Wireless Resource Optimization in Semantic Communication-Based Cellular Networks

Le Xia

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Engineering
College of Science and Engineering
University of Glasgow

Sep 2024

Abstract

Recent advances in artificial intelligence (AI) have made semantic communication (SemCom) a promising solution that can yield significant benefits in guaranteeing high spectrum resource utilization, information interaction efficiency, and transmission reliability. Compared with conventional bit communication (BitCom), which guarantees the precise reception of transmitted bits, the accurate delivery of semantics implied in desired messages becomes the cornerstone of SemCom. Nevertheless, the unique semantic coding and background knowledge matching mechanisms make it challenging to achieve efficient wireless resource optimization for multiple mobile users (MUs) in SemCom-enabled cellular networks. To this end, the objectives of this thesis are to investigate different optimal wireless resource management strategies in different SemCom network scenarios. Specifically, a total of four differing scenarios are taken into account here, i.e., general SemCom-enabled networks (SC-Nets), energy efficient SemCom-enabled networks (EE-SCNs), hybrid semantic/bit communication networks (HSB-Nets), and SemCom-enabled vehicular networks (SCVNs).

For the general SC-Net scenario, we address two fundamental problems of user association (UA) and bandwidth allocation (BA) on the downlink side, where two different knowledge-matching states of all MUs in the SC-Net are identified. Most importantly, a concept of bit-rate-to-message-rate (B2M) transformation is developed along with a new metric, namely system throughput in message (STM), to measure the overall network performance in a semantic manner. By formulating a joint STM-maximization problem for each SC-Net case, the corresponding optimal solution is then proposed. As for the EE-SCN scenario, we focus on jointly addressing the power allocation and spectrum reusing problems involving the device-to-device (D2D) SemCom approach, in which the energy efficiency model of SemCom is dedicatedly defined. To maximize the average energy efficiency of all cellular users (CUEs) and D2D users (DUEs), a fractional-to-subtractive problem transformation method, a heuristic algorithm, and a Hungarian method are employed together to obtain the optimal solutions. In terms of the HSB-Net scenario, the UA, mode selection (MS), and BA problems are jointly optimized

on the uplink side, where two modes of SemCom and BitCom are available for all MUs' selection. By leveraging the B2M method, the unified performance metric of both modes is identified. Then, we specially develop a knowledge matching-aware two-stage tandem packet queuing model and theoretically derive the average packet loss ratio and queuing latency. Based on the corresponding formulated problem, an optimal resource management strategy is proposed by utilizing a Lagrange primal-dual transformation method and a preference list-based heuristic algorithm with polynomial-time complexity. Finally, in line with the next-generation ultra-reliable and low-latency communication (xURLLC) requirements, we identify and jointly tackle two inevitable problems of knowledge base construction (KBC) and vehicle service pairing (VSP) in the SCVN scenario. In this case, we first derive the knowledge matching based queuing latency specific for semantic data packets, and then formulate a latency-minimization problem subject to several KBC and VSP related reliability constraints. Afterward, a SemCom-empowered Service Supplying Solution (S^4) is proposed along with the theoretical analysis of its optimality guarantee and computational complexity. Numerical results in each of the four scenarios demonstrate significant superiority and reliability of our proposed solutions in terms of various performance metrics compared with multiple benchmarks. All the works presented in this thesis can serve as pioneers in exploring the potential of applying SemCom to wireless cellular networks while ensuring optimal resource management.

University of Glasgow

College of Science & Engineering

Statement of Originality

Name: Le Xia

Registration Number: 2603039

I certify that the thesis presented here for examination for a PhD degree of the University of Glasgow is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it) and that the thesis has not been edited by a third party beyond what is permitted by the University's PGR Code of Practice.

The copyright of this thesis rests with the author. No quotation from it is permitted without full acknowledgement.

I declare that the thesis does not include work forming part of a thesis presented successfully for another degree.

I declare that this thesis has been produced in accordance with the University of Glasgow's Code of Good Practice in Research.

I acknowledge that if any issues are raised regarding good research practice based on the review of the thesis, the examination may be postponed pending the outcome of any investigation of the issues.

Signature: .....

Date: Sep 6 2024 .....

List of Publications

Journals

J1. **Le Xia**, Yao Sun, Dusit Niyato, Xiaoqian Li, and Muhammad Ali Imran, "Joint User Association and Bandwidth Allocation in Semantic Communication Networks," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 2, pp. 2699–2711, 2024.

J2. **Le Xia**, Yao Sun, Dusit Niyato, and Muhammad Ali Imran, "Resource Allocation for D2D Semantic Communication Underlying Energy Efficiency-Driven Cellular Networks," under review by *IEEE Transactions on Communications*.

J3. **Le Xia**, Yao Sun, Dusit Niyato, Lan Zhang, and Muhammad Ali Imran, "Wireless Resource Optimization in Hybrid Semantic/Bit Communication Networks," *IEEE Transactions on Communications*, 2024.

J4. **Le Xia**, Yao Sun, Dusit Niyato, Daquan Feng, Lei Feng, and Muhammad Ali Imran, "xURLLC-Aware Service Provisioning in Vehicular Networks: A Semantic Communication Perspective," *IEEE Transactions on Wireless Communications*, vol. 23, no. 5, pp. 4475–4488, 2024.

J5. Runze Cheng, Yao Sun, Yijing Liu, **Le Xia**, Daquan Feng, and Muhammad Ali Imran, "Blockchain-Empowered Federated Learning Approach for an Intelligent and Reliable D2D Caching Scheme," *IEEE Internet of Things Journal*, vol. 9, no. 1, pp. 7879–7890, 2022.

Magazines

M1. **Le Xia**, Yao Sun, Chengsi Liang, Runze Cheng, Yang Yang, and Muhammad Ali Imran, "WiserVR: Semantic Communication Enabled Wireless Virtual Reality Delivery," *IEEE Wireless Communications*, vol. 30, no. 2, pp. 32–39, 2023.

M2. **Le Xia**, Yao Sun, Rafiq Swash, Lina Mohjazi, Lei Zhang, and Muhammad Ali Imran, "Smart and Secure CAV Networks Empowered by AI-Enabled Blockchain: The Next Frontier for Intelligent Safe Driving Assessment," *IEEE Network*, vol. 36, no. 1, pp. 197–204, 2022.

M3. **Le Xia**, Yao Sun, Chengsi Liang, Lei Zhang, and Muhammad Ali Imran, "Generative AI for Semantic Communication: Architecture, Challenges, and Outlook," *IEEE Wireless Communications*, 2024.

Conference Proceedings

C1. **Le Xia**, Yao Sun, Xiaoqain Li, Gang Feng, and Muhammad Ali Imran, "Wireless Resource Management in Intelligent Semantic Communication Networks," in *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2022, pp. 1–6.

C2. **Le Xia**, Yao Sun, Dusit Niyato, Lan Zhang, Lei Zhang, and Muhammad Ali Imran, "Hybrid Semantic/Bit Communication Based Networking Problem Optimization," *GlobeCom 2024 - IEEE Global Communications Conference*.

C3. **Le Xia**, Yao Sun, Dusit Niyato, Kairong Ma, Jiawen Kang, and Muhammad Ali Imran, "Knowledge Base Aware Semantic Communication in Vehicular Networks," in *ICC 2023 - IEEE International Conference on Communications*, 2023, pp. 3989–3994.

C4. Chengsi Liang, Xiangyi Deng, Yao Sun, Runze Cheng, **Le Xia**, Dusit Niyato, and Muhammad Ali Imran, "VISTA: Video Transmission over A Semantic Communication Approach," in *2023 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2023, pp. 1777–1782.

C5. Runze Cheng, Yao Sun, Yijing Liu, **Le Xia**, Sanshan Sun, and Muhammad Ali Imran, "A Privacy-preserved D2D Caching Scheme Underpinned by Blockchain-enabled Federated Learning," in *2021 IEEE Global Communications Conference (GLOBECOM)*, 2021, pp. 1–6.

Book Chapters

B1. **Le Xia**, Yao Sun, and Muhammad Ali Imran, "Joint Cell Association and Spectrum Allocation in Semantic Communication Networks," to appear in the book titled: *Wireless Semantic Communications: Concepts, Principles and Challenges*, Wiley.

Contents

List of Tables

List of Figures

List of Acronyms

2D	2-Dimensional
3GPP	3rd Generation Partnership Project
6G	Sixth Generation
AI	Artificial Intelligence
AIGC	Artificial Intelligence-Generated Content
B2M	Bit-Rate-to-Message-Rate
BA	Bandwidth Allocation
BitCom	Bit Communication
BLEU	Bilingual Evaluation Understudy
BS	Base Station
CDF	Cumulative Distribution Function
CNN	Convolutional Neural Network
CUE	Cellular User
CV	Computer Vision
D2D	Device-to-Device
DL	Deep Learning
DNN	Deep Neural Network
DUE	Device-to-Device User
E2E	End-to-End
EE-SCN	Energy-Efficient Semantic Communication-Enabled Network
FOV	Field of Views
HetNet	Heterogeneous Network
HSB-Net	Hybrid Semantic/Bit Communication Network
IKM	Imperfect Knowledge Matching
JSCC	Joint Source-and-Channel Coding
KB	Knowledge Base
KBC	Knowledge Base Construction
MEC	Mobile Edge Computing

MS	Mode Selection
MU	Mobile User
NLP	Natural Language Processing
PC	Power Control
PDF	Probability Density Function
PKM	Perfect Knowledge Matching
PMF	Probability Mass Function
PTQ	Packet-Transmission Queue
QoS	Quality of Service
RSU	Road Side Unit
S^4	Semantic Communication-Empowered Service Supplying Solution
SC-Net	Semantic Communication-Enabled Networks
SCQ	Semantic-Coding Queue
SCVN	Semantic Communication-Enabled Vehicular Network
SD	Semantic Decoder
SE	Semantic Encoder
SemCom	Semantic Communication
SINR	Signal-to-Noise-plus-Interference Ratio
SLG	Semantic Location Graph
SOTA	State-of-the-Art
STM	System Throughput in Message
TS	Tabu Search
TSP	Throughput in Semantic Packets
UA	User Association
V2V	Vehicle-to-Vehicle
V2X	Vehicle-to-Everything
VR	Virtual Reality
VSP	Vehicle Service Pairing
VUE	Vehicular User
xURLLC	Next-generation Ultra-Reliable and Low-Latency Communication

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my principal supervisor, Dr. Yao Sun, for his professional and patient guidance as well as all valuable discussions and feedback on all my publications. I am so grateful to him for always believing in me and giving me the freedom to explore my research to the greatest extent. He provided me with indispensable help whenever I needed it, not only as the supervisor but also as a friend. His wide knowledge, strong research enthusiasm, and hard-working attitude have inspired me during my Ph.D. period and will still have a profound effect on my future career. Most importantly, it is he who gave me the opportunity to receive the full funding in 2020 so as to sponsor three years of my doctoral study. Honestly, without this funding, it might have been quite difficult for me to finish my studies over here so well, hence it is very lucky to have Dr. Yao Sun as my supervisor.

Then, I would like to thank Prof. Lei Zhang, who is my second supervisor, and Prof. Muhammad Ali Imran, who is my third supervisor, for their very insightful suggestions on each of my publications. I could not finish all of these works without their assistance.

Meanwhile, I would like to acknowledge the support and help of all my friends, without them, this journey would have been much harder and arduous. Also, thank you for sharing your knowledge in this field in and out of the university, which reallt inspired me a lot.

Moreover, special thanks to my lovely pet cat, Lucky, who has been accompanying me throughout this lonely journey in a foreign country and has given me tremendous happiness and luck.

Last but not least, I would like to express my heartfelt gratitude to my family, especially my mother, Yan Xie, and my grandparents, for their endless support and care, and for their constant motivation, patience, and encouragement to make me a better researcher.

Chapter 1

Introduction

Current wireless networks are witnessing tremendous traffic demands to accommodate the upcoming pervasive application intelligence alongside a variety of high-quality, large-capacity, and multimodal content delivery services, including typical multimedia content (e.g., text [1], image [2], and video streaming [3]) and artificial intelligence-generated content (AIGC) [4]. In response to the ever-increasing data rates and stringent requirements for low latency and high reliability, it is foreseeable that available communication resources like spectrum or energy will gradually become scarce. Combined with the almost insurmountable Shannon limit, these destined bottlenecks are, therefore, motivating us to hunt for bold changes in the new design of future networks, i.e., making a paradigm shift from bit-based traditional communication to context-based *semantic communication* (SemCom) [5–15].

The concept of SemCom was first introduced by Weaver in his landmark paper [5], which explicitly categorizes communication problems into three levels, including the technical problem at the bit level, the semantic problem at the semantic level, and the effectiveness problem at the information exchange level. Nowadays, the technical problem has been thoroughly investigated in the light of classical Shannon information theory [16], while the evolution toward SemCom is just beginning to take shape. Different from the conventional bit communication (BitCom) mode that aims at the precise reception of transmitted bits, SemCom focuses more on the accurate delivery of the true meanings implied in source messages. This is mainly benefited from prosper advancement in deep learning (DL) technologies that can drive SemCom models to achieve efficient and high-quality semantic refinement on desired meaning with low spectrum consumption. Moreover, through equipping both ends of the transceiver with equivalent background knowledge, i.e., provisioning massive data samples to serve diverse artificial intelligence (AI) learning and prediction tasks [3], the implicit meaning in conveyed

content can be recovered with ultra-low semantic errors even under harsh channel conditions. With this remarkable semantic resilience, adequate transmission reliability is also guaranteed. Therefore, SemCom is believed to be a preeminent technology in driving the future wireless network a dramatic leap forward.

