achievable capacities of all users in all RRBs across all eRRHs can be represented by the set as follows $\mathcal{R} = \bigotimes_{(b_n, z_m, u_i) \in\ \mathcal{B}\times\mathcal{Z}\times\mathcal{U}} R^{u_i}_{b_n z_m}$, where the symbol $\bigotimes$ represents the product of the set of the achievable capacities.

3.3.2 Opportunistic Network Coding in the Network-Layer

In [145], the authors introduced the ONC graph as a design to represent all users' demands and their possible ONC combinations. The generation of the ONC file combinations κ relies on the fact that the multiplexed-users τ can be served by one combined (encoded) file. Therefore, two distinct vertices v_{u_i, f_l} and $v_{u_{i'}, f_{l'}}$ are linked with a coding conflict edge if the following connectivity condition (CC) is satisfied.

- **CC**: $f_l \neq f_{l'}$ and $(f_l, f_{l'}) \notin \mathcal{H}_{u_i} \times \mathcal{H}_{u_{i'}}$. This condition represents the fact that both users do not want the same file and at least one of the users inducing either vertices does not have what other user wants.

Clearly, this condition ensures that any file combinations κ generated by the vertices of any independent set $\mathcal{I}$ is always decodable to the targeted users τ having vertices in the ONC graph. Thus, each $\mathcal{I}$ in that graph consists of two components, i.e., $\mathcal{I} = (\tau, \kappa)$.

Example 3: We provide an example for a better illustration of the construction of both ONC and color graphs that will be used throughout this chapter. As shown in Figure 3.3, a system containing of 1 eRRH, 1 RRB, and 5 users.

Example 4: To illustrate the concept of CLNC and the performance metrics that are considered in this chapter, let us consider the same example in Figure 3.3. Given a certain power allocation to the eRRH and the CBS, assume that the achievable capacities of the 5 users to the RRB of the eRRH and to the CBS are given in the right side of Figure 3.3. In order to maximize throughput and offload the CBS for this example, there are many possible solutions.

One possible solution is that the eRRH targets the set of users $\tau_{b_1 z_1} = u_2$, u_3 and u_4 in one RRB by XORing f_1, f_2 and f_3 into an encoded combination $\kappa_{b_1 z_1} = f_1 \oplus f_2 \oplus f_3$ and sends κ_{11} with an achievable transmission rate of 2 bits/s. Therefore, the encoded combination κ_{11} is instantly decodable by the set of targeted users τ_{11} and both components are represented by an independent set, i.e., $\mathcal{I}_{b_1 z_1} = (\tau_{b_1 z_1}, \kappa_{b_1 z_1})$. The decoding operation at the users' side can be performed as follows: u_2 retrieves f_1 by XORing κ_{11} with f_2 and f_3, u_3 retrieves f_2 by XORing κ_{11} with f_1 and f_3, and u_4 retrieves f_3 by XORing κ_{11} with f_1 and f_2. The achievable overall throughput in this scenario is 6 bits/s as each of $u2$, $u3$, and $u4$ simultaneously can receive 2 bits/s worth of throughput. Such throughput improves upon the 2.5 bits/s throughput achievable without coding. Indeed, the eRRH targets the user that has the maximum achievable capacity among all other users, i.e., u_4 with a rate of 2.5 bits/s. Next, given this smart user-scheduling and NC process by the eRRH, the CBS schedules the remaining users to its channels so as to minimize its physical-resources consumed time. Thus, CBS targets u_1 by sending f_2 with transmission rate of 2.5 bits/s, which gives a consumed time of $\frac{N}{2.5}$ seconds, and targets u_5 by sending f_3 with transmission rate of 5 bits/s, which gives a consumed time of $\frac{N}{5}$ seconds. As a result, the CBS reserves channel 1 and channel 2 for the duration of sending file $\kappa_1 = f_2$ and $\kappa_2 = f_3$ to $\tau_1 = u_1$ and $\tau_2 = u_5$, respectively. Therefore, the overall CBS total consumed time is the sum of its channel consumed times $\frac{N}{2.5} + \frac{N}{5} = \frac{3N}{5}$

seconds that used to serve u_1 and u_5. Clearly, the smallest this expression, the more CBS's channels can be used to serve other users, thus yielding to a better CBS offloading improvement.

Another possible solution is that the eRRH targets u_1, u_2, and u_3 in one RRB by sending the encoded file $f_1 \oplus f_2$ with a transmission rate of 0.5 bits/s (the lowest rates of u_1, u_2 and u_3), and the CBS targets u_4 and u_5 in one channel by sending the file f_3 with transmission rate of 1.25 bits/s. Scheduling u_1, u_2, and u_3 to the eRRH, which can be performed from the coding/decoding perspective, would achieve an even worse throughput and CBS consumed time results than the above solution. Indeed, it will not only limit the rate sent by the eRRH to 0.5 bits/s that results to an overall throughput of 1.5 bits/s, but will also require $\frac{N}{1.25}$ seconds to complete the file delivery from the CBS, longer than the above solution that needs two transmissions from the CBS. This clearly exhibits the effectiveness of CLNC in maximizing the throughput and minimizing the CBS actual-consumed time. This result can be further improved by optimizing the power assignment and the corresponding rate selection for each RRB in each eRRH, as will be used in this work through these two important and novel features of CLNC.

3.3.3 Problem Formulation

Given the aforementioned configurations, we are now ready to formulate the joint CBS offloading and QoS guarantee optimization problem in an F-RAN setup as follows. Let, $\mathcal{I}_{b_n z_m}$, $b_n \in \mathcal{B}$ and $z_m \in \mathcal{Z}$, be an independent set, whose encoded XOR file $\kappa_{b_n z_m}$ will be combined and transmitted by RRB z_m in eRRH b to serve the users in $\tau_{b_n z_m}$. These independent sets represent the transmissions by eRRHs/RRBs that serve users with the required QoS. Our problem is then reduced to finding the set of independent sets $\mathcal{P} = \mathcal{I}_{c_h}, \cdots, \mathcal{I}_{c_C}$, so as to target the remaining users with the minimum consumed CBS's resources. In other words, $\mathcal{P}$ is the set of all independent sets of the ONC graph after removing $\mathcal{I}_{b_n z_m}, \ \forall \ b_n, \ z_m$, i.e., $\left(\mathcal{G} \setminus \bigcup_{n=1}^{B} \bigcup_{m=1}^{Z} \mathcal{I}_{b_n z_m}\right)$. Now, let us define P as a subset in $\mathcal{P}$, i.e., $P \subseteq \mathcal{P}$,

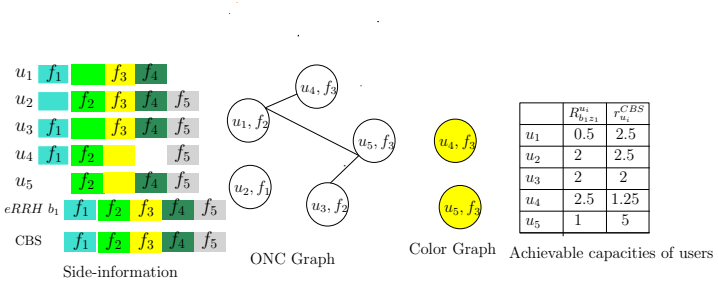


Figure 3.3: A system containing 5 users, in which each user requests one file. Each independent set in the ONC graph represents a feasible ONC combination. For example, one possible independent set $\mathcal{I}$ is: $\{(u_1, f_2), (u_2, f_1), (u_3, f_2)\}$ that corresponds to transmitting the combination $\kappa_{b_1 z_1} = f_1 \oplus f_2$ to the targeted users $\tau_{b_1 z_1} = u_1, u_2$ and u_3, i.e., $\mathcal{I}_{b_1 z_1} = \{(u_1, u_2, u_3), (f_1 \oplus f_2)\}$. The color graph corresponds to the minimum number of coded transmissions to target the remaining users. Each coded transmission is represented by vertices having the same color (yellow) and can be sent by one CBS channel. In particular, both vertices $\{(u_4, f_3), (u_5, f_3)\}$ have the same color and correspond to transmitting file f_3 to both users via one CBS channel.

and $\mathcal{S}(P)$ is a set of independent sets in P. Thus, the joint CBS offloading and QoS guarantee optimization problem in F-RAN settings is then expressed as

$$\min_{\mathcal{I}_{b_n z_m}, \cdots, \mathcal{I}_{b_B z_Z}, P \in \mathcal{P}} \sum_{\mathcal{I}_{c_h} \in \mathcal{S}(P)} \frac{N}{\min\limits_{u_i \in \tau_{c_h}} r^{\mathrm{CBS}}_{u_i}} \tag{3.2a}$$

$$\text{s. t.} \begin{cases} \mathcal{I}_{b_n z_m} = \arg\max_{\kappa_{b_n z_m} \in \mathcal{H}_{b_n}} \sum_{u_i \in \tau_{b_n z_m}} \log_2(1 + \mathrm{SINR}^{u_i}_{b_n z_m}(\mathbf{P})), \\ \bigcup_{m=1}^{Z} \tau_{b_n z_m} \cap \bigcup_{m=1}^{Z} \tau_{b_{n'} z_m} = \phi, \ \forall \ (b_n, b_{n'}) \ \in \mathcal{B}, \\ \tau_{b_n z_m} = \left\{ u_i \in \mathcal{U} \ \middle| |\kappa_{b_n z_m} \cap \mathcal{W}_{u_i}| = 1 \ \& \ R_{b_n z_m} \leqslant R^{u_i}_{b_n z_m} \right\}, \\ T. \sum_{z_m \in \mathbf{Z}_{u_i}} R_{b_n z_m} \leqslant N, \ \forall \ u_i \in \mathcal{U}, \\ \bigcup_{m=1}^{Z} \kappa_{b_n z_m} \subseteq \mathcal{H}_{b_n}, \ \forall \ b_n \in \mathcal{B}, \\ R_{b_n z_m} \geq R_{\mathrm{th}}, \ \forall \ (b_n, z_m) \in \mathcal{B} \times \mathcal{Z}, \\ 0 \leq P_{b_n z_m} \leq P^{\max}_{b_n z_m}, (b_n, z_m) \in \mathcal{B} \times \mathcal{Z}, \\ \kappa_{b_n z_m}, \kappa_{c_h} \ \in \mathcal{P}(\mathcal{F}), (b_n, z_m) \in \mathcal{B} \times \mathcal{Z}, \end{cases} \tag{3.2b}$$

where the optimization is carried over the variables $\kappa_{b_n z_m}$, κ_{c_h}, $R_{b_n z_m}$ and $P_{b_n z_m}$. Second constraint states that the sets of targeted users in all RRBs in eRRH b_n and $b_{n'}$ are disjoint, where $b_n \neq b_{n'}$. In other words, the same user cannot be served from two different eRRHs. Third constraint states that all users belonging to the targeted sets $\tau_{b_n z_m}$ $\forall$ $b_n \in \mathcal{B}$ and $z_m \in \mathcal{Z}$ must receive an instantly decodable transmission. To do so, each RRB should adopt its transmission rate to satisfy the successful reception of all its assigned users. However, this adopted transmission rate may not be able to deliver the whole file/content to the assigned users, given the fixed duration T of the RRB. Thus, such assigned users need to be scheduled to different RRBs, denoted by $\mathbf{Z}_{u_i}$, in the same eRRHs so as to guarantee the successful download of their requested file, given the total number of RRBs in the eRRH's frame. In other words, user u_i will be scheduled to a set of RRBs $\mathbf{Z}_{u_i}$ as long as its wanted file had not delivered yet, i.e., the fourth constraint. If one RRB is capable of delivering such request to user u_i, this user will not be scheduled to any other RRBss in the same eRRH. Fifth constraint ensures that all files/requests to be combined using XOR and served by all RRBs Z in eRRH b_n are stored in its cache. Sixth constraint guarantees the minimum transmission rate R_{th} required to achieve the QoS requirements. Finally, the last constraint accounts for the fact that the transmission power of each RRB in each eRRH is bounded.

The objective function (3.2a) of the optimization problem (3.2) represents the outer CBS consumed time reduction problem, and the constraints (3.2b) represent the inner throughput maximization problem. Finding the global optimal solution to the joint optimization problem (3.2) requires a search over all the sets of targeted users, their possible ONC combinations, and the optimal power levels and the corresponding transmitted rates of each RRB and each CBS channel. As such, finding an optimal solution to (3.2) is challenging and not feasible. Furthermore, finding the global solution to only (3.2b), i.e., joint user-scheduling and power optimization problem is challenging, and it is NP-hard problem in general [154], [155]. In order to seek an efficient solution to (3.2), we propose to use a graph representation. As such, (3.2b) is shown to be equivalent to a maximum-weight independent

set problem over a novel FRAN-CLNC graph, and (3.2a) is minimum coloring problem over a color graph. In order to construct the FRAN-CLNC graph, the rest of this subsection introduces the used variables as follows.

Let $\mathcal{S}_{b_n}$ be the set of all possible associations between users and files that cached by eRRH b_n, i.e., $\mathcal{S}_{b_n} = \mathcal{U} \times \mathcal{H}_{b_n}$. Let φ_{u_i} and φ_{f_l} be a family of mappings from the set $\mathcal{S}_{b_n}$ to the set of users $\mathcal{U}$ and the set of files $\mathcal{H}_{b_n}$. The mathematical definition of these mappings given the element $y = (u_i, f_l) \in \mathcal{S}_{b_n}$ are $\varphi_{u_i}(y) = u_i$ and $\varphi_{f_l}(y) = f_l$. The set of all possible ONC file combinations $\mathcal{S}_{\text{IDNC},b_n}$ in eRRH b_n is created based on the fact that the associations s in $\mathcal{S}_{b_n}$ can be served by one XOR encoded file. Intuitively, s, s' in $\mathcal{S}_{b_n}$ are combined if the non-instant decodability condition (**CC**) not satisfied. The ONC file combinations set of the whole F-RAN system is simply the union of ONC file combinations of all eRRHs, i.e., $\mathcal{S} = \bigcup_{b_n \in \mathcal{B}} \mathcal{S}_{\text{IDNC},b_n}$. Let $\mathcal{A}_{b_n}$ be the set of all possible associations between RRBs, ONC file combinations, and achievable capacities of the b_n-th eRRH, i.e., $\mathcal{A}_{b_n} = \mathcal{Z} \times \mathcal{S}_{\text{IDNC},b_n} \times \mathcal{R}$. Let φ_{z_m}, $\varphi_{\mathtt{s}}$, and φ_r be a family of mappings from the set $\mathcal{A}_b$ to the set of RRBs $\mathcal{Z}$, the set of ONC file combinations $\mathcal{S}_{\text{IDNC},b_n}$, and the set of achievable capacities $\mathcal{R}$. Given the element $y = (z, \mathtt{s}, R) \in \mathcal{A}$, the mathematical definition of the mappings are $\varphi_z(y) = z$, $\varphi_{\mathtt{s}}(y) = \mathtt{s}$, i.e., $\mathtt{s} = (\kappa, \tau)$, and $\varphi_r(y) = R$, respectively. Similarly, the set of all possible associations of the whole system is simply the union of possible associations of all eRRHs, i.e., $\mathcal{A} = \bigcup_{b_n \in \mathcal{B}} \mathcal{A}_{b_n}$. Let $\mathcal{A}_{z_m}$ represent the associations relative to the z_m-th RRBs across all eRRHs, i.e., $y \in \mathcal{A}_{z_m} \Rightarrow \varphi_{z_m}(y) = z_m$.

3.4 CBS Offloading and QoS Guarantee: Joint Approach

3.4.1 FRAN-Cross-Layer NC Graph

The FRAN-CLNC graph is introduced as a cross-layer NC design that jointly optimizes NC combinations, users-eRRHs/RRBs assignments, and transmission power of the RRBs in the downlink F-RAN setup. The FRAN-

CLNC graph, denoted by $\mathcal{G}(\mathcal{V}, \mathcal{E})$, is constructed by first generating all the Z RRB subgraphs, in which the vertex set of the FRAN-CLNC graph is the union of vertices of all the subgraphs, i.e., $\mathcal{V} = \bigcup_{z_m \in \mathcal{Z}} \mathcal{V}^{z_m}$.

Let the subgraph of the z_m-th RRB in the network be denoted by $\mathcal{G}^{z_m}(\mathcal{V}^{z_m}, \mathcal{E}^{z_m})$ where $\mathcal{V}^{z_m}$ and $\mathcal{E}^{z_m}$ refer to the set of vertices and edges of this subgraph, respectively. In order to represent the ONC file combinations $\mathtt{s} = (\tau, \kappa)$, the transmission rate R of each eRRH in the RRB z_m, a feasible schedule $\mathbf{S}$ is generated. The generation of $\mathbf{S}$ is done by merging all possible associations $s \in \mathcal{A}_{z_m}$ for the different eRRHs under the constraint that the same user cannot appear more than once in $\mathbf{S}$. Therefore, each vertex $v \in \mathcal{V}^{z_m}$ associated with $\mathbf{S} \in \mathcal{S}_{\text{fs},z_m}$ satisfies $\varphi_{b_n}(s) \neq \varphi_{b_n}(s')$ and $\varphi_{u_i}(s) \neq \varphi_{u_i}(s')$ for all $s \neq s' \in \mathbf{S}$, where $\mathcal{S}_{\text{fs},z_m}$ is is the set of all possible feasible schedules corresponding to the z_m-th RRB. The fundamental step of merging the set of all possible associations $s \in \mathcal{A}_{z_m}$ relies on the fact that users in these associations can be served simultaneously. Therefore, each vertex $v \in \mathcal{V}^{z_m}$ representing the association $\mathbf{S}$ satisfies the following.

- Two associations $s, s' \in \mathbf{S}$ representing the same RRB z and the same eRRH b_n have the same transmission rate, i.e., $\varphi_r(s) = \varphi_r(s')$.
- Two associations $s, s' \in \mathbf{S}$ representing different eRRHs have different set of targeted user(s), i.e., $\tau \cap \tau' = \varnothing$.

In mathematical terms, each vertex $v \in \mathcal{V}^{z_m}$ representing the association $\mathbf{S}$ satisfies the following local conditions (LC).

- **LC1:** $\forall\ (s, s') \in \mathbf{S}$, $\varphi_{b_n}(s) = \varphi_{b_n}(s')$ and $\varphi_r(s) = \varphi_r(s')$.
- **LC2:** $\forall\ (s, s') \in \mathbf{S}$, $\varphi_{b_n}(s) = \varphi_{b_n}(s')$ and $\tau \cap \tau' = \varnothing$.

Clearly, the above vertice's configurations point out the fact that each vertex in the RRB subgraph reflects partial contribution to the network. In other words, it totally represents that specific RRB, and each pair of vertcies v and $v' \in \mathcal{V}^{z_m}$ is connected by a conflict edge.

Given the configuration of the vertices and the edges between them within the same RRB subgraph, in the following, we will describe the remaining edges corresponding to different RRB subgraphs. Two vertices v and v' corresponding to $\mathbf{S}$ and $\mathbf{S}'$ and belonging to two different subgraphs $\mathcal{G}^{z_m}$ and $\mathcal{G}^{z_{m'}}$, respectively, are linked by a conflict edge if the resulting combination satisfies one of the following events.

- At least one user appears in both $s \in \mathbf{S}$ and $s' \in \mathbf{S}'$, i.e., $\tau \cap \tau' \neq \varnothing$.
- The transmission rate of any RRB to the targeted user(s) τ is sufficient to successfully download their encoded file κ from only one RRB.

In mathematical terms, two vertices v and v' are linked by a conflict edge if the associations they represent satisfy one of the following general conditions (GC).

- **GC1:** $\forall\ (s, s') \in \mathbf{S} \times \mathbf{S}'$, $\tau \cap \tau' \neq \varnothing$ and $\varphi_{b_n}(s) \neq \varphi_{b_n}(s')$.
- **GC2:** $\forall\ (s, s') \in \mathbf{S} \times \mathbf{S}'$, $\tau = \tau'$ and $\varphi_b(s) = \varphi_b(s')$ and $T.R_{b_n z_m} \geq N$.

Assuming that the power distribution $\mathbf{P}$ of all RRBs in the system will be computed later, a vertex v representing the associations $\mathbf{S} \in \mathcal{S}_{\text{fs},z_m}$ has a secondary weight that reflects the partial throughput contribution to the network, i.e., the secondary weight of v can be expressed as

$$w_s(v) = \sum_{s \in \mathbf{S}} \min_{\varphi_{u_i}(s) \in \tau_{\varphi_{b_n}(s), \varphi_{z_m}(s)}} \log_2(1 + \text{SINR}^{\varphi_{u_i}(s)}_{\varphi_{b_n}(s)\varphi_{z_m}(s)}(\mathbf{P})). \tag{3.3}$$

Now a vertex v representing the associations $\mathbf{S}$ has a primary weight that reflects the degree of priority of users in $\mathbf{S}$ to be served from RRBs/eRRHs, i.e., the primary weight of v can be expressed as

$$w_p(v) = \sum_{s \in \mathbf{S}} \frac{N}{r^{\text{CBS}}_{u_i}}. \tag{3.4}$$

In particular, a larger primary value of $w_p(v)$ offers a larger download time from the CBS (lower rate from the CBS) of each represented file in $\mathbf{S}$. The weighting design of each vertex can be summarized as follows. A larger value

of $w_s(v)$ and $w_p(v)$ in (3.3) and (3.4), respectively, leads to a cross-layer encoded transmission that offers a better performance of throughput and CBS offloading. In particular, the higher the primary weight (3.4), the larger the physical-layer resource consumed time of the CBS channels. This yields to a smart scheduling of users representing the corresponding vertex with a higher secondary weight (3.3), which in turn maximizes the throughput. Therefore, any maximal independent set in FRAN-CLNC graph represents a set of coded transmissions that satisfies the following criterion: 1) the represented users by vertices in the maximal independent set are experiencing the lowest rates from the CBS, 2) all targeted users satisfy the required minimum rate, and can decode the requested file from the transmission schedule of all RRBs and eRRHs, 3) users are listening to multiple RRBs in the same eRRH as long as the successful download of their requited files is not guaranteed, and 4) each RRB identified by the vertices in a maximal independent set adopts the transmission rate identified by the vertex. Such rate is less than or equal to the channel capacities of all users served by that RRB.

Given the optimal power allocation $\mathbf{P}$ and the weighting design of each vertex, the following theorem characterizes the solution of the inner throughput maximization problem by allocating users to the RRBs of each eRRH and adapting the power levels, such that the user's QoS is guaranteed.

Theorem 3.1. *The inner throughput maximization problem (3.2b) is equivalent to a maximum-weight independent set problem over the FRAN-CLNC graph, and can be written as*

$$\arg\max_{\mathtt{I}\in\mathbf{I}} \sum_{v\in\mathbf{I}} w_s(v), \tag{3.5}$$

where $\mathtt{I}$ *is the maximum-weight independent set in the FRAN-CLNC graph,* $\mathbf{I}$ *is the set of all possible maximal independent sets of degree* Z*, and the weight of a vertex* $v \in \mathcal{V}$ *is defined in (3.3).*

Proof. The proof of this theorem can follow the same steps that used in investigating Theorem 2 in [109]. The main difference is that we multiplex users that have cached files in eRRHs to the RRBs instead of assigning one

user in each RRB. Interested readers can read the proof in Appendix A. □

The set of targeted users and the file combination of the z_m-th RRB across all eRRHs is obtained by combining the vertices of the maximal-weight independent set $\mathcal{I}_{z_m}$ corresponding to the RRB subgraph $\mathcal{G}^{z_m}$, in which $\mathcal{I}_{z_m}$ is the union of all maximal-weight independent sets for that RRB in each eRRH, i.e., $\mathcal{I}_{z_m} = \bigcup_{n=1}^{B} \mathcal{I}_{b_n z_m}$. Therefore, $\mathtt{I} = \left(\bigcup_{n=1}^{B} \bigcup_{m=1}^{Z} \mathcal{I}_{b_n z_m}\right)$.

3.4.2 Greedy Algorithm

It can be noted that the optimization problem (3.2) is formulated using ONC graph. The global solution is, therefore, equivalent to a maximum-wight independent set over FRAN-CLNC graph and a minimum weight coloring over a color graph, which is an NP-hard problem [74], [156], and so is the problem (3.2). However, it is established that it can be near optimally solved with a reduced complexity as compared the $O(|\mathcal{V}|^2.2^{|\mathcal{V}|} + \mathcal{V}\ \log(\mathcal{V}))$ naive exhaustive search methods, e.g., the algorithms in [149], [150], and [156]. The maximum-weight independent set and the color problems can be solved sequentially in a linear time with their sizes in simple heuristic solution as explained in this subsection. While the proposed solution is not necessarily optimum, it works well for solving (3.2) for one particular transmission.

The joint cross-layer CBS offloading and QoS guarantee optimization algorithm is executed at the CBS at each transmission and broken into two phases as follows.

Phase I: This phase solves the inner throughput maximization problem (3.2b) in two steps, the FRAN-CLNC graph design step and the maximum-weight search step.

First step: The FRAN-CLNC graph can be designed as follows. A subgraph $\mathcal{G}^{z_m}$ is generated for each RRB $z_m \in \mathcal{Z}$ using **LC1** and **LC2**. Afterwards, for each RRB $z_m \in \mathcal{Z}$ across all eRRHs, a vertex $v \in \mathcal{G}^{z_m}$ corresponding to the feasible schedule $\mathbf{S} \in \mathcal{S}_{\text{fs},z_m}$ is generated for all possible associations. The optimal PLs of each association are, then, computed by solving the optimization problem (2.10). The vertex in the subgraph $\mathcal{G}^{z_m}$ is generated by

appending the computed PLs and the corresponding rates to that vertex as shown in Section 3.4.1. The same steps above are repeated for all RRBs $z_m \in \mathcal{Z}$. The FRAN-CLNC graph is, then, designed by merging all subgraphs and adding connections according to **GC1** and **GC2**.

Second step: The algorithm finds the maximum independent set `I` among all the maximal independent sets `I` in FRAN-CLNC graph. Each iteration of finding `I` is implemented as follows. The algorithm computes the secondary weight of all generated vertices using (3.3). Now let the primary weights of these generated vertices are computed using (3.4). The vertex with the maximum-primary weight v^* is labeled and find its corresponding vertex that has the maximum-secondary weight v^{**} among all other corresponding vertices. The selected vertex v^{**} is, then, added to `I`, i.e., `I` is assumed to be initially empty. Afterwards, FRAN-CLNC $\mathcal{G}$ graph is updated by removing the selected vertices v^{**} and its adjacent vertices. As such, the next selected vertex is not in coding nor transmission conflict with the already selected ones in `I`. The process continues until no more vertices exist in FRAN-CLNC graph $\mathcal{G}$. Since each RRB subgraph contributes by a single vertex, the number of vertices in `I` is Z.

Phase II: Second phase solves the outer CBS consumed time reduction problem (3.2a) by serving users that are unserved by the eRRHs, so that they can be served by the least consumed time of CBS physical resources. Mathematically, this phase serves $\left(\mathcal{U} \setminus \sum_{b_n \in B} \sum_{z_m \in Z} \tau_{b_n z_m}\right)$ users. Therefore, in each transmission, phase II removes all vertices from the ONC graph that represent these served users by the eRRHs, i.e., ONC graph\`I`. The remaining vertices are, then, sorted in a list $\mathbf{M}$ in descending order based on their corresponding weights, in which the weight of each remaining vertex $v_{u_i,f_l} = \frac{N}{r_{u_i}^{\text{CBS}}}$. Given the weight of each remaining vertex, a greedy coloring scheme is performed as follows. First, it picks the highest weight vertex v^* and colors it with a certain color c_1. Then, it scans all vertices in that list, searching in each step for any vertex not adjacent to v^*, and coloring it with c_1. The scheme continues until no more vertices are not adjacent to all previously colored vertices in c_1. The scheme, then, removes the

selected vertices having color c_1, and repeating the above steps again with another color c_2. This process continues until no vertices have not colored yet in $\mathbf{M}$. The key idea in this scheme is to color the vertices of users satisfying the NC conditions and experiencing the lowest rates by the same color. Each group of same-color vertices will thus be served in one CBS orthogonal channel. The overall two-phase algorithm to the problem (3.2) is summarized in Algorithm 4

3.4.3 Computational Complexity of Joint Approach

In each transmission of FRAN-CLNC, the CBS requires to compute the complexity of phase I and phase II in Algorithm 4.

Phase I Complexity: Each eRRH b requires to find all its file combination possibilities $|\mathcal{S}_{\text{IDNC},b_n}|$ that satisfy **CC** and relate these combinations to its RRBs, which give the total file combinations in the system $\mathcal{S}$. Since each feasible schedule in the network may involve associating different users to different eRRHs (i.e., satisfy **LC1** and **LC2**), the number of feasible schedules can be computed by permutating $\mathcal{S}$ among all eRRHs. Thus, the total number of these schedules is ${}^{\mathcal{S}}P_B$, which is represented by the vertices in the RRB subgraph. Further, calculating the optimal power needs to execute the optimization problem (2.10) on each vertex. Thus, the complexity of obtaining a RRB subgraph is ${}^{\mathcal{S}}P_B c(A)$, where $c(A)$ is the computational complexity of solving the power allocation problem (2.10) with A variables. The total number of RRB subgraphs is Z, thus introducing an FRAN-CLNC graph design complexity of ${}^{\mathcal{S}}P_B c(A)Z$. To connect all the generated vertices ${}^{\mathcal{S}}P_B Z$ in FRAN-CLNC graph using **GC1** and **GC2**, the algorithm needs $({}^{\mathcal{S}}P_B Z)^2$. Now, solving the maximum-weight independent set in FRAN-CLNC graph requires weight calculations and vertex search method of $f({}^{\mathcal{S}}P_B Z)$, where $f(x)$ is the complexity of solving the maximum-independent set in a graph containing x vertices. The overall phase I complexity is then $O({}^{\mathcal{S}}P_B Z[c(A) + {}^{\mathcal{S}}P_B Z] + f({}^{\mathcal{S}}P_B Z))$.

Phase II complexity: Given the solution to the inner problem, the number of remaining users having vertices in ONC graph is $V = \text{ONC graph}\backslash\mathtt{I}$.

Algorithm 4: Joint Cross-Layer CBS Offloading and QoS Guarantee Optimization Algorithm

Data: $\mathcal{U}, \mathcal{F}, \mathcal{B}, \mathcal{Z}, \mathcal{H}_{b_n}, \mathcal{H}_{u_i}, \mathcal{W}_{u_i}$, h_{u_i} and $h^{u_i}_{b_n z_m}, (u_i, f_l) \in \mathcal{U} \times \mathcal{F}$, $(b_n, z_m, \mathtt{s}) \in \mathcal{B} \times \mathcal{Z} \times \mathcal{S}$;
Construct the ONC graph using Section 3.3.2.
Phase I: Solve the inner throughput maximization problem (3.2b).

- Initialize the maximum-weight independent set $\mathtt{I} = \phi$.
  for $z_m = \{z_1, z_2, \cdots, z_Z\}$ **do**
  Initialize the cluster $\mathcal{G}^{z_m} = \phi$.
  for *each* $\mathbf{S} = \{s_1, s_2, \cdots, s_{\mathtt{S}}\} \in \mathcal{S}_{fs,z_m}$ **do**
  Solve (2.10) to compute the optimal power allocations $\mathbf{P} = \{(p^*_{b_1 z_m}, p^*_{b_2 z_m}, \cdots, p^*_{b_B z_m})\}$.
  Create: $v = \{\{(s_1, r^*_{b_1 z_m}, p^*_{b_1 z_m}), \cdots, (s_{|\tau_{b_1 z_m}(\kappa_{b_1 z_m})|}, r^*_{b_1 z_m}, p^*_{b_1 z_m})\}$, $\cdots$, $\{(s_{b_B}, r^*_{b_B z_m}, p^*_{b_B z_m}), \cdots, (s_{|\tau_{b_B z_m}(\kappa_{b_B z_m})|}, r^*_{b_B z_m}, p^*_{b_B z_m})\}$
  Calculate $ws(v)$ using (3.3).
  Set $\mathcal{G}^{z_m} = \mathcal{G}^{z_m} \cup \{v\}$.
  end
  end
- Set $\mathcal{G} = \bigcup_{z_m \in \mathcal{Z}} \mathcal{G}^{z_m}$ and connect vertices of $\mathcal{G}$ using **GC1** and **GC2**
- Solve the maximum-wight independent set problem in $\mathcal{G}$ to find $\mathtt{I}$ as follows
  while $\mathcal{G} \neq \phi$ **do**
  Select the maximum-primary weight $v^* = \max_{v \in \mathcal{G}}\{w_p(v)\}$ and find its corresponding vertex with the maximum-secondary weight $v^{**} = \max_{v \in \mathcal{G}}\{w_s(v)\}$.
  Set $\mathtt{I} = \mathtt{I} \cup v^{**}$ and $\mathcal{G} = \mathcal{G}(v^{**})$ and continue only with vertices not linked to v^{**} in $\mathcal{G}$.
  end

Phase II: Solve the outer CBS consumed time reduction problem (3.2a).

- Set $h = 1$.
- $\forall$ vertex $v_{u_i f_l} \in$ ONC graph$\backslash\mathtt{I}$: calculate $w(v_{u_i f_l}) = \frac{N}{r^{\mathrm{CBS}}_{u_i}}$ and sort them in a list $\mathbf{M}$ in descending order.
  while $\mathbf{M} \neq \phi$ **do**
  Initialize the independent set $\mathcal{I}_{c_h} = \phi$.
  Select the first vertex v^* and color it with C_c.
  Set $\mathcal{I}_c = \mathcal{I}_{c_h} \cup v^*$.
  Continue coloring all vertices not adjacent to v^* with c_h and adding them to $\mathcal{I}_{c_h}$ and set $\mathbf{M} = \mathbf{M} \setminus \mathcal{I}_{c_h}$.
  Set $h = h + 1$.
  end

Result: $\mathtt{I}$ and $\mathcal{I}_1, \mathcal{I}_2, \cdots$

Thus, the greedy coloring step requires to calculate the weights of vertices V and sort them, which has a computational complexity of $O(V + (V \log(V)))$ [30].

Therefore, the overall per-transmission computational complexity of FRAN-CLNC is $O({}^{\mathcal{S}}P_B Z[c(A) + {}^{\mathcal{S}}P_B Z] + f({}^{\mathcal{S}}P_B Z) + V[1 + \log(V)])$. In the uncoded case, the per-transmission complexity reduces to $O({}^{U}P_B Z[c(A) + {}^{U}P_B Z] + f({}^{U}P_B Z) + V)$. We note that the computational complexity of FRAN-CLNC is sensitive to the number of NC possibilities. Consequently, FRAN-CLNC has higher complexity which justifies our resorting to the iterative simple approach. Finally, the transmission overhead that the CBS needs to execute Algorithm 4 can be analyzed as follows. Since the eRRHs report to the CBS about the cached and the non-downloaded files at the users, all side information is known to the CBS. Clearly, this needs a few bytes of control signaling. Thus, compared to the standard LTE uncoded solution, no additional overhead is needed for the coded users-eRRHs/CBS scheduling scheme. However, the coded solution may need overhead for the headers of each eRRHs/CBS encoded transmission (to inform users of the XORed files in any transmission), which is negligible in size compared to the entire file's size.

3.5 CBS Offloading and QoS Guarantee: Iterative Approach

This section recommends that instead of solving (3.2b) jointly which requires high computational complexity, one way to mitigate such high complexity is to solve (3.2b) iteratively. In particular, (3.2b) is solved iteratively in such a way that for a given fixed transmit power of each RRB, the throughput optimization problem is solved as a coordinated scheduling problem. Afterward, the power allocation problem is solved numerically for the resulting NC and user-RRB/eRRH schedule. The process iterates between these two steps until convergence. Given the iterative solution to the inner problem, the outer CBS time reduction problem (3.2a) is solved

using greedy coloring scheme as in the joint approach. Towards that goal, this section first addresses the optimization problem (3.2) as a coordinated scheduling problem only, and can be written as

$$\min_{\substack{\mathcal{I}_{b_n z_m}, \cdots, \mathcal{I}_{b_B z_Z} \\ P \in \mathcal{P}}} \sum_{\mathcal{I}_c \in \mathcal{S}(P)} \frac{N}{\min\limits_{u \in \tau_c} r_{u_i}} \tag{3.6a}$$

$$\text{s. t.} \begin{cases} \mathcal{I}_{b_n z_m} = \arg\max_{\kappa_{b_n z_m} \in \mathcal{H}_b} \sum_{u \in \tau_{bz}} R_{b_n z_m} \\ \bigcup_{m=1}^{Z} \tau_{b_n z_m} \cap \bigcup_{m=1}^{Z} \tau_{b_{n'} z_m} = \phi, \forall (b_n, b_{n'}) \ \in \mathcal{B}, \\ \tau_{b_n z_m} = \left\{ u_i \in \mathcal{U} \ \middle| |\kappa_{b_n z_m} \cap \mathcal{W}_{u_i}| = 1 \ \& \ R_{b_n z_m} \leqslant R_{b_n z_m}^{u_i} \right\}, \\ T. \sum_{z_m \in \mathbf{Z}_u u_i} R_{b_n z_m} \leqslant N, \ \forall u_i \in \mathcal{U}, \\ \bigcup_{m=1}^{Z} \kappa_{b_n z_m} \subseteq \mathcal{H}_{b_n}, \ \forall b_n \in \mathcal{B}, \\ R_{b_n z_m} \geq R_{\text{th}}, \ \forall (b_n, z_m) \in (\mathcal{B}, \mathcal{Z}). \end{cases} \tag{3.6b}$$

The optimization is carried over the variable $R_{b_n z_m}$, $\kappa_{b_n z_m}$. On the other hand, for the resulting NC and user-RRB/eRRH schedule, the inner optimization problem (3.2b) can be considered as a power allocation step and simplifies on a per RRB basis. For each RRB z_m, the optimization problem (3.2b) can be written as in (2.10).

Coordinated Scheduling Graph

Similar to our preliminary work [128] the scheduling graph is formed by generating a vertex $v_{b_n,z_m,u_i,f_l,r}$ for each $b_n \in \mathcal{B}$, each $z_m \in \mathcal{Z}$, for every user u_i requests a file f_l, and for each achievable rate for that user $r \geq R_{th}$. The configuration of the set of edges in the scheduling graph is divided into coding and transmission conflict edges. Two vertices $v_{b_n,z_m,u_i,f_l,r}$ and $v_{b_n,z_m,u_{u'},f_{l'},r'}$ representing the same RRB z_m and the same eRRH b_n are adjacent by a coding conflict link if the transmission is non decodable to one of them (satisfy the condition in Section 3.3.2 or the transmission rate is different (i.e., $r \neq r'$). Similarly, two vertices $v_{b_n,z_m,u_i,f_l,r}$ and $v_{b_{n'},z_{m'},u_{i'},f_{l'},r'}$ representing different RRB z_m and different eRRH b_n are adjacent by a transmission conflict link if one of these conditions is true:

Algorithm 5: Coordinated Scheduling Algorithm

Data: $\mathcal{U}, \mathcal{F}, \mathcal{B}, \mathcal{H}_b, \mathcal{H}_{u_i}, \mathcal{W}_{u_i}, P_{b_n z_m}, \mathcal{R}$ and $h^{u_i}_{b_n z_m}$, $(b_n, z_m, u_i, f_l) \in \mathcal{B} \times \mathcal{Z} \times \mathcal{U} \times \mathcal{F}$.
Initialize the maximum-weight independent set $\mathtt{I} = \phi$.
Construct $\mathcal{G}$ and solve the maximum-weight independent set problem over $\mathcal{G}$ as follows:
while $\mathcal{G} \neq \phi$ **do**
  $\forall$ vertex $v \in \mathcal{G}$: calculate the primary weight $w_p(v)$ and the secondary weight $\tilde{w}_s(v)$ as illustrated in Section 3.5.
  Select the maximum-primary weight $v^* = \max_{v \in \mathcal{G}}\{w_p(v)\}$ and find its corresponding vertex with the maximum-secondary weight $v^{**} = \max_{v \in \mathcal{G}}\{\tilde{w}_s(v)\}$.
  Set $\mathtt{I} = \mathtt{I} \cup v^{**}$ and continue only with the vertices not adjacent to v^{**}
end
Result: I.

- $u_i = u_{i'}$ and $(b_n, z_m) = (b_{n'}, z_{z'})$, we have $T.R_{b_n z_m} \geqslant N$. This condition schedules user to multiple RRBs in the same eRRH as long as its requested file has not received completely.

- $u_i = u_{i'}$ and $(b_n, z_m) \neq (b_{n'}, z_{m'})$. This condition insists that the same user cannot be scheduled to different eRRHs in all RRBs.

Consider the primary weight of each vertex v_{b_n, z_m, u_i, f_l} is defined as $w_p(v) = \frac{N}{r^{\text{CBS}}_{u_i}}$. Thus, the vertex' weight becomes large when its file download time from the CBS is large, thus representing a lower rate from the CBS. Given the above configuration of the scheduling graph and the weighting design of each vertex, it can be established that any maximal independent set in that graph satisfies that all users have vertices in the selected set receive an instantly decodable transmission that meets QoS requirements, and then can decode a new file.

The following theorem characterizes the solution of allocating users, satisfying the required minimum rate and having the worst download rates from the CBS, to the RRBs across all eRRHs, such that the received throughput is maximized.

Theorem 3.2. *The coordinated scheduling problem (3.6) is equivalent to a maximum-weight independent set problem over the scheduling graph, wherein the secondary weight of a vertex $v_{b_n,z_m,u_i,f_l,r}$ is given by*

$$w(v) = r. \tag{3.7}$$

The set of targeted users and the file combination of the z_m-th RRB in the b-th eRRH is obtained by combining the vertices of the maximum-weight independent set `I` *in the coordinated scheduling graph.*

Heuristic Solution

Maximum-weight independent set problems are NP-hard problems where the complexity of solving these problems optimally needs $|\mathcal{V}|^2.2^{|\mathcal{V}|}$ cost of an exhaustive search solvers where $\mathcal{V}$ is the set of vertices of graph $\mathcal{G}$. However, it is established that it can be efficiently solved with a reduced complexity $\alpha^{|\mathcal{V}|}$ where α is the complexity constant [29, 30]. Even though such complexity is less than an exhaustive search, it is still high for moderate size networks. Thus, the maximum-weight independent set problem can be solved here in a linear time with its size in a simple heuristic solution as in Algorithm 5. Let $w(v)$ the raw weight of vertex v in the scheduling graph as expressed in Theorem 3.2. The modified weight $\tilde{w}_v$ of vertex v can be defined as

$$\tilde{w}_s(v) = w(v) \sum_{v' \in \mathcal{V}_v} w(v'), \tag{3.8}$$

where $\mathcal{V}_v$ is the set of vertices not connected to vertex v by neither coding nor transmission conflict edges. The appropriate design of the weights shows that $\tilde{w}_s(v)$ reflects the contribution of the vertex to the network as it has a large raw weight and non-adjacent to a large number of vertices induced by users with high raw weight.

3.6 Numerical Results

In this section, we present some numerical results that compare the performance of both proposed schemes with the baseline algorithms in [109],

[126] and [127]. We consider a downlink F-RAN system where the positions of 3 eRRHs are fixed and users are distributed randomly within a hexagonal cell of radius 500m. The channel gains are generated according to SUI-Terrain type B. The channels are assumed to be perfectly estimated. The noise power and the maximum' RRB/CBS power are set to $\sigma^2 = -168.60$ dBm/Hz and $P_{b_n z_m}^{\max} = P = -42.60$ dBm/Hz, respectively. The bandwidth is 10 MHz, the shadowing variance is 6 dB, and the eRRH caching ratio μ is set to 0.6. To evaluate the performance of the proposed solutions at different thresholds, we simulate various scenarios: number of users, number of popular files, and number of RRBs per eRRH's frame. In these evaluations, the average throughput and the CBS consumed time are computed over a certain number of iterations, and the average values are presented. For the sake of comparison, we implement following schemes.

- **Upper-Layer (denoted by UL):** This scheme jointly optimizes the selection of ONC file combinations for each RRB in each eRRH and each CBS' channel without considering the physical-layer rates of users. This scheme is proposed in [126].
- **Uncoded:** The transmission strategy in this scheme is performed irrespective of the available information at the network layer, i.e., prior download files. The user-RRB association in this scheme is proposed in [109].
- **Cross-Layer QoS-unaware (denoted by CL):** In this scheme, all users experiencing low rates from the CBS are multiplexed based on their requests and targeted from RRBs/eRRHs randomly. This scheme is proposed in [127].
- **Proposed FRAN-CLNC:** This scheme is described in Section 3.4.
- **Proposed Iterative:** This scheme is described in Section 3.5.

In Figure 3.4, we show the total CBS consumed time and the average throughput versus the number of users U for $M = 10$ files, each with a size of 10 MB, and 3 RRBs in each eRRH's frame. As we can see from Figure

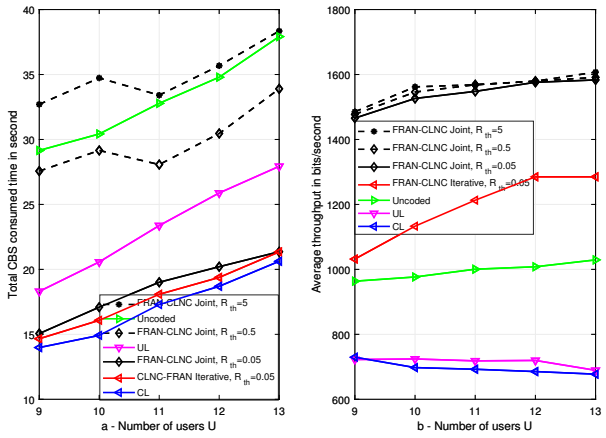


Figure 3.4: Total CBS consumed time and average throughput in bits/second versus number of users U.

3.4-b, both proposed schemes offer improved performance in terms of maximizing the number of received bits as compared to the other schemes. This significant gain is due to the role of the joint FRAN-CLNC and the iterative one that smartly multiplexes users, selects the best transmission rate of each RRB, and controls the power for interference mitigation. In particular, the performance of the CL scheme decreases as the number of users increases. This is due to the fact that it distributes users to RRBs randomly and uses the minimum rate of all assigned users to each RRB. As this number increases, the minimum rate further decreases thus drastically affecting the average throughput of the scheme. On the other hand, the uncoded scheme focuses only on the highest rate at the expense of transmitting at most one file to a single user from each RRB. As the number of users increases, its performance improves slightly. This can be explained by the fact that when the number of users goes beyond the total number of RRBs in the system, i.e., ZB, the scheme does not gain from the increased number of users and benefits only from the diversity of the set of users. Clearly speaking, despite the overloading of users experiencing low rates (their achievable capacities from RRBs are less than the threshold $R_{th} = 0.05$) on the CBS, Figure 3.4-a

marks an appreciable performance improvement by 90% - 93% on average of the CL scheme. However, the gains are more significant when $U = 12, 13$, which are around 97% and 96%, respectively. This is because that proposed schemes focus mainly on the critical set of users who experience low rates from CBS to be served by the eRRHs.

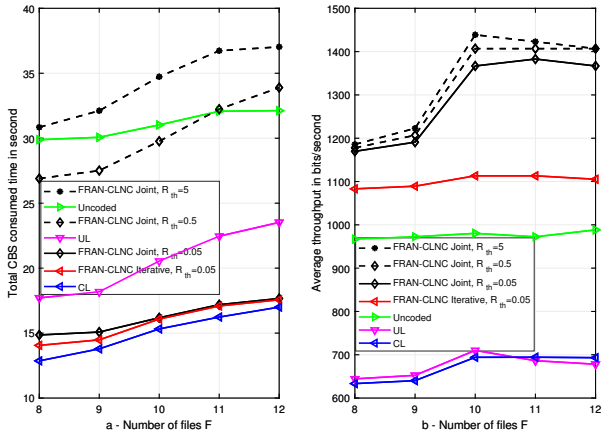


Figure 3.5: Total CBS consumed time and average throughput in bits/second versus number of files F.

In Figure 3.5, we compare the total CBS consumed time and the average throughput versus the number of files F each with a size of 10 MB, $U = 10$ users, and 3 RRBs in each eRRH's frame. Similar to what we have discussed in Figure 3.4, the joint and the iterative schemes show significant gain on the throughput maximization regime in Figure 3.5-b compared to all schemes. Also, both proposed schemes show a certain degradation in the CBS consumed time regime in Figure 3.5-a as compared to the CL scheme. It is worth remarking that the increase in the number of files increases the CBS consumed time for all coded schemes. This is due to the fact that the number of conflict edges between vertices increases as the number of files increases. This results in a smaller ONC opportunities to mix files, thus leading to a large number of colors to cover the CBS graph. For the same reason, the received throughput of all coded schemes improves slightly.

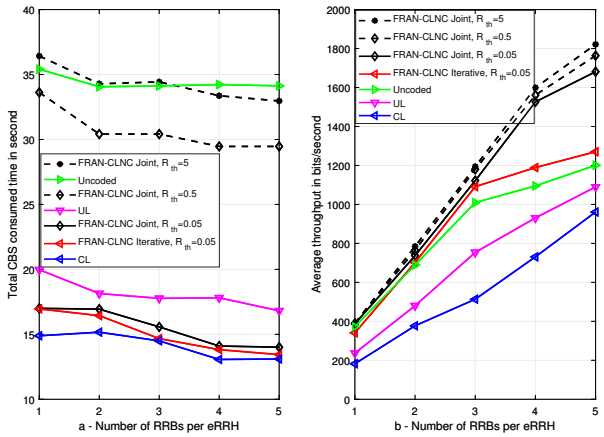


Figure 3.6: Total CBS consumed time and average throughput in bits/second versus number of RRBs Z.

In Figure 3.6, we present the total CBS consumed time and the average throughput versus the number of RRBs Z for $U = 10$ users, $F = 8$ files, with file's size of 10 MB. Clearly, figure 3.6-b shows that the received throughout of all schemes are increased with the increase in the number of RRBs. In fact, all schemes in agreement with serving the same set of users in different RRBs of the same eRRH. The proposed solutions, however, benefit from mixing the user's flows and selecting the best transmission rates which makes the average throughput increase highly noticeable compared to the other solutions. On the other hand, Figure 3.6-a depicts that the CBS consumed time of all solutions decreases with the increase in the number of RRBs. This is due to the fact that as the number of RRBs increases, more users are targeted, which in turn reduces the overload on the CBS to target the remaining users.

3.7 Chapter Summery

In this chapter, we have introduced a cross-layer optimization framework for the immediate file delivery phase with the goal of maximizing the CBS of-

floading while guaranteeing users QoS. Presented numerical results revealed that both proposed schemes achieve significant gains in users throughput compared to the existing solutions. Compared to the QoS un-aware algorithm, our proposed schemes have a certain degradation in CBS consumed time.

Chapter 4

Completion Time Minimization in F-RANs using D2D Communications and Rate-Aware NC

In Chapter 3, we studied the CBS offloading and QoS guarantee problem in F-RAN setting. However, the performance of F-RAN can be further improved by implementing D2D communications. Therefore, we consider in this chapter a more practical 5G mobile network of D2D-aided F-RAN that integrates both F-RAN and D2D communications. In particular, we develop a framework that exploits the cached files at eRRHs, their transmission rates/powers, and previously downloaded files by the different users to deliver the requesting files from both the eRRHs and transmitting users to the requesting users with a minimum completion time in D2D-aided FRAN. This chapter's organization is given as follows. The accomplished works and research contributions are summarized in Section 4.1. The C-RAN and cross-layer NC models are discussed in Section 4.2. The completion time minimization metric is detailed in Section 4.3 and its problem formulation and decorations is provided in Section 4.4. In Section 4.5 and Section 4.6, we develop joint and iterative scheduling and power adaptation approaches. In Section 4.7, we present some selected numerical results, and in Section 4.8, we provide some concluding remarks.

4.1 Accomplished Works and Research Contributions

The contributions of this chapter are summarized as follows. For a D2D-aided F-RAN, we develop a framework where eRRHs and users collaborate to minimize the completion time. In particular, given the intractability of solving the completion time minimization problem over all possible future NC decisions, we reformulate the problem at each transmission with the constraints on user scheduling, their limited coverage zones, transmission rates, maximum power allocations, and QoS rate guarantee. By analyzing the problem, we decompose it into two subproblems. The first subproblem aims to obtain the possible completion time in eRRHs transmissions through minimizing the transmission time. To solve it, we design an interference-aware IDNC (IA-IDNC) graph that efficiently solves the user scheduling and power allocation problem jointly under the completion time constraints. Based on this, the transmission time achieved by eRRHs is revealed for solving the second subproblem. Then, we introduce a new D2D conflict graph to heuristically solve the second subproblem, i.e., maximizing the number of users that can be scheduled on D2D links. The aforementioned graph-based solutions of the corresponding subproblems will be referred as *joint approach*. Since the IA-IDNC graph in the joint approach grows fast with the NC combinations in large network size, we propose an alternative and efficient low-complexity coordinated scheduling approach that solves the completion time problem using graph theoretic method. We compare our proposed schemes with the existing coded and uncoded (i.e., without NC) schemes. Selected numerical results demonstrate that the proposed schemes can effectively improve completion time performance.

4.2 System Model

Network Model Overview

We consider a D2D-aided F-RAN system, shown in Figure 4.1, that consists of one CBS, B single antenna eRRHs, and U users. The sets of eRRHs

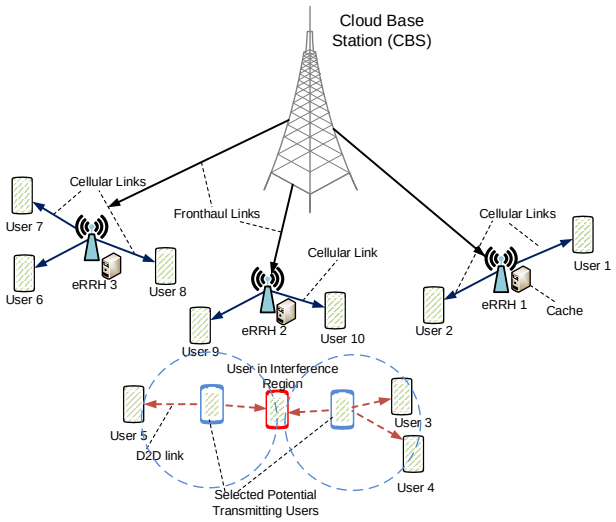


Figure 4.1: Illustration of the D2D-aided F-RAN model with 13 users, 3 eRRHs and 1 CBS.

and users are denoted by $\mathcal{B} = \{b_1, b_2, \cdots, b_B\}$, $\mathcal{U} = \{u_1, u_2, \cdots, u_U\}$, respectively. The CBS is responsible for making the NC decisions, power allocation, delivering the instructions to eRRHs and transmitting users for executions. It also communicates with eRRHs through fronthaul links. Since users are allowed to transmit at a certain amount of power, each device has limited coverage zone, denoted by $\mathcal{C}_{u_i}$, which represents the service area of the u_i-th user to transmit data within a circle of radius R. Note that user and device are used interchangeably throughout this chapter. The set of devices within the transmission range of the u_i-th device is defined by $\mathcal{C}_{u_i} = \{u_j \in \mathcal{U} | d^{d2d}_{u_i,u_j} \leq \mathtt{R}\}$, where $d^{d2d}_{u_i,u_j}$ is the distance between the u_i-th and u_j-th devices. Devices can use the same frequency band and transmit encoded files simultaneously via D2D links. We assume there is a set of F popular files, denoted by $\mathcal{F} = \{f_1, f_2, \cdots, f_F\}$. This data frame constitutes the set of most frequent requested files by the users within a given time duration in a hotspot area. Following the caching model in [129], the b_n-th eRRH caches a subset $\mathcal{H}_{b_n}$ that represents its *cache*, i.e., $|\mathcal{H}_{b_n}| = \mu F, \forall b_n \in \mathcal{B}$,

where $0 \leq \mu \leq 1$ is the fractional cache size. Further, we assume that all eRRHs collectively cache all files in the frame, i.e., $\bigcup_{n=1}^{B} \mathcal{H}_{b_n} = \mathcal{F}$. The distribution of files among eRRHs is assumed to be given, and some common files can be cached in different eRRHs' caches. In this work, each device is assumed to be equipped with single antenna and used half-duplex channel. Thus, each device can access to either a D2D channel or cellular channel, and accordingly, it can either transmit or receive at a given time instant. Moreover, the allocated channels for D2D communications are assumed to be orthogonal (out-of-band) to those used by eRRHs, i.e., an overlay D2D communication model is adopted [24].

Physical Layer Model

The achievable rate at the u_i-th user when receives file from the b_n-th eRRH is given by $R^c_{b_n,u_i} = \log_2(1 + \text{SINR}_{b_n,u_i}(\mathbf{P})), \forall b_n \in \mathcal{B}, \forall u_i \in \mathcal{U}$, where $\text{SINR}_{b_n,u_i}(\mathbf{P})$ is the corresponding signal-to-interference plus noise-ratio experienced by the u_i-th user when it is assigned to the b_n-th eRRH. This SINR is given by $\text{SINR}_{b_n,u_i}(\mathbf{P}) = \dfrac{P_{b_n}|h^c_{b_n,u_i}|^2}{N_0 + \sum_{b_{n'} \in \mathcal{B}, b_{n'} \neq b_n} P_{b_{n'}}|h^c_{b_{n'},u_i}|^2}$, where $h^c_{b_n,u_i}$ denotes the channel gain between the u_i-th user and b_n-th eRRH, N_0 denotes the noise power, P_{b_n} denotes the transmit power of b_n-th eRRH, and $\mathbf{P} = [P_{b_n}], \forall b_n \in \mathcal{B}$ is a row vector containing the power levels of the eRRHs in the considered network. The set of users' rates across all eRRHs can be written as $\mathcal{R} = \bigotimes_{(b_n,u_i) \in \mathcal{B} \times \mathcal{U}} R^c_{b_n,u_i}$, where the symbol $\bigotimes$ represents the product of the set of the achievable rates. Similarly, let $h^{d2d}_{u_k,u_i}$ denote the channel gain for the D2D link between the u_k-th and u_i-th users and P_{u_k} denote the transmit power of the u_k-th user. Then, the achievable rate of D2D pair (u_k, u_i) is given by $r^{d2d}_{u_k,u_i} = \log_2\left(1 + \dfrac{P_{u_k}|h^{d2d}_{u_k,u_i}|^2}{N_0 + \sum_{u_{k'} \in \mathcal{U}_{\text{tra}}, u_{k'} \neq u_k} P_{u_{k'}}|h^{d2d}_{u_{k'},u_i}|^2}\right), \forall u_k, u_{k'} \in \mathcal{U}_{\text{tra}}$, and $u_i \in \mathcal{C}_{u_k} \cap \mathcal{C}_{u_{k'}}$, where $\mathcal{U}_{\text{tra}}$ is the set of transmitting users via D2D links. We assume $h^c_{b_n,u_i}$ and $h^{d2d}_{u_k,u_i}$ to be fixed during a single eRRH and D2D transmissions but change independently from one file transmission to another file transmission. The channel capacities of all pairs of D2D links can be stored in an $U \times U$ *capacity status matrix (CSM)* $\mathbf{r} = [r_{u_k,u_i}]$,

$\forall(u_k, u_i)$. Since u_k-th user does not transmit to itself and cannot transmit to other users outside its coverage zone, $r^{d2d}_{u_k,u_k} = 0$ and $r^{d2d}_{u_k,u_l} = 0, \forall u_l \notin \mathcal{C}_{u_k}$.

4.3 Completion Time Minimization for D2D-aided F-RAN Systems

4.3.1 Network Coding in the Network-Layer

We assume that users are interested in receiving the whole frame $\mathcal{F}$, and they have already acquired some files in $\mathcal{F}$ from prior broadcast transmissions (i.e., without NC) [68]. The previously acquired files by u_i-th user is denoted by the *Has* set $\mathcal{H}_{u_i}$, and its requested files is denoted by the *Wants* set, i.e., $\mathcal{W}_{u_i} = \mathcal{F} \setminus \mathcal{H}_{u_i}$. Taking advantage of the acquired and requested files by different users, each eRRH and D2D transmitter can perform XOR operation on these files and send the combined XORed files to the interested users. As such, the requested files are delivered to requesting users with minimum completion time. We use the subscript t to represent the index of transmission/time slot, e.g., $t = 1$ refers to the first transmission slot. After each transmission, each user feedbacks to the eRRHs and neighboring users an acknowledgment for each received file, and accordingly, the *Has* and *Wants* sets are updated by the CBS [67], [68]. The set of users having *non-empty Wants sets* at the t-th transmission slot is denoted by $\mathcal{U}_{w,t}$, which is defined as $\mathcal{U}_{w,t} = \{u_i \in \mathcal{U} | \mathcal{W}_{u_i,t} \neq \varnothing\}$. When a user receives its requested files, it can act as a D2D transmitter to provide its received files to the interested neighboring users.

Let $\kappa^c_{b_n,t}$ and $\kappa^{d2d}_{u_k,t}$ denote the XOR file combinations to be sent by the b_n-th eRRH and u_k-th D2D transmitter, respectively, to the sets of scheduled users $\tau(\kappa^c_{b_n,t})$ and $\tau(\kappa^{d2d}_{u_k,t})$ at the t-th transmission. For simplicity, the subscript transmission index t is often omitted when it is clear from the context. These file combinations $\kappa^c_{b_n}$ and $\kappa^{d2d}_{u_k}$ are elements of the power sets $\mathcal{P}(\mathcal{H}_{b_n})$ and $\mathcal{P}(\mathcal{H}_{u_k})$, respectively. At every transmission slot t, each scheduled user in $\tau(\kappa^c_{b_n})$ can re-XOR $\kappa^c_{b_n}$ with its previously received files to decode a new file. To ensure successful reception at the users,

the maximum transmission rate of a particular transmitting eRRH/user is equal to the minimum achievable capacity of its scheduled users. For discussion convenience, the term "targeted users" is referred to a set of scheduled users who receives an instantly-decodable transmission. Therefore, the set of targeted users by b_n-th eRRH is expressed as $\tau(\kappa^c_{b_n}) = \left\{u_i \in \mathcal{U}_w \ \middle| |\kappa^c_{b_n} \cap \mathcal{W}_{u_i}| = 1 \text{ and } R^c_{b_n} \leq R^c_{b_n,u_i}\right\}$. Similarly, for D2D transmissions, the set of targeted users by u_k-th D2D transmitter is expressed as $\tau(\kappa^{d2d}_{u_k}) = \{u_j \in \mathcal{U}_w \big| |\kappa^{d2d}_{u_k} \cap \mathcal{W}_{u_j}| = 1 \text{ and } u_j \in \mathcal{C}_{u_k} \text{ and } r^{d2d}_{u_k} \leq r^{d2d}_{u_k,u_j}\}$. Without loss of generality, the set of all targeted users when $|\mathcal{U}_{\text{tra}}|$ D2D transmitters transmit the set of combinations $\kappa^{d2d}(\mathcal{U}_{\text{tra}})$ is represented by $\tau(\kappa^{d2d}(\mathcal{U}_{\text{tra}}))$, where u_k, $\kappa^{d2d}_{u_k}$, $\tau(\kappa^{d2d}_{u_k})$ are elements in $\mathcal{U}_{\text{tra}}$, $\kappa^{d2d}(\mathcal{U}_{\text{tra}})$, and $\tau(\kappa^{d2d}(\mathcal{U}_{\text{tra}}))$, respectively.

4.3.2 Transmission Time Analysis and Expression of the Completion Time

This subsection provides an analysis of the transmission time for sending coded files from the eRRHs and D2D transmitters to a set of scheduled users, which leads to an expression of the completion time in D2D-aided F-RAN.

The transmission time for sending the coded file $\kappa^c_{b_n}$ from the b_n-th eRRH with rate $R^c_{b_n}$ to the set of targeted users $\tau(\kappa^c_{b_n})$ is $T^c_{b_n} = \frac{N}{R_{b_n}}$ seconds, where N is the size of the file in bits. Without loss of generality, let us assume that the b_{n^*}-th eRRH has the minimum rate at the t-th transmission slot that is denoted by $R_{b_{n^*}}$. The corresponding transmission duration is $T^c_{b_{n^*}} = \frac{N}{R_{b_{n^*}}}$ seconds. Since different eRRHs will have different transmission rates, they will have different transmission durations. Thus, the portion of the time that not being used by $b_{n'}$-th eRRH at t-th transmission slot is referred to as the idle time of the $b_{n'}$-th eRRH and denoted by $T^c_{b_{n'}\text{idle}}$. This idle time can be expressed as $T^c_{b_{n'}\text{idle}} = (T^c_{b_{n^*}} - T^c_{b_{n'}})$ seconds. Such idle time can be exploited by the scheduled users of $b_{n'}$-th eRRH via D2D links if it ensures the complete delivery of files, i.e., $T^c_{b_{n'}\text{idle}} \geq T^{\text{d2d}}_{u_m}$, where $T^{\text{d2d}}_{u_m} = \frac{N}{r^{d2d}_{u_k}}$ is the transmission duration for sending $\kappa^{d2d}_{u_m}$ from the u_m-th D2D transmitter with adopted rate r^{d2d}_m, $\forall u_m \in \tau(\kappa^c_{b_{n'}})$. The unscheduled users by the eRRHs

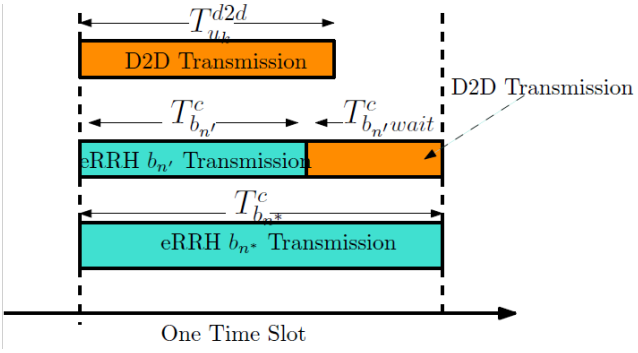


Figure 4.2: Transmission time structure for eRRHs and potential D2D transmitters for one time slot.

can also use D2D links to transmit files, and accordingly, the transmission duration for sending $\kappa^{d2d}_{u_k}$ from the u_k-th D2D transmitter with adopted rate r^{d2d}_{k} is $T^{d2d}_{u_k} = \frac{N}{r^{d2d}_{u_k}}$ seconds, $\forall u_k \notin \tau(\kappa^{c}_{b_n}), \forall b_n \in \mathcal{B}$. Based on the above discussion, u_l-th user experiences one of three possible delays at each transmission, as shown in Figure 4.2, and described below.

1. The time delay for u_l-th user receiving a non-instantly decodable transmission from b_{n^*}-th eRRH, this delay is $T^{c}_{b_{n^*},u_l}$, $\forall u_l \notin \tau(\kappa^{c}_{b_{n^*}})$.

2. The time delay for u_l-th user receiving a non-instantly decodable transmission from $b_{n'}$-th eRRH, this delay is $T^{c}_{b_{n'},u_l}$, $b_{n'} \in \mathcal{B}$ and $u_l \notin \tau(\kappa^{c}_{b_{n'}})$.

3. The time delay for u_l-th user being transmitting or receiving a non-instantly decodable transmission from any D2D transmitter in the set $\mathcal{U}_{\text{tra}}$, this delay is denoted as $T^{d2d}_{u_k,u_l}$, where $(u_l = u_k) \in \mathcal{U}_{\text{tra}}$ or $(u_l \notin \tau(\kappa^{d2d}(\mathcal{U}_{\text{tra}}))$ and $u_k \in \mathcal{U}_{\text{tra}})$.

Note that for u_l-th user, $T^{c}_{b_{n'},u_l}$ is less than $T^{c}_{b_{n^*},u_l}$ and $\max\limits_{u_k \in \mathcal{U}_{\text{tra}}} (T^{d2d}_{u_k,u_l})$. Thus, the maximum delay experienced by u_l-th user, which is not scheduled at the t-th transmission slot, is equal to $T_{\max,t} = \max(T^{c}_{b_{n^*}}, \max\limits_{u_k \in \mathcal{U}_{\text{tra}}} (T^{d2d}_{u_k}))$. Consequently, users that are not scheduled at transmission slot t, experience $T_{\max,t}$ seconds of delay in a cumulative manner defined as follows.

Definition 4.1. A user with non-empty Wants set experiences $T_{\max,t}$ seconds of time delay if it does not receive any requested file at t-th transmission slot. The accumulated time delay of u_l-th user is the sum of $T_{\max,t}$ seconds at each transmission until t-th transmission slot, and denoted by $\mathbb{D}_{u_l,t}$. It can be expressed as

$$\mathbb{D}_{u_l,t} = \mathbb{D}_{u_l,t-1} + \begin{cases} T_{\max,t} & \text{if } u_l \notin \left(\mathtt{u}(\mathtt{f}^c_{b_{n^*},t}) \cup \tau(\kappa^c_{b_{n'},t})\right), \forall b_{n^*}, b_{n'} \in \mathcal{B} \\ T_{\max,t} & \text{if } (u_l = u_k) \in \mathcal{U}_{\text{tra},t} \text{ or } (u_l \notin \tau(\kappa^{d2d}(\mathcal{U}_{\text{tra},t})) \\ \text{and } u_k \in \mathcal{U}_{\text{tra},t}) & \end{cases} \tag{4.1}$$

Definition 4.2. The completion time of u_l-th user, denoted by $\mathtt{T}_{u_l}$, is the total time required in seconds to receive all its requested files. The overall completion time $\mathtt{T}_o$ is the time required to receive all files by all users, and is given by $\mathtt{T}_o = \max_{u_l \in \mathcal{U}_w}\{\mathtt{T}_{u_l}\}$.

Definition 4.3. A transmission schedule $\mathcal{T} = \{(\kappa^c_{b_n,t}, R^c_{b_n}), (u_k, \kappa^{d2d}_{u_k,t}, r^{d2d}_{u_k})\}$ $\forall t \in \{1, 2, ..., |\mathcal{T}|\}, \forall b_n \in \mathcal{B}, \forall u_k \in \mathcal{U}_{\text{tra},t}$ is a collection of transmitting eRRHs/D2D transmitters, their file combinations and adopted rates at every t-th transmission index to receive all files by all users.

The completion time minimization problem in a D2D-aided F-RAN system can be expressed as follows

$$\mathcal{T}^* = \arg\min_{\mathcal{T}\in\mathbf{T}}\{\mathtt{T}_o(\mathcal{T})\} = \arg\min_{\mathcal{T}\in\mathbf{T}}\left\{\max_{u_l\in\mathcal{U}_w}\{\mathtt{T}_{u_l}(\mathcal{T})\}\right\}, \tag{4.2}$$

where $\mathcal{T}^*$ is the schedule that optimally minimizes the overall completion time and $\mathbf{T}$ is the set of all possible transmission schedules. The following theorem expresses the optimal schedule $\mathcal{T}^*$ in terms of time delay defined in definition 1.

Theorem 4.4. *The optimal schedule $\mathcal{T}^*$ that minimizes the overall comple-*

tion time in a D2D-aided F-RAN system can be written as follows

$$\mathcal{T}^* = \arg\min_{\mathcal{T}\in\mathbf{T}} \left\{ \max_{u_l\in\mathcal{U}_w} \left\{ \frac{N.|\mathcal{W}_{u_l,0}|}{\tilde{R}_{u_l}(\mathcal{T})} + \mathbb{D}_{u_l}(\mathcal{T}) \right\} \right\}, \tag{4.3}$$

where $|\mathcal{W}_{u_l,0}|$ is the initial Wants size of u_l-th user, $\mathbb{D}_{u_l}(\mathcal{T})$ is the accumulative time delay of u_l-th user in schedule, and $\tilde{R}_{u_l}(\mathcal{T})$ is the harmonic mean of the transmission rates of transmissions that are instantly decodable for u_l-th user in schedule $\mathcal{T}$.