Given the popularized SemCom paradigm and existing link-level SemCom developments [1, 3, 7, 9–11], we believe that it is time to move forward to the upper layer, i.e., investigating SemCom from a networking perspective. In this thesis, our main task lies in seeking the optimal wireless resource management strategy in the SemCom-enabled wireless cellular network to optimize its overall network performance in a semantics-aware manner. Since the bandwidth budget and transmit power of each base station (BS) are inherently limited, resource competition becomes unavoidable when there are excessive associated mobile users (MUs) requesting SemCom services at the same time. Meanwhile, energy efficiency should be another important indicator in SemCom under the topic of Green Communications [17], motivating us to concentrate on the balance of energy consumption and semantic performance achieved. Especially considering the unique demand for background knowledge matching between multiple MUs and multi-tier base stations (BSs), efficient resource management becomes crucial and indispensable, which can yield a host of benefits, such as ensuring high-quality of SemCom services and strengthening bandwidth utilization.

1.1 Semantic Communication

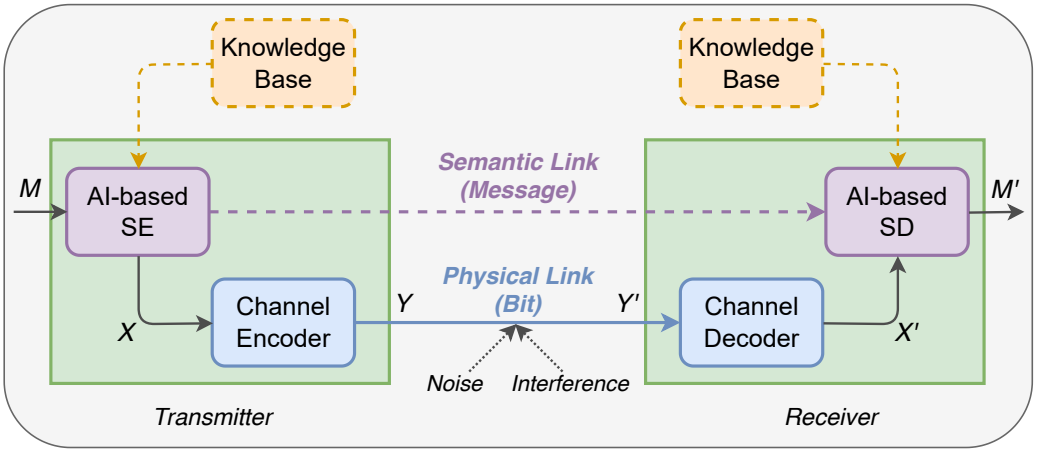

Figure 1.1: The structure diagram of a typical SemCom system.

Compared with the traditional BitCom system [18], a SemCom system generally contains three additional paramount components, including a semantic encoder (SE) and a semantic decoder (SD) equipped at each end of the transceiver,

and knowledge bases (KBs) dedicated to storing the background knowledge of each semantic coding model, as depicted in Fig. 1.1. To be concrete, by embedding cutting-edge sophisticated AI models into terminal devices, a transmitter in SemCom first leverages background knowledge relevant to source messages to filter out irrelevant and redundant content while refining semantic features that only require fewer bits for transmission, the process of which is called semantic encoding. Once the destination receiver has the corresponding background knowledge, the local semantic interpreters are capable of accurately restoring the original meaning from the received bits, even though there may exist severe signal attenuation and distortion due to strong noise and interference, resulting in intolerable bit errors at the syntax level. Taking natural language processing (NLP) models as the exemplification, the Transformer [19] is believed to require fewer bits for encoding a given sentence than using a typical word2vec model [20], and is therefore an ideal candidate for SE and SD in text transmission scenarios.

Moreover, equivalent background knowledge is of critical importance to pursue adequate semantic fidelity and eliminate semantic ambiguity, and the higher the knowledge-matching degree between transceivers, the lower the semantic error in recovered meanings [21]. Consequently, efficient exchanges for the desired information with ultra-low semantic ambiguity can be achieved in SemCom under equivalent background knowledge between source and destination, while significantly alleviating the resource scarcity [6–8]. In the subsequent two subsections, we will showcase the detailed structures of a semantic encoder and a semantic decoder by introducing a specific case about applying SemCom to wireless virtual reality (VR) video delivery. Note that the design of SemCom models should vary depending on the specific application scenario, and it is almost impossible to have a generic semantic coding model that can be applied to any SemCom scenario.

1.1.1 Semantic Encoding Model

To better illustrate the semantic coding models, we present the encoder design from our previous work [3] as an example. Among them, the semantic encoder consists of three different modules, including semantic segmentation module, semantic location graph (SLG) construction module, and channel encoder module, as illustrated in Fig. 1.2. Specifically, the first module is to segment and categorize all objects in each input 2-dimensional (2D) field of views (FOVs) by employing the semantic segmentation technique, which function can be realized by deep convolutional networks [22]. After that, a segmentation map and semantic labels of all objects within the frame can be obtained, respectively, where the segmentation map is a type of high-level image representation with category

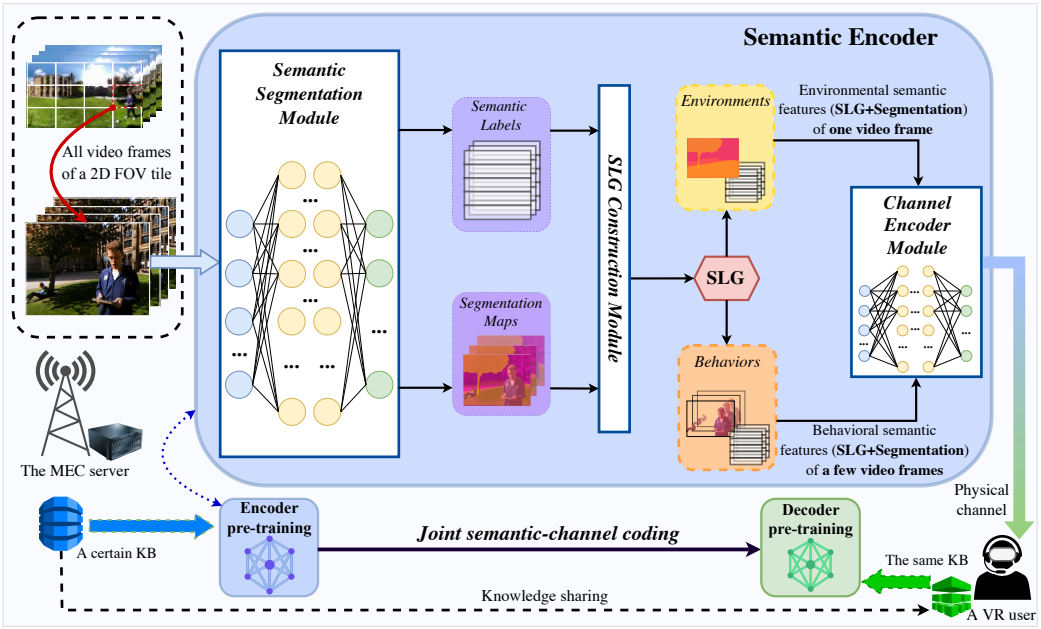

Figure 1.2: The detailed structure of the semantic encoding model in a special case of wireless VR video delivery.

color label assigned to each underlying object, and each semantic label indicates a sequence of word tokens for class labels or natural language descriptions. With these as inputs, the second SLG construction module is able to precisely construct the SLG for each frame, which is a graph containing multiple nodes and edges with corresponding semantic labels attached. To be more specific, each node represents the central point of the segmentation map of each object, and then attaching respective semantic label to form the SLG of each frame. Note that the concept of SLG is different to the knowledge graph [23], where our SLG focuses more on objects' location information (e.g., the *location* of an object "a man" in the tile, referring to the input video frame in Fig. 1.2), their location relationships (e.g., "a man" is *directly below* "a tree"), and their semantic labels (e.g., *"a man" with short blond hair in a blue sweater jacket is sitting on the ground and looking down at a book in his hands*) to provide accurate location and semantic calibration for subsequent video recovery. Accordingly, after comparing the SLG of each object between different frames, both environments and behaviors within each tile can be easily identified in this module. Moreover, semantic features (i.e., the SLG and segmentation map) of each environment are apparently identical in all video frames, which can be generalized as one frame. As for these behaviors, only a few frames' semantic features need to be transmitted to the mobile edge computing (MEC) server. Further, to adapt various physical channel states (e.g., fading, interference, and signal-to-noise ratio), a channel encoder module composed with dense neural network layers is exploited to ensure these environmental

and behavioral features to be accurately transmitted through different channels.

1.1.2 Semantic Decoding Model

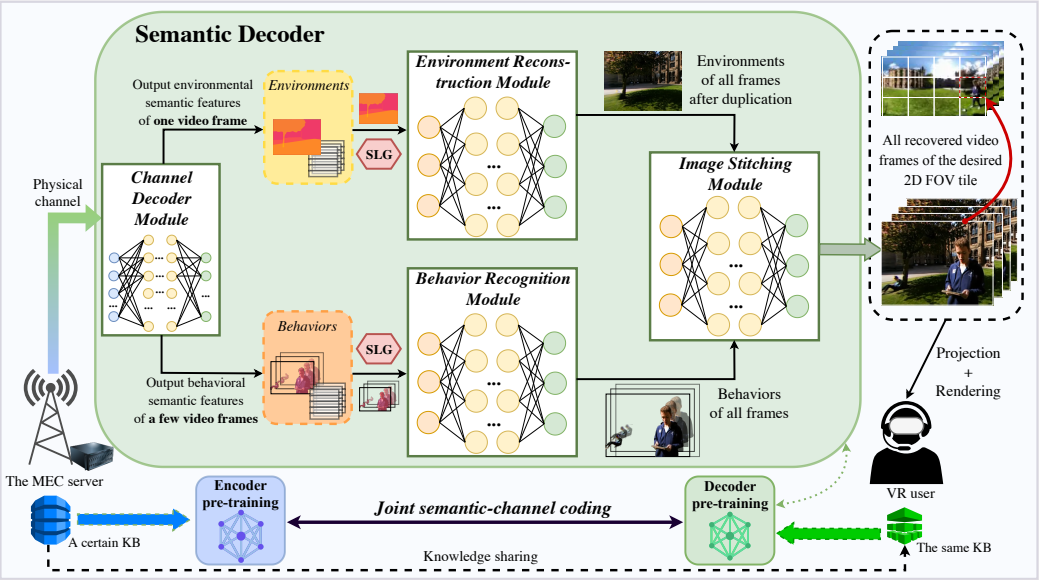

Figure 1.3: The detailed structure of the semantic decoding model in a special case of wireless VR video delivery.

In the semantic decoder, as demonstrated in Fig. 1.3, a channel decoder module with the symmetric structure of the channel encoder module first recovers environmental and behavioral features within each desired 2D FOV tile, respectively. Based on the joint semantic-channel coding design, the channel decoder is capable of greatly preserving semantic features from received data stream. Afterward, the environmental and behavioral features will be input into an environment reconstruction module and a behavior recognition module, respectively. In the environment reconstruction module, the deep convolutional network [22] can still be leveraged to rebuild all static scenes of one frame under image synthesis with multiple visual control, while the behavior recognition function is easily achieved by a Transformer model to predict all missing behaviors of those untransmitted frames [24]. Notably, in both modules above, features related to the segmentation map of each frame are specifically to roughly reconstruct the background image and object profiles, while the SLG is the calibration cornerstone for determining the exact location and status of each object in the tile. Finally, the reconstructed environments are fed into an image stitching module along with the restored behaviors to be merged, thus obtaining all consecutive original information.

1.1.3 Knowledge Base

A key concept of KB is particularly introduced in SemCom, which is deemed a small information entity that stores the background knowledge of one particular application domain (such as music or sports) corresponding to a certain type of SemCom service [8, 25]. This idea originates from the fact that it is unthinkable to represent all knowledge within a single framework given the vastness of knowledge [6]. Concretely, the structure of a KB can roughly cover multiple computational ontologies, facts, rules and constraints associated to a specific domain [26]. In DL-driven SemCom systems, the background knowledge is deemed training data samples serving a certain class of learning tasks [1, 2]. Considering its computation and storage requirements, in the wireless cellular network, BSs and large-capacity user equipment can hold certain amounts and types of KBs. As for these small-capacity users, one potential solution for them is to acquire different SemCom services with required KBs by associating with different BSs. Another viable approach is employing the knowledge sharing method [25] to acquire the desired background knowledge from the adjacent MUs or BSs, which, however, will cause extra preparation delay.

1.2 Wireless Resource Management

Current wireless cellular networks have evolved from homogeneous networks (HomNets) to heterogeneous networks (HetNets), which was introduced in Release 12 of the 3rd generation partnership project (3GPP) [27]. Since the HetNet allows different tiers of BSs (like microBS, picoBS, and femtoBS) to coexist with the macroBS by sharing the same spectrum resources, which can extremely improve spectrum efficiency and reduce uncovered areas [28]. In such a complex network architecture, wireless resource management becomes critical and involves several fundamental yet challenging problems during the networking process, such as user association (UA), bandwidth allocation (BA), power control (PC)/allocation, spectrum reusing, and interference management, etc. These problems take different shapes depending on different system modeling in different scenarios. For instance, the joint UA and BA optimization problem on the uplink side is quite distinct from that on the downlink side, since their interference patterns and objects are diverse. Besides, formulating the wireless resource management problem naturally falls into the scope of integer or mixed integer programming, which can also be an NP-hard assignment problem in most cases [29–39].