Proof. The proof of Theorem 4.4 is omitted in this work as it can follow the same steps that proofed Theorem 1 in [82] for C-RAN networks without D2D communications. Therefore, only a sketch of the proof is given as follows. We first show that the completion time can be expressed as the sum of instantly and non-instantly decodable transmission times from $|\mathcal{B}|$ and $|\mathcal{U}_{\text{tra}}|$ transmitters via cellular and D2D links, respectively. Afterward, we need to proof that the number of instantly decodable transmissions to u_l-th user is equal to the number of its requested files $|\mathcal{W}_{u_l,0}|$ and the number of non-instantly decodable transmissions matches the time delay in definition 1. Finally, we extend the results of the optimal schedule in Theorem 1 in [82] that used in PMP system with a single transmitter to the coordinated D2D-aided F-RAN setting with multiple transmitters. □

Solving the completion time problem in (4.2) optimally is intractable [82]. In fact, the transmission schedule at the current transmission slot does not depend only on the future transmission schedules, but also on users' achievable capacities and eRRHs' transmit powers. Therefore, we pay our special attention to solve such problem at each transmission, where files are transmitted with high transmission rates. If some eRRHs cannot send XOR files to a set of users with the rate threshold R_{th}, these users can be scheduled on D2D links. To this end, our main objective is to minimize the completion time at each transmission, known as the anticipated completion time [115], through minimizing the time delay. This anticipated user's completion time at each transmission in D2D-aided F-RAN system is given in the next corollary.

Corollary 4.5. *The anticipated completion time of u_l-th user at t-th transmission slot is given by*

$$\mathtt{T}_{u_l,t} \approx \frac{N.|\mathcal{W}_{u_l,0}|}{\tilde{R}_{u_l,t}} + \mathbb{D}_{u_l,t}, \tag{4.4}$$

where $\mathbb{D}_{u_l,t}$ is the accumulative transmission delay as given in (4.1), and $\tilde{R}_{u_l,t}$ is the harmonic mean of the transmission rates that are instantly decodable for u_l-th user until t-th transmission.

The anticipated completion time in Corollary 4.5 depends on the number of requested files by u_l-th user, its accumulated time delay and harmonic mean $\tilde{R}_{u_l,t}$. Clearly, this metric is intimately related to the duration of time that all files are delivered to all users, which can be illustrated in the following example.

Example 5: This example considers the model in Figure 4.3 that consists of 2 eRRHs, 6 users, users' received and requested files and their rates. For example, u_2 receives f_1, f_4 and requests f_2, f_3. The sets of files that stored in eRRHs' caches are $\mathcal{H}_{b_1} = \{f_1, f_4, f_3\}$, $\mathcal{H}_{b_2} = \{f_2, f_3, f_4\}$. Each file is assumed to have a size of 10 bits. To minimize the completion time for this example, one possible schedule is given as follows.

First time slot: The b_1-th and b_2-th eRRHs transmit $\kappa^c_{b_1,1} = f_1 \oplus f_4$ and $\mathtt{f}^c_2 = f_3 \oplus f_4$ with rates $R^c_{b_1} = 2.5$ and $R^c_{b_2} = 5$ bits/s, respectively, to the sets $\mathtt{u}(\kappa^c_{b_1,1}) = \{u_4, u_6\}$ and $\tau(\kappa^{c,1}_{b_2}) = \{u_2, u_3\}$. The u_1-th user transmits $\kappa^{d2d}_{u_1,1} = f_4$ with rate $r^{d2d}_{u_1} = 5$ bits/s to the set $\tau(\kappa^{d2d}_{u_1,1}) = \{u_5\}$. Given this, we have the following transmission durations of b_1-th eRRH, b_2-th eRRH, and u_1-th transmitting user, respectively: $T^c_{b_1} = \frac{10}{2.5} = 4, T^c_{b_1} = \frac{10}{5} = 2, T^{d2d}_{u_1} = \frac{10}{5} = 2$ seconds. Since user u_4 receives f_4 from b_1-th eRRH in 4 seconds, it can use the idle time of b_1-th eRRH, i.e., $T^c_{b_1 \text{idle}} = 2$ seconds, to send f_2 to u_6 with rate $r^{d2d}_{u_4} = 5$ bits/s. Therefore, the updated *Wants* sets after the first time slot are: $\mathcal{W}_{u_2,1} = \{f_2\}$, $\mathcal{W}_{u_3,1} = \varnothing, \mathcal{W}_{u_4,1} = \varnothing$, $\mathcal{W}_{u_5,1} = \{f_3\}$, $\mathcal{W}_{u_6,1} = \varnothing$. Note that $T_{\max,1} = \max(T^c_{b_1}, T^c_{b_2}, T^{d2d}_{u_1}) = 4$ seconds.

Second time slot: The b_2-th eRRH transmits $\mathtt{f}^{c,2}_{b_2} = f_2 \oplus f_3$ with rate $R^c_{e_2} = 2.5$ bits/s to the set $\tau(\kappa^{c,2}_{b_2}) = \{u_2, u_5\}$ which requires transmission

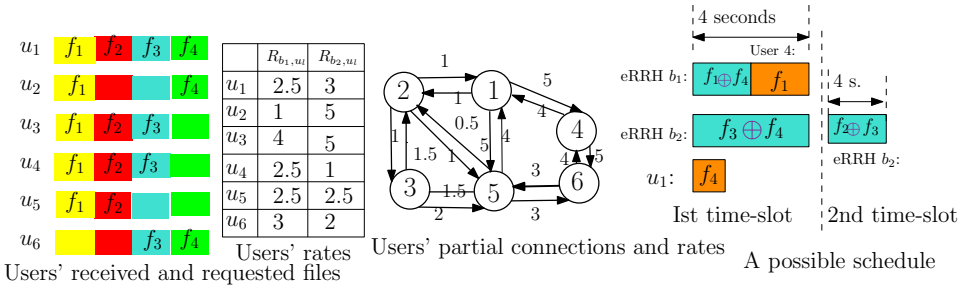


Figure 4.3: D2D-aided F-RAN system containing 2 users and their corresponding requested/received files and rates.

time $T^c_{b_2} = T_{\max,2} = \frac{10}{2.5} = 4$ seconds. By the end of second time slot, all users will have their wanted files. Therefore, the total transmission time is $T_{\max,1} + T_{\max,2} = 8$ seconds.

4.4 Problem Formulation and Problem Decomposition

Problem Formulation

In order to minimize the completion time at each transmission slot, we need to develop a rate-aware network coding framework that decides: i) the adopted transmission rate/power at the e_n-th eRRH, $\{R^c_{b_n}, P_{b_n}\}$, to transmit its XOR combination $\kappa^c_{b_n,t}$ to a set of targeted users $\tau(\kappa^c_{b_n,t})$, $\forall b_n \in \mathcal{B}$, and ii) the set of D2D transmitters $\mathcal{U}_{\text{tra},t}$ for sending $\kappa^{d2d}_{u_k,t}$ to the users $\tau(\kappa^{d2d,t}_{u_k})$, and their adopted transmission rates $r^{d2d}_{u_k}, \forall u_k \in \mathcal{U}_{\text{tra},t}$. As such, all files are delivered to all users with minimum completion time. Therefore, the completion time minimization problem in D2D-aided F-RAN system can be formulated as on the top of next page. The constraints are explained as follows. C1 states that the set of scheduled users to all eRRHs are disjoint, i.e., each user must be scheduled to only one eRRH. C2 makes sure that each user can be scheduled to only one potential D2D transmitter and no user can be scheduled to a D2D transmitter and eRRH at the same time instant. C3 ensures the successful delivery of files from D2D transmissions within the idle

$$\text{P1}: \min_{\kappa^c_{b_n,t},\kappa^{d2d}_{u_k,t},P_{b_n},\mathcal{U}_{\text{tra},t}\in\mathcal{P}(\mathcal{U})} \left\{ \max_{\text{u}_\text{l}\in\mathcal{U}_{\text{w,t}}} \text{T}_{\text{u}_\text{l},\text{t}} \right\}$$

$$\text{subject to} \begin{cases} \text{C1: } \tau(\kappa^c_{b_n,t}) \cap \tau(\kappa^c_{e_{n'},t}) = \varnothing, \forall b_n \neq b_{n'} \in \mathcal{B}; \\ \text{C2: } \tau(\kappa^{d2d}_{u_k,t}) \cap \tau(\kappa^{d2d}_{u_{k'},t}) = \varnothing \ \& \ \tau(\kappa^{d2d}_{u_k,t}) \cap \tau(\kappa^c_{b_n,t}) = \varnothing, \\ \forall u_k \neq u_{k'} \in \mathcal{U}_{\text{tra},t}, b_n \in \mathcal{B}; \\ \text{C3: } r^{d2d}_{u_k}.T^c_{b_{n'}\text{idle}} \geq N, \ \forall u_k \in \mathcal{U}_{\text{tra},t}, \ \forall b_{n'} \in \mathcal{B}; \\ \text{C4: } \kappa^c_{b_n,t} \subseteq \mathcal{P}(\mathcal{H}_{b_n}) \ \& \ \kappa^{d2d}_{u_k,t} \subseteq \mathcal{P}(\mathcal{H}_{u_k,t}), \ \forall (b_n, u_k) \in \mathcal{B} \times \mathcal{U}_{\text{tra},t}; \\ \text{C5: } 0 \leq P_{b_n} \leq P_{\max}, \forall b_n \in \mathcal{B}; \\ \text{C6: } R^c_{b_n} \geq R_{\text{th}}; \text{C7: } r^{d2d}_{u_k} \geq R_{\text{th}}, \forall b_n \in \mathcal{B}, \forall u_k \in \mathcal{U}_{\text{tra},t}. \end{cases}$$

time of the eRRHs. C4 ensures that all files to be combined using XOR operation at all eRRHs and D2D transmitters are stored in their *Caches* and *Has* sets, respectively. C5 bounds the maximum transmit power of each eRRH, and C6 and C7 satisfy the minimum transmission rates required to meet the QoS rate requirement R_{th}. The optimization problem in P1 contains the NC scheduling parameters $\tau(\kappa^c_{b_n,t}), \tau(\kappa^{d2d}_{u_k,t}), \forall b_n \in \mathcal{B}, \forall u_k \in \mathcal{U}_{\text{tra},t}$, power allocations of eRRHs P_{b_n}, $\forall b_n \in \mathcal{B}$, potential set of transmitting users $\mathcal{U}_{\text{tra},t}$ and their transmission rates. We can readily show that problem P1 is NP-hard and intractable [116]. However, by analyzing the problem, we can decompose it into two subproblems and solve them individually and efficiently using graph theory technique.

Problem Decomposition

Since the main objective is to minimize the maximum completion time of users, which depends on the time delay increase at each transmission slot, we can first focus on minimizing the transmission duration for the eRRH-user NC transmissions. In particular, we can get the possible completion time by jointly optimizing the NC user scheduling and power allocations of eRRHs. The mathematical formulation for minimizing the transmission duration for

eRRH-user NC transmissions can be expressed as

$$\mathrm{P2}: \min_{0\leq P_{b_n}\leq P_{\max}} \mathrm{T}^{\mathrm{c}}_{\mathrm{b_n,t}} \tag{4.6a}$$

$$\text{subject to } \begin{cases} \tau(\kappa^c_{b_n,t}) \cap \tau(\kappa^c_{b_{n'},t}) = \varnothing, \forall b_n \neq b_{n'} \in \mathcal{B}; \\ \kappa^c_{b_n,t} \subseteq \mathcal{P}(\mathcal{H}_{b_n});\ R^c_{b_n} \geq R_{\mathrm{th}},\ \forall b_n \in \mathcal{B}. \end{cases}$$

Note that this subproblem contains users' associations and power allocation variables and a joint solution will be developed in Section 4.5. After obtaining the possible transmission duration from eRRH-user NC transmissions, denoted by $T^c_{b_{n^*},t}$ of b_{n^*}-th eRRH, by solving P2, we can now formulate the second subproblem. In particular, we can maximize the number of users Z_t that are not been scheduled to the eRRHs $\mathcal{U}_{w,t}\backslash\tau(\kappa^c_{b_{n^*},t})$ within $T^c_{b_{n^*},t}$ by using D2D communication. In addition, users being scheduled to $b_{n'}$-th eRRH from subproblem P2, have the opportunity to be scheduled on D2D links within the idle times of their corresponding eRRHs at the t-th transmission slot, $\forall b_{n'} \neq b_{n^*} \in \mathcal{B}$. Therefore, the second subproblem of maximizing the number of users to be scheduled on D2D links can be expressed as follows

$$\mathrm{P3}: \max_{\substack{\mathcal{U}_{\mathrm{tra},t}\in\mathcal{P}(\mathcal{U}\backslash\tau(\kappa^c_{b_{n^*},t})) \\ \kappa^{d2d}_{u_k,t}\subseteq\mathcal{P}(\mathcal{H}_{u_k,t})}} \mathrm{Z_t} \tag{4.7a}$$

$$\text{subject to } \begin{cases} (\mathrm{C2}); r^{d2d}_{u_k}.(T^c_{b_{n^*},t} - T^c_{b_{n'},t}) \geq N,\ \forall u_k \in \mathcal{U}_{\mathrm{tra},t},\ \forall (b_{n^*}, b_{n'}) \in \mathcal{B}; \\ T^{d2d}_{u_k} \leq T^c_{b_{n^*},t},\ \forall u_k \in \mathcal{U}_{\mathrm{tra},t}; \\ |\tau(\kappa^{d2d}(\mathcal{U}_{\mathrm{tra},t}))| + |\mathcal{U}_{\mathrm{tra},t}| \leq Z_t. \end{cases}$$

4.5 Joint Solution for Completion Time Minimization Problem

In this section, we propose a joint approach to solve the subproblems in P2 and P3 using designed interference-aware IDNC and new D2D conflict graphs, respectively. Specifically, we design interference-aware IDNC in the fist subsection to solve the sub-problem P2. We then introduce a new D2D

conflict graph to solve the sub-problem P3 as shown in the second subsection.

4.5.1 Solution to Sub-problem P2

Interference Aware-IDNC Graph

Interference-Aware IDNC (IA-IDNC) graph, denoted by $\mathcal{G}_{\text{IA-IDNC}}(\mathcal{V}, \mathcal{E})$, is designed to systematically select an IDNC combination, transmission rate, and power allocation of each eRRH at the t-th transmission slot. Unlike the graph in [82] that resulted in one rate for fixed power eRRHs, our designed IA-IDNC graph leads to different transmission rates/powers from different eRRHs. This gives flexibility to each eRRH to choose its IDNC combination and transmission rate that satisfy a set of scheduled users.

Consider generating all possible associations (pairs) representing users and their corresponding requested files that cached by b_n-th eRRH, denoted by $\mathcal{S}_{b_n} = \mathcal{U}_w \times \mathcal{H}_{b_n}$, i.e., $s \in \mathcal{S}_{b_n} = (u_l, f_h)$ represents the association of u_l-th user and its f_h-th requested file. The corresponding files of a set of associations in $\mathcal{S}_{b_n}$ can be encoded into one IDNC combination if these files are instantly decodable to the corresponding associated users. The set of all IDNC combinations is denoted by $\mathcal{S}_{\text{IDNC},b_n}$. In particular, the corresponding files of any two different associations $s \in \mathcal{S}_{b_n}$ and $s' \in \mathcal{S}_{b_n}$ are encoded if one of the following IDNC conditions is satisfied.

- **IDNC-C1:** $u_{l,s} \neq u_{l',s'}$ and $f_{h,s} = f_{h,s'}$ This condition represents that the same file f_h is requested by two distinct users u_l and $u_{l'}$.
- **IDNC-C2:** $u_{l,s} \neq u_{l',s'}$ and $f_{h',s'} \in \mathcal{H}_{u_l,s}$ and $f_{h,s} \in \mathcal{H}_{u_{l'},s'}$. This condition represents that different files $f_{h'}$ and f_h are requested by two different users $u_{l'}$ and u_l, respectively. Meanwhile, the requested file of each user is in the *Has* set of the user in the other association. We use l, s in $u_{l,s}$ as subscripts to represent u_l-th user in s-th association.

For example, the element $\mathtt{s} = (\kappa^c_{b_n}, \tau(\kappa^c_{b_n})) \in \mathcal{S}_{\text{IDNC},b_n}$ represents the set of scheduled users $\tau(\kappa^c_{b_n})$ that will receive the IDNC combination $\kappa^c_{b_n}$ from b_n-th eRRH. Let $\mathcal{A}_{b_n}$ be the set of all possible associations between the IDNC combinations $\mathcal{S}_{\text{IDNC},b_n}$ and achievable capacities $\mathcal{R}_{b_n} \subset \mathcal{R}$, i.e.,

$\mathcal{A}_{b_n} = \mathcal{S}_{\text{IDNC},b_n} \times \mathcal{R}_{b_n}$. In other words, $\mathbf{S} = (\mathtt{s}, R) \in \mathcal{A}_{b_n}$ is a schedule that consists of a set of associations representing the IDNC combination, set of scheduled users, and rate R of e_n-th eRRH, i.e., $\mathbf{S} = (\mathtt{s}, R) = s_1, s_2, \cdots, s_{|\mathbf{S}|}$, where $|\mathbf{S}|$ is the total number of scheduled users in $\mathbf{S}$. Note that s_1 represents one user, one file, and rate of b_n-th eRRH. Now, any two associations $s_1 \in \mathbf{S}, s_2 \in \mathbf{S}$ representing the b_n-th eRRH should have an equal adopted rate that is greater than or equal to R_{th}. That is, the **Rate Condition (RC)** is satisfied $R_{s_1} = R_{s_2}$ and $R_{s_1} \geq R_{\text{th}}$.

The aforementioned procedures are applied to all eRRHs in the network. Thus, the set of all possible IDNC combinations $\mathcal{S}_{\text{IDNC},b_n}$ and schedules $\mathcal{A}_{b_n}$ in the network are $\mathcal{S}_{\text{IDNC}} = \bigcup_{b_n \in \mathcal{B}} \mathcal{S}_{\text{IDNC},b_n}$ and $\mathcal{A} = \bigcup_{b_n \in \mathcal{B}} \mathcal{A}_{b_n}$, respectively. These schedules $\mathcal{A}$ can be exactly represented by unique vertices $\mathcal{V}$ in $\mathcal{G}_{\text{IA-IDNC}}(\mathcal{V}, \mathcal{E})$ such that we transfer the subproblem P2 to a graph-theory based problem. Therefore, the $\mathbf{S}_i$-th schedule in $\mathcal{S}$ is represented by the V_i-th unique vertex in $\mathcal{G}_{\text{IA-IDNC}}$ $(i = 1, 2, \cdots, |\mathcal{S}|)$. This schedule-to-vertex mapping makes any IDNC combination sent from the b_n-th eRRH with adopted rate to its corresponding associated users is decodable. Two vertices V_i and $V_{i'}$ representing two different schedules $\mathbf{S}_i \in \mathcal{S}_{b_n}$ and $\mathbf{S}_{i'} \in \mathcal{A}_{b_{n'}}$ are adjacent by an edge in $\mathcal{G}_{\text{IA-IDNC}}$, if the associations they represent satisfy the following condition.

- **Transmission Condition (TC):** $\tau \cap \tau' = \phi$, $\forall (s_1, s_2) \in \mathbf{S}_i \times \mathbf{S}_{i'}$. This condition ensures that the same user can be scheduled only to a unique eRRH.

Assuming that the power allocation of the eRRHs in the network will be computed later; then the weight of a given vertex V representing a schedule $\mathbf{S}$ is expressed by

$$w(V) = \sum_{s \in \mathbf{S}} \frac{\min_{u_{l,s} \in \tau(\kappa^c_{b_n})} \log_2(1 + \text{SINR}_{b_n,s,u_{l,s}}(\mathbf{P}))}{N}. \tag{4.8}$$

The weight of each vertex reflects the contribution of each eRRH towards minimizing the completion time of its associated users. Actually, the transmission rate plays a crucial role in minimizing the transmission duration $T^c_{b_n}$.

Thus, a larger value in (4.8) leads to minimize the transmission duration of delivering IDNC files to users, which minimizes their completion times. The design of the IA-IDNC graph makes any maximal weight clique represents a set of transmissions satisfying the following three features: i) each user is scheduled only to a single eRRH that cached one of its requested files, ii) each eRRH delivers an IDNC file with an adopted transmission rate/power that satisfies a lower completion time for a potential set of users. Such adopted rate satisfies the QoS rate guarantee and no larger than the channel capacities of all scheduled users, iii) the weight of each vertex strikes a balance between the adopted rate and the number of scheduled users to each eRRH.

The following theorem characterizes the solution of subproblem P2 based on the designed IA-IDNC graph.

Theorem 4.6. *The transmission duration minimization subproblem P2 is equivalently represented by the maximum weight clique problem in the IA-IDNC graph, and can be expressed as*

$$\arg\max_{\mathtt{C}\in\mathbf{C}} \sum_{V_i\in\mathtt{C}} w(V_i),\ \forall i = 1, 2, \cdots, |\mathtt{C}|, \tag{4.9}$$

where $\mathtt{C}$ *is the maximum weight clique of a maximum degree* $|\mathcal{B}|$ *in the IA-IDNC graph and* $\mathbf{C}$ *is the set of all possible maximal cliques.*

Proof. A sketch of proof of Theorem 4.6 is provided as follows. First, we need to show that there is a unique one-to-one mapping between each schedule $\mathbf{S}_i \in \mathcal{S}$ and each vertex $V_i \in \mathcal{V}$ in $\mathcal{G}_{\text{IA-IDNC}}$ $(i = 1, 2, \cdots, |\mathcal{S}|)$. Then, we emphasize that each maximal clique in the IA-IDNC graph that consists of a set of vertices satisfying all edge conditions represents feasible coded transmissions from the eRRHs. Finally, the proof can be concluded by showing that the contributed weight of the maximum weight clique $\mathbf{C}$ for minimizing the transmission duration is: $w(\mathtt{C}) = \sum_{V_i\in\mathtt{C}} w(V_i) = \sum_{b_n\in\mathcal{B}} \sum_{\mathbf{S}_i\in\mathtt{S}} w(\mathbf{S}_i) = \sum_{b_n\in\mathcal{B}} \sum_{s\in\mathbf{S}_i} \frac{\min_{u_{l,s}\in\tau(\kappa^c_{b_n})} \log_2(1+\text{SINR}_{b_n,s,u_{l,s}}(\mathbf{P}))}{N},\ \forall i = 1, 2, \cdots, |\mathtt{C}|$, where $\mathtt{S}$ is the set of the selected potential schedules. Therefore, the subproblem P2 is equivalent to the maximum weight clique problem among the maximal

cliques in the IA-IDNC graph. □

The problem in Theorem 4.6 is clearly NP-hard problem, and solving it optimally is intractable [74]. However, we heuristically and efficiently solve it in the next subsection.

Development of Greedy Algorithm

We solve the problem in Theorem 4.6 by characterizing the joint solution to the NC user scheduling and power allocation problem while designing the IA-IDNC graph. A proper power allocation for each eRRH leads to suppress the interference in the system, thus a better transmission rate is achieved. As a result, the transmission duration for delivering files to the scheduled users is minimized.

Consider a power-clique (PC) in $\mathcal{G}_{\text{IA-IDNC}}$ that is associated with a network-coded user scheduling $\mathtt{S} = \{\mathbf{S}_1, \mathbf{S}_2, \cdots, \mathbf{S}_{|B|}\}$, where $\mathbf{S}_1$ is the schedule of e_1-th eRRH, which consists of a set of associations $s_1, s_2, \cdots, s_{|\mathbf{S}_1|}$, and $|\mathbf{S}_1| = |\kappa(\tau^c_{b_n})|$. Our goal is to obtain a local optimal eRRH power allocation vector, denoted as $(P^*_{b_1}, \cdots, P^*_{b_B})$ for that PC. The power allocation problem is formulated as an optimization problem of maximizing the weighed sum-rate. As such, all scheduled users receive files sent by their associated eRRHs with minimum transmission duration, which can be expressed as

$$\begin{aligned} \text{P4}: \max_{P_{b_n}} & \sum_{n=1}^{B} \frac{|\tau(\kappa^c_{b_n}(\mathbf{S}_n))|}{N} * \min_{u_l \in \tau(\kappa^c_{b_n}(\mathbf{S}_n))} \left(\log_2(1 + \text{SINR}_{b_n,u_l}(\mathbf{P}))\right), \\ \text{s. t. } & 0 \leqslant P_{b_n} \leqslant P_{\max}, \forall n = 1, 2, \cdots, B, \end{aligned} \tag{4.10}$$

where $\tau(\kappa^c_{b_n}(\mathbf{S}_n))$ is the set of scheduled users in n-th schedule corresponding to b_n-th eRRH. The power allocation problem in P4 is a non-convex optimization problem. Therefore, similar to the works in literature (see for example, [147, 148] and references therein), we focus on finding the local optimal solution.

The proposed solution to the problem in Theorem 4.6 is executed at the CBS at each transmission slot and divided into two stages: designing the

IA-IDNC graph and finding its corresponding maximum PC.

First stage: The IA-IDNC graph is designed as follows. Using **IDNC-C1**, **IDNC-C2**, and **RC** conditions that explained in Section 4.5.1, we generate all schedules $\mathcal{A}$ and represent them by vertices in $\mathcal{V}$. Afterwards, we check the connection between any two pairs V_i and $V_{i'}$ of vertices in $\mathcal{G}_{\text{IA-IDNC}}$ based on the transmission condition **TC** in Section 4.5.1. Any connected vertices result in a feasible network-coded scheduling to the eRRHs. Then, we evaluate the power allocation of such network-coded scheduling $\{\mathbf{S}_1, \mathbf{S}_2, \cdots, \mathbf{S}_B\}$ by solving the optimization problem (4.10). By the computed power allocation and corresponding rate, we compute the weights of the corresponding vertices in $\mathcal{G}_{\text{IA-IDNC}}$ as expressed in (4.8). We repeat the above process to all network-coded schedules in the IA-IDNC graph.

Second stage: The second stage finds the maximum weight PC $\mathtt{C}$ among all other maximal PCs in $\mathcal{G}_{\text{IA-IDNC}}$ graph. In the first step, we select vertex $V_i \in \mathcal{V}, (i = 1, 2, \cdots, |\mathcal{V}|)$ that has the maximum weight $w(V_i^*)$ and add it to $\mathtt{C}$ (at this point $\mathtt{C} = \{V_i^*\}$). Then, the subgraph $\mathcal{G}_{\text{IA-IDNC}}(\mathtt{C})$, which consists of vertices in graph $\mathcal{G}_{\text{IA-IDNC}}$ that are adjacent to vertex V_i^* is extracted and considered for the next vertex selection process. In the next step, a new maximum weight vertex $V_{i'}^*$ (i.e., $V_{i'}^*$ should be in the corresponding PC of V_i^*) is selected from subgraph $\mathcal{G}_{\text{IA-IDNC}}(\mathtt{C})$. Now, $\mathtt{C} = \{V_i^*, V_{i'}^*\}$. We repeat this process until no further vertex is adjacent to all vertices in the maximal weight PC $\mathtt{C}$. The selected $\mathbf{C}$ contains at most B vertices.

4.5.2 Solution to Sub-problem P3

New D2D Conflict Graph

We introduce a new D2D conflict graph, denoted by $\mathcal{G}_{\text{d2d}}(\mathcal{V}, \mathcal{E})$, that considers all possible conflicts for scheduling users on D2D links, such as transmission, network coding, half-duplex conflicts. This leads to feasible transmissions from the potential D2D transmitters $|\mathcal{U}_{\text{tra}}|$, where each $u_k \in \mathcal{U}_{\text{tra}}$ transmits the IDNC combination $\kappa_{u_k}^{d2d}$ to the scheduled users $\tau(\kappa_{u_k}^{d2d})$ with the transmission rate $r_{u_k}^{d2d}$.

Let $\mathcal{U}_{\text{d2d}}$ denote the set of unscheduled users to the eRRHs, i.e., $\mathcal{U}_{\text{d2d}} =$

$\mathcal{U} \setminus \bigcup_{n=1}^{B} \tau(\kappa_{b_n}^c)$, and let $\mathcal{U}_{\text{d2d},w} \subset \mathcal{U}_{\text{d2d}}$ denote the set of users that still wants some files. Hence, the D2D conflict graph is designed by generating all vertices for u_k-th possible D2D transmitter, $\forall u_k \in \mathcal{U}_{\text{d2d}}$. The vertex set $\mathcal{V}$ of the entire graph is the union of vertices of all users. Consider, for now, generating the vertices of u_k-th user. Note that u_k-th D2D transmitter can encode its IDNC file $\kappa_{u_k}^{d2d}$ using its previously received files $\mathcal{H}_{u_k}$. Therefore, each vertex is generated for each single file $f_h \in \mathcal{W}_{u_i} \cap \mathcal{H}_{u_k}$ that is requested by each user $u_i \in \mathcal{U}_{\text{d2d},w} \cap \mathcal{C}_{u_k}$ and for each achievable rate r of u_k-th user that is defined below.

Definition 4.7. The set of achievable rates $\mathcal{R}_{u_k,u_i}^{d2d}$ from u_k-th user to u_i-th user is a subset of achievable rates $\mathcal{R}_{u_k}^{d2d}$ that are less than or equal to channel capacity r_{u_k,u_i}^{d2d}. It can be expressed by $\mathcal{R}_{u_k,u_i}^{d2d} = \{r \in \mathcal{R}_{u_k}^{d2d} | r \leq r_{u_k,u_i}^{d2d}$ and $u_i \in \mathcal{U}_{\text{d2d,w}} \cap \mathcal{C}_{u_k}$ and $r \geq R_{b_{n^*}}^c\}$.

The above definition emphasizes that u_i-th user in the coverage zone $\mathcal{C}_{u_k}$ can receive a file from u_k-th D2D transmitter if the adopted transmission rate r is in the achievable set R_{u_k,u_i}^{d2d} and no less than the minimum transmission rate of b_{n^*}-th eRRH. Therefore, we generate $|\mathcal{R}_{u_k,u_i}^{d2d}|$ vertices for a requesting file $f_h \in \mathcal{H}_{u_k} \cap \mathcal{W}_{u_i}, \forall u_i \in \mathcal{U}_{\text{d2d,w}} \cap \mathcal{C}_{u_k}$. In summery, a vertex $v_{r,i,f}^k$ is generated for each association of a transmitting user u_k, a rate $r \in \mathcal{R}_{u_k,u_i}^{d2d}$, and a requesting file $f_h \in \mathcal{H}_{u_k} \cap \mathcal{W}_{u_i}$ of user $u_i \in \mathcal{U}_{\text{d2d,w}} \cap \mathcal{C}_{u_k}$. Similarly, we generate all vertices for all users in $\mathcal{U}_{\text{d2d}}$.

All possible conflict connections between vertices (conflict edges between circles) in the D2D conflict graph are provided as follows. Two vertices $v_{r,i,h}^k$ and $v_{r',i',h'}^{k'}$ are adjacent by a conflict edge in $\mathcal{G}_{\text{d2d}}$, if one of the following conflict conditions (CC) is true:

- **IDNC (CC1):** $(u_k = u_{k'})$ and $(f_h \neq f_{h'})$ and $(f_h, f_{h'}) \notin \mathcal{H}_{u_{k'}} \times \mathcal{H}_{u_k}$. A conflict edge between vertices is connected as long as the files they represent are not-instantly decodable to a set of scheduled users to the same u_k-th D2D transmitter.

- **Rate (CC2):** $(u_k = u_{k'})$ and $(r \neq r')$. All adjacent vertices correspond to the same u_k-th D2D transmitter should have the same achievable rate.

– **Transmission (CC3):** $(u_k \neq u_{k'})$ and $(u_i = u_{i'})$. The same user cannot be scheduled to two different D2D transmitters u_k and $u_{k'}$.

– **Half-Duplex (CC4):** $(u_k = u_{i'})$ or $(u_{k'} = u_i)$. The same user cannot transmit and receive in the same transmission slot.

Given the aforementioned designed D2D conflict graph, the following theorem reformulates the subproblem P3.

Theorem 4.8. *The subproblem of maximizing the number of scheduled users on D2D links in P_3 at the t-th transmission is equivalently represented by the maximum weight independent set (IS) selection among all the maximal sets in the $\mathcal{G}_{d2d}$ graph, where the weight $\psi(v^k_{r,i,h})$ of each vertex $v^k_{r,i,h}$ is given by*

$$\psi(v^k_{r,i,f}) = |\mathcal{C}_{u_k} \cap \mathcal{U}_{d2d,w}(\mathcal{H}_{u_k})|(\frac{r}{N}). \tag{4.11}$$

The above weight metric shows two potential benefits: i) $|\mathcal{C}_{u_k} \cap \mathcal{U}_{d2d,w}(\mathcal{H}_{u_k})|$ represents that the u_k-th transmitting user is connected to many other users that are requesting files in $\mathcal{H}_{u_k}$ and ii) $\left(\frac{r}{N}\right)$ provides a balance between the transmission rate and the number of scheduled users on D2D links.

Details of Greedy Solution

The proposed solution here greedily maximizes the number of scheduled users on D2D links within $T^c_{b_{n^*}}$ by maximizing the number of vertices in any IS in the D2D conflict graph. In order to maximize the number of vertices in any IS, we update the weight of each vertex. An appropriate design of the updated weights of vertices leads to selection of a large number of vertices and each vertex has high original weight that defined in Theorem 4.8.

Let $\pi_{v,v'}$[3] define the non-adjacency indicator of vertices v and v' in the $\mathcal{G}_{\text{d2d}}$ graph such that:

$$\pi_{v,v'} = \begin{cases} 1 & \text{if } v \text{ is not adjacent to } v' \text{ in } \mathcal{G}_{\text{d2d}}, \\ 0 & \text{otherwise.} \end{cases} \tag{4.12}$$

[3]For notation simplicity, we replace $v^i_{r,k,l}$ by v and $v^{i'}_{r',k',l'}$ by v'.

Now, let the weighted degree n_v of vertex v is defined by $n_v = \sum_{v' \in \mathcal{G}_{\text{d2d}}} \pi_{v,v'}.\psi(v')$, where $\psi(v')$ is the original weight of vertex v' defined in Theorem 4.8. Hence, the modified weight of vertex v is defined as

$$w(v) = \psi(v) n_v = \psi(v) \sum_{v' \in \mathcal{G}_{\text{d2d}}} \pi_{v,v'}.\psi(v'). \tag{4.13}$$

The modified weight of a vertex v in (4.13) points two attractive features: (i) it has a large original weight, and (ii) it is not adjacent to a large number of vertices that have high original weights. Based on this, we iteratively and heuristically execute a greedy vertex search scheme as follows. Initially, a vertex v^* that has the maximum weight $w(v^*)$ is selected and added to the maximal IS $\mathtt{I}$ (i.e., $\mathtt{I} = \{v^*\}$). Then, the subgraph $\mathcal{G}_{\text{d2d}}(\mathtt{I})$, which consists of vertices in graph $\mathcal{G}_{\text{d2d}}$ that are not adjacent to vertex v^*, is extracted and considered for the next process. In the next step, a new maximum weight vertex v'^* is selected from subgraph $\mathcal{G}_{\text{d2d}}(\mathtt{I})$ (at this point $\mathtt{I} = \{v^*, v'^*\}$). We repeat this process for all transmitting users so that no further vertex is not adjacent to all the vertices in $\mathtt{I}$. The selected D2D transmitters in the maximum IS $\mathtt{I}$ generate coded files and broadcast them to all neighboring users on D2D links.

The overall two-solution joint approach that are explained in Section 4.5.1 and Section 4.5.2, respectively, is provided in Algorithm 6.

Example 6: We illustrate in this example how to design the IA-IDNC and D2D conflict graphs that shown in Figure 4.4 of the network presented in Figure 4.3.

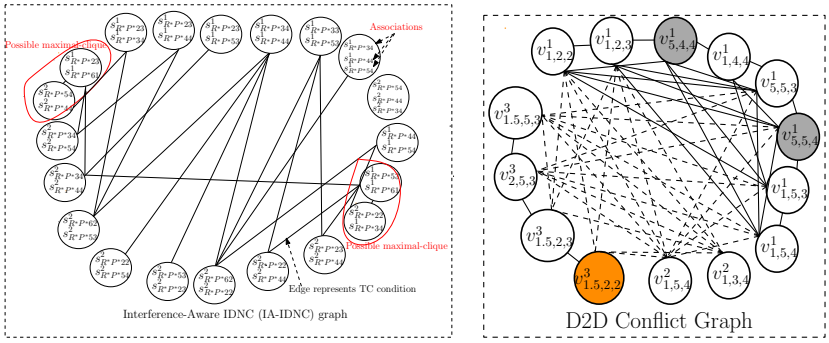
– In Figure 4.4(a), we plot the IA-IDNC graph, where each vertex represents a possible NC combination that consists of combined associations in each eRRH. For plotting simplification, we did not include the vertices that represents one association (i.e., no NC). The connections between vertices (circles) is based on the **TC** condition that explained in Section 4.5.1. There are many possible maximal PCs in $\mathcal{G}_{\text{IA-IDNC}}$ that are represented by connected vertices. Each one represents the potential network-coded scheduling of the eRRHs that minimizes the com-

Algorithm 6: Proposed Joint Approach

Require $\mathcal{N}$, $\mathcal{F}$, $\mathcal{K}$, $\mathcal{C}_{b_n}$, $\mathcal{H}_{u_l,0}$, $\mathcal{W}_{u_l,0}$, $P_{\max}$, P_{u_k}, N, $h^{c}_{b_n,u_l}, h^{d2d}_{u_k,u_l}$, $\forall u_k, u_l \in \mathcal{U}$, $\forall b_n \in \mathcal{B}$.
Initialize $\mathtt{C} = \varnothing$ and $\mathtt{I} = \varnothing$.
Solution of Subproblem P2: eRRH-user Transmission
Design $\mathcal{G}_{\text{IA-IDNC}}$ according to Section 4.5.1.
for each $\mathtt{S}$ **do**
    Calculate $\mathbf{P} = \{P^*_{b_1}, P^*_{b_2}, \cdots, P^*_{b_B}\}$ by solving (4.10).
    Obtain $V_i = \{(R^*_{b_i}, P^*_{b_i}, s_{b_i}), \cdots, (R^*_{b_i}, P^*_{b_i}, s_{|\tau(\kappa^c_{b_i})|})\}$ $(i = 1, \cdots, B)$ according to $\mathbf{P}$.
    Calculate $w(V_i)$ using (4.8).
end for
$\mathcal{G}_{\text{IA-IDNC}}(\mathtt{C}) \leftarrow \mathcal{G}_{\text{IA-IDNC}}$.
while $\mathcal{G}_{\text{IA-IDNC}}(\mathtt{C}) \neq \varnothing$ **do**
    $V^*_i = \arg\max_{V_i \in \mathcal{G}(\mathtt{C})}\{w(V_i)\}$.
    Set $\mathtt{C} \leftarrow \mathtt{C} \cup V^*_i$ and $\mathcal{G}_{\text{IA-IDNC}}(\mathtt{C}) \leftarrow \mathcal{G}_{\text{IA-IDNC}}(V^*_i)$.
end while
Solution of Subproblem P3: D2D Transmission
Design $\mathcal{G}_{\text{d2d}}$ according to Section 4.5.2.
$\mathcal{G}_{\text{d2d}}(\mathtt{I}) \leftarrow \mathcal{G}_{\text{d2d}}$.
while $\mathcal{G}_{\text{d2d}}(\mathtt{I}) \neq \varnothing$ **do**
    $\forall v \in \mathcal{G}_{\text{d2d}}(\mathtt{I})$: calculate $\psi(v)$ and $w(v)$ using (4.11) and (4.13), respectively.
    $v^* = \arg\max_{v \in \mathcal{G}_{\text{d2d}}(\mathtt{I})}\{w(v)\}$.
    Set $\mathtt{I} \leftarrow \mathtt{I} \cup v^*$ and obtain $\mathcal{G}_{\text{d2d}}(\mathtt{I})$.
end while
Obtain $\mathtt{C}$ and $\mathtt{I}$.

pletion time of users. For example, one possible maximal PC shown in red color in Figure 4.4(a) is $\{s^1_{R^*P^*53}, s^1_{R^*P^*61}, s^2_{R^*P^*22}, s^2_{R^*P^*34}\}$. The five indices $b_1, R^*, P^*, 5, 3$ in the first association represent first eRRH, its transmission rate and power level, scheduled user and its requested file, respectively.

– To ease the understanding of the D2D conflict graph, we plot it only for the first three users $\{u_1, u_2, u_3\}$ of the network presented in Figure 4.3 and irrespective to the possible scheduled users to the eRRHs.



(a) The IA-IDNC graph.

(b) The D2D conflict graph.

Figure 4.4: Illustration of both IA-IDNC and D2D conflict graphs of the example presented in Figure 4.3.

The D2D conflict graph is shown in Figure 4.4(b), where the conflict conditions **CC1** and **CC2** are represented by solid lines and conditions **CC3** and **CC4** are represented by dash lines. By Theorem 4.8, one possible maximal IS in this graph is $\{v^1_{5,4,4}, v^1_{5,5,4}, v^3_{1.5,2,2}\}$. The first vertex $v^1_{5,4,4}$ represents the transmitting user u_1, its rate $r = 5$, the scheduled user u_4 and its requested file f_4, respectively.

4.6 Coordinated Scheduling Solution for Completion Time Minimization Problem

In this section, we propose a faster and simpler coordinated scheduling approach. The need for this approach is invoked by the possibly large number of IDNC combinations generated by the joint approach in large network size. In such large networks, we can utilize this alternative approach to obtain fast and efficient solution. Let P_{b_n} be a fixed power level of b_n-th eRRH, $\forall b_n \in \mathcal{B}$. The completion time minimization problem at t-th transmission

slot can be written as a coordinated scheduling problem as follows

$$\text{P5}: \max_{\substack{\kappa^c_{b_n}\subseteq\mathcal{P}(\mathcal{H}_{b_n})\\ \tau(\kappa^c_{b_n})\cap\tau(\kappa^c_{b_{n'}})=\varnothing\\ R_{b_n}\in\mathcal{R}}} Z_t \tag{4.14a}$$

$$\text{subject to} \begin{cases} (\tau(\kappa^{d2d}_{u_k}), r_{u_k}) = \arg\min_{\substack{\mathcal{U}_{\text{tra},t}\in\mathcal{P}(\mathcal{U})\\ \kappa^{d2d}_k\in\mathcal{P}(\mathcal{H}_{u_k})\\ r_{u_k}\in\mathcal{R}_{u_k}}} \left\{\max_{u_l\in\mathcal{U}_{w,t}} \mathtt{T}_{u_l,t}\right\},\ \forall u_k\in\mathcal{U}_{\text{tra},t};\\ \tau(\kappa^{d2d}_{u_k})\cap\tau(\kappa^{d2d}_{u_{k'}})=\varnothing, \forall u_k\neq u_{k'}\in\mathcal{U}_{\text{tra},t};\\ \mathtt{f}^{d2d}_{u_k}\subseteq\mathcal{P}(\mathcal{H}_{u_k}); r_{u_k}\geq R_{\text{th}}. \end{cases} \tag{4.14b}$$

The objective function (4.14b) of problem P5 represents the possible completion time minimization that we can obtain from D2D transmissions and (4.7a) represents maximizing the number of users left to be scheduled to eRRHs.

To solve the problem in P5, we develop a simple and fast approach that first schedules users that have low channel capacities from eRRHs on D2D links, and the remaining unscheduled users (if any) can be scheduled by eRRHs with high transmission rates. This solution not only minimizes the completion time of users, but also offloads the eRRHs' radio resources. Indeed, after D2D transmissions, few users left to be scheduled by eRRHs. This approach is summarized in Algorithm 7, which consists of the following two stages at every transmission slot: i) users experience relatively weak channels from the eRRHs should be scheduled to the potential D2D transmitters on D2D links, and ii) the eRRHs deliver encoded files to a set of users that have not previously scheduled on D2D links. The coordinated scheduling approach is described below.

First stage: Here, we focus on scheduling a set of users that have low rates from the eRRHs to a set of potential transmitters via D2D links as such we solve problem (4.14b) efficiently.

First step: Inspired by Section 4.5.2, we follow the same procedures that construct the new D2D conflict graph $\mathcal{G}_{\text{d2d}}(\mathcal{V},\mathcal{E})$. Thus, we generate a ver-

tex $v^k_{r,l,f}$ for each transmitting user u_k, a transmission rate $r_{u_k} \in \mathcal{R}^{d2d}_{u_k,u_i}$ and a missing file $f_h \in \mathcal{H}_{u_k} \cap \mathcal{W}_{u_l}$ of a user $u_l \in \mathcal{U}_w \cap \mathcal{C}_{u_k}$. Further, the rate of each transmitting user in each generated vertex should be greater than or equal to R_{th}. Similarly, we generate the vertices for U users and then connect them as in Section 4.5.2.

Second step: We design two-layer weights for each generated vertex in the $\mathcal{G}_{\text{d2d}}$ graph, named by *secondary* and *primary* weights. The secondary weight of a vertex $v^k_{r,l,f}$ is defined as $w(v^k_{r,l,f}) = \frac{r_{u_k}}{B}$ that shows a partial contribution of that vertex in terms of reducing the completion time in the network. The primary weigh of a vertex v_{u_l,f_h} is defined as $w(v_{u_i,f_h}) = \frac{B}{\min_{b_n \in \mathcal{K}} R_{b_n,u_l}}, \forall f_h \in \mathcal{H}_{b_n}$. This primary weight characterizes the users based on their channel capacities from the eRRHs to give them priority to be scheduled on D2D links. A vertex with high primary weight (low rate from eRRHs) leads to prolonged file delivery time from eRRHs. Then the corresponding users of such vertices should be scheduled on D2D links with the maximum rate from any possible potential D2D transmitters. As such, the completion time of these users is minimized.

Third step: We propose to iteratively perform maximum weight search to form the set of D2D transmitters and their scheduled users in the maximal IS $\mathcal{I}$ as follows. First, we search for the vertex with the maximum primary weight and find its corresponding maximum secondary weight. If two or more vertices have equal weights, we select one vertex randomly. We continue this process until there are no other available vertices that can be included in the selected IS. At the end, the final IS $\mathcal{I}$ consists of vertices that represent a set of potential D2D transmitters. Each of these D2D transmitters serves users that have low channel capacities from the eRRHs.

Second stage: Here, we schedule users that are not scheduled on D2D links to eRRHs using RA-IDNC. In particular, we solve problem (4.7a) by maximizing the number of scheduled users to the eRRHs. First, we construct the coordinated scheduling graph, denoted by $\mathcal{G}_{\text{cord}}(\mathcal{E}, \mathcal{V})$, by generating a vertex $V_{b_n,u_i,f_h,R_{b_n}}$ for each $b_n \in \mathcal{B}$, for every file f_h is requested by user $u_i \in \mathcal{U}_{w,t} \backslash \mathcal{I}$, and for each achievable rate for that user $R_{b_n} \geq r_{\min}$, where $r_{\min}$ is the minimum selected transmission rate of any transmitting user in Γ.

Algorithm 7: Proposed Coordinated Scheduling Approach

Require $\mathcal{U}$, $\mathcal{F}$, N, $\mathcal{B}$, $\mathcal{H}_{b_n}$, $\mathcal{H}_{u_l,0}$, $\mathcal{W}_{u_l,0}$, P_{b_n}, P_{u_k}, $h^c_{b_n,u_l}, h^{d2d}_{u_k,u_l}$, $\forall u_k, u_l \in \mathcal{U}$, $\forall b_n \in \mathcal{B}$;
First stage
Initialize $\mathcal{I} = \varnothing$ and $\mathcal{G}_{\text{d2d}}(\mathcal{I}) \leftarrow \mathcal{G}_{\text{d2d}}$.
$\forall v^k_{r,l,h} \in \mathcal{G}_{\text{d2d}}$: calculate $w(v^k_{r,l,h}) = \frac{r_{u_k}}{N}$ and its corresponding $w(v_{u_l,f_h}) = \frac{N}{\min_{b_n \in \mathcal{K}} R_{b_n,u_l}}$.
while $\mathcal{G}_{\text{d2d}}(\mathcal{I}) \neq c$ **do**
    Choose the maximum primary weight $v^*_{u_l,f_h}$ and finds its corresponding maximum secondary weight $v^{k^*}_{r,l,h} = \arg\max_{v^k_{r,l,h} \in \mathcal{G}_{\text{d2d}}(\mathcal{I})} \{w(v^k_{r,l,h})\}$.
    Set $\mathcal{I} \leftarrow \mathcal{I} \cup v^{k^*}_{r,l,h}$ and $\mathcal{G}_{\text{d2d}}(\mathcal{I}) \leftarrow \mathcal{G}_{\text{d2d}}(v^{k^*}_{r,l,h})$.
end while
Second stage
Design $\mathcal{G}_{\text{cord}}$ according to Section 4.6.
$\forall V_{e_n,u_i,f_h,R_{e_n}} \in \mathcal{G}_{\text{cord}}$: calculate $w(V_{e_n,u_i,f_h,R_{e_n}}) = \dfrac{R_{e_n}}{B}$.
Obtain the maximum IS $\mathtt{I}$ as follows.
Initialize $\mathtt{I} = \varnothing$.
for each $V_{b_n,u_i,f_h,R_{e_n}} \in \mathcal{G}_{\text{cord}}$ non-conflicting with $\mathtt{I}$ **do**
    Select $V^*_{b_n,u_i,f_h,R_{e_n}} = \arg\max_{V_{b_n,u_i,f_h,R_{b_n}} \in \mathcal{G}_{\text{cord}}} \{w(V_{b_n,u_i,f_h,R_{b_n}})\}$.
    $\mathtt{I} \leftarrow \mathtt{I} \cup V^*_{b_n,u_i,f_h,R_{b_n}}$.
end for

The configuration of the set of edges in the scheduling graph is divided into coding (NC and rate edges) and transmission edge. Two vertices $V_{b_n,u_i,f_h,R_{b_n}}$ and $V_{b_n,u_{i'},f_{h'},R_{b_{n'}}}$ representing the same eRRH are adjacent by a conflict edge if they do not satisfy the IDNC and rate conditions in Section 4.5.1. Similarly, two vertices $V_{b_n,u_i,f_h,R_{b_n}}$ and $V_{b_{n'},u_{i'},f_{h'},R_{b_{n'}}}$ representing different eRRHs are adjacent by a conflict transmission edge if the same user u_i is scheduled to different eRRHs, i.e., $u_i = u_{i'}$ and $b_n \neq b_{n'}$. Then, a maximum search process is executed in $\mathcal{G}_{\text{cord}}$ to obtain the maximum IS $\mathtt{I}$ as presented in Algorithm 7.

4.7 Numerical Results

This section presents selected simulation results that compare the completion time performances of our proposed two schemes with baseline algorithms. We consider a downlink D2D-aided F-RAN system where the eRRHs have fixed locations and users are distributed randomly at every transmission within a hexagonal cell of radius 900m. We set the radius of the users' coverage zone `R` to 50m and the number of eRRHs B to 3. We consider the SUI-Terrain type B model in which the channel model of both F-RAN and D2D communications is mostly affected by the location of the users within the cell. Path loss is calculated as $148 + 40\log_{10}(\text{distance[km]})$. We consider that the channels are perfectly estimated. The noise power and the maximum' eRRH/user power are assumed to be -174 dBm/Hz and $P_{\max} = P = -42.60$ dBm/Hz, respectively. The bandwidth is 1 MHz and the eRRH caching ratio μ is 0.6. As discussed in Section 4.2, at the beginning of the D2D-aided F-RAN transmission, each user already has about 45% and 55% of F files. To assess the performances of our proposed schemes with different thresholds ($R_{\text{th1}} = 0.05$, $R_{\text{th2}} = 0.5$, and $R_{\text{th3}} = 5$), we simulate various scenarios with different number of users, number of files, and file sizes.

For the sake of comparison, our proposed schemes are compared with the baseline RLNC and Classical IDNC schemes. For completeness of our work, we also compare our proposed schemes with the uncoded schemes.

- **Uncoded Unicast**: This scheme schedules only one user to each eRRH from which it receives an uncoded file with its maximum transmission rate. In addition, the untargeted users by the eRRHs is served by implementing uncoded D2D transmissions.

- **Uncoded Broadcast-F-RAN**: The eRRHs broadcast uncoded files sequentially as such all users are served. In this scheme, each eRRH transmits with the lowest transmission rate of all scheduled users.

- **Uncoded Broadcast-D2D**: In this scheme, set of transmitting users is selected randomly to broadcast uncoded files from their *Has* sets

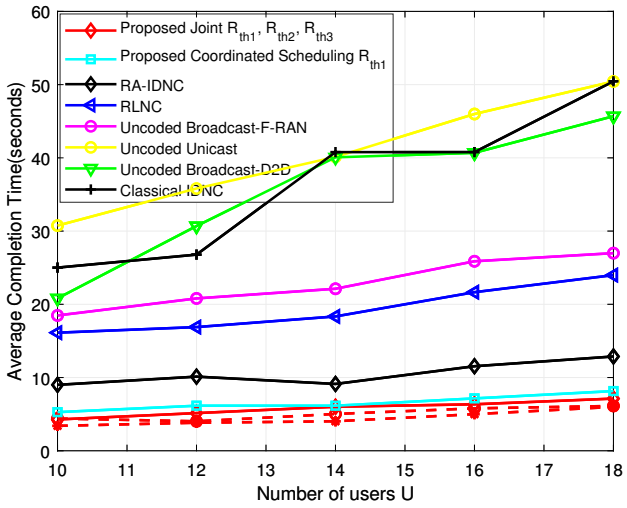


Figure 4.5: Average completion time versus the number of users U.

that are missing at the largest number of their neighboring users. The transmission rate is selected based on the minimum transmission of all scheduled users.

Recently, RA-IDNC scheme is studied in [82] where all eRRHs use the same transmission rate that corresponds to the minimum transmission rate of all scheduled users. In addition, the unscheduled users by the eRRHs are scheduled from transmitting users over D2D links with the same rate that is used by the eRRHs. Thus, we include the RA-IDNC and compare it with the proposed schemes.

In Figure 4.5, we depict the average completion time versus the number of users U. We consider a D2D-aided F-RAN model with a frame of 15 files and a file size of 1 Mbits. From this figure, it can be seen that the proposed schemes offer improved performance in terms of completion time reduction as compared to the other schemes. This improved performance is due to the joint and coordinated schemes that (i) judiciously schedule users, adopt the transmission rate of each eRRH and optimize the transmission power of each eRRH, and (ii) select potential users for transmitting coded

files over D2D links. In particular, the uncoded unicast suffers from targeting few users that have relative strong channel qualities. As a result, a higher number of transmissions, at least $(UF)/(B + |\mathcal{U}_{\text{tra}}|)$ transmissions, is needed for frame delivery completion, and it leads to a high completion time. Uncoded broadcast schemes suffer from serving all users at the cost of adopting the transmission rate of all eRRHs and transmitting users with the minimum transmission rate of the served users. Furthermore, uncoded broadcast D2D scheme offers a poor completion time performance as all transmitting users do not benefit from the transmission, i.e., they cannot transmit and receive at the same time. RLNC is a rate-less scheme that targets all users by sending encoded file with the lowest rate of all users. On the other hand, RA-IDNC scheme offers an improved performance compared to uncoded, RLNC, and classical IDNC schemes as mentioned in [82]. This is because the coding decisions in RA-IDNC scheme not only depends on the file combinations, but also on the channel qualities of the scheduled users. This effectively balances between the number of scheduled users and the transmission rate of eRRHs/transmitting users. However, selecting one transmission rate (the minimum rate) for all eRRHs and transmitting users degrades the completion time performance of the RA-IDNC scheme. This is a clear limitation of the RA-IDNC scheme in [82], as it does not fully exploit the typical variable channel qualities of the different eRRHs/transmitting users to their scheduled users. Our proposed joint and coordinated schemes fully utilize the eRRHs and transmitting users' resources to choose their own transmission rates, XOR combinations, and scheduled users. Consequently, a better performance of our proposed schemes compared to the RA-IDNC scheme is achieved. Moreover, the joint scheme optimizes the employed rates using power control on each eRRH. Thus, it works better than our proposed coordinated scheme. Note that the completion time performances of the classical IDNC and uncoded broadcast D2D schemes are of orders 10^5 and 10^3, respectively. Thus, we omit them from all the remaining figures.

In Figure 4.6, we show the average completion time versus the number of files F. The simulated D2D-aided F-RAN system composed of 15 users and file size of 1 Mbits. Again, for the above-mentioned reasons in Figure

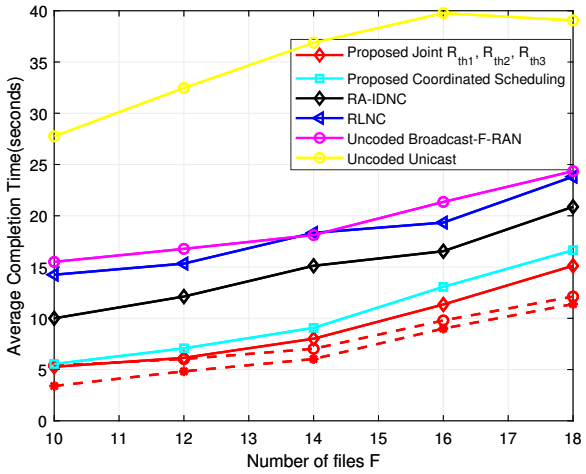


Figure 4.6: Average completion time versus the number of files F.

4.6, our proposed schemes outperform the other schemes. It can be observed from the figure that increasing the frame size leads to an increasing in the completion time of all schemes. The opportunities of mixing files using NC in the RA-IDNC and proposed schemes are limited with few files. Therefore, all NC schemes have roughly similar performances. As the number of files increases, the increase in the completion time with our proposed schemes is low. This is in accordance with our results in Theorem 4.6 and Theorem 4.8, where it is shown that our proposed schemes judiciously allow each eRRH and each transmitting user to decide on a set of files to be XORed. As such, they are beneficial to a significant set of users that have relatively good channel qualities. Even though uncoded broadcast and RLNC schemes completes file transmissions in fewer transmissions (F transmissions) than our proposed schemes, each of their transmission durations is longer than a single transmission of the proposed schemes. Thanks to their optimized higher rate/power and users' transmission that result in less completion time.

In Figure 2.6, we illustrate the impact of increasing the file size N on the average completion time. In this figure, we simulate the D2D-aided F-

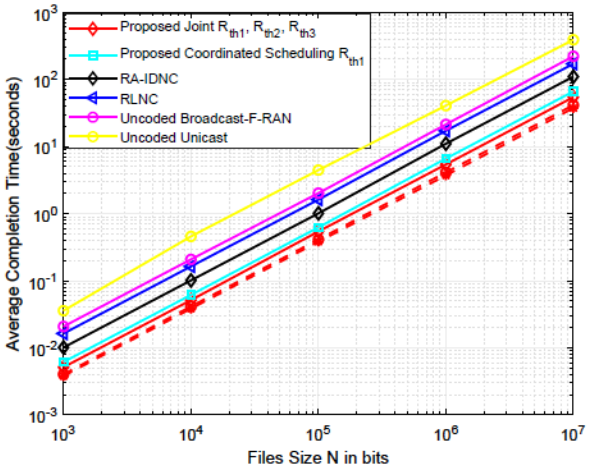


Figure 4.7: Average completion time versus file size N.

RAN system composed of 12 users and 15 files. We can observe that the performances of all schemes increase linearly with the file size. This is in accordance with the completion time expression in Corollary 4.5, where it was emphasized that $\mathtt{T}_o$ increases linearly with N. From physical-layer view, as N increases, more bits are needed for delivering files. Thus, time delay is increased to receive files from eRRHs/transmitting users.

Finally, some insights from our presented numerical results are given as follows. First, it is advantageous to serve many users with NC files as in the classical IDNC algorithm, but selecting the minimum transmission rate of all scheduled users degrades its performance. Thus, this scheme is impractical from physical-layer perspective. Second, although the uncoded unicast scheme uses the maximum transmission rate of each user, it needs a large number of transmissions for completion. Thus, its completion time performance is degraded. Third, RA-IDNC scheme overcomes the limitations of the aforementioned schemes but suffers from selecting the lowest-rate of the fixed-power eRRHs in the system. This limitation further degrades the completion time performance of the RA-IDNC scheme in large network sizes with large number of eRRHs. This due to the fact that RA-IDNC always

selects the lowest-rate of all eRRHs. Conversely, our transmission framework is more practically relevant as it enables different transmission rates from different eRRHs/transmitting users and optimizes the employed rates using power control on each eRRH.

4.8 Chapter Summary

In this chapter, we have developed two frameworks that exploit the cached contents at eRRHs, their transmission rates, and previously received contents by different users to deliver the requesting contents to users with a minimum completion time in D2D-aided F-RAN. Simulation results have shown that our proposed frameworks can effectively minimize the frame delivery time as compared to conventional algorithms.

Chapter 5

Coalition Formation Game for Cooperative Content Delivery in Network Coding Assisted D2D Communications

In Chapter 4, we considered a centralized system of D2D-aided F-RAN, where there is a centralized computing unit called the CBS. This CBS is responsible for making the NC decisions, rate adaptation, delivering the instructions to transmitting users and eRRHs for executions. The CBS is assumed to be within the transmission range of all users and has perfect knowledge of the network topology. Further, in order to make NC decisions for the transmitting eRRHs/users, the CBS has to track the downloading history of all users after each transmission which requires a high computation capability at the CBS. This is expensive for service providers to implement fixed infrastructures of many distributed CBS units. Importantly, in some scenarios of an extremely high user density, the CBS is fully loaded. To this end, in this chapter, we consider D2D networks to develop a distributed and decentralized framework that overcomes all the aforementioned limitations of the centralized system. In fact, one of the major directions towards 5G networks is D2D communications and cooperative packet recovery among users. This will offload traffic from the CBS can thus free up some spectrum for it to serve more users. The organization of this chapter is given as fol-

lows. The accomplished works and research contributions are summarized in Section 5.1. The D2D system model and problem formulation are discussed in Section 5.2. In this section, we provide an example for illustrating the completion time minimization. In Section 5.3, we model the completion time minimization problem as a coalition game. Section 5.4 proposes a fully distributed coalition formation game solution using merge and split rules. In Section 5.5, we analyze the convergence, stability, complexity, and communication overhead of the proposed algorithm. In Section 5.6, we evaluate the completion time and game performances of our proposed solution, and in Section 5.7, we conclude this chapter.

5.1 Accomplished Works and Research Contributions

The contributions of this chapter are summarized as follows. First, for partially D2D networks, we develop a decentralize IDNC-assisted D2D game theoretical framework to minimize the completion time while offloading the CBSs. In particular, the completion time minimization problem is formulated and modeled as a coalition game. Given the difficulty of expressing the problem as a coalition game with non-transfer function (NTU), we relax it to a coalition formation game (CFG). Second, given the relaxed CFG, we express the completion time metric as the utility function, which is transferred to each player's payoff in each coalition. We, then, derive the rules for associating players, selecting the transmitting player, and finding its packet combination that is beneficial for a set of interested UDs in each disjoint altruistic coalition. Afterward, we develop a coalition formation distributed algorithm in each transmission slot based on merge-and-split rules. Finally, we prove that the proposed coalition formation algorithm is converged to a Nash-stable equilibrium, and we theoretically analyze its complexity and communication overhead. We validate our theoretical finding using comprehensive numerical simulations, which reveal that our distributed scheme can significantly outperform existing centralized and fully distributed methods.

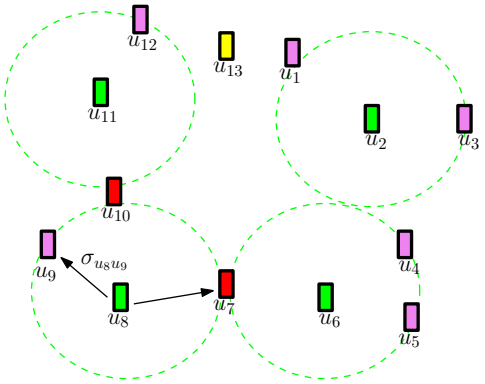


Figure 5.1: Illustration of a partially connected D2D network with 13 UDs. For simplicity, only the coverage zones of the possibly transmitting UDs u_2, u_6, u_8, and u_{11} are drawn.

For the presented network setups, our proposed decentralized scheme offers almost the same performance as the centralized F-RAN D2D scheme.

5.2 System Overview and Problem Formulation

5.2.1 System Overview

Consider a D2D-enabled wireless network, shown in Figure 5.1, that consists of U UDs, denoted by the set $\mathcal{U} = \{u_1, u_2, \cdots, u_U\}$, that are interested in receiving a delay-sensitive content representing a frame of M packets, denoted by the set $\mathcal{P} = \{p_1, p_2, \cdots, p_M\}$. The size of the frame $\mathcal{P}$ depends on the sizes of the packet and content. This work assumes that UDs have already acquired some packets in $\mathcal{P}$ from previous transmissions, i.e., $\mathcal{H}_{u_k}$ denotes the *Has* set of packets acquired by u_k-th UD. We assume UDs request a set of packets, which is not included in their *Has* sets, from the frame $\mathcal{P}$. This requested set will be referred as *Wants* set, i.e., $\mathcal{W}_{u_k} = \mathcal{P} \setminus \mathcal{H}_{u_k}$ is set of wanted packets for u_k-th UD. The *Has* and *Wants* sets of all UDs can be summarized in a binary $U \times M$ *state matrix* $\mathbf{S} = [s_{u_k,p_l}]$ wherein the entry $s_{u_k,p_l} = 0$ represents that the p_l-th packet is successfully received at u_k-th UD and 1 otherwise. In order for all UDs to obtain the whole frame $\mathcal{P}$ from

D2D communications, we assume that each packet $p_l, 1 \leq l \leq M$ is received by at least one UD, i.e., $\sum_{u_k \in \mathcal{U}} s_{u_k,p_l} \geq 1$ for all packets $p_l \in \mathcal{P}$. At each transmission slot, each UD can be packets requester (receiver) or packets holder (transmitter) that providing its received packets to other requesting UDs via D2D links. When any UD receives all its requested packets, it most likely acts as a transmitter to transmit its received packets to interested UDs.

This work considers a realistic partial network topology, where UDs can only target the subset of UDs in their coverage zone, denoted by $\mathcal{C}_{u_k}$ of u_k-th UD. The network topology can be captured by a unit diagonal symmetric $U \times U$ adjacency matrix $\mathbf{C}$ represents the connectivity of the UDs such that $\mathbf{C}_{u_k u_l} = 1$ if and only if $u_l \in \mathcal{C}_{u_k}$. The matrix $\mathbf{C}$ is updated after each transmission based on the connectivity index C, where C is a fractional connection ratio. We assume that no part of the network is disjoint, i.e., the matrix $\mathbf{Z}$ is connected. Otherwise, the proposed algorithm is separately applied to each independent part of the network. Upon successful reception of a packet, each UD sends an error-free acknowledgment (ACK) to all UDs in its coverage zone to update their *Has* and *Wants* sets. Therefore, maintaining a feedback matrix of neighboring UDs required low complexity and low signaling overhead than the fully connected model.

We focus only on the upper layer view of the network, where network coding is performed at the network layer, and a memory-less erasure channel abstracts the physical-layer. This abstraction is widely used in network coding literature, where a packet is either perfectly received or completely lost with certain average probability [44], [61], [62], [63], [64], [65], [66], [67], [68], [69]. Therefore, the physical channel between the u_k-th and u_l-th UDs is modeled by a Bernoulli random variable whose mean $\sigma_{u_k u_l}$ indicates the packet erasure probability from u_k-th UD to u_l-th UD. We assume that these probabilities remain constant during the transmission of a single packet $p_l \in \mathcal{P}$, and they are known to all UDs. Due to the channel's asymmetry between u_k-th and u_l-th UDs, the equality of $\sigma_{u_k u_l}$ and $\sigma_{u_l u_k}$ is not guaranteed. To clarify the system model and its most variables, Figure 5.1 shows a partially connected D2D network with 13 UDs. The coverage zones of the possibly

transmitting UDs and their targeted UDs are shown in Figure 5.1.

We consider a slowly changing network topology, in which UDs have fixed locations during the IDNC packet transmission and change from one transmission to another. However, after one transmission, the UDs can move, and all the network variables will be updated. Our model, i.e., the coalition formation solution, can be used with updated network parameters. It is important to note that in fully connected networks, each UD is connected to all other UDs in the network, and hence, it precisely knows the *Has* and *Wants* sets of all UDs. To avoid any collision in the network, only one UD is allowed to transmit an encoded packet at any transmission slot. In partial connected D2D networks, single-interface UDs can receive two transmissions from different transmitting UDs, but this causes collisions. This collision is due to two or more UDs transmit simultaneously using the same radio resource block, e.g., frequency. As such, the receiving UD cannot decode these collided packets correctly, see for examples [20], [21], [22], [23],, [24] and their references. Therefore, UDs are grouped in disjoint and independent coalitions to avoid any collisions or conflict transmissions. Considering the interference of transmissions caused by other UDs to the set of transmitting UDs can be pursued in future work.

5.2.2 Instantly Decodable Network Coding Model

IDNC XOR combination enables the complete packet reception at every time slot when each scheduled UD has at most one packet in that XOR combination. Let $\mathtt{p}_{u_k}$ be an XOR combination sent by the u_k-th UD to the set of scheduled UDs $\mathtt{u}(\mathtt{p}_{u_k})$. Note that the packet combination $\mathtt{p}_{u_k} \subset \mathcal{H}_{u_k}$ is an element of the power set $\mathcal{X}$ of the received packets at the u_k-th UD. A transmission from the u_k-th UD is instantly decodable at the u_l-th UD if: i) it contains only one packet from the *Wants* set of the u_l-th UD, and ii) the u_l-th UD is in the coverage zone of the u_k-th UD. Therefore, the set of targeted[4] UDs by u_k-th UD is expressed as $\mathtt{u}(\mathtt{p}_{u_k}) = \left\{u_l \in \mathcal{U} \,\middle|\, |\mathtt{p}_{u_k} \cap \mathcal{W}_{u_l}| = 1 \text{ and } u_l \in \mathcal{C}_{u_k}\right\}$. In that case, u_l-th UD

[4]The term "targeted UDs" is given for a set of scheduled UDs who receives an instantly-decodable transmission.

can re-XOR the combination $\mathtt{p}_{u_k}$ with $\mathcal{H}_{u_l}$ to retrieve one of its requesting packets. Hence, we say that u_k-th UD is a transmitting UD that provides a set of its received packets to a set of targeted UDs.