1.2.1 User Association

The UA problem occurs more often in the multi-BS cellular network, where multiple MUs exist to acquire the assignment of bandwidth resources from their associated BSs to support their respective communication services. Without loss of generality, one MU can be only associated with one BS while one BS can serve multiple MUs at a time [29]. However, considering the limited bandwidth budget of each BS, some BSs can be fully loaded if they have excessive MUs asking for UA, but some BSs are idle if they have only a few associated MUs, which is obviously not a perfect UA solution. Moreover, for each MU, associating with different BSs indicates different space distances for communication, which can greatly affect the rendered signal-to-interference-plus-noise ratio (SINR) of its link. Especially when targeting at optimizing the overall network performance, no matter from the bit-level throughput or the semantic-level throughput, achieving load balancing among multiple BSs through the UA optimization should be very meaningful and valuable.

1.2.2 Bandwidth Allocation

It is known that each BS has limited bandwidth resources ready to be assigned to its associated MUs. When there are multiple associated MUs with different SINRs, how to allocate the specific amount of bandwidth resources to each MU becomes an inevitable optimization problem. Note that the BA problem is generally investigated jointly with the UA, since differing UA schemes can result in different numbers of associated MUs of each BS. Besides, the uplink BA should be distinguished from the downlink BA problem due to their different spectrum divisions and communication modes. In some standards such as LTE, the bandwidth resources of each BS can also be distributed among MUs in the form of resource blocks (RBs), and each RB spans over a certain frequency range and time duration. Based on the total system bandwidth available and the scheduling interval of the scheduler, the number of RBs at different BSs can be different.

1.2.3 Power Control

Power control in the realm of wireless communication is also known as power allocation, which is to assign the optimal transmit power for each MU on the uplink side and for each BS on the downlink side. The transmit power output of each entity is constrained between a prescribed range, and the higher the transmit power, the greater the received signal power. However, this issue is generally considered in the case of multiple MUs or multiple BSs, where higher transmit

power also increases the interference power, and thus power control is sometimes closely related to interference management. Particularly, if we further take into account the energy efficiency factor, the higher transmit power represents the greater power consumption as well. Accordingly, how to comprehensively devise the optimal power allocation scheme has become an issue of great concern in the process of networking.

1.2.4 Spectrum Reuse

The spectrum reuse technique is to use the same frequency bands in different geographic locations within a cellular network. This can be either completed by dividing the service area into smaller cells and each is served by a base station or done by different types of wireless linking, such as cellular links and D2D links. By reusing frequencies in multiple cells or links, more MUs can be served within the same spectrum, greatly enhancing network capacity and resource utilization. However, one of the primary challenges here is how to manage the interference between cells or links using the same spectrum band, which generally requires sophisticated interference mitigation techniques. Another challenge is how to design a network with an optimal frequency reuse pattern involves complex planning to balance capacity and interference. In recent years, spectrum reuse has been complemented by other techniques like small cells, massive multiple input multiple output, and beamforming, etc.

1.2.5 Interference Management

Most specifications of current cellular wireless networks are based on reuse-one deployment to achieve the most utilization of limited spectrum resources. Besides, improving the network density is deemed an efficient approach to enhance the traffic capacity and user throughput. Nevertheless, as the density and load grow, receivers in the cellular network simultaneously suffer from increased co-channel interference, particularly at the boundaries of cells. Hence, co-channel interference has become one of the major problems that should be optimized in next-generation cellular systems, and thus efficient interference management methods are indispensable. Classical solutions can be divided into two categories of using advanced receivers with interference joint detection/decoding on the user side and employing joint scheduling on the BS side.

1.3 Motivation

In next-generation SemCom-enabled cellular networks, the wireless resource management problem is recognized as a unique and challenging one due to some inevitable changes in both the network architecture and the communication system as well as other limitations associated with SemCom. These prevent the resource management solutions in conventional BitCom-enabled networks from being directly extended to the SemCom-enabled network scenarios.

Specifically, since SemCom focuses on the successful delivery of meanings rather than bits compared with BitCom, a proper semantic-related metric should be defined to assess the network from the overall semantic-level performance perspective. In the meantime, multiple MUs should be correctly associated with specific BSs to match their required background knowledge bases and the appropriate channel conditions in the UA phase. Apart from this, different MUs generally have different knowledge-matching degrees with their respective communication counterparts, which can greatly affect the BA strategy. This is because the higher knowledge-matching degree can be deemed as the more powerful reasoning capability of the semantic coding model, which is equivalent to the more resilient and robust semantic recovery. Intuitively, the MU with the high knowledge-matching degree demands fewer bandwidth resources to convey the same meaning than the MU with the low knowledge-matching degree. Most importantly, since the information source is generally modeled as a stochastic process [13], the specific amount of its generated messages corresponding to the matched knowledge or the mismatched knowledge becomes uncertain, even given the knowledge-matching state. Consequently, there is always only a random proportion of messages that can be correctly interpreted in the SemCom, which indicates that the knowledge-matching degree is actually a random variable and the corresponding BA problem in the SemCom-enabled network becomes a stochastic optimization problem. In parallel, the SemCom-based power control problem is also quite distinct from that in BitCom. When conventional BitCom scenarios care about the energy efficiency in units of Bits/Joule, SemCom is foreseeable to concentrate upon the one in units of semantic-level performance per Joule. Moreover, the circuit power consumption mode becomes unique in SemCom, since the transmitted data can be classified into two categories of knowledge-matching state and knowledge-mismatching state. In the DL-driven semantic coding models, the knowledge-matching data are envisioned to consume less power than the knowledge-mismatching data, which is justified since the knowledge-mismatching content necessarily requires more computing power and processing time for accurate contextual reasoning and in-

terpretation due to the use of more sophisticated semantic-coding networks or the knowledge-sharing method [8].

Besides, it is worth pointing out that there is still a missing investigation on wireless resource optimization in a more practical scenario, i.e., a hybrid semantic/bit communication network, where both SemCom and BitCom modes are available for multiple MUs. The necessity of emphasizing the hybrid scenario lies in the current colossal infrastructures and user groups in BitCom that cannot be completely replaced at one time, while taking into account the unprecedented potential and superiorities of SemCom in terms of efficient information exchanges and wireless resource savings. Therefore, the hybrid semantic/bit communication network will be an inevitable and long-lasting intermediate network paradigm during the future evolution of wireless cellular networks and is also expected to yield a host of benefits, such as flexible and targeted service provisioning, adequate resource utilization, and satisfactory user experience on semantic performance.

Finally, notice that ubiquitous intelligence is expected to emerge in next-generation vehicular networks to accommodate diverse smart on-board applications and large-capacity vehicle-to-everything (V2X) services (e.g., multimodal artificial intelligence generated content offered by ChatGPT or Dall-E), which poses tremendous demands on high data rates along with stringent requirements for reliability and latency [40, 41]. The application of SemCom into large-scale vehicle-to-vehicle (V2V) communications is also a very interesting and forward-looking topic. To be concrete, due to the varying practical KB sizes, personal KB preferences, and limited vehicular storage capacities, the first problem is how to devise an optimal knowledge base construction (KBC) policy not only proactively but also collaboratively for all vehicle users (VUEs) to construct their respective appropriate KBs for better service provisioning. In the meantime, when considering different types of KBs equipped on numerous VUEs and unstable wireless link quality, it can be challenging to well solve the service provisioning-driven VUE pairing problem to meet the knowledge matching restriction, thus shaping the second problem namely vehicle service pairing (VSP). Solving the KBC and VSP issues is believed to yield a bunch of benefits, such as improving V2V information interaction efficiency, reducing data traffic congestion, and ensuring high-quality vehicular services.

1.4 Objectives

According to the motivation, the main goal of this thesis is to seek optimal wireless resource management solutions for DL-driven semantic communication networks.

For better illustration, our research involves four different objectives: 1) Investigating the optimal joint UA and BA schemes for the SemCom-enabled cellular network; 2) Exploring the optimal joint power allocation and spectrum reusing strategy for device-to-device (D2D) SemCom underlying energy efficient-driven cellular networks; 3) Optimizing the wireless resources in hybrid semantic/bit communication networks with the awareness of reliability and latency; 4) Guaranteeing the best SemCom service provisioning in SemCom-enabled V2V networks for next-generation ultra-reliable and low-latency communications (xURLLC).

The first objective aims to investigate SemCom from a networking perspective by considering varying knowledge-matching states between MUs and associated BSs, since the unique demand for background knowledge matching makes it challenging to achieve efficient wireless resource management for multiple users in SemCom-enabled networks. The relationship between message rate and bit rate needs to be elucidated for developing the new semantic-level performance metric. Then the UA and BA issues then need to be addressed jointly to maximize the overall performance of the network, subject to practical limitations and SemCom-related constraints. Among them, the stochasticity of the knowledge-matching factors should be taken into account to comprehensively characterize SemCom. After proposing the optimization solution, the corresponding validation needs to be realized by conducting numerical simulations.

In terms of our second objective, one urgent demand is to build the energy efficiency-based SemCom model, and the semantic performance measurement can follow our previous work. The power consumption model can incorporate the knowledge-matching degree, and combined with the D2D SemCom, the spectrum reusing problem can also be jointly considered. In line with the formulated problem, the corresponding optimal solution can be studied and validated through numerical simulations.

As for our third objective, wireless resource management in hybrid semantic/bit communication networks is envisioned to be rather complicated and challenging, given the unique background knowledge matching and time-consuming semantic coding requirements in SemCom. To that end, a novel problem of mode selection (MS) needs to be tackled, which is about how to determine the best communication mode for each MU with the joint consideration of UA and BA to optimize overall network performance. Proceeding the semantic performance metric in our previous work, another goal is to mathematically characterize the unique semantic-coding process in SemCom when combined with bit transmission in such a hybrid scenario. Based on the correspondingly formulated joint UA, MS, and BA problem, an optimal resource management solution needs to be

designed, which should also be tested under a sufficiently large number of trials.

The last objective is to investigate the potential of applying SemCom to vehicular networks with the awareness of xURLLC. Notably, the unique background knowledge matching mechanism in SemCom makes it challenging to realize efficient vehicle-to-vehicle service provisioning for multiple users at the same time. In this case, the objective is to propose an efficient SemCom service provisioning policy for multiple VUEs. In particular, constructing appropriate KBs at each VUE and selecting the best vehicle node for each VUE from multiple candidate neighbors become quite important for executing SemCom. On this basis, the xURLLC-driven joint optimization problem needs to be formulated, followed by the corresponding optimal solution and numerical verification.

1.5 Research Contributions

The objectives mentioned above indicate that this thesis aims at deriving the optimal wireless resource management solution for different SemCom-enabled networks. In full view of our research progress, the main contributions of this thesis are summarized as follows:

- Considering varying knowledge matching states between MUs and associated BSs, we identify two general SemCom-enabled network scenarios, namely perfect knowledge matching-based SemCom-enabled network and imperfect knowledge matching-based SemCom-enabled network. Afterward, in each case, we describe its distinctive semantic channel model from the semantic information theory perspective, whereby a concept of bit-rate-to-message-rate transformation is developed along with a new semantics-level metric, namely system throughput in message (STM), to measure the overall network performance. In this way, we then formulate a joint STM-maximization problem of UA and BA for each SemCom-enabled network scenario, followed by a corresponding optimal solution proposed. Numerical results in both scenarios demonstrate significant superiority and reliability of our solutions in the STM performance compared with two benchmarks.

- We specially consider a D2D SemCom scenario, where multiple D2D users (DUEs) can reuse the spectrum resources of cellular users (CUEs) for SemCom. By taking into account the knowledge-matching conditions, the SemCom-based energy efficiency model is built at a semantic data packet level. With the aim of maximizing the overall energy efficiency of all CUEs and DUEs, a joint power allocation and spectrum reusing problem is for-

mulated, followed by an optimal solution proposed via the fractional-to-subtractive problem transformation method, a heuristic algorithm, and the Hungarian method. Numerical results show the energy efficiency superiority of our solution in comparison to two baselines.

- We jointly investigate UA, MS, and BA problems in a hybrid semantic/bit communication network scenario. Concretely, a unified performance metric of message throughput for both SemCom and BitCom links is first identified. Next, we specially develop a knowledge matching-aware two-stage tandem packet queuing model and theoretically derive the average packet loss ratio and queuing latency. Combined with practical constraints, we then formulate a joint optimization problem for UA, MS, and BA to maximize the overall message throughput of hybrid semantic/bit communication network. Afterward, we propose an optimal resource management strategy by utilizing a Lagrange primal-dual transformation method and a preference list-based heuristic algorithm with polynomial-time complexity. Numerical results not only demonstrate the accuracy of our analytical queuing model, but also validate the performance superiority of our proposed strategy compared with different benchmarks.

- We identify and jointly address two fundamental problems of KBC and VSP inherently existing in SemCom-enabled vehicular networks in alignment with the xURLLC requirements. Concretely, we first derive the knowledge matching based queuing latency specific for semantic data packets, and then formulate a latency-minimization problem subject to several KBC and VSP-related reliability constraints. Afterward, a SemCom-empowered Service Supplying Solution (S^4) is proposed along with the theoretical analysis of its optimality guarantee and computational complexity. Numerical results demonstrate the superiority of S^4 in terms of average queuing latency, semantic data packet throughput, user knowledge matching degree and knowledge preference satisfaction compared with two benchmarks.

1.6 Thesis Outline

The remainder of this thesis is organized as follows. Chapter 2 starts with the related works of the current research development of SemCom and wireless resource management. Besides, the state-of-the-art (SOTA) DL techniques are presented that can be employed in the field of SemCom.

Chapter 3 is produced on top of "Joint User Association and Bandwidth

Allocation in Semantic Communication Networks" (i.e., the journal paper **J1** and its conference version **C1** in **List of Publications**). It starts with presenting the semantic channel models of both PKM-based and IKM-based SC-Nets in Section 3.2. Then for the two different SC-Nets, Sections 3.3 and 3.4 formulate their joint UA and BA optimization problems and propose the corresponding solutions, respectively. Numerical results are demonstrated and discussed in Section 3.5, followed by conclusions in Section 3.6.

Chapter 4 is produced on top of "Resource Allocation for D2D Semantic Communication Underlying Energy Efficiency-Driven Cellular Networks" (i.e., the journal paper **J2** in **List of Publications**). It starts with the system model of EE-SCN and the corresponding problem formulation in Section 4.2. Then the optimal power allocation and spectrum reusing solutions are presented in Section 4.3, followed by the numerical results in Section 4.4 and conclusions in Section 4.5.

Chapter 5 is produced on the top of "Wireless Resource Optimization in Hybrid Semantic/Bit Communication Networks" (i.e., the journal paper **J3** and its conference version **C2** in **List of Publications**). Section 5.2 first introduces the system model of HSB-Net. Then, the queuing analysis for both SemCom and BitCom cases is presented, and the corresponding joint resource management problem is formulated in Section 5.3. In Section 5.4, we illustrate the proposed optimal UA, MS, and BA strategy. Numerical results are demonstrated and discussed in Section 5.5, followed by the conclusions in Section 5.6.

Chapter 6 is produced on the top of "xURLLC-Aware Service Provisioning in Vehicular Networks: A Semantic Communication Perspective" (i.e., the journal paper **J4** and its conference version **C3** in **List of Publications**). Section 6.2 first introduces the system model of SemCom-enabled vehicular networks. Then a joint service provisioning problem is identified and formulated in Section 6.3. In Section 6.4, we illustrate the proposed solution S^4. Numerical results are presented and discussed in Section 6.5, followed by conclusions in Section 6.6.

Finally, Chapter 7 concludes the thesis and discusses the future trends associated with wireless resource management in intelligent SemCom networks.

Chapter 2

Background and Literature Review

2.1 Overview in Semantic Communications

The history of SemCom can be traced back to the seminal work done by Weaver in the 1950s [5], in which he proposed communication problems at three levels:

- *How accurately can the symbols of communication be transmitted?*
- *How precisely do the transmitted symbols convey the desired meaning?*
- *How effectively does the received meaning affect conduct in the desired way?*

Fortunately, the first technical problem is believed to be fully covered by Shannon's classical information theory [16], which has served as proven guidance in communication system design for more than seven decades. Recently, with the explosive growth of AI-related research, the second semantic-level problem has gradually attracted widespread attention, which concentrates upon how to successfully convey semantics, rather than bits, implied in the source information. As an interdisciplinary technology, SemCom involves multiple different research areas including linguistics, computer science, and wireless communications. Especially benefiting from the powerful inference and interpretation capabilities of current AI models, a variety of semantic-aware communication techniques are emerging, making SemCom a promising next-generation communication paradigm. As for the third effectiveness communication, it should be a future communication difficulty that needs to be considered after solving the lower two levels. In the subsequent three subsections, we will present recent advances in semantic information theory, physical-layer semantic transmission, and SemCom-enabled network management, respectively. Note that research in the SemCom field is still

in an infancy stage, a comprehensive and consistent theory regarding SemCom has not yet been established.

2.1.1 Development of Semantic Information Theory

The concept of semantics originated from the field of semiotics [42], in which syntactic signs, semantic signs, and pragmatic signs were proposed by Morris in [43]. Soon after that, Weaver argued that Shannon's classical information theory is general enough to be extended to consider other two-level problems. Then, Carnap and Bar-Hillel [12] were among the first to introduce the concept of "Semantic Information Theory" and used truth tables and logical probability to define semantic entropy in 1953. On this basis, Barwise and Perry further proposed a definition of situational logic to extend semantic information theory in their pioneering work [44]. In [45, 46], Floridi developed a theory of Strongly Semantic Information and tried to solve the contradictions in semantic information measurement. D'Alfonso aimed to quantify semantic information by employing a "value aggregate" method to support a broader range of use cases in [47].

Over the last two decades, modern semantic information theory has gone beyond the above classical one. Bao *et al.* in [7, 11] summarized the existing works on quantifying semantic information and quantitatively measured semantic entropy by putting forward a semantic channel coding theorem. Then, Zhong *et al.* in [48] introduced the information trinity concept and proved that semantic information is the unique representative of the trinity. Besides, Kolchinsky and Wolpert in [49] identified the syntactic information between a system and its environment from the physical perspective, while Kountouris and Pappas in [50] defined a multi-granularity-based semantic information measured by Rényi entropy [51]. Moreover, Jiang *et al.* [52] reckoned that the limitations of the current communication systems are due to the lack of semantic information awareness. More recently, Liu *et al.* [13] studied the semantic rate-distortion function of information source on the basis of its intrinsic state and extrinsic observation in the memoryless source case.

2.1.2 Physical-Layer Semantic Transmission

In the semantic-transceiver-design related works, Farsad *et al.* in [53] first used DL approaches for joint source-and-channel coding (JSCC) to realize text SemCom in the end-to-end (E2E) communication system. Bourtsoulatze *et al.* in [54] proposed another JSCC technique for wireless image transmission that does not rely on explicit codes for either compression or error correction. Then, inspired

by advanced NLP algorithms, Xie *et al.* in [1] and [9] developed a Transformer-based text sentence similarity metric to measure the semantic performance in E2E SemCom systems. In parallel, two speech distortion ration-related semantic metrics are employed in [10] for testing the speech signals received via SemCom. Weng *et al.* in [10] considered a SemCom system for speech signals based on an attention-driven DL model called squeeze-and-excitation networks. Moreover, [3] sought the possibility of applying SemCom to wireless virtual reality video delivery to realize high-performance feature extraction and semantic recovery. Zhang *et al.* [14] integrates a concept of semantic base with next-generation communication systems to enable intelligent interactions among various communication objects in 6G. Other existing works [55–57] have proved that physical-layer transmission efficiency can be ensured by utilizing semantic encoding and decoding models in representative application scenarios.

2.1.3 Semantic Communication-Enabled Networks

Several other preliminary studies related to SemCom have further investigated the wireless resource management issue from a networking perspective. Powered by deep reinforcement learning algorithms, Zhang *et al.* [2] adopted a dynamic resource allocation scheme to maximize the long-term transmission efficiency in task-oriented SemCom networks. In [58], Yang *et al.* exploited a probability graph and a rate-splitting method to achieve energy-efficient SemCom networks on both transmission and computation. Likewise, a quantum key distribution-secured resource management framework was considered by Kaewpuang *et al.* [59] for the edge devices communicating semantic information. Apart from these, Xia *et al.* [21] specially developed a bit-rate-to-message-rate transformation function along with a new semantic-aware metric called system throughput in message to jointly optimize UA and BA problems in SemCom-enabled cellular networks. Yan *et al.* [60] exploited the semantic spectral efficiency optimization-based channel assignment.

2.2 Overview in Wireless Resource Optimization

In this section, the existing related works pertinent to wireless resource management in conventional BitCom are reviewed and categorized according to a diverse range of three different performance metrics, including UA and BA optimization, spectrum efficiency optimization, energy efficiency optimization, URLLC-aware

network optimization, and outage probability optimization. Most of these research results are investigated in the HetNet scenarios, since HetNet has become the dominant theme in current and next-generation wireless network architecture.

2.2.1 UA and BA Optimization

It is known that the co-channel transmissions will lead to severe inter-cell interference, and one viable solution is to optimize the resource allocation among multiple BSs to maximize the system performance. In [61], Lopez-Perez *et al.* proposed a dynamic algorithm to jointly allocate bandwidth and power to mitigate intercell interference. Apart from resource allocation, UA is considered another efficient factor in dealing with intercell interference. Qian *et al.* in [62] devised an algorithm through the classic Benders' decomposition to tackle with the optimization problem of joint UA and power control. Besides, Li et al. in [63] proposed an asymptotically optimal solution for the resource allocation problem in heterogeneous cellular networks with cooperative relay nodes. In [64] and [65], the joint optimization on UA and BA was formulated, and the performance of different schemes was investigated.

2.2.2 Spectrum Efficiency Optimization

Apart from the outage probability, spectrum efficiency is another widely accepted network performance metric in traditional wireless networks. Corroyin *et al.* in [66] derived an upper bound on the downlink spectrum efficiency in HetNets a heuristic dynamic UA scheme with low complexity to approach this upper bound. Besides, the joint optimization of UA and channel allocation between macroBSs and microBSs was explored by Fooladivanda *et al.* in [67] to maximize the data rate. Proceeding as [67], Ghimire *et al.* in [64] developed an optimal joint UA, transmission coordination, and channel allocation solution with the aim of maximizing the data rate-based utility. In line with the spectrum efficiency maximization, the authors in [68–70] employed similar solutions in an iterative manner of simultaneously updating the UA and the power control factors until convergence for the downlink HetNets. Contrary to this, the authors in [71, 72] utilized another solution where they optimized the UA factors by first fixing power/channel allocation factors and then optimizing the power/channel allocation factors with the aid of the fixed UA factors.

2.2.3 Energy Efficiency Optimization

As a result of the international community's concern for environmental protection, green communication has gained tremendous attention from both industry and academia [73,74]. In [75], the UA issue for the downlink of HetNets was specially optimized by maximizing the ratio between the total data rate of all users and the total energy consumption. In parallel, Zhu *et al.* in [76] proposed an energy-efficient UA solution by minimizing the total power consumption while satisfying the users' traffic demand. A Benders' decomposition method [77] was employed in [62] for joint UA and power control to maximize the downlink throughput and minimize the total transmit power consumption. Su *et al.* in [78] jointly considered the optimization of long-term BS sleep-mode operation, UA, and sub-carrier allocation for maximizing the energy efficiency and minimizing the total power consumption subject to constraints of average sum rates and rate fairness. Another energy efficiency optimization algorithm was developed in [79] for minimizing the energy consumption by beneficially adjusting both the UA and the BS sleep-mode operations with the awareness of the dependence of the energy consumption on both the spatio-temporal variations of traffic demands and the internal hardware components of BSs. In addition, the coverage probability and energy efficiency of K-tier HetNets were derived in [80] together under different sleep-mode operations using a stochastic geometry-based method.

2.2.4 URLLC-Aware Network Optimization

Achieving URLLC brings new challenges in optimizing the average performance for current cellular networks, e.g., overall throughput, communication reliability, and average latency [81]. For instance, the authors in [82, 83] aimed to reach high performance on average metrics in vehicular networks with the awareness of strict URLLC requirements at the occurrence of extreme events. In [84], a probabilistic limitation has been identified and imposed on the optimization problem to shorten the queue length at each V2V pair. Besides, [85] investigated the distribution of queue length by extreme value theory, which is a powerful tool to characterize the occurrence probability of extreme events. The authors in [86] proposed a Lyapunov-based distributed resource allocation algorithm to reduce the queuing latency by employing both extreme value theory and federated learning. Moreover, the authors in [87] employed the age of information as the latency metric and modeled its tail distribution using the extreme value theory.

2.2.5 Outage Probability Optimization

As a matter of fact, the outage/coverage probability is the primary performance metric for UA and BA analysis in conjunction with stochastic geometry. Dhillon *et al.* in [88, 89] analyzed the system performance in K-tier downlink HetNets with the aid of stochastic geometry and incorporated a flexible notion of BS load by introducing a new idea of conditionally thinning the interference field. Similarly, Cheung *et al.* in [90] first derived the success probability for each tier BS under different BA and femtoBS access policies by introducing a tractable model, and then formulated the throughput maximization problem subject to several quality-of-service (QoS) constraints in terms of both coverage probabilities and per-tier minimum rates. In [91], Lin *et al.* obtained the optimal UA bias and bandwidth partitioning ratios theoretically for maximizing the proportionally fair utility based on the outage probability in both downlink and uplink of HetNets.

2.3 Advanced DL Technologies Enabling SemCom

Recent advances in SOTA DL technologies have created great opportunities to develop sophisticated SemCom systems, providing a viable path for undertaking next-generation semantic service provisioning. Many researchers have tried to leverage powerful feature learning and feature representation capabilities of DL models to extract and recover the semantics implied in the source information, and some of them have reached excellent performance in SemCom [92, 93]. In terms of the text-based SemCom, the Transformer [19], GPT [94], and BERT [95] models have proven significant success in many prediction and inference tasks in the field of NLP. On this basis, diverse DL model structures like encoding-autoencoding, decoding-autoregression, and encoding-decoding are proposed to further enhance the word and sentence representation. When it comes to the image-based SemCom, the convolutional neural network (CNN) [96, 97] is of paramount importance to greatly support image semantic extraction and restoration-related tasks, which has also been widely used in the realm of computer vision (CV) for classical image classification and object recognition tasks. As for other multi-modal data such as speech and videos, squeeze-and-excitation networks [98] and deep neural network (DNN)-powered semantic segmentation models [99] have become the best candidates in the pertinent tasks. Overall, the proliferation of these SOTA DL technologies has led to the convergence of ubiquitous intelligence and next-generation communication systems, providing a

promising approach for intelligent SemCom design. This allows not only the true information of interest of MUs for communications, rather than raw data but also alleviates the bandwidth pressure and strengthens resilience by reducing the redundant data.

Chapter 3

Joint User Association and Bandwidth Allocation in Semantic Communication Networks

3.1 Introduction

AI has been widely regarded as an indispensable component in future networking paradigms. Benefited from a variety of SOTA DL techniques, many sophisticated computation tasks can be well accomplished. Moreover, due to the limited wireless resources, traditional communication systems are becoming gradually insufficient to process diversified service requirements under various application scenarios. This destined bottleneck is, therefore, motivating us to hunt for a bold change in new designs on AI-enabled 6G networks, for a paradigm revolution from traditional BitCom to *intelligent SemCom* [1, 9, 10, 56].