Let $\mathcal{U}_{\text{tra},t} \subset \mathcal{U}$ denote the set of transmitting UDs at the t-th transmission and $\underline{\mathtt{p}}(\mathcal{U}_{\text{tra},t}) = (\mathtt{p}_{u_1}, \cdots, \mathtt{p}_{u_{|\mathcal{U}_{\text{tra},t}|}})$ denote the packet combinations to be sent by UDs in $\mathcal{U}_{\text{tra},t}$. For notation simplicity, the transmission index t is often omitted when it is clear from the context. Let $\underline{\mathtt{u}}(\underline{\mathtt{p}}(\mathcal{U}_{\text{tra}})) = (\mathtt{u}(\mathtt{p}_{u_k}), \cdots, \mathtt{u}(\mathtt{p}_{u_{|\mathcal{U}_{\text{tra}}|}}))$ denote the set of targeted UDs by the transmitting UDs wherein $u_l \in \mathtt{u}(\mathtt{p}_{u_k}(\mathcal{U}_{\text{tra}}))$ implies that $|\mathcal{W}_{u_l} \cap \mathtt{p}_{u_k}(\mathcal{U}_{\text{tra}})| = 1$ and $\{u_l\} \cap \mathcal{C}_{u_k} \cap \mathcal{C}_{u_m} = \delta_{u_k u_m}\{u_l\}$ for all transmitting UDs $u_m \in \mathcal{U}_{\text{tra}}$ wherein $\delta_{u_k u_m}$ is the Kronecker symbol. Mathematically, $\delta_{u_k u_m}\{u_l\} = u_l$ if $u_k = u_m$ and 0 otherwise.

Definition 5.1. *The completion time of u_k-th UD, denoted by $\mathtt{T}_{u_k}$, is the number of D2D transmissions required to get all its packets in $\mathcal{W}_{u_k}$. The overall completion time $\mathtt{T} = \max_{u_k \in \mathcal{U}}\{\mathtt{T}_{u_k}\}$ represents the time required for all UDs to get all the packets.*

5.2.3 Completion Time Minimization Problem Formulation

In this subsection, we formulate the distributed completion time reduction problem in IDNC-enabled D2D network. Let $\underline{\mathtt{U}}_w$ be a binary vector of size U whose u_k-th index is 1 if u_k-th UD has non-empty *Wants* set, i.e., $\mathcal{W}_{u_k} \neq \varnothing$ and 0 otherwise, and let $\underline{\bar{\mathtt{u}}}(\underline{\mathtt{p}}(\mathcal{U}_{\text{tra}})) = \underline{1} - \underline{\mathtt{u}}(\underline{\mathtt{p}}(\mathcal{U}_{\text{tra}}))$ be the set of the non-targeted UDs by the encoded packets $\underline{\mathtt{p}}(\mathcal{U}_{\text{tra}})$. The different erasure occurrences at the t-th transmission slot are denoted by $\boldsymbol{\omega} : \mathbb{Z}_+ \to \{0,1\}^{U \times U}$ with $\boldsymbol{\omega}_t = [Y_{u_k u_l}]$, for all $(u_k, u_l) \in \mathcal{U}^2$, where $Y_{u_k u_l}$ is a Bernoulli random variable equal to 0 with probability $\sigma_{u_k u_l}$.

Let $\underline{\mathtt{n}}_t = (n^{[u_1]}, n^{[u_2]}, \cdots, n^{[u_u]})$ be a binary vector of length U whose $n^{[u_k]}$-th element is equal to 1 if u_k-th UD is transmitting, i.e., $\|\underline{\mathtt{n}}\|_1 = |\mathcal{U}_{\text{tra}}|$. Likewise, let $\underline{\mathcal{D}}(\underline{\mathtt{n}}_t)$ be the decoding delay experienced by all UDs in the t-th transmission slot. In particular, $\underline{\mathcal{D}}(\underline{\mathtt{n}}_t)$ is a metric quantifies the ability of the transmitting UDs to generate innovative packets for all the targeted UDs. This metric increases by one unit for each UD that still wants packets

and successfully receives a nonuseful transmission from any transmitting UD in $\mathcal{N}_{\text{tra}}$ or for a transmitting UD that still wants some packets. Let $\underline{\mathcal{I}} = (\mathcal{I}^{[u_1]}, \mathcal{I}^{[u_2]}, \cdots, \mathcal{I}^{[u_U]})$ be a binary vector of size N whose $\mathcal{I}^{[u_l]}$ entry is 1 if u_l-th UD is hearing more than one transmission from the set $\mathcal{U}_{\text{tra}}$, i.e., $u_l \in \mathcal{C}_{u_k} \cap \mathcal{C}_{u_m}$ where $u_k \neq u_m \in \mathcal{U}_{\text{tra}}$ and 0 otherwise, and let $\underline{\mathcal{O}} = (\mathcal{O}^{[u_1]}, \mathcal{O}^{[u_2]}, \cdots, \mathcal{O}^{[u_U]})$ be a binary vector of size U whose $\mathcal{O}^{[u_l]}$ element is 1 if u_l-th UD is out of transmission range of any transmitting UD in $\mathcal{U}_{\text{tra}}$, i.e., $u_l \notin \mathcal{C}_{u_k}, \forall u_k \in \mathcal{U}_{\text{tra}}$ and 0 otherwise.

Given the above configurations, the overall decoding delays $\underline{\mathbb{D}}(\underline{\mathtt{n}}_t)$ experienced by all UDs, since the beginning of the delivery phase until the t-th transmission, can be expressed as follows.

$$\underline{\mathbb{D}}(\underline{\mathtt{n}}_t) = \underline{\mathbb{D}}(\underline{\mathtt{n}}_{t-1}) + \begin{cases} \underline{U} & \text{if } \|\underline{\mathtt{n}}_t\|_1 = 0 \\ \underline{\mathcal{I}} + \underline{\mathcal{O}} + \underline{\mathtt{n}}_t + \underline{\mathcal{D}}(\underline{\mathtt{n}}_t) & \text{otherwise.} \end{cases} \tag{5.1}$$

As mentioned, the completion time is a difficult and intractable metric to optimize. However, in network coding literature, such metric is approximated by the *anticipated* completion time which can be computed at each transmission using the decoding delay. The anticipated completion time that uses the decoding delay in (5.1) is defined as follows.

Definition 5.2. *The anticipated completion time of the u_k-th UD is defined by the following expression*

$$\mathtt{T}_{u_k}(\underline{\mathtt{n}}_t) = \frac{|\mathcal{W}_{u_k,0}| + \mathbb{D}_{u_k}(\underline{\mathtt{n}}_t) - \mathbb{E}[\sigma_{u_k}]}{1 - \mathbb{E}[\sigma_{u_k}]}, \tag{5.2}$$

where $|\mathcal{W}_{u_k,0}|$ is the number of the requested packets in the Wants set of u_k-th UD at the beginning of delivery phase and $\mathbb{E}[\sigma_{u_k}]$ is the expected erasure probability linking u_k-th UD to the other UDs in its coverage.

Clearly, (5.2) represents the number of D2D transmissions that are required for u_k-th UD to receive all packets in $\mathcal{W}_{u_k}$. In this context, completion time is intimately related to the throughput of the system. Throughput is measured as the number of cooperative D2D transmission rounds required by

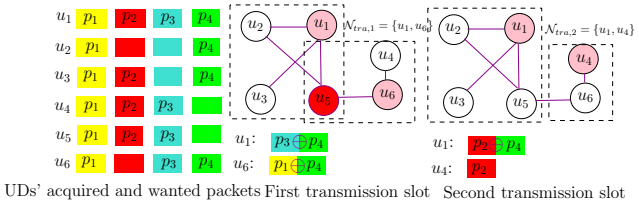


Figure 5.2: A partially connected D2D network containing 6 UDs and 4 packets.

the UDs to receive all their requested packets. The overall anticipated completion time can be written as $\mathtt{T}(\underline{\mathtt{n}}_t) = \max_{u_k}(\mathtt{T}_{u_k}(\underline{\mathtt{n}}_t)) = \|\underline{\mathtt{T}}(\underline{\mathtt{n}}_t)\|_\infty$. Therefore, the anticipated completion time minimization problem at the t-th transmission slot in IDNC-enabled partially D2D network can be written as follows.

$$\min_{\substack{\underline{\mathtt{n}}_t \in \{0,1\}^U \\ \underline{\mathtt{p}}(\mathcal{U}_{\text{tra}}) \in \{0,1\}^M}} \|\underline{\mathtt{T}}(\underline{\mathtt{n}}_t)\|_\infty. \tag{5.3}$$

Unlike a fully connected model that only requires optimization over a single transmitting UD and its corresponding packet combination, a partially connected model needs to select the set of transmitting UDs $\mathcal{U}_{\text{tra}}$ and the encoded packets $\underline{\mathtt{p}}(\mathcal{U}_{\text{tra}})$. As such, the probability of increasing the anticipated completion time is minimized.

Example of IDNC Transmissions in a Partially Connected D2D Network

We illustrates the aforementioned definitions and concepts with a simple example. Consider a simple partially connected D2D network containing 6 UDs and a frame $\mathcal{P} = \{p_1, p_2, p_3, p_4\}$ as illustrated in Figure 5.2. The acquired and wanted packets of all UDs are given on the left part of Figure 5.2, and the coverage zone of each UD is represented by edges. For ease of analysis, we assume error-free transmissions. Assume that u_1 transmits the combination $\mathtt{p}_{u_1} = p_3 \oplus p_4$ to u_2, u_3, u_5 and u_6 transmits the combination $\mathtt{p}_{u_6} = p_1 \oplus p_4$ to u_4, u_5 in the first transmission slot. In the second trans-

mission slot, assume that u_4 transmits $\mathtt{p}_{u_4} = p_2$ to u_6, and u_1 transmits $\mathtt{p}_{u_1} = p_2 \oplus p_4$ to u_2, u_5. The decoding delay experienced by the different UDs is given as follows.

- The u_5-th UD experiences one unit delay as it is in the intersection of the coverage zone of transmitting UDs u_1 and u_6. In other words, u_5 is in collision, i.e., $u_5 \in \underline{\mathcal{I}}$. Thus, u_5-th UD would not be able to decode packet $\mathtt{p}_6$ transmitted by the u_6-th UD.

- The u_6-th UD experiences one unit of delay as it transmits in the first transmission slot.

Under this scenario, we have the following assumption.

- **First transmission slot:** $\underline{\mathrm{U}}_w = (0\ 1\ 1\ 1\ 1\ 1)$, the set of transmitting UDs $\mathcal{U}_{\text{tra},1} = \{u_1, u_6\} = \underline{\mathrm{n}}_1 = (1\ 0\ 0\ 0\ 0\ 1)$, the corresponding encoded packets $\underline{\mathtt{p}}(\mathcal{U}_{\text{tra},1}) = (\mathtt{p}_{u_1}, \mathtt{p}_{u_6})$, and the set of targeted UDs $\underline{\mathtt{u}}(\underline{\mathtt{p}}(\mathcal{U}_{\text{tra},1})) = (\mathtt{u}(\mathtt{p}_{u_1}), \mathtt{u}(\mathtt{p}_{u_6})) = \{(u_2, u_3), (u_4)\}$. The set of UDs that hearing more than one transmission $\underline{\mathcal{I}} = (0\ 0\ 0\ 0\ 1\ 0)$, and the set of UDs that out of transmission range of any UD in $\mathcal{U}_{\text{tra},1}$ is $\underline{\mathcal{O}} = \underline{0}$. The decoding delay experienced by all UDs is $\underline{\mathcal{D}}(\underline{\mathrm{n}}_1) = (0\ 0\ 0\ 0\ 1\ 1)$. The accumulative decoding delay is $\underline{\mathbb{D}}(\underline{\mathrm{n}}_1) = (0\ 0\ 0\ 0\ 1\ 1)$.

- **Second transmission slot:** $\underline{\mathrm{N}}_w = (0\ 1\ 0\ 0\ 1\ 1)$, the set of transmitting UDs $\mathcal{U}_{\text{tra},2} = \{u_1, u_4\} = \underline{\mathrm{n}}_2 = (1\ 0\ 0\ 1\ 0\ 0)$, the corresponding encoded packets $\underline{\mathtt{p}}(\mathcal{U}_{\text{tra},2}) = (\mathtt{p}_{u_1}, \mathtt{p}_{u_4})$, and the set of targeted UDs $\underline{\mathtt{u}}(\underline{\mathtt{p}}(\mathcal{U}_{\text{tra},2})) = (\mathtt{u}(\mathtt{p}_{u_1}), \mathtt{u}(\mathtt{p}_{u_4})) = \{(u_2, u_5), (u_6)\}$. The set of UDs hearing more than one transmission $\underline{\mathcal{I}} = \underline{0}$, and the set of UDs that out of transmission range of any UD in $\mathcal{U}_{\text{tra},2}$ is $\underline{\mathcal{O}} = \underline{0}$. The decoding delay is $\underline{\mathcal{D}}(\underline{\mathrm{n}}_2) = \underline{0}$ and the accumulative decoding delay $\underline{\mathbb{D}}(\underline{\mathrm{n}}_2) = (0\ 0\ 0\ 0\ 1\ 1)$.

- The completion time of all UDs after the second transmission is $\mathtt{T} = (0\ 2\ 1\ 1\ 2\ 2)$. Thus, the maximum completion time is 2 transmission slots which represents the overall completion time for all UDs to get their requested packets, i.e., $\underline{\mathrm{U}}_w = \underline{0}$.

5.3 Distributed Completion Time Minimization as a Coalition Game

Our main goal is to develop a distributed framework that models the collaborations among the UDs of IDNC-enabled partially D2D networks. In order to do that, we use game theory because it involves a set of players that interact with each other to form a coalition without any coordination from the CP. The key idea of forming coalitions is to study the cooperative behavior of UDs in coalitions. In particular, UDs are formed coalitions so that they minimize their completion times compared to when they do not form coalitions as will be discussed in Section 5.4. Through coalitions and IDNC content delivery among players, the overall completion time can be reduced.

In this section, we first model the completion time problem in IDNC-enabled partially connected D2D networks using coalition games [83]. Afterward, we define and provide fundamental concepts in coalition games. These concepts are used in Section 5.4 to derive the proposed distributed completion time reduction solution.

Completion Time Minimization as a Coalition Game

We use a coalition game to formulate the completion mentioned above time problem. In particular, our problem is modeled as a coalition game with a non-transferable utility (NTU) [83].

Definition 5.3. *A coalition game with a non-transferable utility is defined as a pair $(\mathcal{N}, \phi)$, where $\mathcal{N}$ is the set of players consisting of N UDs and ϕ is a real function such that for every coalition $\mathcal{S}_s \subseteq \mathcal{N}$, $\phi(\mathcal{S}_s)$ is the payoff that coalition $\mathcal{S}_s$ receives which cannot be arbitrarily apportioned between its players.*

For the considered D2D completion time problem with colaition games, packets are transmitted simultaneously from a set of cooperative transmitting UDs $\mathcal{N}_{\text{tra}}$. Each transmitting UD serves a group of interested and nearby UDs that constitutes a coalition. In the coalition game, each UD

acts as a game player and aims to join a coalition such that: i) it can receive packets and minimize its completion time, ii) it does not increase the completion time of its alliances in the coalition. Therefore, for each coalition, we need to determine the transmitting player and its packet combination that is beneficial to a set of interested players. As such, we minimize the increasing of the completion time.

Let $\phi(\mathcal{S}_s) = (\phi_{u_1}(\mathcal{S}_s), \cdots, \phi_{u_{|\mathcal{S}_s|}}(\mathcal{S}_s))$ define a tuple wherein element $\phi_{u_k}(\mathcal{S}_s)$ represents the payoff of u_k-th UD in coalition $\mathcal{S}_s$, where $\mathcal{S}_s \subseteq \mathcal{N}$, and $|\mathcal{S}_s|$ is the total number of players in $\mathcal{S}_s$. We use s as the subscript to identify coalitions. The $|\mathcal{S}|$-dimensional vector represents the family of real vector payoffs of coalition $\mathcal{S}_s$, which is denoted by $\underline{\phi}(\mathcal{S}_s)$. Consequently, by adopting the cooperative D2D completion model described in the previous section, the total payoff of any coalition $\mathcal{S}_s \subseteq \mathcal{U}$, $\forall s = \{1, \cdots, m\}$ is given by

$$\phi(\mathcal{S}_s) = \max_{u_k}(\phi_{u_k}(\mathcal{S}_s)) = \|\underline{\phi}(\mathcal{S}_s)\|_\infty, \tag{5.4}$$

where $\phi_{u_k}(\mathcal{S}_s)$ is the payoff of u_k-th UD in a coalition $\mathcal{S}_s$ which is in our problem given by

$$\phi_{u_k}(\mathcal{S}_s) = -\|\mathtt{T}_{u_k}(\underline{\mathtt{n}}_t)\|_\infty - \|\mathbb{D}_{u_k}(\underline{n}_t) - \mathbb{D}_{u_k}(\underline{n}_{t-1})\|_1. \tag{5.5}$$

The payoff function in (5.4) represents the total payoff that a coalition receives due to self-organize players. For a player $u_k \in \mathcal{S}_s$, the first term in (5.5) represents the maximum anticipated completion time among players in $\mathcal{S}_s$ that is defined in (5.2). Similarly, the second term in (5.5) represents the augmentation of the sum decoding delay that is defined in (5.1). Therefore, players in coalitions prefer to increase the payoff in (5.5) by minimizing the anticipated completion time through controlling the decoding delay. Note that the negative sign indicates that the payoff of a coalition is inversely proportional to the completion time.

Property 1: The proposed D2D completion time cooperative problem is modeled as a coalition game with NTU $(\mathcal{U}, \phi)$ where $\mathcal{U}$ is the set of players

and ϕ is the payoff function given by (5.4).

Proof. From the nature of definition 1 and definition 2, the u_k-th player has its own unique anticipated completion time and decoding delay, and, thus, it has a unique payoff $\phi_{u_k}(\mathcal{S}_s)$ within a coalition $\mathcal{S}_s$. Therefore, the payoff function in (5.4) cannot be arbitrarily apportioned between coalition's players. Thus (5.4) is considered as an NTU. Further, the overall completion time is the maximum individual completion times of the players regardless of the coalition. In other words, the dependency of $\phi(\mathcal{S}_s)$ in any coalition structure is not only on packet delivery of players inside $\mathcal{S}_s$, but also on packet delivery outside $\mathcal{S}_s$, which concludes that the proposed game model is NTU game. □

Although cooperation generally reduces the payoffs of players [38], it is limited by inherent information exchange cost that needs to be paid by the players when acting cooperatively. Consequently, for any coalition $\mathcal{S}_s \subseteq \mathcal{U}$, players need to exchange information for cooperation, which is an increasing function of the coalition size. The problem becomes severe when all players are in the same coalition, i.e., grand coalition (GC). However, given the realistic scenario of a partially connected network where each UD has limited coverage, it is highly likely that when attempting to form the GC, one of these scenarios will hold:

- There exists a pair of u_k-th, u_l-th players in $\mathcal{U}$ that are distant enough to receive packets within the GC. Thus, they have no incentive to join the grand coalition.
- There exists a player $u_k \in \mathcal{U}$ with a payoff in GC $\phi_{u_k}(\mathcal{U})$ that is less than its payoff in any coalition $\phi_{u_k}(\mathcal{S}_s)$. Hence, this player has an incentive to deviate from the GC.

Further, the limited coverage players in the considered partially connected D2D model would most likely form small coalitions' sizes, not large coalitions' sizes. Therefore, the GC of all players is *seldom* formed, and the cooperation cost due to forming small coalitions' sizes would not have a significant impact on the payoff functions.

The aforementioned discussion emphasizes that the proposed $(\mathcal{N}, \phi)$ game is classified as a coalition formation game (CFG) [134], where players form several independent disjoint coalitions. Hence, classical solution concepts for coalition games, such as the core [83], may not apply to our problem. In summery, the proposed coalition game $(\mathcal{U}, \phi)$ is a CFG, where the objective is to develop an algorithm for forming coalitions.

Coalition Formation Concepts

This section provides some concepts of coalition formation games that are used in the next subsection. CFG, a subclass of coalition games, has been a topic of high interest in game theory research [134], [135], [138]. The fundamental approach in coalition formation games is to allow players in the formation set to join or leave a coalition based on a well-defined and most suitable *preference* for NTU games, i.e., *Pareto Order*. Pareto Order is the basis of many existing coalition formation concepts, e.g., the merge-and-split algorithm [136].

Definition 5.4. *A coalition structure, denoted as* Ψ*, is defined as* $\Psi = \{\mathcal{S}_1, \cdots, \mathcal{S}_m\}$ *for* $1 < |\mathcal{S}_m| < |\mathcal{U}|$ *independent disjoint coalitions* $\mathcal{S}_m$ *of* Ψ*.*

One can see from definition 5 that different coalition structures may lead to different system payoffs as each coalition structure Ψ has its unique payoff $\phi(\Psi)$. These differences in Ψ and their corresponding payoffs $\phi(\Psi)$ are usually ordered through a comparison relationship. In the coalition game literature, e.g., [136], comparison relationships based on orders are divided into individual value orders, and coalition value orders. Individual order implies that comparison is performed based on the players' payoffs. This is referred to as the Pareto Order. In particular, in such order, no player is willing to move to another coalition when it has a negative effect on the payoff of that coalition. In other words, the payoff of players would be worse off after the new player joins. This is known as selfish behavior. Coalition order implies that two coalition structures are compared based on the payoff of the coalitions in these coalition structures. This is known as a utilitarian order and is denoted by $\triangleright$. In other words, the notation $\Psi_2 \triangleright \Psi_1$ means that

$\phi(\Psi_1) > \phi(\Psi_2)$. Subsequently, the definition of the preference operator that considered in this work is given as follows.

Definition 5.5. *A preference operator $\triangleright$ is defined for comparing two coalition structures $\Psi_1 = \{\mathcal{S}_1, \cdots, \mathcal{S}_m\}$ and $\Psi_2 = \{\mathcal{R}_1, \cdots, \mathcal{R}_n\}$ that are partitions of the same set of players $\mathcal{U}$. The notation $\Psi_2 \triangleright \Psi_1$ denotes that players in $\mathcal{U}$ are preferred to be in Ψ_2 than Ψ_1.*

5.4 Proposed Fully Distributed Solution

Here, we derive the constraints of forming a coalition that represents the players' associations, transmitting player, and its packet combination in a coalition. By the given constraints, we aim to propose a distributed coalition formation algorithm using merge-and-split rules.

5.4.1 Coalition Formation Constraints

Let $\mathcal{U}_s$ be the set of all associated players in $\mathcal{S}_s$-th coalition and $\mathcal{U}_{s,w}$ the subset of $\mathcal{U}_s$ that have non-empty *Wants* set. Let $\mathcal{M}_s$ be the subset of packets that in the *Has* set of each player in $\mathcal{U}_s$, which defined as $\mathcal{M}_s = \bigcup_{u_k \in \mathcal{U}_s} \mathcal{H}_{u_k}$. Let $\mathtt{S}_s$ denote the set of all neighboring coalitions to $\mathcal{S}_s$-th coalition. For the $\mathcal{S}_s$-th coalition, the $u_{n_s^*}$-th transmitting UD is the one that can achieve the least expected increase in the completion time. According to the analysis available in [69], a transmitting UD $u_{n_s^*}$ and its packet combination $\mathtt{p}_{u_{n_s^*}}$ sent to the set of users $\mathtt{u}(\mathtt{p}_{u_{n_s^*}})$ can be obtained by solving the following problem

$$u_{n_s^*} = \underset{u_n \in \mathcal{U}_{\mathrm{tra},s} \setminus \mathcal{L}_s}{\arg\max} |\mathcal{C}_{u_n} \cap \mathcal{N}_{s,w}| + \underset{\mathtt{p}_{u_n} \in \underline{\mathtt{p}}(\mathcal{U}_{\mathrm{tra,s}})}{\arg\max} \sum_{u_l \in \mathcal{L}_s \cap \mathtt{u}(\mathtt{p}_{u_n})} \log \frac{1}{\sigma_{u_n u_l}}, \quad (5.6)$$

where $\mathcal{U}_{\mathrm{tra},s}$ is the set of players in coalition $\mathcal{S}_s$ that are not in any coverage zone of all other players in $\mathtt{S}_s$ and $\mathcal{L}_s$ is the set of critical players that can potentially increase the overall payoff of the coalition $\mathcal{S}_s$. This set characterizes the players based on their anticipated completion times to give them priority to be targeted in the next transmission. In other words, $\mathcal{L}_{s,t}$

contains players that would potentially increase the maximum anticipated completion time if they are not targeted in the t-th transmission. It can be defined as

$$\mathcal{L}_{s,t} = \Big\{ u_k \in \ \mathcal{U} \cap \mathcal{U}_{s,w} \big| \mathtt{T}_{u_k}(\underline{n}_{t-1}) + \frac{1}{1 - \mathbb{E}[\sigma_{u_k}]} \geqslant \|\underline{\mathtt{T}}(\underline{n}_{t-1})\|_\infty \Big\}. \tag{5.7}$$

The set of targeted players in coalition $\mathcal{S}_s$ when device $u_{n_s^*}$ transmits the combination $\mathtt{p}_{u_{n_s^*}}$ is

$$\mathtt{u}(\mathtt{p}_{u_{n_s^*}}) = \Big\{ u_k \in \mathcal{S}_s \ \big| |\mathtt{p}_{u_{n_s^*}} \cap \mathcal{W}_{u_k}| = 1 \text{ and } \mathbf{C}_{u_{n_s^*} u_k} = 1 \Big\}.$$

With the aforementioned variable definitions, we can reformulate the completion time minimization problem in IDNC-based partially connected D2D network per coalition at each transmission slot as follows

$$\min_{\substack{\underline{n}_t \in \{0,1\}^{|\mathcal{U}_s|} \\ \underline{\mathtt{p}} \in \{0,1\}^{|\mathcal{M}_s|}}} \phi(\mathcal{S}_s) \tag{5.8a}$$

$$\text{s. t. } |\mathtt{u}(\mathtt{p}_{u_{n_s^*}})| \geqslant 1, \tag{5.8b}$$

$$\mathtt{u}(\mathtt{p}_{u_{n_s^*}}) \cap \mathtt{u}(\mathtt{p}_{n_v^*}) = \varnothing, \forall u_{n_s^*} \neq u_{n_v^*} \in \mathtt{S}_s. \tag{5.8c}$$

Constraint (5.8b) says that the number of targeted players in each coalition must be more than one to ensure that at each transmission at least one player is benefiting. Constraint (5.8c) states that all targeted players should not experience any collision. To find the optimal solution to the problem in (5.8), we need to search over all the sets of optimal player-coalition associations, their different erasure patterns, players' actions, and their optimal IDNC packets. As pointed out in [67] for a centralized fog system, this is an intractable problem. Further, the solution to (5.8) must go through the players' decisions to join/leave a coalition at each stage of the game. To seek an efficient solution to (5.8) that is capable of achieving significant completion time reduction, a distributed coalition formation algorithm is developed.

5.4.2 A Distributed Coalition Formation Algorithm

We develop a distributed coalition formation algorithm to obtain the minimum completion time of UDs. The key mechanism is to allow players in the coalition formation process to make individual decisions for selecting potential coalitions at any transmission slot. We first define two rules of merge-and-split that allow the modification of Ψ of the set $\mathcal{N}$ players as follows.

Definition 5.6. (**Merge Operation**). *Any set of coalitions* $\{\mathcal{S}_1, \cdots, \mathcal{S}_m\}$ *in* Ψ_1 *can be merged if and only if* $(\bigcup_{s=1}^{m} \mathcal{S}_s, \Psi_2) \triangleright (\{\mathcal{S}_1, \cdots, \mathcal{S}_m\}, \Psi_1)$*, where* $\bigcup_{s=1}^{m} \mathcal{S}_s$ *and* Ψ_2 *are the new set of coalitions and the new coalition structure after the merge operation, respectively.*

Definition 5.7. (**Split Operation**). *Any set of coalitions* $\bigcup_{s=1}^{m} \mathcal{S}_s$ *in* Ψ_1 *can be split if and only if* $(\{\mathcal{S}_1, \cdots, \mathcal{S}_m\}, \Psi_2) \triangleright (\bigcup_{s=1}^{m} \mathcal{S}_s, \Psi_1)$*, where* $\{\mathcal{S}_1, \cdots, \mathcal{S}_m\}$ *and* Ψ_2 *are the new set of coalitions and the new coalition structure after the split operation, respectively.*

The merge rule means that two coalitions merge if their merger would benefit not only the players in the united alliance but also benefit the overall coalition structure value, i.e., the whole completion time. On the other hand, a coalition split into smaller ones if its splitter coalitions enhance at least the payoff of one player in that coalition. Therefore, using these two known rules, we present a distributed algorithm to solve the completion time minimization problem in (5.3) as follows.

Step 1 (Coalition Members Discovery): Players in the initialized random coalition structure Ψ_{ini} discover their neighbors by utilizing one of different known neighbor discovery schemes, e.g., those used in wireless networks [157]. For example, the u_k-th player broadcasts a message consisting of two segments; each segment consists of one byte. While the first byte indicates the number of players in its coverage zone $\mathcal{C}_{u_k}$, the second byte indicates the u_k's completion time. Further, players use the collected aforementioned

information to decide on the one who: i) has a large *Has* set that can serve a large number of players in its coverage zone, and ii) not in the coverage zone of any player in any other coalitions in Ψ_{ini}. However, if such a player does not exist, the size of the coalition is increased until that player exists. To summarize, the $u_{n_s^*}$-th selected transmitting player in $\mathcal{S}_s$-th coalition should satisfy (5.8b) and (5.8c) and can be obtained by solving problem (5.6). Afterward, each player evaluates its potential payoff as in (2.6) to make an accurate decision in step 2. The $u_{n_s^*}$ selected player in $\mathcal{S}_s$-th coalition ($\forall \mathcal{S}_s \in \Psi_{\text{ini}}$) will do the analysis in step 2. This step significantly reduces the search space for associating the players to the coalitions.

Step 2 (Coalition Formation for Players' Association): This step optimizes the selection of the transmitting players and their packet combinations in step 1 through many successive split-and-merge rules between coalitions. Therefore, step 2 is to associate players to potential neighboring coalitions, select the transmitting player, and find its packet combination, which can be accomplished by the following. In this step, the index is updated to $\tau = \tau + 1$. The merge rules are implemented by checking the merging possibilities of each pair of neighboring coalitions $\mathcal{S}_s$ and $\mathcal{S}_v$. Initially, $\Psi_\tau = \Psi_{\text{ini}}$. Thus, a coalition $\mathcal{S}_s \in \Psi_\tau$ can decide to merge with another coalition $\mathcal{S}_v \in \mathbb{S}_s$ to form a new coalition $\mathcal{S}_j$. As such, the resulting structure guarantees both merge conditions (MC).

- MC1: There exists at least one player satisfies (5.8b) and (5.8c).
- MC2: At least one player in the merged coalition can enhance its individual payoff without negatively affecting the payoffs of all the remaining players.

After all the coalitions have made their merge decisions based on the players' preferences, the merge rules end. This results in the updated coalition structure Ψ_τ. Similarly, the split rules performed on the players that do not benefit from being a member of that coalition. In other terms, coalition $\mathcal{S}_s \in \Psi_\tau$ can be split into coalitions of smaller sizes as long as the splitter coalitions guarantee both split conditions (SC).

- SC1: At least one player can strictly enhance its payoff without affecting the payoffs of all the remaining players.
- SC2: In each split coalition, there exists at least one player satisfying (5.8b) and (5.8c).

At the end of the split rules, the coalition structure Ψ_1 is updated. The index τ is updated along with a sequence of merge-and-split rules which take place in a distributed manner. Such sequence continues based on the resulting payoff of each player and coalition. It ends when there are no further merge-and-split rules required in the current coalition structure Ψ_τ, which is the final converged coalition structure Ψ_{fin}.

Step 3 (IDNC Packet Transmissions:) Each transmitting UD in each coalition broadcasts its IDNC packet to all UDs in its coverage zone. Accordingly, each targeted UD by any transmitter sends ACK to all its neighboring UDs indicating its packet reception status. Thus, the *Has*, and *Wants* sets of all UDs in the system are updated.

The coalition formation algorithm is distributively executed at each transmission slot and summarized in Algorithm 8. It is repeated until all packets are delivered to all UDs, as presented in Algorithm 9.

Some remarks on executing Algorithm 8 are given below.

- The merge-and-split rules enumerate only neighboring coalitions, and this does not necessarily need significant computations. To further reduce the computations, the players of a coalition $\mathcal{S}_s$ can avoid merging with other neighboring coalition $\mathcal{S}_v$ if the payoffs of the players in both coalitions are equal $\phi_{u_k}(\mathcal{S}_s) = \phi_{u_l}(\mathcal{S}_v)$, $\forall u_k \in \mathcal{S}_s$ and $\forall u_l \in \mathcal{S}_v$.
- The connectivity constraints of UDs are incorporated in the coalition formations by: i) forming disjoint coalitions of nearby UDs based on their preferences, and ii) selecting the potential transmitting UD that can only target its connected players in its coalition. As such, we make sure that connected UDs are grouped into a single coalition.
- Forming coalitions only one time, i.e., at the beginning of the transmission round, is not guaranteed to deliver all packets to all players.

Algorithm 8: Coalition Formation Algorithm

Input: $\mathcal{U}$, $\mathcal{P}$, $\mathcal{H}_{u_k}$, $\mathcal{W}_{u_k}$, $\mathcal{C}_{u_k}$, $\forall u_k \in \mathcal{U}$.;
Initialization: $\Psi_{\text{ini}} = \{\mathcal{S}_1, \cdots, \mathcal{S}_m\}$, $\tau = 0$, $\Psi_\tau = \Psi_{\text{ini}}$.;
Step 1: Coalition Members Discovery

– Each player discovers its neighboring players.
for *each* $\mathcal{S}_s \in \Psi_{ini}, \forall s = \{1, 2, \cdots, m\}$ **do**
Select the transmitting players $\mathcal{U}_{\text{tra},s}$ that satisfying (5.8b) and (5.8c).
Solve (5.6) to find $u_{n_s^*}$ and its packet combination $\mathtt{p}_{u_{n_s^*}}$.
Calculate $\phi_{u_k}(\mathcal{S}_s)$ as in (2.6), $\forall u_k \in \mathcal{U}_s$.
end

Step 2: Coalition Formation

– The optimization target in coalition $\mathcal{S}_s$ is $\min\limits_{\substack{\underline{n}_t \in \{0,1\}^{|\mathcal{U}_s|} \\ \underline{\mathtt{p}} \in \{0,1\}^{|\mathcal{M}_s|}}} \phi(\mathcal{S}_s)$.

– Obtain player's assignments based on the two main rules of merge and split:
repeat
Update $\tau = \tau + 1$.
for *each* $\mathcal{S}_s \in \Psi_{\tau-1}, \forall s = \{1, 2, \cdots, m\}$ **do**
$u_{n_s^*}$ analyzes all possible merge rules with $\mathtt{S}_s$.
If a merge occurs:

1. Update $\Psi_{\tau-1}$.
2. Solve (5.6) to update $\mathcal{U}_{\text{tra},s}$ and $u_{n_s^*}$.

Set $\Psi_\tau = \Psi_{\tau-1}$.
end
for *each* $\mathcal{S}_s \in \Psi_\tau, \forall s = \{1, 2, \cdots, m\}$ **do**
$u_{n_s^*}$ analyzes all possible split rules.
If a split occurs:

1. Update Ψ_τ.
2. Update $\mathcal{U}_{\text{tra},s}$ and update $u_{n_s^*}$ by solving (5.6).

end
until *No further merge nor split rules*;

Output The convergence coalition structure $\Psi_{\text{fin}} = \Psi_\tau$.
Step 3: IDNC Packet Transmission

– Each $u_{n_s^*}$ in each coalition $\mathcal{S}_s$ broadcasts $\mathtt{p}_{u_{n_s^*}}$ to all players in its coverage zone, $\forall s = \{1, 2, \cdots, m\}$.