As a matter of fact, there have been several noteworthy related works paving ways for the development of SemCom. Powered by advanced natural language processing (NLP) algorithms, the authors in [1] and [9] developed a Transformer-based text sentence similarity metric to measure the semantic performance in end-to-end SemCom systems. In parallel, two speech distortion ration-related semantic metrics are employed in [10] for testing the speech signals received via SemCom. Moreover, [3] sought the possibility of applying SemCom into wireless virtual reality video delivery to realize high-performance feature extraction and semantic recovery. Apart from these semantic-transceiver-design related works, some researches on information-theoretic characterization for SemCom is also

of paramount importance. The authors in [7] and [11] quantitatively measured semantic entropy by putting forward a semantic channel coding theorem, which is based on the logical probability of messages proposed by Carnap and Bar-Hillel in [12]. Besides, [13] recently studied the semantic rate-distortion function of information source on the basis of its intrinsic state and extrinsic observation in the memoryless source case.

Considering the novel paradigm of intelligent SemCom-enabled networks (SC-Nets), we are encountering three fundamental networking challenges as follows:

- *Challenge 1: How to mathematically construct a reasonable semantic channel model in view of the characteristics of SemCom?* Different from the traditional bit-based channel models, the first priority in the semantic channel model is to mathematically characterize semantic information delivered from a source to its destination [13]. In particular, mismatched background knowledge between the semantic encoder and decoder can cause a certain degree of semantic ambiguity as well as information distortion [7]. Hence, the first challenging problem is how to sketch a reasonable semantic channel model based on different knowledge-matching degrees from a semantic information theory perspective.

- *Challenge 2: How to define a proper metric to measure the SemCom-related network performance?* Since the meaning of delivered messages, rather than transmitted bits, becomes the sole focus of SemCom, traditional performance metrics based on Shannon's legacy, such as system throughput in bit, are no longer applicable to measure the network performance of SC-Net. Given the unique semantic channel model, how to define a proper SemCom-related metric should be another challenge.

- *Challenge 3: How to determine an optimal resource management strategy to maximize the SemCom-related performance of SC-Net?* In the cellular network architecture, UA and BA are two key mechanisms to realize resource management [39]. When it comes to SC-Net, besides practical constraints like limited bandwidth resources and single-BS association, varying degrees of knowledge matching between MUs and BSs should also impose new stringent criteria on the UA and BA. Especially noting that the SemCom-related network performance is linked with the stochasticity of source information generation, how to efficiently devise a joint optimal UA and BA strategy forms the third challenge.

To the best of our knowledge, no paper has addressed all these challenges before. In this chapter, we mainly investigate the resource management problem in

the downlink of SC-Net. By taking into account the unique knowledge-matching mechanism in SemCom, two different SC-Net scenarios are identified along with their respective joint optimization problems in terms of UA and BA. Correspondingly, two effective solutions are proposed to achieve the optimal semantics-level performance of SC-Net. In a nutshell, the novelty and the main contributions of this chapter are summarized as follows:

- We first identify and formally define two general SC-Net scenarios based on all possible knowledge matching states between MUs and BSs, namely perfect knowledge matching (PKM)-based SC-Net and imperfect knowledge matching (IKM)-based SC-Net. We then mathematically describe the distinctive semantic channel capacity model for the PKM-based SC-Net scenario from a semantic information-theoretical perspective. Taking this as the baseline case, the semantic channel model of IKM-based SC-Net is systematically constructed. The above addresses the aforementioned *Challenge 1*.

- Given the unique semantic channel models of SC-Net, we leverage a bit-rate-to-message-rate (B2M) transformation function to measure the message rate of each SemCom-enabled link, whereby a new metric, namely system throughput in message (STM), is effectively developed to accurately characterize the overall network performance at the semantic level. This corresponds to the aforementioned *Challenge 2*. Moreover, two joint STM-maximization problems of UA and BA are formulated for the two SC-Net scenarios, respectively.

- Resource management solutions are derived separately under the two scenarios. For the deterministic optimization problem in the PKM-based SC-Net, we directly employ a primal-dual decomposition method with a Lagrange-multiplier method to obtain the optimal UA and BA strategy. Notably, for the case of IKM-based SC-Net with a stochastic optimization problem, we particularly devise a two-stage solution to tackle with it. The first stage exploits a chance-constrained model to transform the primal stochastic problem into a deterministic one by introducing a given semantic confidence level, followed by the second stage solution using an interior-point method and a heuristic algorithm to finalize the joint optimal solution of UA and BA. Hence, *Challenge 3* is also well addressed.

- Extensive simulations are conducted for both SC-Net scenarios to evaluate the performance of proposed solutions. Compared with two baselines, numerical results demonstrate significant superiority of our solutions in terms

of STM performance. Moreover, the importance of adequate knowledge matching is also revealed, which can ensure low semantic ambiguity and high message rates in SemCom.

3.2 Semantic Communication Model

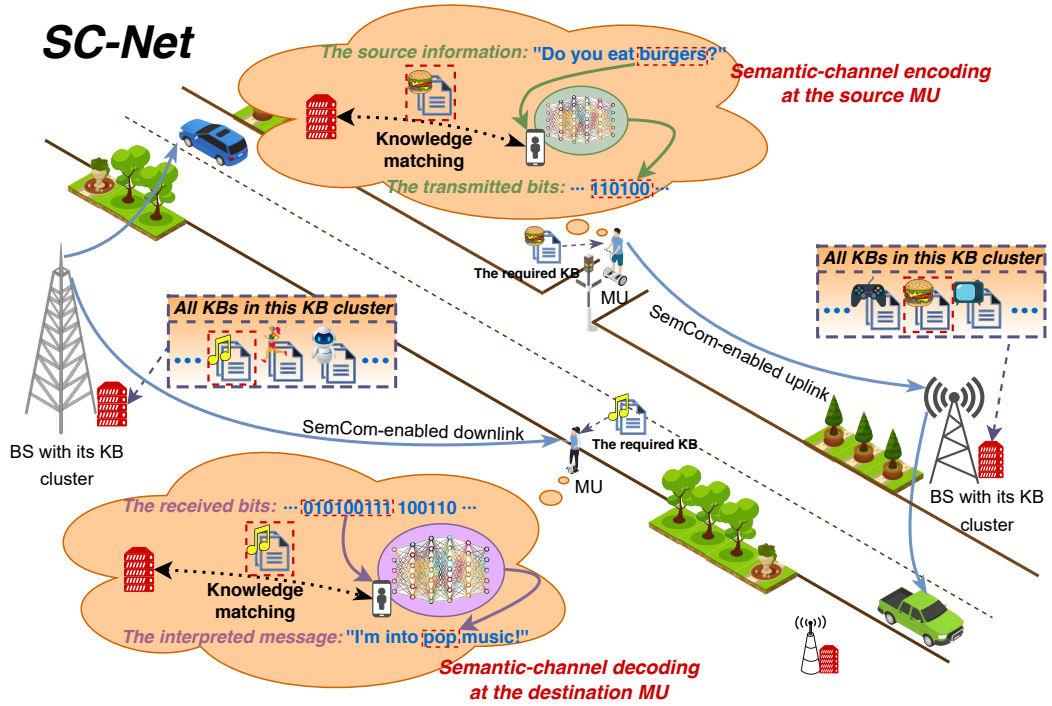

Figure 3.1: An overview of SC-Net.

3.2.1 Background Knowledge Matching in SemCom

Consider an SC-Net scenario as shown in Fig. 3.1, where all communication parties (i.e., BSs and MUs) are capable of performing SemCom with each other. Recall that the accuracy of SemCom strongly relies on the matching degree of correct background knowledge between the transceiver (i.e., each pair of interrelated BS and MU), and the better knowledge matching degree is believed to guarantee lower semantic ambiguity and more efficient information interaction [1,6,7]. Taking a single downlink in Fig. 3.1 as an example, when a message related to the personal favorite music genre is delivered from a BS to an associated MU, they must have the same background knowledge in the musical domain so as to achieve accurate SemCom. In other words, the MU should ensure that its associated BS has the background knowledge that matches its own as closely as possible before requesting its desired SemCom services.

On this basis, a key concept of auxiliary *KB* is introduced in SemCom, which is deemed a small information entity that stores the background knowledge of one particular application domain (such as music or sports) corresponding to a certain type of SemCom service [6, 8, 25]. Combined with the powerful computation and storage ability of the BS, we further assume that each BS holds random amounts and types of KBs and name them a KB cluster, thus the MUs can acquire different SemCom services with required KBs by associating with different BSs. Nevertheless, it should be noted that messages received by each MU may cover differing background knowledge at the same time, leading to varying degrees of knowledge mismatch between the MU and its associated BS in the UA process. In this respect, we give our first definition as follows.

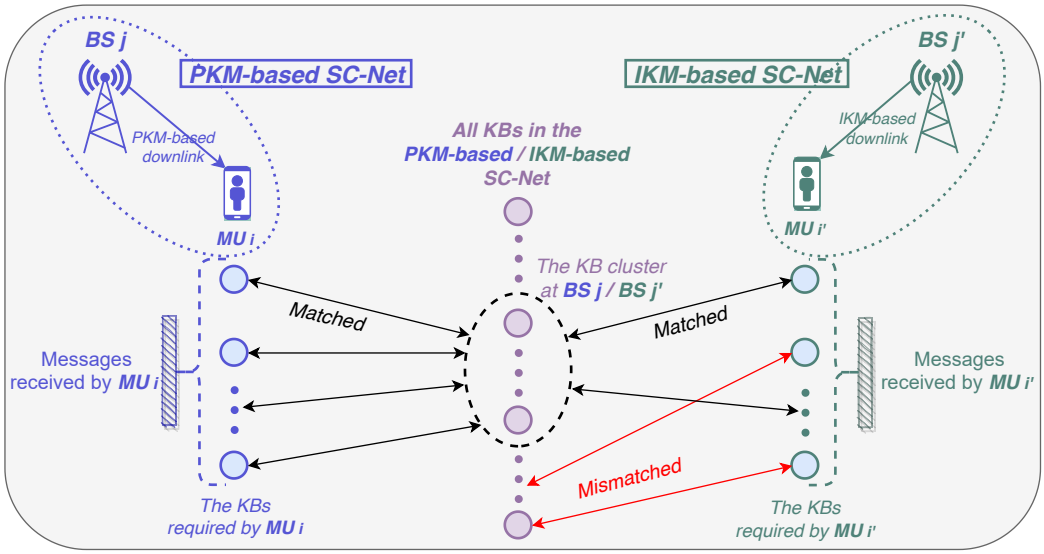

Figure 3.2: Example illustration of the PKM-based SC-Net (on the left) and the IKM-based SC-Net (on the right) with respect to a single SemCom-enabled link.

Definition 1. *According to all possible knowledge matching cases, we define two different SC-Net scenarios, as illustrated in Fig. 3.2.*

- *Perfect knowledge matching (PKM)-based SC-Net: For each MU in this network, there is at least one available BS holding all its required KBs to achieve perfect knowledge matching for SemCom.*
- *Imperfect knowledge matching (IKM)-based SC-Net: For each MU in this network, no BS holds all its required KBs, but its different associated BS may achieve varying degrees of (imperfect) knowledge matching for SemCom.*

From Fig. 3.2, it can be observed that MU i on the left is categorized into the PKM case, as its associated BS j precisely holds all KBs coherent with its

received messages. As for MU i' on the right, its associated BS j' possesses only some of the required KBs, thereby only part of its received messages can be successfully interpreted, the case of which is defined as IKM. In the following two subsections, we will elaborate on the above two different SC-Net scenarios and their corresponding semantic channel models, respectively.

3.2.2 Semantic Channel Model in the PKM-based SC-Net

Let us first consider a SemCom diagram as depicted in Fig. 3.3. Without loss of generality, the source information (i.e., the meaning desired to be conveyed) is modeled as a random variable W and the generated observable message[1] (e.g., a sentence or a speech signal representing the desired meaning) is denoted as X, which are defined over an appropriate product alphabet $\mathcal{W}\times\mathcal{X}$. Correspondingly, $\hat{X}\in\hat{\mathcal{X}}$ is the received message (e.g., the reconstructed sentence or speech), and $\hat{W}\in\hat{\mathcal{W}}$ is the interpreted information from $\hat{X}$ at the destination side. Among them, one bit encoder and one bit decoder are connected via a bit pipe (e.g., the wireless physical channel in traditional communications) to transmit the codeword $Y\in\mathcal{Y}$ at a certain code rate [13].

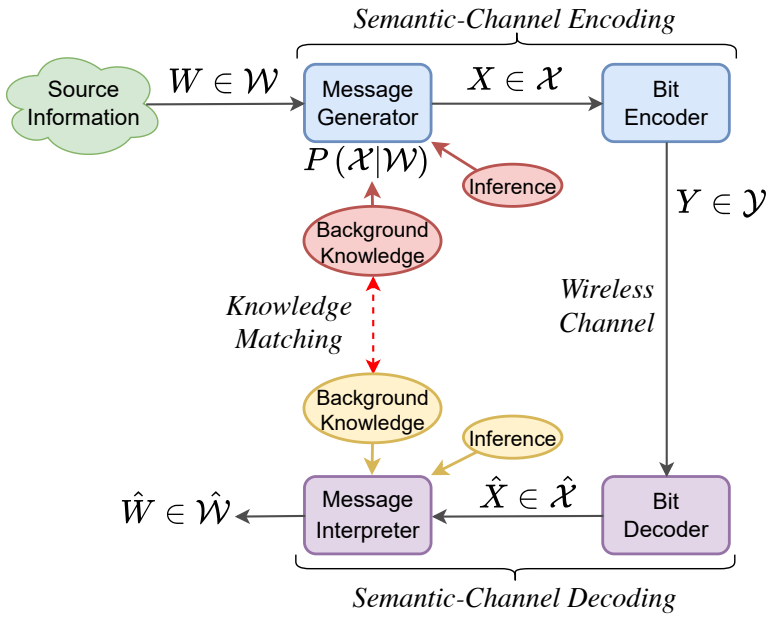

Figure 3.3: A SemCom diagram of information source and destination.

It is worth pointing out that the message generator of the source in Fig. 3.3 is to generate X from W based on a specific semantic encoding strategy, and here we

[1]The observable message here indicates a sequence of symbols syntactically (extrinsically) expressed in the language of the source, but actually contains specific information wished to be shared with the destination [7].

model the semantic encoding strategy as a conditional probabilistic distribution $P(\mathcal{X}|\mathcal{W})$ as that in [7] and [13]. Meanwhile, we notice that different coding strategies $P(\mathcal{X}|\mathcal{W})$ can incur different degrees of semantic ambiguity from a statistical level [6], since a given observable message may semantically have more than one meaning while only some of them are true with respect to (w.r.t.) the source.[2] In order to guarantee adequate efficiency and accuracy for semantic coding, the background knowledge and the inference capability of each coding model become crucial in SemCom, as elucidated in [7] and [11]. Herein, the inference capability can be understood as semantic coding models' feature compression and meaning interpretation abilities in the application case of deep learning-driven SemtCom, which are strongly correlated with the specific structures and composition of used neural networks [6].[3]

Further proceeding as in [7], if the message generator and the message interpreter are assumed to have the identical inference capability and the perfectly matched background knowledge, the condition of Theorem 3 (semantic-channel coding theorem) proposed in [7] can be fully met, stating that the semantic channel capacity (in units of messages per unit time, *msg/s*) of a discrete memoryless channel should be

$$C^s = \sup_{P(\mathcal{X}|\mathcal{W})} \left\{ \sup_{P(\mathcal{Y}|\mathcal{X})} \left\{ I(\mathcal{X};\hat{\mathcal{X}}) \right\} - H(\mathcal{W}|\mathcal{X}) + \overline{H_s(\hat{\mathcal{X}})} \right\}. \tag{3.1}$$

Here, $I(\mathcal{X};\hat{\mathcal{X}})$ is the mutual information between $\mathcal{X}$ and $\hat{\mathcal{X}}$ under the traditional bit encoding strategy (modeled as $P(\mathcal{Y}|\mathcal{X})$), while $H(\mathcal{W}|\mathcal{X})$ measures semantic ambiguity of coding at the source (w.r.t. $P(\mathcal{X}|\mathcal{W})$), both expressed in the form of classical Shannon entropy. Specially, unlike the above two terms, $\overline{H_s(\hat{\mathcal{X}})}$ measures the *semantic entropy* of received messages calculated by the *logical probability* (denoted as $P_s(\hat{X})$) [7], where $\overline{H_s(\hat{\mathcal{X}})} = -\sum_{\hat{X}\in\hat{\mathcal{X}}} P(\hat{X}) \log_2 P_s(\hat{X})$.[4]

In this case, if denoting the optimal semantic encoding strategy as $P^*(\mathcal{X}|\mathcal{W})$,

[2]This is obvious and has been sufficiently demonstrated with examples in many existing studies [6–8].

[3]For instance, in NLP-driven SemCom, attention-based models generally have a better inference capability than traditional recurrent (e.g., LSTM) or convolutional models (e.g., TextCNN) in the face of context prediction or sequence transduction related tasks [19].

[4]The concept of logical probability was first introduced by Carnap and Bar-Hillel in [12], determining how likely it is for an observable message to be true, which is quite different from the common statistical probability $P(\hat{X})$. More technical details of $P_s(\hat{X})$ can refer to [6,7,11], and [12].

we can substitute it into (3.1) and obtain

$$\begin{aligned} C^s &= \sup_{P(\mathcal{Y}|\mathcal{X})} \left\{ I^*(\mathcal{X};\hat{\mathcal{X}}) \right\} - H^*(\mathcal{W}|\mathcal{X}) + \overline{H_s^*(\hat{\mathcal{X}})} \\ &\triangleq \sup_{P(\mathcal{Y}|\mathcal{X})} \left\{ I^*(\mathcal{X};\hat{\mathcal{X}}) \right\} + H^s \\ &\triangleq C^b + H^s, \end{aligned} \tag{3.2}$$

where C^b characterizes the traditional Shannon channel capacity (in units of bits per unit time, *bit/s*), and $H^s = \overline{H_s^*(\hat{\mathcal{X}})} - H^*(\mathcal{W}|\mathcal{X})$ is a semantic-relevant term (can be positive or negative) depending on the background knowledge (i.e., the aforementioned KBs) and inference capability of specific semantic coding models adopted at the source and the destination.

Keeping (3.2) in mind and let us now consider a single downlink channel between MU i and BS j in the PKM-based SC-Net. According to Definition 1, we first know that MU i and BS j must have the perfectly matched KBs. If further assuming BS j's semantic encoder (i.e., message generator) has equal inference ability to MU i's semantic decoder (i.e., message interpreter), it is seen that Theorem 3 in [7] can be applied to this link. In line with (3.2), let C_{ij}^s be its achievable message rate, let C_{ij}^b be its achievable bit rate, and let H_{ij}^s be its given semantic-relevant term of this link, thus we can formalize their relationship by giving the following definition.

Definition 2. *In the PKM-based SC-Net, we define $S_{ij}^P(\cdot)$ as the Bit-rate-to-Message-rate (B2M) transformation function of the physical link between MU i and BS j, such that*

$$S_{ij}^P\left(C_{ij}^b\right) \triangleq C_{ij}^s = C_{ij}^b + H_{ij}^s. \tag{3.3}$$

In the light of Shannon theorem, we understand that C_{ij}^b can be directly calculated based on the bandwidth and the SINR of the link. Hence, given the channel condition and the semantic coding models, we are able to adjust the bandwidth (i.e., the input of $S_{ij}^P(\cdot)$) allocated to this link so as to optimize the corresponding achievable message rate (i.e., the output of $S_{ij}^P(\cdot)$). Moreover, it can be observed that $S_{ij}^P(\cdot)$ is linear with bit rate C_{ij}^b, which renders a clear path towards the solution to the later PKM-based resource optimization problem.

3.2.3 Semantic Channel Model in the IKM-based SC-Net

Note that the knowledge-matching mechanism is the key in SemCom as claimed in [14], each communication link no longer satisfies the perfect knowledge matching condition so that the aforementioned semantic-channel coding theorem be-

comes inapplicable to any IKM-based case. More unfortunately, to the best of our knowledge, no work has proposed a SemCom-related information theory with rigorous derivations to declare the semantic channel capacity under mismatched background knowledge between the transceiver (i.e., the IKM case). Nevertheless, we note in this work that for a given IKM-based link, there still exists an explicit relationship between the achievable message rate and the knowledge matching degree, i.e., the better the knowledge matching between source and destination, the more messages the destination can correctly interpret, and vice versa [6]. The rationale behind this is quite speculative. For instance, we know from Definition 2 that the message rate is capable of reaching an upper bound C_{ij}^{s} if MU i and BS j are in the PKM state, and conversely, no source information can be correctly interpreted if they have no matched KBs [7].

In view of the above, the following definition describes our semantic channel modeling for the IKM case.

Definition 3. *In the IKM-based SC-Net, we define $S_{ij}^{I}(\cdot)$ as the B2M transformation function of the physical link between MU i and BS j, which is correlated with its PKM-based B2M function $S_{ij}^{P}(\cdot)$ in the following manner*

$$S_{ij}^{I}(\cdot) = \beta_{ij} \cdot S_{ij}^{P}(\cdot). \tag{3.4}$$

Here, β_{ij} is named knowledge matching coefficient modeled as a random variable with the value ranging from 0 to 1, where $\beta_{ij} = 0$ represents a completely mismatched state between MU i and BS j, and $\beta_{ij} = 1$ represents a perfectly matched state.

Clearly, the upper bound of the message rate in the IKM case must be the message rate obtained in its PKM case (i.e., $\beta_{ij} = 1$), while it is also able to reach zero when there is no common background knowledge between the transceiver (i.e., $\beta_{ij} = 0$), as mentioned earlier. More importantly, since the information source is generally modeled as a stochastic process [13], the specific amount of its generated messages corresponding to the matched KBs or the mismatched KBs becomes uncertain, even given the knowledge matching state. As a result, compared to the PKM case, there is always only a random proportion β_{ij} of messages that can be correctly interpreted in the IKM case, eventually rendering a random message rate w.r.t. $S_{ij}^{I}(\cdot)$. In accordance with the above, we further make the following proposition.

Proposition 1. *Given the knowledge matching degree, denoted as τ_{ij}, between MU i and BS j in the IKM case, the random knowledge matching coefficient*

β_{ij} obeys a Gaussian distribution with mean τ_{ij} and variance σ_{ij}^2, i.e., $\beta_{ij} \sim \mathcal{N}\left(\tau_{ij}, \sigma_{ij}^2\right)$, where $\sigma_{ij}^2 = \tau_{ij}\left(1-\tau_{ij}\right)$.

Proof. Please see Appendix A. □

From Proposition 1, it is observed that in the IKM-based SC-Net case, each BS j is capable of only getting the deterministic information of τ_{ij} (i.e., the distribution of β_{ij}) from its link associated with MU i, which always leads to a stochastic optimization problem for IKM-based resource management. Therefore, the IKM problem is inevitable, and its solution should be quite distinct from that of the PKM problem due to the stochasticity of each β_{ij}.

3.2.4 Basic Network Topology of SC-Net

Let us now consider the network topology of both PKM-based and IKM-based SC-Nets. As shown in Fig. 3.1, suppose that there are a total of U MUs randomly located within the coverage of B BSs, in which each MU $i \in \mathcal{U} = \{1, 2, \ldots, U\}$ can only be associated with one BS $j \in \mathcal{B} = \{1, 2, \ldots, B\}$ at a time. Specially, in alignment with Definition 1, first let $\mathcal{B}_i^P$ ($\mathcal{B}_i^P \subseteq \mathcal{B}$, $\forall i \in \mathcal{U}$) denote the set of BSs holding all the KBs required by MU i. As for the case of IKM-based SC-Net, assuming there is a minimum threshold for the knowledge matching degree, denoted as τ_0, to guarantee the minimum quality of SemCom. That way, let $\mathcal{B}_i^I$ denote the set of BSs that MU i is eligible for (user association) UA in the IKM-based SC-Net, where $\mathcal{B}_i^I = \{j \mid j \in \mathcal{B}, \tau_{ij} \geqslant \tau_0\}$, $\forall i \in \mathcal{U}$. Based on the above, if we define the binary UA indicator for both scenarios as $x_{ij} \in \{0, 1\}$, where $x_{ij} = 1$ means that MU i is associated with BS j and $x_{ij} = 0$ otherwise, the UA constraints for MU i in the PKM-based and IKM-based SC-Nets are defined as follows:

$$\sum_{j \in \mathcal{B}_i^P} x_{ij} = 1 \qquad \text{and} \qquad \sum_{j \in \mathcal{B}_i^I} x_{ij} = 1, \quad \forall i \in \mathcal{U}, \tag{3.5}$$

respectively.

In the meantime, the total budget for bandwidth allocation (BA) of BS j is denoted as N_j, and the amount of bandwidth that the BS j assigns to MU i is denoted as n_{ij}. Let γ_{ij} be the SINR experienced by the link, so the achievable bit rate can be found by $C_{ij}^b = n_{ij} \log_2\left(1 + \gamma_{ij}\right)$. Further according to Definition 2 and Definition 3, the corresponding achievable message rate is $S_{ij}^P\left(C_{ij}^b\right)$ in the PKM-based SC-Net and $S_{ij}^I\left(C_{ij}^b\right)$ in the IKM-based SC-Net. With these, considering the uniqueness and significance of message rate in SemCom (i.e., the conveyed message itself becomes the sole focus of correct reception in SemCom rather than traditional transmitted bits [6, 7]), we define a new performance metric herein,

namely *system throughput in message* (STM), to specifically measure the overall message rates obtained by all MUs in the network. Consequently, the STM of PKM-based SC-Net is given as

$$STM^P = \sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{B}} x_{ij} S_{ij}^P\left(C_{ij}^b\right) = \sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{B}} x_{ij} S_{ij}^P\left(n_{ij}\log_2\left(1+\gamma_{ij}\right)\right). \tag{3.6}$$

Likewise, the STM of IKM-based SC-Net is

$$STM^I = \sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{B}} x_{ij} S_{ij}^I\left(C_{ij}^b\right) = \sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{B}} x_{ij}\beta_{ij} S_{ij}^P\left(n_{ij}\log_2\left(1+\gamma_{ij}\right)\right). \tag{3.7}$$

Based on STM^P and STM^I, we are now able to jointly optimize the UA and BA from the SemCom perspective so as to respectively maximize the overall network performance for the two SC-Net scenarios.