Algorithm 9: Overall Completion Time Approach for Solving Problem (5.3)

Data: $\mathcal{U}$, $\mathcal{P}$, $\mathcal{H}_{u_k}$, $\mathcal{W}_{u_k}$, $\mathcal{C}_{u_k}$, $\mathtt{T}_{u_k} = 0$, $\mathcal{D}_{u_k} = 0$, $\forall u_k \in \mathcal{U}$ and $\boldsymbol{\epsilon}$.
Initialize: Transmission slot $t = 1$
Repeat:

- Execute Algorithm 8 to obtain $\mathcal{U}_{\text{tra},t}$, $\underline{\mathtt{p}}(\mathcal{U}_{\text{tra},t})$, and $\underline{\mathtt{u}}(\underline{\mathtt{p}}(\mathcal{U}_{\text{tra},t}))$ in Ψ_{fin}.
- **For** $\forall u_{n_s^*} \in \mathcal{U}_{\text{tra},t}$ **do**
  - Each $u_k \in \mathtt{u}(\mathtt{p}_{u_{n_s^*}})$ re-XOR $\mathtt{p}_{u_{n_s^*}}$ with $\mathcal{H}_{u_k}$.
  - Calculate the completion time of each $u_k \in \mathtt{u}(\mathtt{p}_{u_{n_s^*}})$ as in (5.2).
  - Each $u_k \in \mathtt{u}(\mathtt{p}_{u_{n_s^*}})$ broadcasts one bit ACK to all players in $\mathcal{C}_{u_k}$.
- **End for**
- $t = t + 1$.
- Update $\mathcal{H}_{u_k}$, $\mathcal{W}_{u_k}$, $\forall u_k \in \mathcal{U}_w$.

Until $\mathcal{W}_{u_k} = \phi$, $\forall u_k \in \mathcal{U}_w$.
Output: Transmission slot t.

This is because each formed coalition has only some portion of packets and does not have the requested packets of other players in other coalitions. For packet delivery completion, each coalition is formed, at each transmission slot, based on the individual preference of its alliances and irrespective of the *Has* sets of its alliances. Thus, each transmitting player has delivered some packets to each visited coalition in previous transmissions.

In the considered game, each player has two actions to take either to transmit a packet combination $\mathtt{p}$ or to listen to a transmission. Thus, the action of u_k-th player at t-th game stage is $\mathcal{AC}_{u_k,t} = \{\text{transmit } \mathtt{p}_{u_k}, \text{ remain silent}\}$. The asymmetry of the side information at each player generates a different packet combination to be sent by each player at each transmission slot. This causes the asymmetry of the action space of each player. Also, in each trans-

mission, different players are associated with each coalition. All these make the payoff of each coalition unique.

5.5 Theoretical Analysis of the Proposed Game

Convergence and Stability

The stability of the coalition structures in coalition formation games corresponds to an equilibrium state known as Nash-equilibrium. Here, we prove that the resulting coalition structure from Algorithm 8 is converged, and it is a Nash-stable. The following theorem demonstrates that Algorithm 8 terminates in a finite number of iterations.

Theorem 5.8. *Given any initial coalition structure Ψ_{ini}, the coalition formation step of Algorithm 8 maps to a sequence of merge-and-split rules which converges, in a finite number of iterations, to a final coalition structure Ψ_{fin} composed of a number of disjoint coalitions.*

Proof. To proof this theorem, we need to show that for any merge or split rule, there exists a new coalition structure which results from the coalition formation step of Algorithm 8. Starting from any initial coalition structure Ψ_{ini}, the coalition formation step of Algorithm 8 can be mapped to a sequence of merge/split rules. As per definition 8 and definition 9, every merge or split rule transforms the current coalition structure into another coalition structure, hence we obtain the following sequence of coalition structures

$$\Psi_{\text{ini}} \rightarrow \Psi_1 \rightarrow \Psi_2 \rightarrow \ \cdots \ \Psi_{\text{fin}} \tag{5.9}$$

where $\Psi_{i+1} \triangleright \Psi_i$, and $\rightarrow$ indicates the occurrence of a merge-and-split rule. Since the Pareto Order introduced in definition 6 is irreflexive, transitive, and monotonic, a coalition structure cannot be revisited. Given the fact that the number of merge and split rules of a finite set is *finite* and the merge/split operations-coalition structure mapping, the number of coalition structure sequences in (5.9) is finite. Therefore, the sequence in (5.9) always terminates and converge to a final coalition structure Ψ_{fin}. □

Definition 5.9. *A coalition structure* $\Psi = \{\mathcal{S}_1, \cdots, \mathcal{S}_m\}$ *is Nash-stable if players have no incentive to leave* Ψ *through merge-and-split operations.*

This definition implies that any coalition structure Ψ is considered as a Nash-stable coalition structure if and only if no player has an incentive to move from its current coalition and join another alliance or make an individual decision by performing any merge/split rules. Further, the alliances in the final coalition structure, Ψ_{fin} have no incentive to do more merge and split operations. A Nash-stable coalition structure is also an individually stable coalition structure. In general, in a coalition formation game, Nash-stability is a subset of individual stability [158]. Specifically, no player leaves its current coalition through a split rule and form an empty coalition, i.e., no singleton coalition is created if the following property holds.
Property: There exists at least one coalition structure Ψ that satisfies both Nash-stability and individual stability if and only if $\forall \mathcal{S}_s \in \Psi$ such that $|\mathcal{S}_s| > 1$.

Proof. This property states that forming a singleton coalition cannot happen. Indeed, since each player cannot send an encoded packet to itself, it believes that a better payoff can be obtained by being a member of any coalition. Further, since the payoff of a non-targeted player in any coalition and a single player-coalition is the same, our proposed algorithm, as mentioned in the previous section, avoids making any merge-and-split rules for equal payoff values. Thus, according to Algorithm 8, a Nash-stable and individual stable coalition structure can be obtained. □

As a consequence of Property 2, the final coalition structure Ψ_{fin} that results from Algorithm 8 is $\mathbb{D}_{\text{hp}}$ stable as the coalitions have no incentive to do further merge-and-split operations. $\mathbb{D}_{\text{hp}}$ stable is also known as merge-and-split proof [158]. Furthermore, Ψ_{fin} can be considered as $\mathbb{D}_{\text{c}}$ stable. This is because players have no incentive to leave Ψ_{fin} and form any other coalitions [136].

Coalition Formation Example

In order to better illustrate the proposed coalition formation algorithm and the aforementioned stability concepts, consider the example presented in Figure 5.2. For ease of analysis, we assume error-free transmissions. Given the coverage zone of the players and their side information as in Figure 5.2, the resulting coalition structure Ψ_{fin} from Algorithm 8 for the network presented in Figure 5.2 consists of two disjoint coalitions $\mathcal{S}_1, \mathcal{S}_2$ where only one player transmits in each coalition. In particular, in coalition $\mathcal{S}_1$, player u_4 transmits packet p_1 to player u_6, and in coalition $\mathcal{S}_2$, player u_1 transmits the combination $p_3 \oplus p_4$ to players u_2, u_3, u_5. The transmitting player in each coalition is shown in a red circle; their targeted players and the packet combinations are shown in Figure 5.3. Given the resulting coalition structure $\Psi_{\text{fin}} = \{\mathcal{S}_1, \mathcal{S}_2\}$ that shown in Figure 5.3, we now analyze its Nash stability. The coalition structure Ψ_{fin} is Nash-stable as no player has an incentive to leave its current coalition. For example, player u_5 has a payoff of $\phi_{u_5}(\mathcal{S}_2) = -2$ when being part of the coalition $\mathcal{S}_2 = \{u_1, u_2, u_3, u_5\}$. The payoff $\phi_{u_5}(\mathcal{S}_2)$ is calculated as follows. Since player u_5 receives an instantly-decodable combination from player u_1, it does not experience any decoding delay increases. Thus, by (5.2), its anticipated completion time is $\mathtt{T}_{u_5}(\underline{\mathrm{n}}_t) = \frac{|\mathcal{W}_{u_5,0}| + \mathbb{D}_{u_5}(\underline{\mathrm{n}}_t) - \mathbb{E}[\sigma_{u_5}]}{1 - \mathbb{E}[\sigma_{u_5}]} = 2$, and, by (5.5) its payoff is -2. If player u_5 switches to act non-cooperatively and joins $\mathcal{S}_1$, player u_6 would be the new transmitting player in $\mathcal{S}_1$. In this case, player u_5 will be in the coverage zone of both transmitting players u_1 in $\mathcal{S}_2$ and u_6 in $\mathcal{S}_1$. Consequently, the payoff of player u_5 decreases to $\phi_{u_5}(\mathcal{S}_1) = -3$, and the payoff of player u_6 decreases from $\phi_{u_6}(\mathcal{S}_1) = -3$ to $\phi_{u_6}(\mathcal{S}_1) = -4$. Thus, player u_5 does not deviate form its current coalition $\mathcal{S}_2$ and join $\mathcal{S}_1$. Similarly, if players u_2 and u_3 act non-cooperatively by leaving $\mathcal{S}_2$ and forming a singleton coalition for each, i.e., $\mathcal{S}_3$ and $\mathcal{S}_4$, their payoffs decrease from $\phi_{u_2}(\{2\}) = -2$ and $\phi_{u_3}(\{3\}) = -1$ to $\phi_{u_2}(\mathcal{S}_3) = -3$ and $\phi_{u_3}(\mathcal{S}_4) = -2$, respectively. Clearly, Ψ_{fin} is an individual Nash-stable as it does not have any singleton coalition. Further, it is both $\mathbb{D}_{\text{hp}}$ and $\mathbb{D}_{\text{c}}$ stable as no further merge-and-split operations can be performed by the coalitions and no player has incentive to deviate from Ψ_{fin}, respectively.

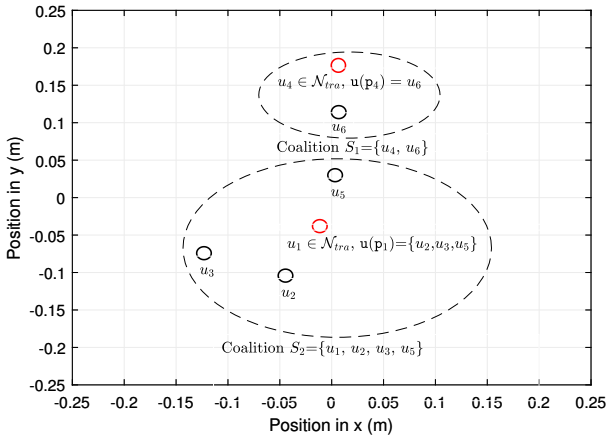


Figure 5.3: A resulting coalition structure $\Psi_{\text{fin}} = \{\mathcal{S}_1, \mathcal{S}_2\}$ from Algorithm 8 for a partially connected D2D network that is presented in Figure 5.2.

Computational Complexity and Communication Overhead

Here, we analyze the complexity and communication burden of Algorithm 8.

Complexity: Each UD at any transmission slot generates its packet combination, which depends on the packets that it previously received. Further, since a game with incomplete information, i.e., each UD knows only the received and requested packets of UDs in its coverage zone, every UD generates its packet combinations of all other UDs in its coverage zone. This allows every UD to calculate the payoff function (5.5) of all other UDs in its coverage zone. The complexity of generating the packet combination by the CBS using a maximum weight search method is given as follows. First, the CBS generates all the vertices that represent all users and their requested packets, which needs $O(UM)$ operations. Note that a vertex is generated for each file p_l is requested by each UD u_k, $\forall p_l \in \mathcal{P}, \forall u_k \in \mathcal{N}$. The CBS, then, connects these vertices (circles) by edges that represent network coding conditions of $O(N^2M)$. Afterward, the CBS executes the maximum weight search method that computes the weight of $O(UM)$ vertices and selects a maximum of N UDs. Hence, the overall complexity of finding the IDNC

packet is $O(UM) + O(U^2M) + O(U^2M) = O(U^2M)$ operations [44]. In our case, the complexity for generating the packet combination at each UD is bounded by $O(U^2M)$. This is because the number of UDs in the coverage zone of each UD is less than the total number of UDs in the network.

Communication Overhead: The communication overhead of Algorithm 8 is related to perform the members' discovery step, transmitting player selection, and the analysis of merge-and-split rules. It is explained as follows.

Step 1: Similar to many algorithms in the literature, e.g., [157], the member discovery step needs $|U|$ 2-byte messages, in which each message is being sent by every UD to all its neighbors which is denoted by $\mathbf{U}$. Thus, the total communication overhead for discovering the neighbor UDs is $|2U\mathbf{U}|$ bytes.

Step 2: Selecting the transmitting player in each coalition can be performed in many different strategies, e.g., based on players' attributes [159], [160]. In Algorithm 8, coalition's members initially exchange an advertisement message among them, and the one that satisfies the conditions in Section 5.4.2 would be chosen as a transmitting player in that coalition. The same process is applied for selecting/updating the transmitting player in step III. Since it is connected to most players in the coalition, the transmitting player is responsible for ensuring that the rest of the coalition's members received an acknowledgment (ACK). As such, they update their *Has* and *Wants* sets after each D2D transmission.

Step 3: The communication overhead due to forming coalitions is based on the number of merge-and-split operations. This is mainly related to the total number of decisions made by each of the U players. As previously mentioned in Section 5.4.2, the merge-and-split operations enumerate only the neighboring coalitions $\mathtt{S}_s$. Thus, two extreme cases can occur:

- If all coalitions' alliances decide to leave their current coalitions and join other coalitions. In this case, u_k-th player in $\mathcal{S}_s$-th coalition would make $|\mathtt{S}_s|$ decisions (u_k-th player has an $|\mathtt{S}_s|$ possibilities to join any of the neighbor coalitions). Consequently, the total number of players' decisions is $Q_{\text{worst}} = U|\mathtt{S}_s|$.

- If players did not make any decisions. Since players make no decision,

the overhead, in this case, is only $Q_{\text{best}} = U$ (due to the initial player-coalition associations as in step 1).

In practical, the number of players' decisions is between the above two cases, i.e., $Q_{\text{best}} \leq Q \leq Q_{\text{worst}}$. Hence, if L average decisions are made by players, then $Q = NL$ decisions that perform split-and-merge rules in Algorithm 8. Therefore, combining all the overhead signaling components, the total overhead in each transmission is $U(2\mathbf{U} + L)$. Such signaling cost will add only a few bytes, which are negligible in size compared to the entire packet's size. Furthermore, to update the *Has* and *Wants* sets of UDs, only the indices of packets need to be exchanged between the UDs, not their contents. Hence, we ignore the overhead signaling factor because it is first constant (independent on the completion time and decoding delay) and that its size is negligible.

5.6 Numerical Results

This section evaluates the performance of our proposed coalition formation game (denoted by CFG partially-connected D2D) to demonstrate its capability of reducing the completion time compared to the baseline schemes. We first give the simulation setup and the implemented schemes. Then, the completion time and game performances are comprehensively evaluated. We consider a partially connected D2D network where every UD is connected to some other UDs within its coverage zone based on the connectivity index C. The considered model is shown in Figure 5.1 in the system model section. The connectivity index C is defined as the ratio of the average number of neighboring UDs to the total number of UDs U. Further, the UDs in the considered model are uniformly re-positioned after each transmission in a 500m×500m cell. The system setting in this work follows the setup studied in [64], [67]. As mentioned in the system model, each UD has already acquired some packets and wants some other packets from the set $\mathcal{P}$. Such initial *Has* and *Wants* sets $\mathcal{H}_{u_k}$ and $\mathcal{W}_{u_k}$, $\forall u_k \in \mathcal{N}$ of UDs is independently drawn based on their average erasure probability. The short-range D2D

links are more reliable than the CBS-UD communications [62], [63]. Hence, unless specified, we assume that the player-to-player erasure probability σ is half the CBS-to-UD erasure ϵ in all simulations, i.e., $\sigma = 0.5\epsilon$. Our simulations were implemented using Matlab on a Windows 10 laptop 2.5 GHz Intel Core i7 processor and 8 GB 1600 MHz DDR3 RAM. For the sake of comparison, we implement the following schemes.

- **The fully-connected D2D scheme:** In this scheme, a single UD who has the largest number of received packets transmits an IDNC packet at each transmission slot.
- **PMP:** The CBS in this scheme is responsible for the transmissions. It provides all the requested packets to all the UDs. This scheme was proposed in [115].
- **The one coalition formation game:** This scheme is denoted by OCF partially-connected D2D. Only one coalition is formed in this scheme, and accordingly, a single UD transmits an IDNC packet at each transmission. The transmitting UD is selected based on its number of received packets as well as on the maximum number of UDs in its coverage zone.
- **The FRAN partially-connected D2D scheme:** In this scheme, a fog central unit is responsible for determining the set of transmitting UDs and their corresponding packet combinations. This scheme was proposed in [67].

5.6.1 Completion Time Performance Analysis

The completion time measures the total number of transmissions that the network needs until all UDs receive all their requested packets. Thus, all simulated schemes are executed until all UDs receive all their requested packets. The presented average values of the completion time are computed over a certain number of iterations. To study the completion time performance of the proposed solution by changing the number of players, packets, connectivity index, and the packet's erasure probability.

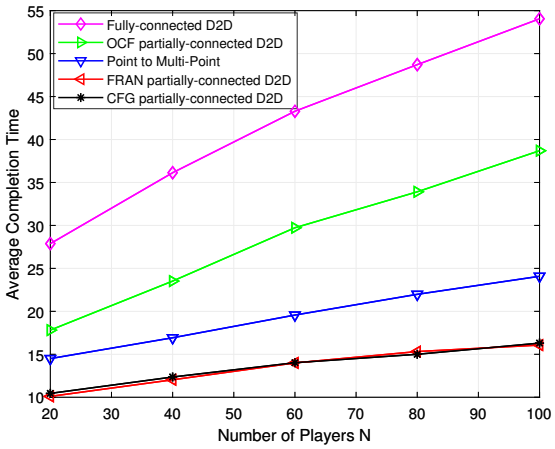


Figure 5.4: Average completion time versus the number of players U.

In Figure 5.4, we depict the average completion time versus the number of players U for a network composed of $M = 30$ packets, $\epsilon = 0.25$, $\sigma = 0.12$, and connectivity index $C = 0.4$. It is observed from Figure 5.4 that the proposed CFG partially-connected D2D algorithm outperforms the PMP, fully-connected D2D, and OCF partially-connected D2D schemes for all simulated numbers of players. This is because of the simultaneous IDNC packet transmissions from cooperating players at the same time. In particular, the fully-connected D2D system only considers the size of the *Has* set as a metric to select a single UD for transmission at each time slot, i.e., $u_{n^*} = \max_{u_n \in \mathcal{U}} \mathcal{H}_{u_n}$. The OCF partially-connected D2D scheme focuses on the maximum number of connected UDs to be formed as well as on the size of the *Has* set of the transmitting UD. On the other hand, although the transmitter in the PMP scheme can encode all the IDNC combinations and schedule a certain number of UDs, the PMP scheme sacrifices the utility of the simultaneous transmissions by considering only one transmission. Our proposed algorithm strikes a balance between these aspects by jointly considering the number of scheduled UDs and the *Has* set size of each transmitting UD. Despite the gain achieved by the FRAN partially-connected D2D solution

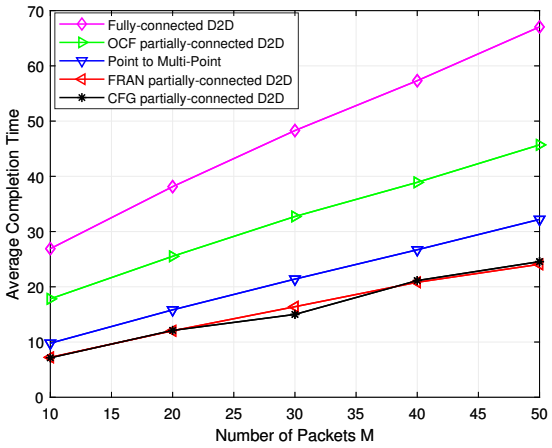


Figure 5.5: Average completion time versus the number of packets M.

with the presence of a fog that executes the whole process, our decentralized solution reaches the same performance. Clearly, due to the philosophy of D2D simultaneous transmissions that considers in the centralized FRAN and our proposed schemes, their performances are roughly the same.

We observe from Figure 5.4 that, for a small number of players, the PMP system is close to both the CFG partially-connected D2D and FRAN partially-connected D2D schemes. This is because, for a small amount of players ($U \leq 60$), the certainty that the whole frame M is distributed between players in the initial transmissions is low, thus decreasing the probability of delivering potential IDNC packets to UDs. This makes the overall completion time performance of the partial D2D scenarios close to the PMP scheme. As the number of players increases ($U \geq 80$), the bigger the certainty that the union of their *Has* sets is equal to M. This results in delivering more potential D2D IDNC packets, thus increasing the gap between the PMP performance and both the FRAN partially-connected D2D and proposed schemes.

In Figure 5.5, we illustrate the average completion time as a function of the number of packets M for a network composed of $U = 30$ UDs, $\epsilon = 0.25$,

$\sigma = 0.12$, and connectivity index $C = 0.4$. The figure shows that the proposed scheme outperforms the fully connected, one coalition game, and PMP schemes. For a few packets, the IDNC combinations are limited, which affects the ability of the proposed scheme to generate coded packets that satisfy several scheduled UDs. With increasing the number of packets, the number of transmissions needed for the completion of the aforementioned schemes is remarkably increasing. Therefore, as the number of packets increases, the proposed scheme outperforms largely the fully connected and one coalition game schemes. We see from Figure 5.5 that the completion time of all schemes linearly increases with the number of packets. This is expected as the number of packets increases, a high number of transmissions is required towards the complete delivery of packets. This results in increasing the average completion time.

In Figure 5.6, we plot the average completion time as a function of the average player-player erasure probability σ for a network composed of $U = 60$, $M = 30$, $\epsilon = 2\sigma$, and $C = 0.4$. Similar to what we have discussed in Figure 5.4 and Figure 5.5, the average completion time of the partial D2D schemes is noticeable compared to the fully-connected D2D and OCF partially-connected D2D schemes, as shown in Figure 5.6. We clearly see that the completion time of the partial D2D schemes is better than the PMP one because of their multiple UDs' transmissions at each time slot. Moreover, as the player-to-player erasure probability increases, the BS-player erasure probability increases two-fold ($\epsilon = 2\sigma$), thus slightly affecting the performance of the PMP scheme. The partial D2D settings, however, benefit from short D2D links, which provide much better UDs reachability and IDNC packet successful delivery compared to the PMP setting.

In Figure 5.7, we investigate the average completion time as a function of the connectivity index C for a network composed of $U = 60$, $M = 30$, $\epsilon = 0.25$, and $\sigma = 0.12$. It can clearly be seen that for a low connectivity index ($C \leq 0.4$), the proposed CFG partially-connected D2D approach noticeably outperforms the fully-connected D2D and OCF partially-connected D2D approaches. In such poorly connected networks ($C \leq 0.4$), multiple simultaneous players' transmissions are exploited in partially D2D algo-

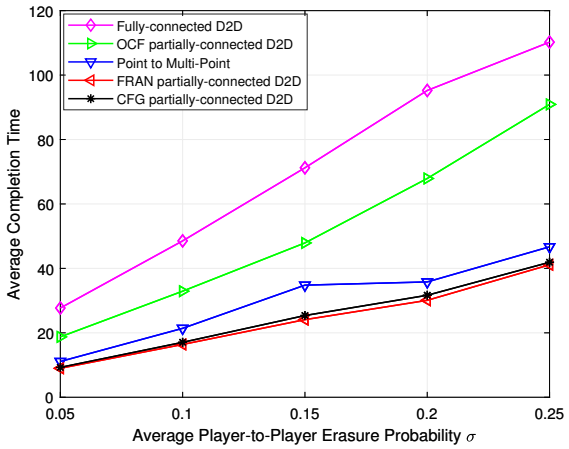


Figure 5.6: Average completion time versus the average player-player erasure probability σ.

rithms. However, as the connectivity index increases ($C \geq 0.6$), the number of formed disjoint coalitions in our proposed solution is drastically reduced, thus reducing the number of transmitting UDs. This results in a performance agreement with the fully-connected D2D scheme. Being independent of the coverage zones of the transmitting UDs and the delay created by those UDs, the PMP scheme is not affected by the changes to C. Thus, the PMP scheme has a constant average completion time.

To conclude this section, we study the influence of the setting $\sigma = 0.5\epsilon$ on the completion time performance of our proposed scheme. In Table 5.1, we summarize the completion time performance for different values of σ. The considered network setup has 30 UDs, 20 packets, $\epsilon = 0.5$, and $C = 0.1$. From Table 5.1, we note that the completion time of our proposed solution still outperforms the PMP scheme for $\sigma = 0.7\epsilon$ and approximately reaches the same performance as for the PMP scheme for $\sigma = 0.9\epsilon$. This is due to the simultaneous transmissions and cooperative decisions by the transmitting UDs, which show the potential of the proposed CFG solution in minimizing the completion time of UDs.

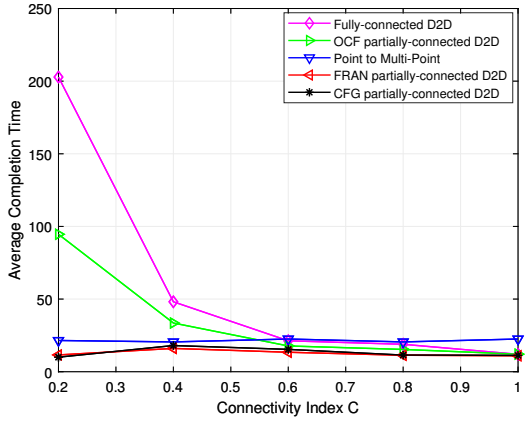


Figure 5.7: Average completion time versus the connectivity index C.

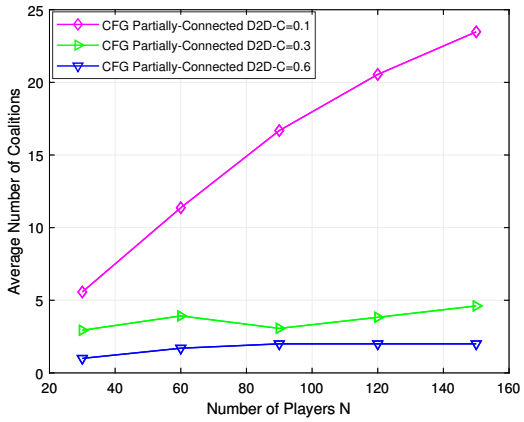


Figure 5.8: Average number of coalitions versus the number of players U.

5.6.2 Proposed CFG Performance Analysis

To quantify the analysis of the proposed formation coalition solution, we plot in Figure 5.8 the average number of coalitions as a function of the number of players U for a network composed of $M = 30$, a different connectivity index ($C = 0.6$, $C = 0.3$, and $C = 0.1$), and $\sigma = 0.12$. Figure 5.8 shows that the average coalition size increases with the increase in the number of

players. This is because, as N increases, the number of cooperating players increases, thus increasing the average size of the formed coalitions. We can conclude from Figure 5.8 that the resulting coalition structure Ψ_{fin} from Algorithm 8 is composed of a small number of relatively large coalitions when $C = 0.6$. When $C = 0.1$, this number of formed coalitions increases, and the resulting coalition structure Ψ_{fin} is composed of a large number of small coalitions' sizes.

In Table 5.2, we evaluate the complexity of the proposed coalition game solution as a function of the algorithmic running time. In particular, Table 5.2 lists the consumed time of MATLAB to execute all schemes in different network setups since starting the algorithms until all players receive their wanted packets. The considered small network setup has 30 players, 20 packets, $\epsilon = 0.5$, $\sigma = 0.25$, and $C = 0.1$. The considered large network setup has 100 players, 70 packets, $\epsilon = 0.5$, $\sigma = 0.25$, and $C = 0.1$. It can clearly be seen from the table that the proposed CFG-partially D2D scheme needs low consumed time than all other solutions for both network setups. Although the completion time achieved by the CFG partially-connected D2D scheme is roughly the same as the centralized FRAN partially-connected D2D, the computing time required by our developed scheme is slightly higher than that required by the FRAN partially-connected D2D. This is because our proposed scheme needs time to converge before generating the output. The centralized FRAN scheme has low execution time due to the presence of the fog entity.

Finally, to evaluate the convergence rate analysis of the proposed scheme, the average number of merge-and-split rules before Algorithm 8 converges to the final coalition structure is listed in Table 5.3. To achieve the stable coalition with our proposed CFG scheme, network setup 1 requires, on average 16 iterations, and network setup 2 needs, on average 22 iterations. These results show that our proposed distributed algorithm is robust to different network setups. In summary, these results show that our proposed algorithm allows UDs to quickly form converged and stable coalition structures, which confirms the theoretical findings in Section 5.4.

Table 5.1: The influence of changing σ on the completion time performance of the proposed scheme

Solution	$\sigma = 0.6\epsilon$	$\sigma = 0.7\epsilon$	$\sigma = 0.9\epsilon$	$\sigma = \epsilon$
Point to Multi-Point	30.2900	30.2800	30.3100	30.4800
CFG partially-connected D2D	20.1800	23.4702	30.4500	33.9300

Table 5.2: Average Running Times of the different schemes

Solution	Time(s)- Small network	Time(s)- Large network
FRAN partially-connected D2D	0.561893	15.98450
Point to Multi-Point	1.994500	1103.020716
Fully-connected D2D	0.756420	128.772580
OCF partially-connected D2D	0.783575	28.726515
CFG partially-connected D2D	0.736737	21.725739

Table 5.3: Average Number of Coalitions and Split/merge rules of the proposed scheme in the first iteration

Network Setup	Number of Coalitions	Split-and-merge rules
Setup 1: $N = 100$ and $C = 0.1$	16.34	8.12
Setup 2: $N = 160$ and $C = 0.1$	23.67	12.76

5.7 Chapter Summery

In this chapter, we have developed a distributed game-theoretical framework for a partially connected D2D network using coalition game and IDNC optimization. As such, the distribution of contents among UDs is minimized. A comprehensive completion time and game performance evaluation have been carried out for the proposed distributed solution. In particular, our performance evaluation results comprehensively demonstrated that the proposed distributed solution offers almost the same completion time performance similar to the centralized F-RAN D2D network.

Chapter 6

Conclusions

In this chapter, we provide the concluding remarks on the accomplished works in this thesis. We also discuss some potential future research topics.

This thesis exploited the potential of NC for an effective allocation of radio resources in cache-enabled 5G mobile networks. State-of-the-art literature usually focuses on allocating a single user to each RRB, which needs large numbers of RRBs to satisfy the tremendous increase in the number of users. The developed NC schemes in this thesis overcome this issue by mixing users' contents and simultaneously allocating different users to the same RRB. To this end, we developed NC schemes to efficiently deliver requested contents to users while satisfying user' QoS in terms of throughput and completion time. The concluding remarks on these developed schemes in each chapter is provided in the next section.

6.1 Concluding Remarks

Some concluding remarks on the accomplished works of this thesis are presented as follows.

1. In Chapter 2, we introduced a cross-layer framework in downlink C-RAN to optimize throughput by streaming a set of popular files to users. For any arbitrary transmission, we considered optimizing the received throughput, subject to users' side information, the cached/stored files at each RRH, NC constraints, and the channel qualities. We have proposed two schemes, namely, joint and iterative schemes. Specifically, in the joint scheme, we used a novel graph theoretical representation to solve the joint coordinated scheduling and power control problem, while in the iterative scheme, we solved the problem iteratively.

Numerical results showed that both proposed schemes offer improved throughput performances compared to the existing solutions. Numerical results provided the following observation. When a large number of users is used, the RA-IDNC schemes significantly outperform the uncoded scheme. This is due to the fact that uncoded scheme always serves one user per RRB regardless of the network's size, which limits the overall throughput gain. On the other hand, the number of served users in the RA-IDNC scheme increases with the increase in the number of users in the system due to NC. Thus, the overall throughput gain of the RA-IDNC scheme is increased.

2. In Chapter 3, we studied the joint design of CBS and edge processing for F-RAN setting in which the eRRHs are equipped not only with the functionalities of standard RRHs in C-RAN, but also with local cache. For any arbitrary fixed pre-fetching strategy, we considered the optimization of the immediate file delivery phase with the goal of maximizing the CBS offloading while guaranteeing user's QoS, subject to the eRRHs cache contents, the required minimum rate, power control, and network coding constraints. By using a graph theory technique, we proposed joint and iterative efficient approaches that use cross-layer NC and coordinated scheduling graphs, respectively. Presented numerical results revealed that as the rate threshold increases, the throughput improvement increases as well. This is because as the rate threshold increases, certain number of users is targeted and the transmission rate of each RRB in each eRRH becomes high. As a result, number of users is left to be targeted by the CBS, which degrades its consumed time. Consequently, compared to the QoS un-aware algorithm, our proposed schemes have a certain degradation in CBS consumed time.

3. In Chapter 4, we developed frameworks that exploit the cached contents at eRRHs, their transmission rates, and previously received contents by different users to deliver the requesting contents to users with a minimum completion time in D2D-aided F-RAN. Simulation results

shown that our proposed schemes can effectively minimize the frame delivery time as compared to the conventional schemes. Such an improved performance has been achieved due to the joint and coordinated schemes that: (i) judiciously schedule users, adopt the transmission rate of each eRRH and optimize the transmission power of each eRRH, and (ii) select potential users for transmitting coded files over D2D links.

4. In Chapter 5, we developed a distributed game-theoretical framework for a partially connected D2D network using coalition game and IDNC optimization. As such, the delivery of contents to UDs is minimized. In particular, our proposed model is formulated as a coalition formation game with non-transferable utility, and a fully distributed coalition formation algorithm is proposed. The proposed distributed algorithm is converged to a Nash-stable coalition structure using split-and-merge rules while accounting for the altruistic player's preferences. With such a distributed solution, each UD has to maintain a partial feedback matrix only for the UDs in its coverage zone instead of the global feedback matrix required in the fully connected D2D networks. Some observations from the accomplished works in Chapter 5 are given as follows. First, the OCF partially-connected D2D scheme focuses on the maximum number of connected UDs to be formed as well as on the size of the Has set of the transmitting UD. On the other hand, although the transmitter in the PMP scheme can encode all the IDNC combinations and schedule a certain number of UDs, the PMP scheme sacrifices the utility of the simultaneous transmissions by considering only one transmission. Our proposed algorithm strikes a balance between these aspects by jointly considering the number of scheduled UDs and the Has set size of each transmitting UD. Second, despite the gain achieved by the FRAN partially-connected D2D solution with the presence of a fog that executes the whole process, our decentralized solution reaches the same performance. This is due to the fact that both schemes follow he same philosophy of D2D simultaneous transmissions which results

in similar completion time performances. Third, although the completion time achieved by the CFG partially-connected D2D scheme is roughly the same as the centralized FRAN partially-connected D2D, the computing time required by our developed scheme is slightly higher than that required by the FRAN partially-connected D2D. This is due to the fact that our proposed scheme needs time to converge before generating the output. The centralized FRAN scheme has low execution time due to the presence of the fog entity.

6.2 Suggested Future Work

This thesis investigates NC schemes for effective resource allocation in cache-enabled 5G systems. Still, there are some interesting research topics for efficient resource allocation utilizations using NC need to be investigated. In what follows, we summarize some potential future extensions of the research conducted in this thesis.

Low complexity power allocation for network-coded scheduling in Cloud-RANs

In Chapter 2, we have developed a joint cross-layer NC method for maximizing the cross-layer throughput in C-RAN setting. The proposed method provides significant performance gain as compared to existing schemes. However, its implementation complexity highly increases with the increase in the number of NC combinations, RRHs, and RRBs, especially for dense networks. Notably, the proposed method requires generating multiple RRB subgraphs whose union gives the CRAN-CLNC graph. One possible extension is that we develop a low-complexity solution that involves the construction of a single RRB subgraph. To do that, we can take advantage of the realistic assumption that channels can typically be approximated by a constant within each RRH's frame to propose a simple solution consists of a single RRB subgraph. Indeed, for a scheduling epoch in a short duration of time, all RRBs in the same RRH's frame do not change [110]. In particular, we consider slow-varying channels, meaning that if the RRB represents a

frequency block, the channel is non-frequency selective. While if the RRB represents a time slot, the duration of the transmission frame is within the channel coherence time. This will make the channel gain of users are constant in a short period of RRH's frame. To further reduce the complexity of the future proposed approach, one can consider generating only potential set of NC combinations that aims to serve users with the maximum throughput.