3.3 Resource Management for PKM-based SC-Net

3.3.1 Problem Formulation

In order to empower very high quality of SemCom services for all MUs in the PKM-based SC-Net, it is of paramount importance to achieve the optimality of STM^P subject to several SemCom-related and practical system constraints. To that end, we formulate an STM-maximization problem in a joint optimization manner of the UA variable x_{ij} and the BA variable n_{ij}. For ease of illustration, hereafter we define a matrix $\boldsymbol{x} = \{x_{ij} \mid i \in \mathcal{U}, j \in \mathcal{B}\}$ and a matrix $\boldsymbol{n} = \{n_{ij} \mid i \in \mathcal{U}, j \in \mathcal{B}\}$ consisting of all variables related UA and BA, respectively. Note that both $\boldsymbol{x}$ and $\boldsymbol{n}$ are strongly correlated to semantic components, for example, $\boldsymbol{x}$ is strictly constrained by PKM-based links and $\boldsymbol{n}$ determines the upper bound of not only the bit-based channel capacity but also the semantic channel capacity. To be specific, the joint optimization problem of PKM-based

SC-Net is given as follows:

$$\mathbf{P1}: \max_{\boldsymbol{x},\boldsymbol{n}} \quad \sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{B}} x_{ij} S_{ij}^{P}\left(n_{ij}\log_2\left(1+\gamma_{ij}\right)\right) \tag{3.8}$$

$$\text{s.t.} \quad \sum_{j\in\mathcal{B}_i^P} x_{ij} = 1,\ \forall i\in\mathcal{U}, \tag{3.8a}$$

$$\sum_{i\in\mathcal{U}} x_{ij} n_{ij} \leqslant N_j,\ \forall j\in\mathcal{B}, \tag{3.8b}$$

$$x_{ij}\in\{0,1\},\ \forall (i,j)\in\mathcal{U}\times\mathcal{B}. \tag{3.8c}$$

Constraint (3.8a) refers to the aforementioned single-BS constraint for UA, which also ensures that only the BSs in $\mathcal{B}_i^P$ can associate with MU i to achieve the PKM state. Constraint (3.8b) represents that the total bandwidth allocated to MUs cannot exceed the BA budget of each BS, and constraint (3.8c) characterizes the binary property of $\boldsymbol{x}$.

3.3.2 Optimal Solution for UA

Since the main difficulty of solving **P1** lies on the 0-1 constraint in (3.8c), we first relax $\boldsymbol{x}$ into the continuous variable between 0 and 1. Notably, although we can directly solve the relaxed problem after the slack to $\boldsymbol{x}$, the most nontrivial point on recovering the binary property of $\boldsymbol{x}$ with low performance compromise is still intractable. To avoid this obstacle, in our solution, we assume that there is a minimum bandwidth amount that BS j should allocate to its associated MU i, denoted as n_{ij}^T, to guarantee a basic quality of signal under the given channel condition. As such, by fixing each n_{ij} as n_{ij}^T, **P1** can be rephrased as

$$\mathbf{P1.1}: \max_{\boldsymbol{x}} \quad \sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{B}} x_{ij}\xi_{ij}^T \tag{3.9}$$

$$\text{s.t.} \quad \sum_{j\in\mathcal{B}_i^P} x_{ij} = 1,\ \forall i\in\mathcal{U}, \tag{3.9a}$$

$$\sum_{i\in\mathcal{U}} x_{ij} n_{ij}^T \leqslant N_j,\ \forall j\in\mathcal{B}, \tag{3.9b}$$

$$0\leqslant x_{ij}\leqslant 1,\ \forall (i,j)\in\mathcal{U}\times\mathcal{B}, \tag{3.9c}$$

where

$$\xi_{ij}^T \triangleq S_{ij}^P\left(n_{ij}^T\log_2\left(1+\gamma_{ij}\right)\right). \tag{3.10}$$

Notably, ξ_{ij}^T is deemed a constant in the objective function (3.9), since n_{ij}^T, γ_{ij}, and H_{ij}^s in the B2M function $S_{ij}^P(\cdot)$ are all constants for the given link between

MU i and BS j.

In the context of **P1.1**, we employ the Lagrange dual method [100] to obtain its dual optimization problem herein. By associating a Lagrange multiplier $\boldsymbol{\mu} = \{\mu_j \mid j \in \mathcal{B}\}$, the inequality constraint (3.9b) can be incorporated into (3.9), thereby its Lagrange function should be

$$L(x, \mu) = \sum_{i \in \mathcal{U}} \sum_{j \in \mathcal{B}} x_{ij} \xi_{ij}^T + \sum_{j \in \mathcal{B}} \mu_j \left(N_j - \sum_{i \in \mathcal{U}} x_{ij} n_{ij}^T \right). \tag{3.11}$$

Hence, the Lagrange dual problem of **P1.1** becomes

$$\mathbf{D1.1}: \min_{\boldsymbol{\mu}} \quad D(\boldsymbol{\mu}) = g_{\boldsymbol{x}}(\boldsymbol{\mu}) + \sum_{j \in \mathcal{B}} \mu_j N_j \tag{3.12}$$

$$\text{s.t.} \quad \mu_j \geqslant 0, \ \forall j \in \mathcal{B}, \tag{3.12a}$$

where we have

$$\begin{aligned} g_{\boldsymbol{x}}(\boldsymbol{\mu}) \ &= \ \sup_{\boldsymbol{x}} \ \sum_{i \in \mathcal{U}} \sum_{j \in \mathcal{B}} x_{ij} \left(\xi_{ij}^T - \mu_j n_{ij}^T \right) \\ &\text{s.t.} \ \ (3.9a), \ (3.9c). \end{aligned} \tag{3.13}$$

It is worth pointing out that strong duality holds in such primal-dual transformation, since the objective function (3.9) of **P1.1** is convex and all its constraints are linear and affine inequalities, thus satisfying the Slater's condition [101].

Given the initial dual variable $\boldsymbol{\mu}$, we first determine the optimal $\boldsymbol{x}$ (denoted as $\boldsymbol{x}^* = \left\{ x_{ij}^* \mid i \in \mathcal{U}, j \in \mathcal{B} \right\}$), and then leverage a gradient descent method [101] in charge of updating $\boldsymbol{\mu}$ to solve **D1.1** in an iterative fashion. Carefully examining (6.12), it is easily derived that based on the fixed n_{ij}^T, MU i can be served by its optimal BS j if and only if it satisfies the following condition

$$x_{ij}^* = \begin{cases} 1, & \text{if } j = \arg\max\limits_{j \in \mathcal{B}_i^P} \left(\xi_{ij}^T - \mu_j n_{ij}^T \right) \\ 0, & \text{otherwise} \end{cases}, \ \forall i \in \mathcal{U}. \tag{3.14}$$

After getting $\boldsymbol{x}^*$, the gradient w.r.t. $\boldsymbol{\mu}$ in the objective function $D(\boldsymbol{\mu})$ are calculated and set as the gradient in each iteration, whereby μ_j ($\forall j \in \mathcal{B}$) is updated as

$$\mu_j(t+1) = \left[\mu_j(t) - \delta(t) \cdot \left(N_j - \sum_{i \in \mathcal{U}} x_{ij}(t) n_{ij}^T \right) \right]^+. \tag{3.15}$$

The operator $[\cdot]^+$ here is to output the maximum value between its argument and zero, ensuring that $\boldsymbol{\mu}$ must be non-negative as constrained in (3.12a). $\delta(t)$ is the

stepsize in iteration t and generally, convergence of the gradient descent method can be guaranteed with the proper stepsize [35]. Finally, to further ensure that the BA constraint (3.9b) is not violated, the total amount of bandwidth consumed at each BS j needs to be checked based on the obtained $\boldsymbol{x}^*$. For each BS that violates (3.9b), we choose to reallocate its associated MUs who are consuming the most bandwidth to other BSs according to (6.3), until meeting the bandwidth budget requirements of all BSs. In summary, by alternatively updating $\boldsymbol{x}$ and $\boldsymbol{\mu}$ until convergence, the UA problem can be well solved in the PKM-based SC-Net.

3.3.3 Optimization Solution for BA

Given the obtained UA solution $\boldsymbol{x}^*$ and the fixed bandwidth threshold n_{ij}^T, we can directly formulate the BA problem for each BS j ($\forall j \in \mathcal{B}$) as follows:

$$\mathbf{P1.2^{(j)}}: \max_{\boldsymbol{n}} \quad \sum_{i\in\mathcal{U}_j^P} S_{ij}^P \left(n_{ij}\log_2\left(1+\gamma_{ij}\right)\right) \tag{3.16}$$

$$\text{s.t.} \quad \sum_{i\in\mathcal{U}_j^P} n_{ij} = N_j, \tag{3.16a}$$

$$n_{ij} \geqslant n_{ij}^T,\ \forall i \in \mathcal{U}_j^P, \tag{3.16b}$$

where

$$\mathcal{U}_j^P \triangleq \left\{ i \mid x_{ij}^* = 1 \right\}. \tag{3.17}$$

Here, $\mathcal{U}_j^P$ stands for the set of MUs associated with BS j in the previous UA phase. Owing to the linear property of $S_{ij}^P(\cdot)$, it is seen that for each $\mathbf{P1.2^{(j)}}$, the objective function as well as all constraints are convex, thereby some efficient optimization toolboxes such as CVPXY [102] can be applied to directly finalize the optimal BA solution of PKM-based SC-Net.

3.4 Resource Management for IKM-based SC-Net

3.4.1 Problem Formulation

Similar to the rationale behind **P1**, in the IKM-based SC-Net, achieving the optimality of STM^I is also necessary for optimizing the overall SemCom-related network performance. Based on the UA indicator $\boldsymbol{x}$ and the BA indicator $\boldsymbol{n}$,[5]

[5]In order to avoid unnecessary redundant notations, in the IKM-based SC-Net, we use the same notations (e.g., $\boldsymbol{x}$ and $\boldsymbol{n}$, etc.) as in the PKM-based SC-Net, which have exactly the same

the joint optimization problem of IKM-based SC-Net can be formulated as

$$\mathbf{P2}: \max_{\boldsymbol{x},\boldsymbol{n}} \quad \sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{B}} x_{ij}\beta_{ij}S_{ij}^{P}\left(n_{ij}\log_2\left(1+\gamma_{ij}\right)\right) \tag{3.18}$$

$$\text{s.t.} \quad \sum_{j\in\mathcal{B}_i^I} x_{ij}=1,\ \forall i\in\mathcal{U}, \tag{3.18a}$$

$$\sum_{i\in\mathcal{U}} x_{ij}n_{ij}\leqslant N_j,\ \forall j\in\mathcal{B}, \tag{3.18b}$$

$$x_{ij}\in\{0,1\},\ \forall\,(i,j)\in\mathcal{U}\times\mathcal{B}. \tag{3.18c}$$

Different from **P1**, the UA constraint (3.18a) in **P2** ensures that only the IKM-enabled BSs in $\mathcal{B}_i^I$ can associate with MU i, which is determined by the minimum knowledge matching threshold τ_0 in the network. Likewise, constraints (3.18b) and (3.18c) represent the bandwidth budget limitation of each BS and the binary nature of $\boldsymbol{x}$, respectively.

It is worth noting that the introduction of the random knowledge matching coefficient β_{ij} leads to the biggest distinction between **P1** and **P2**, where **P1** is clearly a deterministic optimization problem and **P2** is a stochastic optimization problem. That is, the solution $(\boldsymbol{x},\boldsymbol{n})$ to **P1** directly determines the numerical value of STM^P, while these two variables in **P2** actually affect the probability density function (PDF) of STM^I (w.r.t. β_{ij}). Hence, the main difficulty of solving **P2** lies on how to cope with the stochasticity of β_{ij}. In this work, we dedicatedly develop a two-stage method to determine the optimal $\boldsymbol{x}$ and $\boldsymbol{n}$. Specifically, the first stage is to convert the nondeterministic problem **P2** into a deterministic one by leveraging a chance-constrained optimization model. Afterward, we devise an effective heuristic algorithm in the second stage to finalize the solution of UA and BA for the IKM-based SC-Net.

3.4.2 Problem Transformation with Semantic Confidence Level

Carefully examining **P2**, it is seen that $\boldsymbol{\beta}=\{\beta_{ij}\mid i\in\mathcal{U},j\in\mathcal{B}\}$ only exists in its objective function (3.18). By taking into account the distribution of (3.18), in our first-stage solution, we employ Kataoka's model [103] to introduce a new objective function along with an extra constraint to make the primal problem suitable for stochastic optimization without altering the original intention. Denoting the new objective function as $\bar{F}(\boldsymbol{x},\boldsymbol{n})$ (which expression will be given later), according

physical meaning.

to [103], **P2** can be equivalently transformed into

$$\mathbf{P2.1}: \max_{\boldsymbol{x},\boldsymbol{n}} \quad \bar{F}(\boldsymbol{x},\boldsymbol{n}) \tag{3.19}$$

$$\text{s.t.} \quad \Pr\left\{STM^I \geqslant \bar{F}(\boldsymbol{x},\boldsymbol{n})\right\} \geqslant \alpha, \tag{3.19a}$$

$$(3.18a),\ (3.18b),\ (3.18c). \tag{3.19b}$$

Constraint (3.19a) is the newly introduced probabilistic (chance) constraint by a prescribed confidence level α ($0 < \alpha < 1$, large in practice [104]). To be more explicit, due to the randomness of β_{ij}, the goal of **P2.1** becomes to reach the optimality of $(\boldsymbol{x},\boldsymbol{n})$ to determine the optimal PDF of STM^I, whereby its lower bound $\bar{F}(\boldsymbol{x},\boldsymbol{n})$ can be maximized based on the given confidence level α. In this case, we name α as a semantic confidence level preset for the IKM-based SC-Net.[6]

Besides, it can be observed that in the case of the optimal solution to **P2.1**, STM^I has a nondegenerate distribution (i.e., it does not reduce to a constant [105]), which means

$$\Pr\left\{STM^I \geqslant \bar{F}(\boldsymbol{x},\boldsymbol{n})\right\} = \alpha \tag{3.20}$$

should have the same bound effect as constraint (3.19a) to reach the optimality of **P2.1**. According to our Proposition 1, the sufficient condition of Theorem 10.4.1 proposed in [105] is fully satisfied, which is to determine the specific expression of $\bar{F}(\boldsymbol{x},\boldsymbol{n})$ from (3.20). As such, we obtain

$$\begin{aligned}\bar{F}(\boldsymbol{x},\boldsymbol{n}) = &\sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{B}} x_{ij}\tau_{ij}S_{ij}^P\left(n_{ij}\log_2\left(1+\gamma_{ij}\right)\right) \\ &- \Phi^{-1}(\alpha)\sqrt{\sum_{i\in\mathcal{U}}\left(\sum_{j\in\mathcal{B}} x_{ij}\sigma_{ij}S_{ij}^P(n_{ij}\log_2\left(1+\gamma_{ij}\right))\right)^2},\end{aligned} \tag{3.21}$$

where $\Phi^{-1}(\cdot)$ is the inverse function of the standard normal probability distribution. In view of (3.20) and (3.21), we see that even the biggest value of $\bar{F}(\boldsymbol{x},\boldsymbol{n})$ (or in other words, all $(\boldsymbol{x},\boldsymbol{n})$) satisfies the confidence constraint in (3.19a). Therefore, (3.19a) can now be eliminated in **P2.1**.