Cross-layer network codes for file delivery in cache-enabled D2D networks

In the recent literature, most works on file delivery problem with IDNC focused on modeling the status of physical channels by file erasure probabilities and integrated such probability into aiding the coding decisions, e.g., works listed in the survey paper [68]. Further, they considered minimum transmission of all scheduled UDs. This results in prolonged transmission duration and thus, consumes the time resources of network. On the contrary, a more practically relevant scheme is to involve the dynamic nature of wireless channels in the coding decisions. This allows the transmitting UDs to intelligently select their transmission rates and IDNC files. In this direction, the authors of [70] proposed to incorporate the transmission rates in the NC decisions in D2D network. As such, all files are delivered to all users with a minimum completion time. The main drawback is that only one UD is allowed to deliver files in each transmission slot. Thus they ignored the interference caused by different transmitting UDs to the scheduled UDs. Actually, in D2D networks, UDs are spatially separated in a region. This creates an opportunity to intelligently select multiple transmitting UDs that schedule a large number of other UDs. Therefore, as an interesting future work, one can study a realistic scenario that consider multiple transmitting UDs selection and optimization of the employed transmission rates using power control on each transmitting UD. To this end, we aim to develop a novel optimization framework taking network coding and rate/power optimization into account. This framework will be referred to D2D cross-layer network coding (D2D-CLNC). In the proposed D2D-CLNC framework, network-coded transmissions from potential UDs are developed

to deliver all files to all requesting UDs in the least amount of time. The main consideration of this interesting future work can be summarized as follows.

- We show that the completion time minimization problem over power control, transmitting UDs and their file combination selections is computationally intractable due to the interdependent among variables such as the UDs' acquired and requested files, channel qualities and transmitting UDs.
- We derive a lower bound on the completion time for each UD. Using these lower bounds, we iteratively propose a heuristic algorithm that selects a set of transmitting UDs only if the interference caused by the transmissions of the newly selected UDs does not significantly degrade the transmission rates and completion time performance. In other words, the main idea of the heuristic algorithm is iteratively including more transmitting UDs and allocating transmission powers subject to the reduction in the completion time.

Rate-aware network codes for distributed content sharing in cache-enabled D2D network

In Chapter 5, we have developed a distributed and decentralized game theoretical framework for a partially connected D2D network such that we minimize the distribution of contents among UDs. The developed framework focused only on the upper layer view of the network, where network coding is performed at the network layer, and a memory-less erasure channel abstracts the physical-layer. The main drawback of the proposed framework is that it did not consider the interference of transmissions caused by other UDs to the set of transmitting UDs. Consequently, UDs at the intersection region of any transmitting UDs will experience collisions and no packets will be decoded. Further, we only consider D2D communications to deliver all files to all UDs. Even with proactive caching of popular files in the file-holder (FH) UDs, the immediate file delivery to file-requester (FR) UDs may necessitate the help of the edge-servers (ESs) in serving the FR UDs that have not been served by the FH UDs.

Motivated by the aforementioned limitation, developing a rate-aware NC framework that involves immediate transmissions from both ESs and FH UDs. To this end, our goal is to develop a novel distributed file sharing mechanism, by leveraging rate-aware network codes and coalition game theory, to efficiently deliver the requested files within the lowest possible network delay to all the file-requesters (FRs) in a distributed caching network. Therefore, a smart scheduling of file delivery from the FH users and (possibly) the ESs is required to: i) offload the ESs (i.e., minimize the amount of ESs' resources consumed for helping the FH users), and ii) guarantee the required QoS of the UEs. The main consideration of this work can be summarized as follows.

- We first model the file combination that should be transmitted from both the ES and FH users to their corresponding served FR users. Then, the duration of transmission over an edge channel, D2D channel, and the file fetching delay are defined. Given these delays, we formulate completion time minimization problem in cache-enabled D2D network.

- We develop a coalition formation/coalition switch (CF/CS) algorithm that allows the UDs to switch to any coalition for increasing their utility. The switch-operations among coalitions may take long-time to converge. As a result, we aim to develop a low-complexity algorithm.

Bibliography

[1] *Cisco visual networking index: global mobile data traffic forecast update,* 2017-2022, White Paper, Feb., 2019. → pages 1, 23

[2] A. Nordrum, "The internet of fewer things [news]," *IEEE Spectrum, pp.* 12-13, Oct. 2016. → pages 1

[3] J. G. Andrews et al., "What Will 5G Be?," in *IEEE Jou. on S. Areas in Commu.,* vol. 32, no. 6, pp. 1065-1082, Jun. 2014. → pages 1, 4, 23

[4] M. Peng, Y. Sun, X. Li, Z. Mao, and C. Wang, "Recent advances in cloud radio access networks: System architectures, key techniques, and open issues," *IEEE Commun. Surveys and Tutorials,* vol. 18, no. 3, pp. 2282-2308, Third Quarter 2016. → pages 2

[5] B. Dai and W. Yu, "Sparse beamforming for limited-backhaul network MIMO system via reweighted power minimization," in *Proc. of IEEE Global Telecom. Conf. (GLOBECOM' 2013),* Atlanta, GA, USA, Dec. 2013, pp. 1962-1967. → pages 2

[6] Y. Shi, J. Zhang, and K. Letaief, "Group sparse beamforming for green Cloud-RAN," *IEEE Trans. on Wireless Commun.,* vol. 13, no. 5, pp. 2809-2823, May 2014. → pages 2

[7] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, "Joint precoding and multivariate backhaul compression for the downlink of cloud radio access networks," *IEEE Trans. on Signal Proc.,* vol. 61, no. 22, pp. 5646-5658, Nov. 2013. → pages 2, 15, 16, 19

[8] A. Douik, H. Dahrouj, T. Y. Al-Naffouri, and M.-S. Alouini, "Coordinated scheduling for the downlink of cloud radio-access networks," in

Proc. of IEEE Intern. Conf. on Commun. (ICC' 2015), London, UK., 2015. → pages 2

[9] H. Dahrouj, W. Yu, T. Tang, J. Chow, and R. Selea, "Coordinated scheduling for wireless backhaul networks with soft frequency reuse," in *Proc. of the 21st Europea Signal Proc. Conf. (EUSIPCO' 2013),* Marrakech, Morocco, Sep. 2013, pp. 1-5. → pages 2

[10] T. Wang, Y. Sun, L. Song and Z. Han, "Social data offloading in D2D-enhanced cellular networks by network formation games," in *IEEE Trans. on Wireless Commun.,* vol. 14, no. 12, pp. 7004-7015, Dec. 2015. → pages 2, 3

[11] K. Shanmugam, N. Golrezaei, A. Dimakis, A. Molisch, and G. Caire, "Femtocaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. on Inf. Theory,* vol. 59, no. 12, pp. 8402-8413, Dec. 2013. → pages 3, 23

[12] L. Li, G. Zhao, and R. S. Blum, "A survey of caching techniques in cellular networks: Research issues and challenges in content placement and delivery strategies," *IEEE Commun. Surveys and Tutorials,* vol. 20, no. 3, pp. 1710-1732, 3rd Quart., 2018. → pages 3

[13] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, and H. Dai, "A survey on low latency towards 5G: RAN, core network and caching solutions, *IEEE Commun. Surveys and Tutorials,* vol. 20, no. 4, pp. 3098-3130, 4th Quart., 2018. → pages 3

[14] Q. Li, H. Niu, A. Papathanassiou, and G. Wu, "Edge cloud and underlay networks: Empowering 5g cell-less wireless architecture," in *Proc. of European Wireless 2014; 20th European Wireless,* pp. 1-6, May 2014. → pages 3

[15] S. H. Park, O. Simeone, and S. S. Shitz, "Joint optimization of cloud and edge processing for fog radio access networks," *IEEE Trans. on Wireless Commun.,* vol. 15, no. 11, pp. 7621-7632, Nov. 2016. → pages 3

[16] M. Peng and K. Zhang, "Recent advances in fog radio access networks: Performance analysis and radio resource allocation," in *IEEE Access,* vol. 4, pp. 5003-5009, Aug. 2016. → pages 3

[17] R. Tandon and O. Simeone, "Harnessing cloud and edge synergies: toward an information theory of fog radio access networks," *IEEE Commun. Magazine,* vol. 54, no. 8, pp. 44-50, Aug. 2016. → pages 3, 18

[18] C. Fang, F. R. Yu, T. Huang, J. Liu, and Y. Liu, "Energy-efficient distributed in-network caching for content-centric networks," in *Computer Commun., Workshops (INFOCOM WKSHPS),* 2014 IEEE Conf. on. IEEE, 2014, pp. 91-96. → pages 3

[19] A. Asadi, Q. Wang, and V. Mancuso, "A survey on device-to-device communication in cellular networks," *IEEE Commun. Surveys and Tutorials,* vol. 16, no. 4, pp. 1801-1819, 4th Quart., 2014. → pages 3, 23

[20] B. Bangerter, S. Talwar, R. Are, and K. Stewart, "Networks and devices for the 5G era," *IEEE Commun. Mag.,* vol. 52, no. 2, pp. 90-96, Feb. 2014. → pages 3, 23, 134

[21] F. Jameel, Z. Hamid, F. Jabeen, S. Zeadally, and M. A. Javed, "A survey of device-to-device communications: Research issues and challenges," *IEEE Commun. Surveys and Tutorials,* vol. 20, no. 3, pp. 2133-2168, 3rd Quart., 2018. → pages 3, 23, 134

[22] L. Lei, Z. Zhong, C. Lin, and X. Shen, "Operator controlled device-to-device communications in LTE-advanced networks," *IEEE Wireless Commun.,* vol. 19, no. 3, pp. 96-104, Jun. 2012. → pages 3, 23, 74, 134

[23] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.,* vol. 52, no. 2, pp. 74-80, Feb. 2014. → pages 4, 134

[24] Roy Karasik, O. Simeone, and S. Shamai, "How much can D2D communication reduce content delivery latency in Fog networks with edge

caching?," *IEEE Trans. on Commun.,* vol. 68, no. 4, pp. 2308-2323, Apr. 2020. → pages 4, 21, 101, 134

[25] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inf. Theory,* vol. 60, no. 5, pp. 2856-2867, May 2014. → pages 4

[26] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless D2D networks," *IEEE Trans. Inf. Theory,* vol. 62, no. 2, pp. 849-869, Feb. 2016. → pages 4

[27] R. Ahlswede, N. Cai, S.-Y. R. Li, and R. W. Yeung, "Network information ow," *IEEE Trans. Inf. Theory,* vol. 46, no. 4, pp. 1204-1216, Jul. 2000. → pages 5, 6

[28] J. K. Sundararajan, D. Shah, and M. Medard, "Online network coding for optimal throughput and delayThe three-receiver case, in Proc. *IEEE Int. Symp. Inf. Theory Appl. (ISITA),* Auckland, New Zealand, Dec. 2008, pp. 1-6. → pages 5

[29] D. Nguyen, T. Nguyen, and X. Yang, "Multimedia wireless transmission with network coding, in *Proc. Int. Packet Video Workshop (PV),* Lausanne, Switzerland, Nov. 2007, pp. 326-335. → pages 5, 92

[30] D. Nguyen and T. Nguyen, "Network coding-based wireless media transmission using POMDP, in *Proc. Int. Packet Video Workshop (PV),* Seattle, WA, USA, May 2009, pp. 1-9. → pages 5, 89, 92

[31] X. Li, A. Gani, R. Salleh, and O. Zakaria, "The future of mobile wireless communication networks," in *Proc. Int. Conf. Commun. Softw. Netw. (ICCSN),* Chengdu, China, 2009, pp. 554-557. → pages 5

[32] S. Rayanchu, S. Sen, J. Wu, S. Banerjee, and S. Sengupta, "Loss-aware network coding for unicast wireless sessions: Design, implementation, and performance evaluation," in *Proc. ACM Int. Conf. Meas. Model. Comput. Syst. (SIGMETRICS),* Annapolis, MD, USA, 2008, pp. 85-96. → pages 6

[33] S. Rayanchu, S. Sen, J. Wu, S. Banerjee, and S. Sengupta, "Loss-aware network coding for unicast wireless sessions: Design, implementation, and performance evaluation," *SIGMETRICS Perform. Eval. Rev.,* vol. 36, no. 1, pp. 85-96, 2008. → pages 6

[34] T. Ho and D. Lun, Network Coding: An Introduction. New York, NY, USA: Cambridge Univ. Press, 2008. → pages 6

[35] C. Fragouli, D. Lun, M. Medard, and P. Pakzad, "On feedback for network coding," in *Proc. 41st Annu. Conf. Inf. Sci. Syst. (CISS),* Baltimore, MD, USA, pp. 248-252, Mar. 2007. → pages 6

[36] L. Keller, E. Drinea, and C. Fragouli, "Online broadcasting with network coding," in *Proc. IEEE 4th Workshop Netw. Coding Theory Appl. (NetCod),* Hong Kong, Jan. 2008, pp. 1-6. → pages 6

[37] A. Eryilmaz, A. Ozdaglar, M. Medard, and E. Ahmed, "On the delay and throughput gains of coding in unreliable networks," *IEEE Trans. Inf. Theory,* vol. 54, no. 12, pp. 5511-5524, Dec. 2008. → pages 7

[38] P. Sadeghi and M. Yu, "Instantly decodable versus random linear network coding: A comparative framework for throughput and decoding delay performance," *Arxiv e-prints,* vol. abs/1208.2387, 2012. → pages 7, 141

[39] T. Ho et al., "A random linear network coding approach to multicast, *IEEE Trans. Inf. Theory,* vol. 52, no. 10, pp. 4413-4430, Oct. 2006. → pages 7, 24

[40] D. E. Lucani, M. Medard, and M. Stojanovic, "Broadcasting in timedivision duplexing: A random linear network coding approach," in *Proc. Workshop Netw. Coding Theory Appl. (NetCod), Lausanne, Switzerland,* Jun. 2009, pp. 62-67. → pages 7

[41] L. Lima, M. Medard, and J. Barros, "Random linear network coding: A free cipher," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT), Nice, France,* Jun. 2007, pp. 546-550. → pages 7

[42] S.-Y. R. Li, R. W. Yeung, and N. Cai, "Linear network coding," *IEEE Trans. Inf. Theory, vol.* 49, no. 2, pp. 371-381, Feb. 2003. → pages 7

[43] S.-Y. R. Li, N. Cai, and R. W. Yeung, "On theory of linear network coding," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT),* Adelaide, SA, Australia, Sep. 2005, pp. 273-277. → pages 7

[44] S. Sorour and S. Valaee, "Completion delay minimization for instantly decodable network codes," *IEEE/ACM Trans. on Net.,* vol. PP, no. 99, pp. 1553 - 1567, Oct. 2015. → pages 7, 16, 17, 21, 61, 133, 154

[45] D. Traskov, M. Medard, P. Sadeghi, and R. Koetter, "Joint scheduling and instantaneously decodable network coding," in *Proc. IEEE Glob. Telecom. Conf. (GLOBECOM),* Honolulu, HI, USA, Nov. 2009, pp. 1-6. → pages

[46] S. Sorour and S. Valaee, "On minimizing broadcast completion delay for instantly decodable network coding," in *Proc. of IEEE Int. Conf. on Commun. (ICC' 2010),* Cape Town, South Africa, May 2010, pp. 1-5. → pages 7

[47] S. Sorour and S. Valaee, "Minimum broadcast decoding delay for generalized instantly decodable network coding," in *Proc. of IEEE Global Telecom. Conf. (GLOBECOM' 2010),* Miami, Florida, USA, Dec. 2010, pp. 1-5. → pages 7, 9

[48] S. Sorour and S. Valaee, "Completion delay reduction in lossy feedback scenarios for instantly decodable network coding," in *Proc. of IEEE 22nd Int. Symposium on Personal Indoor and Mobile Radio Commun. (PIMRC' 2011),* Toronto, Canada, Sep. 2011, pp. 2025-2029. → pages 7, 16, 17, 21

[49] S. Sorour and S. Valaee, "Adaptive network coded retransmission scheme for wireless multicast," in *Proc. of IEEE Int. Symposium on Inf. Theory (ISIT' 2009),* Seoul, Korea, Jun. 2009, pp. 2577-2581. → pages 7

[50] S. Sorour and S. Valaee, "On densifying coding opportunities in instantly decodable network coding graphs," in *Proc. of IEEE Int. Symposium on Inf. Theory (ISIT' 2012),* Cambridge, MA, USA, Jul. 2012, pp. 2456-2460. → pages 7

[51] X. Li, C.-C. Wang, and X. Lin, "Optimal immediately-decodable intersession network coding (IDNC) schemes for two unicast sessions with hard deadline constraints," in *Proc. 49th Annu. Allerton Conf. Commun. Control Comput. (Allerton),* Monticello, IL, USA, Sep. 2011, pp. 784- 791. → pages 8

[52] L. Lu, M. Xiao, and L. K. Rasmussen, "Design and analysis of relay-aided broadcast using binary network codes," *J. Commun.,* vol. 6, no. 8, pp. 610-617, 2011. → pages 8, 24

[53] E. Drinea, C. Fragouli, and L. Keller, "Delay with network coding and feedback," in *Proc. IEEE ISIT, Seoul,* South Korea, Jun. 2009, pp. 844-848. → pages 8, 24

[54] L. Lu, M. Xiao, and L. K. Rasmussen, "Design and analysis of relay-aided broadcast using binary network codes," *J. Commun.,* vol. 6, no. 8, pp. 610-617, 2011. → pages 8, 24

[55] X. Li, C.-C. Wang, and X. Lin, "On the capacity of immediately decodable coding schemes for wireless stored-video broadcast with hard deadline constraints," *IEEE J. Sel. Areas Commun.,* vol. 29, no. 5, pp. 1094-1105, May 2011. → pages 8

[56] X. Li, C.-C. Wang, and X. Lin, "Optimal immediately-decodable intersession network coding (IDNC) schemes for two unicast sessions with hard deadline constraints," in *Proc. 49th Annu. Allerton Conf. Commun., Control, Comput.,* Monticello, IL, USA, Sep. 2011, pp. 784-791. → pages 8

[57] M. Muhammad, M. Berioli, G. Liva, and G. Giambene, "Instantly decodable network coding protocols with unequal error protection," in *Proc. IEEE Int. Conf. Commun.,* Jun. 2013, pp. 5120-5125. → pages 8

[58] Y. Liu and C. W. Sung, "Quality-aware instantly decodable network coding," *IEEE Trans. Wireless Commun.,* vol. 13, no. 3, pp. 1604-1615, Mar. 2014. → pages 8

[59] Y. Keshtkarjahromi, H. Seferoglu, R. Ansari, and A. Khokhar, "Contentaware instantly decodable network coding over wireless networks," in *Proc. Int. Conf. Comput., Netw. Commun. (ICNC),* Feb. 2015, pp. 803-809. → pages 8

[60] M. S. Karim, P. Sadeghi, S. Sorour, and N. Aboutorab, "Instantly decodable network coding for real-time scalable video broadcast over wireless networks," *EURASIP J. Adv. Signal Process.,* vol. 2016, no. 1, p. 1, Jan. 2016. → pages 8, 21

[61] N. Aboutorab, P. Sadeghi, and S. Tajbakhsh, "Instantly decodable network coding for delay reduction in cooperative data exchange systems," in *Proc. IEEE Int. Symp. Inf. Theory,* Jul. 2013, pp. 3095-3099. → pages 8, 21, 133

[62] N. Aboutorab and P. Sadeghi "Instantly decodable network coding for completion time or delay reduction in cooperative data exchange systems," *IEEE Trans. on Vehicular Tech.* vol. 65, no. 3, pp. 1212-1228, Mar. 2016. → pages 8, 133, 156

[63] S. E. Tajbakhsh and P. Sadeghi, "Coded cooperative data exchange for multiple unicasts," in *Proc. IEEE Inf. Theory Workshop,* Sep. 2012, pp. 587-591. → pages 8, 133, 156

[64] A. Douik, S. Sorour, T. Y. Al-Naffouri, H.-C. Yang, and M.-S. Alouini, "Delay reduction in multi-hop device-to-device communication using network coding," in *Proc. IEEE Int. Symp. Netw. Coding,* 2015, pp. 6-10. → pages 8, 133, 155

[65] S. E. Tajbakhsh and P. Sadeghi, "Coded cooperative data exchange for multiple unicasts," in *Proc. IEEE Inf. Theory Workshop,* Sep. 2012, pp. 587-591. → pages 8, 133

[66] A. Douik, S. Sorour, T. Y. Al-Naffouri, H.-C. Yang, and M.-S. Alouini, "Delay reduction in multi-hop device-to-device communication using network coding," *IEEE Trans. Wireless Commun.,* vol. 17, no. 10, Oct. 2018. → pages 8, 133

[67] A. Douik, and S. Sorour, "Data dissemination using instantly decodable binary codes in fog radio access networks," *IEEE Transactions on Commun.,* vol. 66, no. 5, pp. 2052-2064, May 2018. → pages 8, 21, 24, 102, 133, 144, 155, 156

[68] A. Douik, S. Sorour, T. Y. Al-Naffouri, and M.-S. Alouini, "Instantly decodable network coding: From centralized to device-to-device communications," *IEEE Commun. Surveys and Tutorials,* vol. 19, no. 2, pp. 1201-1224, 2nd Quart., 2017. → pages 8, 10, 11, 14, 16, 17, 24, 35, 102, 133, 169

[69] A. Douik, S. Sorour, H. Tembine, T. Y. Al-Naffouri, and M.-S. Alouini "A game-theoretic framework for decentralized cooperative data exchange using network coding," *IEEE Trans. Mobile Comput.,* vol. 16, no. 4, pp. 901-917, Apr. 2017. → pages 8, 133, 143

[70] M. S. Karim, A. Douik, S. Sorour, and P. Sadeghi, "Rate-aware network codes for completion time reduction in device-to-device communications," in *IEEE Int. Conf. on Commu. (ICC),* 2016, pp. 1-7. → pages 8, 13, 169

[71] S. Y. El Rouayheb, M. A. R. Chaudhry, and A. Sprintson, "On the minimum number of transmissions in single-hop wireless coding networks," in *Proc. IEEE Inf. Theory Workshop (ITW),* Bergen, Norway, Sep. 2007, pp. 120-125. → pages 9, 21

[72] M. A. R. Chaudhry and A. Sprintson, "Efficient algorithms for index coding," in *Proc. IEEE 27th Conf. Comput. Commun. Workshops (INFOCOM),* Phoenix, AZ, USA, Apr. 2008, pp. 1-4. → pages 9

[73] M. Mdard and A. Sprintson, Network Coding: *Fundamentals and Applications*. Amsterdam, The Netherlands: Academic Press, 2012. → pages 10

[74] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness.* New York, NY, USA: Freeman, 1990. → pages 11, 85, 114

[75] G. Ausiello et al., *Complexity and Approximation: Combinatorial Optimization Problems and Their Approximability Properties,* 1st ed. New York, NY, USA: Springer-Verlag, 1999. → pages 11

[76] J. Harris, J. L. Hirst, and M. Mossingho, "Combinatorics and graph theory," in *Undergraduate Texts in Mathematics and Technology.* New York, NY, USA: Springer, 2008. → pages 11

[77] I. M. Bomze, M. Budinich, P. M. Pardalos, and M. Pelillo, "The maximum clique problem," in *Handbook of Combinatorial Optimization.* Boston, MA, USA: Kluwer, 1999, pp. 1-74. → pages 11

[78] A. Douik, S. Sorour, T.-Y. Al-Naffouri, and M.-S. Alouini, "Rate aware instantly decodable network codes," *IEEE Trans. on Wireless Commun.,* vol. 16, no. 2, pp. 998-1011, Feb. 2017. → pages 13, 17, 21, 22

[79] M. S. Karim, A. Douik, and S. Sorour, "Rate-aware network codes for video distortion reduction in point-to-multipoint networks," *IEEE Trans. on Vehicular Tech.,* vol. 66, no. 8, pp. 7446-7460, Aug. 2017. → pages 13, 17, 21, 22

[80] M. Saif, A. Douik and S. Sorour, "Rate aware network codes for coordinated multi base-station networks," *2016 IEEE Int. Conf. on Commun. (ICC),* Kuala Lumpur, 2016, pp. 1-7. → pages 13, 17, 21, 22

[81] X. Wang, C. Yuen, and Y. Xu, "Coding based data broadcasting for time critical applications with rate adaptation," *IEEE Trans. on Vehicular Tech.,* vol. 63, no. 5, pp. 2429-2442, Jun. 2014. → pages 13, 17, 21, 22

[82] M. S. Al-Abiad, A. Douik, and S. Sorour, "Rate aware network codes for cloud radio access networks," *IEEE Trans. on Mobile Comp.,* vol. 18, no 8, pp 1898-1910, Aug. 2019. → pages 13, 17, 21, 22, 62, 63, 72, 106, 111, 125, 126

[83] R. B. Myerson, "Game Theory, Analysis of Conflict," Cambridge, MA, USA: Harvard University Press, Sep. 1991. → pages 14, 25, 139, 142

[84] E. Koutsoupias and C. Papadimitriou, "Worst-case equilibria," in *Proc. of the 16th Annual Symposium On Theoretical Aspects Of Computer Science (STACS' 1999),* Trier, Germany, 1999, pp. 404-413. → pages 14

[85] O. Simeone, A. Maeder, M. Peng, O. Sahin, and W. Yu, "Cloud radio access network: Virtualizing wireless access for dense heterogeneous systems," *J. Commun. Netw.,* vol. 18, no. 2, pp. 135-149, Apr. 2016. → pages 15

[86] R. Zakhour and D. Gesbert, "Optimized data sharing in multicell MIMO with finite backhaul capacity," *IEEE Trans. Signal Process.,* vol. 59, no. 12, pp. 6102-6111, Dec. 2011. → pages 15

[87] B. Dai and W. Yu, "Sparse beamforming and user-centric clustering for downlink cloud radio access network," in *IEEE Access,* vol. 2, pp. 1326-1339, Oct. 2014. → pages 15

[88] S. Smirani, M. Kamoun, M. Sarkiss, A. Zaidi, and P. Duhamel, "Achievable rate regions for two-way relay channel using nested lattice coding," *IEEE Trans. on Wireless Commun.,* vol. 13, no. 10, pp. 5607-5620, Oct. 2014. → pages 15

[89] I.-E. Aguerri and A. Zaidi, "Lossy compression for compute-andforward in limited backhaul uplink multicell processing," in *IEEE Trans. on Commun.,* vol. 64, no. 12, pp. 5227-5238, Dec. 2016. → pages 15

[90] N. Liu and W. Kang, "A new achievability scheme for downlink multicell processing with finite backhaul capacity," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT),* Honolulu, HI, USA, Jun./Jul. 2014. → pages 15

[91] C.-Yi Wang, M. Wigger, and A. Zaidi, "On achievability for downlink cloud radio access networks with base station cooperation," *IEEE Trans. on Inf. Theory,* vol. 64, no. 8, pp. 5726-5742, Aug. 2018. → pages 15

[92] L. Liu and W. Yu. "Cross-layer design for downlink multihop cloud radio access networks with network coding," *IEEE Trans. Signal Process.,* vol. 65, no. 7, pp. 1728- 1740, Apr. 2017. → pages 15

[93] Y. Wu, P.-A. Chou, Q. Zhang, K. Jain, W. Zhu, and S.-Y. Kung, "Network planning in wireless Ad Hoc networks: a cross-layer approach," *IEEE journal on Selected Areas in Commun.,* vol. 23, no. 1, pp. 136-150, Jan. 2005. → pages 15

[94] Y. E. Sagduyu and A. Ephremides, "On joint MAC and network coding in wireless Ad Hoc networks," *IEEE Trans. on Inf. Theory,* vol. 53, no. 10, pp. 3697-3713, Oct. 2007. → pages 15

[95] Y. E. Sagduyu and A. Ephremides, "Cross-layer optimization of MAC and network coding in wireless queueing tandem networks," *IEEE Trans. on Inf. Theory,* vol. 54, no. 2, pp. 554-571, Feb. 2008. → pages 15

[96] A. Khreishah, C.-C.Wang, and N.-B. Shro, "Cross-layer optimization for wireless multihop networks with pairwise intersession network coding," *IEEE journal on Selected Areas in Commun.,* vol. 27, no. 5, pp. 606-621, Jun. 2009. → pages 15

[97] J. Huang, V.-G. Subramanian, R. Agrawal, and R. Berry, "Downlink scheduling and resource allocation for OFDM systems," in *Process of Conf. Info. Science Sys. (CISS),* Mar. 2006. → pages 15, 16

[98] J. Huang, V.-G. Subramanian, R. Agrawal, and R. Berry, "Joint scheduling and resource allocation in uplink OFDM systems for broadband wireless access networks," in *IEEE J. Sel. Top. Signal Proc.,* vol. 27, no. 2, pp. 226-234, Feb. 2009. → pages 15, 16

[99] W. Yu, T. Kwon, and C. Shin, "Joint scheduling and dynamic power spectrum optimization for wireless multicell networks," in *Proc. of 2010*

44th Annual Conf. on Inf. Sciences and Systems (CISS 2010), Princeton, New Jersey, USA, pp. 1-6, Mar. 2010. → pages 15, 16

[100] L. Venturino, N. Prasad, and X. Wang, "Coordinated scheduling and power allocation in downlink multicell OFDMA networks," *IEEE Trans. on Vehicular Tech.,* vol. 58, no. 6, pp. 2835-2848, Jul. 2009. → pages 15

[101] Y. Shi, J. Zhang, and K.-B. Letaief, "Group sparse beamforming for green cloud-ran," *IEEE Trans. on Wireless Commun.,* vol. 13, no. 5, pp. 2809-2823, May 2014. → pages 15

[102] S. Kiani and D. Gesbert, "Optimal and distributed scheduling for multicell capacity maximization," *IEEE Trans. on Wireless Commun.,* vol. 7, no. 1, pp. 288-297, Jan. 2008. → pages 15

[103] W. Yu, T. Kwon, and C. Shin, "Multicell coordination via joint scheduling, beamforming, and power spectrum adaptation," *IEEE Trans. Wireless Commun.,* vol. 12, no. 7, pp. 1-14, Jul. 2013. → pages 15, 16

[104] A. Stolyar and H. Viswanathan, "Self-organizing dynamic fractional frequency reuse for best-effort traffic through distributed inter-cell coordination," in *Proc. 28th IEEE Conf. Comput. Commun. (INFOCOM09),* Rio de Janeiro, Brazil, pp. 1287-1295, Apr. 2009. → pages 15, 16

[105] R. Bendlin, Y.-F. Huang, M. Ivrlac, and J. Nossek, "Fast distributed multicell scheduling with delayed limited-capacity backhaul links," in *Proc. IEEE Int. Conf. Commun. (ICC09),* Dresden, Germany, pp. 1-5, Jun. 2009. → pages 15, 16

[106] L. Jiang, S. Parekh, and J. Walrand, "Base station association game in multi-cell wireless networks (special paper)," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC08),* Las Vegas, Nevada, USA, pp. 1616-1621, Mar. 2008. → pages 15, 16

[107] R. Sun, M. Hong, and Z.-Q. Luo, "Optimal joint base station assignment and power allocation in a cellular network," in *Proc. IEEE 13th Int.*

Workshop Signal Process. Adv. Wireless Commun. (SPAWC12), Cesme, Turkey, pp. 234-238, Jun. 2012. → pages 15, 16

[108] J.-W. Lee, R. Mazumdar, and N. Shro, "Joint resource allocation and base-station assignment for the downlink in CDMA networks," *IEEE/ACM Trans. Netw.,* vol. 14, no. 1, pp. 1-14, Feb. 2006. → pages 15, 16

[109] A. Douik, H. Dahrouj, T.-Y. Al-Naffouri, and M.-S. Alouini, "Coordinated scheduling and power control in cloud-radio access networks," *IEEE Trans. on Wireless Commun.,* vol. 15, no. 4, pp. 2523-2536, Apr. 2016. → pages 15, 16, 19, 37, 53, 62, 63, 72, 84, 92, 93, 192, 196

[110] A. Douik, H. Dahrouj, T.-Y. Al-Naffouri, and M.-S. Alouini, "Low-complexity scheduling and power adaptation for coordinated cloud-radio access networks," *IEEE Commun. Letters,* vol. 21, no. 10, pp. 2298-2301, Oct. 2017. → pages 15, 16, 168

[111] S. Katti, D. Katabi, W. Hu, H. Rahul, and M. Medard, "The importance of being opportunistic: practical network coding for wireless environments," in *Proc. of Annual Allerton Conf. on Commun., Control and Computing (Allerton 2005),* Monticello, Illinois, USA, 2005. → pages 16, 17

[112] A. Afshin, M. Khabbazian, M. Ardakani, and G. Bansal "Blind instantly decodable network codes for wireless broadcast of real-time multimedia," *IEEE Trans. on Wireless Commun.,* vol. 17, no. 4, pp. 2276-2288, Apr. 2018. → pages 16, 17

[113] L. Gou, G. Zhang, Z. Bian, D. Bian, and Z. Xie, "Minimizing completion time for relay-assisted multicast with instantly decodable network coding", *IEEE Commun. Letters,* vol. 20, pp. 434-437, Mar. 2016. → pages 16, 17

[114] L. Gou, G. Zhang, D. Bian, W. Zhang, and Z. Xie, "Data dissemination in wireless sensor networks with instantly decodable network

coding," *Journal of Commun. and Networks,* vol. 18, pp. 846-856, Oct. 2016. → pages 16, 17

[115] A. Douik, S. Sorour, M. S. Alouini, and T. Y. Al-Naffouri, "Completion time reduction in instantly decodable network coding through decoding delay control," in *Proc. of IEEE Global Telecom. Conf. (GLOBECOM 2014),* Austin, Texas, USA, pp. 5008-5013, Dec. 2014. → pages 16, 17, 106, 156

[116] A. Le, A.-S. Tehrani, A.-G. Dimakis, and A. Markopoulou, "Instantly decodable network codes for real-time applications," in *Proc. of 2013 Int. Symp. on Network Coding (NetCod),* Calgary, AB, Canada, pp 1-6, Jun. 2013. → pages 16, 17, 44, 109

[117] M. S. Al-Abiad, A. Douik, S. Sorour, and Md. J. Hossain, "Throughput maximization in cloud radio access networks using network codes," in *Proc. of IEEE Int. Conf. on Commun. Workshops (ICC 2018),* Kansas, USA, May, 2018, pp 1-6. → pages 22, 32

[118] M. S. Al-Abiad, A. Douik, S. Sorour, and Md. J. Hossain, "Throughput maximization in cloud-radio access networks using rate-aware network Coding," *IEEE Trans. Mobile Comput,* Early Access, Aug. 2020. → pages 22, 32

[119] R. Tandon and O. Simeone, "Cloud-aided wireless networks with edge caching: Fundamental latency trade-offs in fog radio access networks," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT),* Barcelona, Spain, Jul. 2016, pp. 2029-2033. → pages 19

[120] A. Sengupta, R. Tandon, and O. Simeone. (May 2016). "Cloud and cache-aided wireless networks: Fundamental latency trade-offs." [Online]. Available: *https://arxiv.org/abs/1605.01690.* → pages 19

[121] S. Gitzenis, G. S. Paschos, and L. Tassiulas, "Asymptotic laws for joint content replication and delivery in wireless networks," *IEEE Trans. on Inf. Theory,* vol. 59, no. 5, pp. 2760-2776, May 2013 → pages 19

[122] K. Shanmugam, N. Golrezaei, A. Dimakis, A. Molisch, and G. Caire, "Femtocaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. on Inf. Theory,* vol. 59, no. 12, pp. 8402-8413, Dec. 2013. → pages 19, 62

[123] J. Hachem, N. Karamchandani, and S. Diggavi, "Multi-level coded caching," in *proc. of IEEE ISIT,* pp. 56-60, Jun. 2014. → pages 19

[124] W. Chen, K. Letaief, and Z. Cao, "Opportunistic network coding for wireless networks," in Proc. of *IEEE ICC,* pp. 4634-4639, Jun. 2007. → pages 19

[125] Ahmed A. Al-Habob, Y. Shnaiwer, S. Sorour, N. Aboutorab, and P. Sadeghi, "Multi-client file download time reduction from cloud/fog storage servers," *IEEE Trans. Mobile Comput.,* vol. 17, pp. 1924-1937, Dec. 2017. → pages 19

[126] Y. Shnaiwer, S. Sorour, P. Sadeghi, N. Aboutorab, and T. Y. Al-Naffouri, "Network-coded macrocell offloading in femtocaching-assisted cellular networks," *IEEE Trans. on Vehicular Technology,* vol. PP, no. 99, pp. 1-1, Nov. 2017. → pages 19, 20, 93

[127] Y. Shnaiwer, S. Sorour, T. Y. Al-Naffouri, S. Al-Ghadhban, "Cross layer cloud offloading using fog radio access networks and network coding," in *Proc. of IEEE ICC18,* Kansas City, MO, USA, May 2018. → pages 20, 93

[128] M. S. Al-Abiad, S. Sorour, and Md. J. Hossain, "Cloud offloading with QoS provisioning using cross-layer network coding," in *Proc. of IEEE Globecom18,* Abu Dhabi, UAE, 2018. → pages 22, 33, 90

[129] M. S. Al-Abiad, M. J. Hossain, and S. Sorour, "Cross-layer cloud offloading with quality of service guarantees in Fog-RANs," in *IEEE Trans. on Commun.,* vol. 67, no. 12, pp. 8435-8449, Jun. 2019. → pages 22, 33, 100

[130] C. Xu, C. Gao, Z. Zhou, Z. Chang, and Y. Jia, "Social network-based content delivery in device-to-device underlay cellular networks using matching theory," in *IEEE Access,* vol. 5, pp. 924-937, Nov. 2016. → pages 23

[131] A. Shokrollahi, "Raptor codes," *IEEE Trans. Inf. Theory,* vol. 52, no. 6, pp. 2551-2567, Jun. 2006. → pages 24

[132] A. Douik, S. Sorour, H. Tembine, T. Y. Al-Naffouri, and M.-S. Alouini, "A game-theoretic framework for decentralized cooperative data exchange using network coding," *IEEE Trans. Mobile Comput.,* vol. 16, no. 4, pp. 901-917, Apr. 2017. → pages 24

[133] A. Douik, S. Sorour, H. Tembine, T. Y. Al-Naffouri, and M.-S. Alouini, "A game theoretic approach to minimize the completion time of network coded cooperative data exchange," *IEEE Global Commun. Conference,* Austin, TX, 2014, pp. 1583-1589. Apr. 2017. → pages 24

[134] W. Saad, Z. Han, M. Debbah, Are H., and T. Basar, "Coalition game theory for communication networks: A tutorial," *IEEE Signal Processing Mag., Special issue on Game Theory in Sig. Pro. and Com.,* vol. 26, no. 5, pp. 77-97, Sep. 2009. → pages 25, 142

[135] W. Saad, Z. Han, M. Debbah, and Are Hjrungnes, "A distributed coalition formation framework for fair user cooperation in wireless networks," *IEEE Trans. Wireless Commun.,* vol. 8, no. 9, pp. 4580-4593, Sep. 2009. → pages 25, 142

[136] K. Apt and A. Witzel, "A generic approach to coalition formation (extended version)," in *Int. Game Theory Rev.,* vol. 11, no. 3, pp. 347-367, Mar. 2009. → pages 25, 142, 151

[137] K. Akkarajitsakul, E. Hossain, and D. Niyato, "Coalition-based cooperative packet delivery under uncertainty: A dynamic bayesian coalitional Game," *IEEE Trans. Mobile Comput.,* vol. 12, no. 2, pp. 371-385, Feb. 2013. → pages 25

[138] L. Militano et al., "A constrained coalition formation game for multihop D2D content uploading," *IEEE Trans. Wireless Commun.,* vol. 15, no. 3, pp. 2012-2024, Mar. 2016. → pages 25, 142

[139] C. Xu, J. Feng, Z. Zhou, J. Wu, and C. Perera, "Cross-layer optimization for cooperative content distribution in multihop device-to-device networks," in *IEEE Internet of Things Journal,* vol. 6, no. 1, pp. 278-287, Feb. 2019. → pages 25

[140] Zhao, Y., Li, Y., Ding, and Z. et al. "A coalitional graph game framework for network coding-aided D2D communication," *EURASIP J. Adv. Signal Process.,* 2016, 2 (2016). → pages 25

[141] C. Gao, Y. Li, Y. Zhao, and S. Chen, "A two-level game theory approach for joint relay selection and resource allocation in network coding assisted D2D communications", in *IEEE Trans. Mobile Comput.,* vol. 16, no. 10, pp. 2697-2711, 1 Oct. 2017. → pages 25

[142] Y. Dong, Md. J. Hossain, and J. Cheng, "Joint power control and subchannel allocation for D2D communications underlaying cellular networks: A coalitional game perspective", in *Proc. GameNets 2016,* Kelowna, BC, CA, pp. 1–14, May 2016. → pages 25

[143] A. Antonopoulos and C. Verikoukis, "Multi-player game theoretic MAC strategies for energy efficient data dissemination," in *IEEE Trans. Wireless Commun.,* vol. 13, no. 2, pp. 592-603, Feb. 2014. → pages 25

[144] A. Antonopoulos, J. Bastos, and C. Verikoukis, "Analogue network coding-aided game theoretic medium access control protocol for energy-efficient data dissemination," in *IET Science, Measurement and Technology,* vol. 8, no. 6, pp. 399-407, Nov. 2014. → pages 25

[145] M. Zayene, O. Habachi, V. Meghdadi, T. Ezzedine, and J. P. Cances, "A coalitional game-theoretic framework for cooperative data exchange using instantly decodable network coding," in *IEEE Access,* vol. 7, pp. 26752-26765, Feb. 2019. → pages 25, 76

[146] M. S. Al-Abiad and M. J. Hossain, "Completion time minimization in Fog-RANs using D2D communications and rate-aware network coding," *IEEE Trans. Wireless Commun.,* submitted for publication. → pages 33

[147] M. S. Al-Abiad, A. Douik, and M. J. Hossain, "Coalition formation game for cooperative content delivery in network coding assisted D2D communications," in *IEEE Access*, Early Access, pp.1-1, Sept. 2020. → pages 34, 114

[148] N. Bourgeois, B. Escoffier, V.-T. Paschos, and J-M.-M. van Rooij, "A bottom-up method and fast algorithms for max independent set," in *Proc. of the 12th Scandinavian Conf.* on Algorithm Theory (SWAT' 2010), Bergen, Norway. → pages 48, 56, 114

[149] K. Yamaguchi and S. Masuda, "A new exact algorithm for the maximum weight clique problem," in *Proc. Of the 23rd Int. Technical Conf. on Circuits/Systems, Computers and Commun. (ITCCSCC 2008),* Yamaguchi, Japan. → pages 48, 56, 85

[150] P.-R.-J. Ostergard, "A fast algorithm for the maximum clique problem," *Discrete Appl. Math,* vol. 120, pp. 197-207. → pages 48, 56, 85

[151] H. Dahrouj, W. Yu, and T. Tang, "Power spectrum optimization for interference mitigation via iterative function evaluation," *EURASIP Journal on Wireless Commun. and Net.,* vol. 2012, no. 1, Dec. 2012. → pages 49, 59

[152] L.-P. Qian, Y. Zhang, and J. Huang, "Mapel: Achieving global optimality for a non-convex wireless power control problem," *IEEE Trans. on Wireless Commun.,* vol. 8, no. 3, pp. 1553-1563, Mar. 2009. → pages 49

[153] W. Yu, "Multiuser water-filling in the presence of crosstalk," in *Proc. of Inf. Theory and Applications Workshop (ITA 2007),* San Diego, CA, USA, Jan. 2007, pp. 414-420. → pages 59

[154] G. Sharma, R. R. Mazumdar, and N. B. Shro, "On the complexity of scheduling in wireless networks," in *ACM Int. Conf. Mobile Comput. Netw. (MobiCom),* pp. 227-238, Sep. 2006. → pages 80

[155] Z.-Q. Luo and S. Zhang, "Dynamic spectrum management: Complexity and duality," *IEEE Trans. Signal Processing,* vol. 2, no. 1, pp. 57-73, Feb. 2008. → pages 80

[156] V. Estivill-Castro and D. Wood, "A survey of adaptive sorting algorithms," *ACM Computing Surveys (CSUR),* vol. 24, no. 4, pp. 441-476, Dec. 1992. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0012365X88901148. → pages 85

[157] Z. Han and K. J. Liu, Resource Allocation for Wireless Networks: Basics, Techniques, and Applications. Cambridge University Press, 2008. → pages 145, 154

[158] K. Apt and T. Radzik, "Stable partitions in coalitional games," 2006. Available: *http://arxiv.org/abs/cs/0605132* → pages 151

[159] B. P. Deosarkar, N. S. Yadav, and R. P. Yadav, "Cluster-head selection in clustering algorithms for wireless sensor networks: a survey," *Proc. 2008 IEEE Int. Conf. on Computing, Commun. and Net.,* pp. 1-8, Feb. 2009. → pages 154

[160] A. A. Abbasi and M. Younis, "A survey on clustering algorithms for wireless sensor networks," *Computer Commun.,* vol. 30, pp. 2826-2841, 2007. → pages 154

Appendix A

Proof of Theorem 2.1

To prove Theorem 2.1, we first need to show that for any feasible schedule S, there exists a unique clique $\mathtt{C}$ in CRAN-CLNC graph which resulting from Algorithm 1 and a solution to problem (2.5). Then, for each resulting clique $\mathtt{C}$ in CRAN-CLNC graph based on Algorithm 1, there exist a unique feasible schedule $\mathtt{S}$. In other words, we prove that there is a one-to-one mapping between the set of cliques of a maximum size of Z that Algorithm 1 can generate and the set of the feasible schedules. Finally, the proof can be concluded by showing that the contributed weight of the clique $\mathtt{C}$ is the objective function (2.5) to be maximized.

Let $\mathtt{S} = \{\{\mathbf{S}_{z_1}, \cdots, \mathbf{S}_{z_Z}\} = \{\{s_1, s_2, \cdots, s_{|\tau_{b_1 z_1}|}\}, \cdots, \{s_B, \cdots, s_{|\tau_{b_B z_Z}|}\}\}$ be a feasible schedule (i.e., $\mathtt{S} \in \mathcal{S}_{\text{fs}}$).[5] Since each RRB $z_m \in \mathcal{Z}$ across all RRHs contributes with only one feasible schedule, $|\mathtt{S}| = Z$. Proving that $\mathtt{S}$ can be represented by a maximal clique in the CRAN-CLNC graph equivalents to show that $\mathtt{S}$ can be decomposed into Z schedules each representing a vertex in a local power subgraph. Afterwards, we show that the vertices satisfy **GC** condition that explained in Section 2.3.

Define $\mathbf{S}_{z_m}$ as a specific schedule that represents the set of associations for which RRBs is indexed by z_m. This schedule can be mathematically expressed as $\mathbf{S}_{z_m} = \{s \in \mathtt{S} | z_m(s) = z_m\}$. It is important to note that all associations in $\mathtt{S}$ represent coded combinations in the first place (refer to Section 2.2.2 for more details). By the assumption that $\mathbf{S}_{z_m} \in \mathtt{S}$ is a feasible schedule, thus it satisfies **LC2**. Therefore, $\mathbf{S}_{z_m}$ satisfy **LC1** and **LC2** and represents a vertex in the local power subgraph $\mathcal{G}_{z_m}$. Now, we show that the vertices across all subgraphs are connected by **GC**. The general condition

[5] This proof considers only the coded case as the uncoded case can follow the same steps that used in [109].

GC of the CRAN-CLNC graph insists that each user cannot be targeted by distinct RRHs, which represents the system network **CC1**. Hence, all vertices in the CRAN-CLNC graph are connected as long as the associated users are distinct. Since each RRB's subgraph contributes by one unique vertex, the union of all these connected vertices given **GC** is a unique clique. Therefore, $\mathtt{C} = \{v_{z_1}, v_{z_2}, \cdots, v_{z_Z}\}$ is a maximal clique of size Z in the CRAN-CLNC graph. The feasible schedule $\mathtt{S}$ is decomposed into a set of $\mathbf{S}_{z_m}$, $z_m \in \mathcal{Z}$ in which each $\mathbf{S}_{z_m}$ consists of unique associations. Clearly, the clique $\mathtt{C}$ representing the schedule $\mathtt{S}$ is unique. In the following, we prove the converse.

Define $\mathtt{C}$ as the set of all possible maximal cliques of degree Z in the CRAN-CLNC graph. Let $\mathtt{C} = \{v_{z_1}, v_{z_2}, \cdots, v_{z_Z}\} \in \mathtt{C}$ be a maximal clique. From Section 3.2, it can be noted that each $v_{z_m} \in \mathcal{G}_{z_m}$ represents at least B associations (i.e., uncoded scenario) or at most U associations. By the configuration of the vertices $v_{z_m} \in \mathcal{G}_{z_m}$, we have $v_{z_m} \in \mathcal{S}_{z_m,\mathrm{fs}}$. Thus, vertices in any subgraph satisfy **LC1** and **LC2** conditions. Since each subgraph is configured for each RRB, then by extension $\mathtt{C}$ satisfies **GC** condition. Next, we analyze the vertices included in the clique $\mathtt{C}$. In particular, we aim to show that there exists a one-to-one mapping between vertices and the subgraphs $\mathcal{G}_{z_m}$, $z_m \in \mathcal{Z}$. In other terms, each subgraph contributes by only one unique vertex to $\mathtt{C}$, and two different vertices belong to two different subgraphs. These unique configurations of the vertices in the subgraph $\mathcal{G}_{z_m}$ leads to the uniqueness of the schedules. From the above analysis and the general connectivity condition **GC** for any two different subgraphs, clearly, $\mathtt{C}$ satisfy **CC1**. Consequently, the resulting policy of Algorithm 1, i.e., $\mathtt{C}$ represents a feasible schedule. Thus, a one-to-one mapping between them exists.

To finalize the proof, we now prove that the weight of the maximal clique is the objective function in (2.5) to be maximized. Let $\mathtt{C} = \{v_{z_1}, \cdots, v_{z_Z}\} = \{\{s_1, s_2, \cdots, s_{|\tau_{b_1 z_1}|}\}, \cdots, \{s_B, \cdots, s_{|\tau_{b_B z_Z}|}\}\} = \mathtt{S}$, $v_{z_m} \in \mathcal{G}_{z_m}$. Let a vertex $v_{z_m} \in \mathcal{V}_{z_m}$ is associated with $\mathbf{S}_{z_m} = \{s_1, \cdots, s_{|\mathbf{S}_{z_m}|}\} \in \mathcal{S}_{z_m,\mathrm{fs}}$, where $\mathbf{S}_{z_m}$ represents the total targeted users of $\sum_{b_n \in B} |\tau_{b_n z_m}(\kappa_{b_n z_m})|$. The weight of the

maximal clique can be written as

$$w(\mathtt{C}) = \sum_{v_{z_m} \in \mathtt{C}} w(v_{z_m}) = \sum_{z_m \in \mathcal{Z}} \sum_{\mathbf{S}_{z_m} \in \mathcal{S}_{z_m,\mathrm{fs}}} w(\mathbf{S}_{z_m}) = \sum_{z_m \in \mathcal{Z}} \sum_{\mathbf{S}_{z_m} \in \mathcal{S}_{z_m,\mathrm{fs}}} \sum_{s \in \mathbf{S}_{z_m}} \min_{u_i(s) \in \tau_{b_n(s) z_m(s)}} \log_2(1 + \mathrm{SINR}^{u_i(s)}_{b_n(s) z_m(s)}(\mathbf{P})) \tag{A.1}$$

Therefore, the problem of maximizing the throughput (2.5) in C-RAN setting is equivalent to the maximum weight clique problem among the maximal cliques in the CRAN-CLNC graph.

Table B.1: All possible IDNC combinations $\mathcal{S}_{\text{IDNC}}$

i	$\mathbf{s}_i = (\kappa_i, \tau_i)$
1	$((f_1 \oplus f_2), (u_1, u_2))$
2	$((f_1), (u_1))$
3	$((f_1 \oplus f_3), (u_1, u_3))$
4	$((f_2), (u_2))$
5	$((f_2 \oplus f_3), (u_2, u_3))$
6	$((f_3), (u_3))$
7	$((f_1 \oplus f_2 \oplus f_3), (u_1, u_2, u_3))$

Appendix B

Illustration of Algorithm 1

We illustrate here how to use Algorithm 1 to design the CRAN-CLNC graph shown in Figure 2.3 of the C-RAN model presented in Example 1. In Example 1, a simple C-RAN composed of 2 RRHs, 1 RRB, 3 users, and 3 files. A step-by-step illustration is given as follows.

First step: We generate the set of all possible associations between users and their requested files, i.e., $\mathcal{S} = \{(u_1, f_1), (u_2, f_2), (u_3, f_3)\}$. Then, based on the instant decodability conditions **C1** and **C2** that explained in Section 2.3, we generate all IDNC file combinations $\mathcal{S}_{\text{IDNC}}$ as summarized in Table B.1.

Second step: We generate feasible schedules $\mathbf{S}_i \in \mathcal{S}_{z_1,\text{fs}}, \forall i = 1, \cdots, |\mathcal{S}_{z_1,\text{fs}}|$, such that each schedule $\mathbf{S}$ consists of many associations s, in which they satisfy **CC1**. Thus, $\mathbf{S}_i$ is represented by a vertex v_i in the power control subgraph as described in Section 2.3. Table B.2 summarizes all these feasible schedules $\mathcal{S}_{z_1,\text{fs}}$.

Third step: We solve the power problem for each vertex v_i as explained in Section 2.3. Then, we use the optimal power distribution $\mathbf{P}$ to calculate the weight of that vertex v_i as in (2.10). This weight reflects the total contri-

Table B.2: All feasible schedules $\mathcal{S}_{z_1,\text{fs}}$

i	$\mathbf{S}_i = \{s_1, \cdots, s_{\lvert\mathbf{S}_i\rvert}\}$	i	$\mathbf{S}_i = \{s_1, \cdots, s_{\lvert\mathbf{S}_i\rvert}\}$
1	$\{b_1 z_1 \mathbf{s}_1 R_{b_1 z_1}, b_2 z_1 \mathbf{s}_6 R_{b_2 z_1}\}$	9	$\{b_2 z_1 \mathbf{s}_1 R_{b_2 z_1}, b_1 z_1 \mathbf{s}_6 R_{b_1 z_1}\}$
2	$\{b_1 z_1 \mathbf{s}_5 R_{b_1 z_1}, b_2 z_1 \mathbf{s}_2 R_{b_2 z_1}\}$	10	$\{b_2 z_1 \mathbf{s}_7 R_{b_2 z_1}\}$
3	$\{b_1 z_1 \mathbf{s}_7 R_{b_1 z_1}\}$	11	$\{b_1 z_1 \mathbf{s}_2 R_{b_1 z_1}, b_2 z_1 \mathbf{s}_4 R_{b_2 z_1}\}$
4	$\{b_2 z_1 \mathbf{s}_2 R_{b_2 z_1}, b_1 z_1 \mathbf{s}_6 R_{b_1 z_1}\}$	12	$\{b_1 z_1 \mathbf{s}_6 R_{b_1 z_1}, b_2 z_1 \mathbf{s}_4 R_{b_2 z_1}\}$
5	$\{b_1 z_1 \mathbf{s}_6 R_{b_1 z_1}, b_2 z_1 \mathbf{s}_1 R_{b_2 z_1}\}$	13	$\{b_1 z_1 \mathbf{s}_4 R_{b_1 z_1}, b_2 z_1 \mathbf{s}_2 R_{b_2 z_1}\}$
6	$\{b_1 z_1 \mathbf{s}_4 R_{b_1 z_1}, b_2 z_1 \mathbf{s}_3 R_{b_2 z_1}\}$	14	$\{b_2 z_1 \mathbf{s}_6 R_{b_2 z_1}, b_1 z_1 \mathbf{s}_2 R_{b_1 z_1}\}$
7	$\{b_1 z_1 \mathbf{s}_3 R_{b_1 z_1}, b_2 z_1 \mathbf{s}_4 R_{b_2 z_1}\}$	15	$\{b_2 z_1 \mathbf{s}_6 R_{b_2 z_1}, b_1 z_1 \mathbf{s}_4 R_{b_1 z_1}\}$
8	$\{b_2 z_1 \mathbf{s}_5 R_{b_2 z_1}, b_1 z_1 \mathbf{s}_6 R_{b_1 z_1}\}$	16	$\{b_1 z_1 \mathbf{s}_6 R_{b_1 z_1}, b_2 z_1 \mathbf{s}_2 R_{b_2 z_1}\}$

bution of the vertex to the network. These vertices and their corresponding weights are illustrated in Table B.3. A vertex v_i in Table B.3 contains a set of associations each one labeled by the subscripts of RRH b_n, RRB z_m, user u_i, file f_k, achievable capacity $R^*_{b_n,z_m}$, and PL $P^*_{b_n z_m}$. For example, the association $1111R^*_{11}P^*_{11}$ represents the subscripts of b_1-th RRH, z_1-th RRB, u_1-th user and its f_1-th requested file, adopted transmission rate $R^*_{b_1 z_1}$, and power level $P^*_{b_1 z_1}$. The first vertex has two associated users to RRB z_1 in RRH b_1 and one associated user to RRB z_1 in RRH b_2. Thus, the corresponding weight is $2 * R^*_{11} + R^*_{21}$ bits/s. The CRAN-CLNC graph that contains all configured vertices is shown in Figure B.1. Since the C-RAN model that presented in Example 1 contains only one RRB, the constructed CRAN-CLNC graph in Figure B.1 contains only one power control subgraph and does not contain any edge connection.

Fourth step: We execute the maximum-weight clique algorithm in the CRAN-CLNC graph to find the best clique that provides an efficient solution. It is important to note that the proposed solution does guarantee the uncoded [109] solution. For example, if the best schedule that maximizes the received throughput assigns a single user to each RRB/RRH, then the proposed algorithm would provide such schedules, i.e., red color circles in Figure B.1.

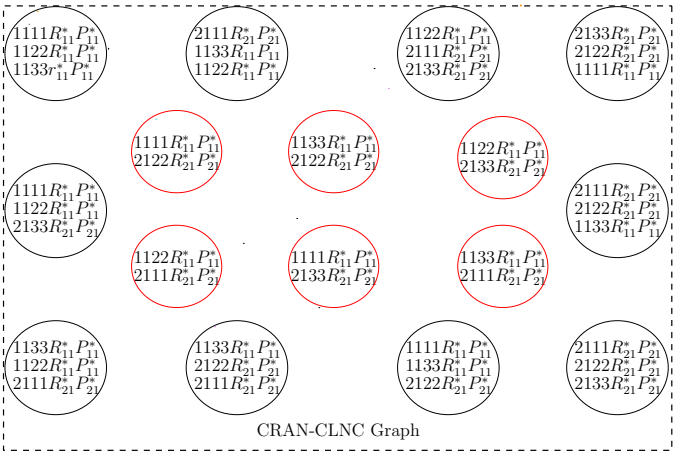


Figure B.1: CRAN-CLNC graph of Example 1 that presented in Figure 2.3 based on Algorithm 1.

Table B.3: The generated vertices and their corresponding weights

i	v_i	w_i
1	$\{1111R_{11}^*P_{11}^*, 1122R_{11}^*P_{11}^*, 2133R_{21}^*P_{21}^*\}$	$2*R_{11}^*+R_{21}^*$
2	$\{1122R_{11}^*P_{11}^*, 1133R_{11}^*P_{11}^*, 2111R_{21}^*P_{21}^*\}$	$2*R_{11}^*+R_{21}^*$
3	$\{1111R_{11}^*P_{11}^*, 1122R_{11}^*P_{11}^*, 1133R_{11}^*P_{11}^*\}$	$3*R_{11}^*$
4	$\{2111R_{21}^*P_{21}^*, 1122R_{11}^*P_{11}^*, 1133R_{11}^*P_{11}^*\}$	$R_{21}^*+2*R_{11}^*$
5	$\{1133R_{11}^*P_{11}^*, 2122R_{21}^*P_{21}^*, 2111R_{21}^*P_{21}^*\}$	$R_{11}^*+2*R_{21}^*$
6	$\{1122R_{11}^*P_{11}^*, 2111R_{21}^*P_{21}^*, 2133R_{21}^*P_{21}^*\}$	$R_{11}^*+2*R_{21}^*$
7	$\{1111R_{11}^*P_{11}^*, 1133R_{11}^*P_{11}^*, 2122R_{21}^*P_{21}^*\}$	$2*R_{11}^*+R_{21}^*$
8	$\{2122R_{21}^*P_{21}^*, 2133R_{21}^*P_{21}^*, 1111R_{11}^*P_{11}^*\}$	$2*R_{21}^*+R_{11}^*$
9	$\{2111R_{21}^*P_{21}^*, 2122R_{21}^*P_{21}^*, 1133R_{11}^*P_{11}^*\}$	$2*R_{21}^*+R_{11}^*$
10	$\{2111R_{21}^*P_{21}^*, 2122R_{21}^*P_{21}^*, 2133R_{21}^*P_{21}^*\}$	$3*R_{21}^*$
11	$\{1111R_{11}^*P_{11}^*, 2122R_{21}^*P_{21}^*\}$	$R_{11}^*+R_{21}^*$
12	$\{1133R_{11}^*P_{11}^*, 2122R_{21}^*P_{21}^*\}$	$R_{11}^*+R_{21}^*$
13	$\{1122R_{11}^*P_{11}^*, 2111R_{21}^*P_{21}^*\}$	$R_{11}^*+R_{21}^*$
14	$\{1111R_{11}^*P_{11}^*, 2133R_{21}^*P_{21}^*\}$	$R_{11}^*+R_{21}^*$
15	$\{1122R_{11}^*P_{11}^*, 2133R_{21}^*P_{21}^*\}$	$R_{11}^*+R_{21}^*$
16	$\{1133R_{11}^*P_{11}^*, 2111R_{21}^*P_{21}^*\}$	$R_{11}^*+R_{21}^*$

Appendix C

Illustration of Algorithm 4

To illustrate Algorithm 4, a simple example is presented in this section, where a network consists of 2 eRRHs caching subsets of a library of 5 popular files that are all in possession of the CBS as shown in the left side of Figure C.1. Two eRRHs are assumed to collectively know the whole library. Each user requests one file, which has a fixed size of 1 Mbits.

First, Algorithm 4 generates the set of all possible associations in the two eRRHs. i.e., $\mathcal{S}_{b_1} = \{u_1 f_1, u_2 f_2, u_4 f_4\}$ and $\mathcal{S}_{b_2} = \{u_1 f_1, u_3 f_3, u_5 f_5\}$, respectively. Then, it generates all ONC file combinations $\mathcal{S}_{\text{IDNC},b_1}$ and $\mathcal{S}_{\text{IDNC},b_2}$ of the two eRRHs as summarized in Table C.1. Afterwards, it generates all feasible schedules $\mathbf{S} \in \mathcal{S}_{\text{fs}}$, where each feasible schedule $\mathbf{S}_i$ is represented by a vertex (circle) v_i in the RRB subgraphs based on **LC1** and **LC2** that are described in Section 3.4. Table C.2 summarizes all feasible schedules $\mathbf{S}$. The power optimization problem (2.10) is then applied for each feasible schedules $\mathbf{S}$ in Table C.2. Given the optimal power distribution $\mathbf{P}$, a vertex v has a secondary weight $w(v)$ as defined in (3.3). Each vertex has a primary weight as defined in (3.4). The generated vertices and their weights are illustrated in Table C.3, in which each vertex contains a set of associations each one labeled by $bzufrp$, where b, z, u, f, r, and p represent the indices of eRRHs, RRBs, users, files, achievable capacities, and PLs, respectively. The constructed FRAN-CLNC graph is shown in Figure C.1.

Second, given all vertices and their corresponding weights that were constructed previously, the maximum-weight independent set algorithm is run to find the maximal independent set that provides the best solution. Since we only consider one RRB per eRRH's frame, the vertices of the resulted FRAN-CLNC graph are adjacent, thus a maximum-weight independent set contains only one vertex, which reflects a partial contribution of that RRB to

Table C.1: All Possible File Combinations $\mathcal{S}_{\text{IDNC},b_1}$ and $\mathcal{S}_{\text{IDNC},b_2}$

i	eRRH b_1 $\mathtt{s}_i^1(\kappa,\tau)$	eRRH b_2 $\mathtt{s}_i^2(\kappa,\tau)$
1	$((f_1),(u_1))$	$((f_1),(u_1))$
2	$((f_1 \oplus f_4),(u_1,u_4))$	$((f_1 \oplus f_3),(u_1,u_3))$
3	$((f_2),(u_2))$	$((f_3),(u_3))$
4	$((f_2 \oplus f_4),(u_2,u_4))$	$((f_5),(u_5))$
5	$((f_4),(u_4))$	

Table C.2: All feasible Schedules $\mathbf{S}_{\text{fs}}$

i	$\mathbf{S}_i$	i	$\mathbf{S}_i$
1	$\{11\mathtt{s}_1^1R, 21c_3^2R\}$	9	$\{11\mathtt{s}_4^1R, 21c_1^2R\}$
2	$\{11\mathtt{s}_1^1R, 21c_4^2R\}$	10	$\{11\mathtt{s}_4^1R, 21c_2^2R\}$
3	$\{11\mathtt{s}_2^1R, 21\mathtt{s}_3^2R\}$	11	$\{11\mathtt{s}_4^1R, 21c_3^2R\}$
4	$\{11\mathtt{s}_2^1R, 21\mathtt{s}_4^2R\}$	12	$\{11\mathtt{s}_4^1R, 21c_4^2R\}$
5	$\{11\mathtt{s}_3^1R, 21\mathtt{s}_1^2R\}$	13	$\{11\mathtt{s}_5^1R, 21c_1^2R\}$
6	$\{11\mathtt{s}_3^1R, 21\mathtt{s}_2^2R\}$	14	$\{11\mathtt{s}_5^1R, 21c_2^2R\}$
7	$\{11\mathtt{s}_3^1R, 21\mathtt{s}_3^2R\}$	15	$\{11\mathtt{s}_5^1R, 21c_3^2R\}$
8	$\{11\mathtt{s}_3^1R, 21\mathtt{s}_4^2R\}$	16	$\{11\mathtt{s}_5^1R, 21c_4^2R\}$

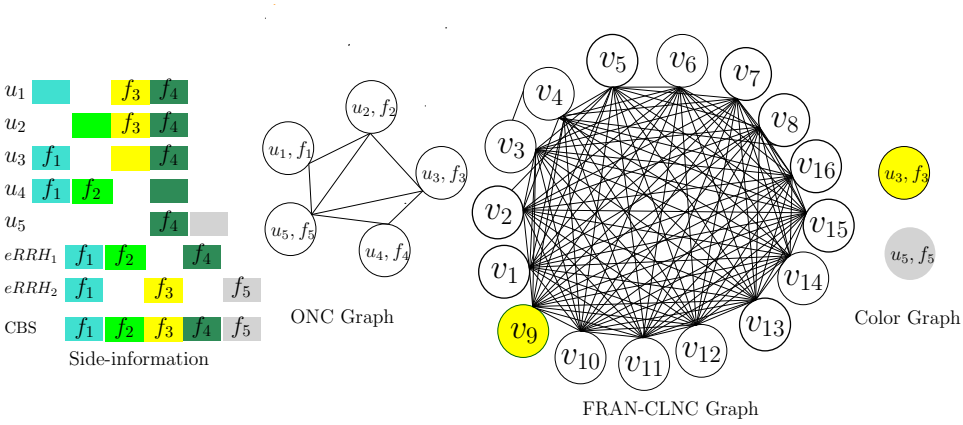


Figure C.1: F-RAN system containing 5 users, 2 eRRHs, 1 RRB in each eRRH's frame, and their side information is given on the left part of the figure. This figure also shows the ONC, FRAN-CLNC, and color graphs of example 3, respectively.

the network. From Table C.3, it is clear that the user association $\{u_1, u_2, u_4\}$ has the maximum-primary weight of 1.3333 seconds. This association represents the vertex v_9 in the FRAN-CLNC graph in Figure C.1. Therefore, the maximum-weight independent set algorithm selects this vertex, which gives the maximum possible throughput. Then, Phase II is performed. The algorithm removes the vertices associated to $\{u_1, u_2, u_4\}$ from the ONC graph and runs the greedy coloring scheme. Since the requested files of the unserved users (u_3 and u_5) cannot be combined, each vertex has a different color as shown in the right part of Figure C.1, thus the CBS would serve these users in two different orthogonal channels.

Table C.3: The represented vertices and their corresponding weights

i	v_i	w_i secondary weight(bps)	w_i primary weight(s)
1	$\{1111r^*_{11}p^*_{11}, 2133r^*_{21}p^*_{21}\}$	$r^*_{11} + r^*_{21}$	0.25
2	$\{1111r^*_{11}p^*_{11}, 2155r^*_{21}p^*_{21}\}$	$r^*_{11} + r^*_{21}$	0.5
3	$\{1111r^*_{11}p^*_{11}, 1144r^*_{11}p^*_{11}, 2133r^*_{21}p^*_{21}\}$	$2 * r^*_{11} + r^*_{21}$	1.2
4	$\{1111r^*_{11}p^*_{11}, 1144r^*_{11}p^*_{11}, 2155r^*_{21}p^*_{21}\}$	$2 * r^*_{11} + r^*_{21}$	1.25
5	$\{1122r^*_{11}p^*_{11}, 2111r^*_{21}p^*_{21}\}$	$r^*_{11} + r^*_{21}$	0.5833
6	$\{1122r^*_{11}p^*_{11}, 2111r^*_{11}p^*_{11}, 2133r^*_{21}p^*_{21}\}$	$r^*_{11} + 2 * r^*_{21}$	0.7833
7	$\{1122r^*_{11}p^*_{11}, 2133r^*_{21}p^*_{21}\}$	$r^*_{11} + r^*_{21}$	0.5333
8	$\{1122r^*_{11}p^*_{11}, 2155r^*_{21}p^*_{21}\}$	$r^*_{11} + r^*_{21}$	0.5833
9	$\{1122r^*_{11}p^*_{11}, 1144r^*_{11}p^*_{11}, 2111r^*_{21}p^*_{21}\}$	$2 * r^*_{11} + r^*_{21}$	1.3333
10	$\{1122r^*_{11}p^*_{11}, 1144r^*_{11}p^*_{11}, 2111r^*_{21}p^*_{21}, 2133r^*_{21}p^*_{21}\}$	$2 * r^*_{11} + 2 * r^*_{21}$	1.25
11	$\{1122r^*_{11}p^*_{11}, 1144r^*_{11}p^*_{11}, 2133r^*_{21}p^*_{21}\}$	$2 * r^*_{11} + r^*_{21}$	1.2
12	$\{1122r^*_{11}p^*_{11}, 1144r^*_{11}p^*_{11}, 2155r^*_{21}p^*_{21}\}$	$2 * r^*_{11} + r^*_{21}$	1.25
13	$\{1144r^*_{11}p^*_{11}, 2111r^*_{21}p^*_{21}\}$	$r^*_{11} + r^*_{21}$	1
14	$\{1144r^*_{11}p^*_{11}, 2111r^*_{21}p^*_{21}, 2133r^*_{21}p^*_{21}\}$	$r^*_{11} + 2 * r^*_{21}$	1.2
15	$\{1144r^*_{11}p^*_{11}, 2133r^*_{21}p^*_{21}\}$	$r^*_{11} + r^*_{21}$	1.2
16	$\{1144r^*_{11}p^*_{11}, 2155r^*_{21}p^*_{21}\}$	$r^*_{11} + r^*_{21}$	1.25