On this basis, here we adopt the same strategy as in (3.10) to make **P2.1** tractable, where each n_{ij} in $\boldsymbol{n}$ is fixed by the given bandwidth threshold n_{ij}^T in the IKM-based SC-Net. Meanwhile, the UA variable $\boldsymbol{x}$ is relaxed into continuous as well to deal with the NP-hard obstacle. Consequently, we can obtain a

[6] An expected value of optimization goal seems to be also applicable for the measure of optimality criterion [104]. However, the dispersion of random variables' distribution leads to a greater risk of getting a very low profit under the given expectation, as explained in [103]. Hence, we set a given probability instead of the expected value in order to seek a higher practicality.

deterministic optimization problem as

$$\mathbf{P2.2}: \max_{\boldsymbol{x}} \quad \bar{F}(\boldsymbol{x}) \triangleq \bar{F}(\boldsymbol{x}, \boldsymbol{n})|_{n_{ij}=n_{ij}^T} \tag{3.22}$$

$$\text{s.t.} \quad \sum_{j \in \mathcal{B}_i^I} x_{ij} = 1, \ \forall i \in \mathcal{U}, \tag{3.22a}$$

$$\sum_{i \in \mathcal{U}} x_{ij} n_{ij}^T \leqslant N_j, \ \forall j \in \mathcal{B}, \tag{3.22b}$$

$$0 \leqslant x_{ij} \leqslant 1, \ \forall (i,j) \in \mathcal{U} \times \mathcal{B}. \tag{3.22c}$$

As declaimed in [103] and [104], the convexity of the objective function $\bar{F}(\boldsymbol{x})$ can be guaranteed if assuming $\alpha > 1/2$, i.e., $\Phi^{-1}(\alpha) > 0$. Such an assumption is quite reasonable and practical, since a too small α means a very high-level limit on solution space $(\boldsymbol{x}, \boldsymbol{n})$ according to constraint (3.19a), which may even cause the nonexistence of feasible solutions combined with other constraints in **P2.1**. In addition, it should be noted that n_{ij}^T, τ_{ij}, σ_{ij}, and α in **P2.2** should all be treated as known constants related to the link between MU i and BS j when solving this problem.

3.4.3 Solution Finalization for UA and BA

In our second-stage solution, we first utilize the interior-point method [106], to approximately formulate the inequality-constrained problem **P2.2** into an equality-constrained problem so as to efficiently approach the optimality. To be concrete, let $\varphi(\boldsymbol{x})$ be a logarithmic barrier associated with the BA constraint (3.22b), where

$$\varphi(\boldsymbol{x}) = \sum_{j \in \mathcal{B}} \log(N_j - \sum_{i \in \mathcal{U}} x_{ij} n_{ij}^T). \tag{3.23}$$

That way, **P2.2** can be rephrased as

$$\mathbf{P2.3}: \max_{\boldsymbol{x}} \quad \bar{F}(\boldsymbol{x}) + r \cdot \varphi(\boldsymbol{x}) \tag{3.24}$$

$$\text{s.t.} \quad (3.22a), \ (3.22c). \tag{3.24a}$$

Here, r is a small positive scalar that sets the accuracy of the approximation, and as r decreases to zero, the maximum of the new objective function as in (3.24) is able to converge to the optimal solution to the primal problem [106]. It is important to mention that (3.24) still holds the convexity since both $\bar{F}(\boldsymbol{x})$ and $\varphi(\boldsymbol{x})$ are convex. As such, we can easily find a set $\boldsymbol{x}(r)$ that contains all the optimal x_{ij} w.r.t. a given r for **P2.3**. Furthermore, according to the sequential unconstrained minimization mechanism [107], the optimal solution $\boldsymbol{x}$ to **P2.2**

(denoted as $\hat{\boldsymbol{x}}$) can be eventually obtained by iteratively updating the descent value of r until convergence.[7]

Nevertheless, such $\hat{\boldsymbol{x}}$ cannot guarantee the binary value for each $\hat{x}_{ij}$. Therefore, we devise a heuristic algorithm herein to finalize the optimal solution to **P2** (i.e., $\boldsymbol{x}^*$) based on the given $\hat{\boldsymbol{x}}$. Specifically, each x_{ij}^* is determined according to the following rule

$$x_{ij}^* = \begin{cases} 1, & \text{if } j = \arg\max\limits_{j \in \mathcal{B}_i^I} \hat{x}_{ij} \\ 0, & \text{otherwise} \end{cases}, \ \forall i \in \mathcal{U}. \tag{3.25}$$

An implicit interpretation to (3.25) is that each MU in the IKM state has multiple potentially associated BSs along with the corresponding optimal weights, i.e., $\hat{x}_{ij}$s, which are strongly correlated with the performance of STM. Therefore, each user can select the BS with the maximum weight for UA to pursue the highest overall network performance for the SC-Net.

Nevertheless, the resulted bandwidth consumption may still exceed some BSs' budget after executing (3.25). In this regard, we utilize the same countermeasure as in the PKM case by reassigning these MUs who consume the most bandwidth of these BSs to other BSs, based on their weight list given in $\hat{\boldsymbol{x}}$, until the BA constraint (3.22b) is satisfied at all BSs. Afterward, similar to the rationale of solving $\mathbf{P1.2}^{(\mathrm{j})}$, we can further formulate the BA optimization problem for each BS j ($j \in \mathcal{B}$) based on the obtained $\boldsymbol{x}^*$ and the fixed n_{ij}^T. That is,

$$\mathbf{P2.4}^{(\mathrm{j})}: \max_{\boldsymbol{n}} \quad \bar{F}(\boldsymbol{x}^*, \boldsymbol{n}) \tag{3.26}$$

$$\text{s.t.} \quad \sum_{i \in \mathcal{U}_j^I} n_{ij} = N_j, \tag{3.26a}$$

$$n_{ij} \geqslant n_{ij}^T, \ \forall i \in \mathcal{U}_j^I, \tag{3.26b}$$

where

$$\mathcal{U}_j^I = \left\{ i \mid i \in \mathcal{U}, x_{ij}^* = 1 \right\}. \tag{3.27}$$

In each $\mathbf{P2.4}^{(\mathrm{j})}$, the objective function (3.26) is clearly convex as $\bar{F}(\boldsymbol{x}, \boldsymbol{n})$ is convex, and both constraints (3.26a) and (3.26b) are linear, to which the toolbox CVPXY can be applied as well [102]. Finally, both the UA and BA problems have been well optimized in the IKM-based SC-Net, even with the intervention of the random knowledge matching coefficient $\boldsymbol{\beta}$. In terms of the computational complexity of the proposed solution, the interior-point method has a complexity of

[7]The value of r and its update rule will be well initialized at the beginning of the barrier method. More technical details can be found in [101].

$\mathcal{O}((UB)^{3.5}\log(1/r))$. Combined with the linear programming method for solving each $\mathbf{P2.4^{(j)}}$, the overall complexity should be still $\mathcal{O}((UB)^{3.5}\log(1/r))$.

3.5 Numerical Results and Discussions

In this section, we evaluate the performance of our proposed UA and BA solutions for PKM-based and IKM-based SC-Nets, respectively. In the basic network settings, we randomly drop 5 pico BSs (PBS), 10 femto BSs (FBS), and 200 MUs in a circular area with a radius of 500 meters, where a macro BS (MBS) is placed at the circle center. Meanwhile, the transmit power of the MBS, PBSs, and FBSs is set to 43 dBm, 35 dBm, and 20 dBm, respectively, each of which has a bandwidth budget of 2 MHz. For the wireless propagation model, we use $L(d) = 34 + 40\log(d)$ and $L(d) = 37 + 30\log(d)$ as the path loss model of the MBS/PBSs and FBSs, respectively, while supposing there is a fixed noise power of -111.45 dBm [29].

As for the SemCom-related model, we simulate a general text transmission-enabled SC-Net environment to examine the proposed solutions for accurate demonstration purposes. Note here that the transmission scenarios for other types of content (e.g., image or video) can also be simulated for performance test, and the reason we choose the text-based scenario is because that there already exist well-established NLP-driven SemCom models. To be specific, the Transformer with the same structure as proposed in [1] is adopted as a unified semantic coding model for all SemCom-enabled links, and the PyTorch-based Adam optimizer is applied for network training with an initial learning rate of 1×10^{-3}. Apart from this, all the source information used for transmission is based on a public dataset from the proceedings of European Parliament [108], where all initial sentences are pruned into a given word-counting range from 4 to 30 to facilitate subsequent computing efficiency and avoid potential gradient vanishing or explosion. With these, the corresponding PKM-based B2M function $S_{ij}^{P}(\cdot)$ can be approximated from model testing and will be shown later in the results. In the solution simulation of the PKM-based case, we set a dynamic stepsize of $\delta(t) = 0.8/t$ to update the Lagrange multipliers in (3.15), where the convergence of each trial can always be guaranteed. In the IKM-based case, the knowledge matching degree τ_{ij} (w.r.t. β_{ij} in Proposition 1, $\forall\,(i,j) \in \mathcal{U} \times \mathcal{B}$) is unified to 0.5 for all possible links in the SC-Net, and hereafter we omit the subscript ij from τ_{ij} for expression brevity. Moreover, the semantic confidence level is set to $\alpha = 95\%$ in the two-stage solution of the IKM case.

For comparison purposes, we utilize two baselines of UA and BA algorithms for

both the PKM-based and IKM-based SC-Nets: 1) A *max-SINR plus water-filling* algorithm [109], in which each MU is associated with the BS that can provide the strongest SINR in its UA phase with the water-filling BA method; 2) A *max-SINR plus evenly-distributed* algorithm [30] that adopts the same max-SINR strategy for UA and an evenly-distributed BA method. Furthermore, a bit rate threshold (w.r.t. n_{ij}^T) of 0.01 Mbit/s is fixed in both the proposed and baseline solutions to ensure a basic quality of SemCom services for all MUs. Notably, all the above parameter values are set by default unless otherwise specified, and all subsequent simulation results are obtained by averaging over a significantly large number of trials.

3.5.1 Performance Evaluations in the PKM-based SC-Net

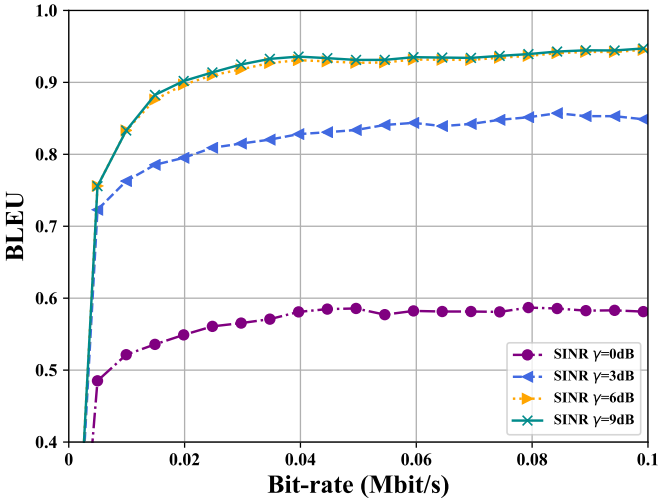

Figure 3.4: The BLEU score (1-gram) vs. bit rates under four different SINRs of 0, 3, 6, and 9 dB in the PKM-based SC-Net.

We first examine the performance of bilingual evaluation understudy (BLEU) in the PKM-based SC-Net, which is a classical metric in the NLP field with a value between 0 and 1 [110]. To be more concrete, it is scored via counting the word difference between the source and restored texts, and the closer its score is to 1, the better the text recovery. By testing the BLEU, the accuracy of semantic interpretation can be observed, which performance is also strongly related to the amount of messages MUs can correctly interpret in the network. As such, we first present the BLEU scores (1-gram) with different bit rates (i.e., C_{ij}^b) in Fig. 3.4, where four different SINRs are considered in the link. In this figure, it is seen that the BLEU under each SINR first grows as the bit rate improves, and soon stays at a stable score after about 0.03 Mbit/s. Besides, we can observe a higher BLEU under a higher SINR, and when the SINR is larger than 6 dB, the obtained BLEU

scores are almost the same. This trend is predictable since the received bits can suffer different degrees of signal attenuation from different channel conditions, and obviously, the more correct bits the MU receives, the lower semantic ambiguity it achieves. Particularly, the above phenomena indicate a necessity of providing a minimum bit rate for MUs under good channel conditions in the SC-Net to achieve high-quality SemCom.

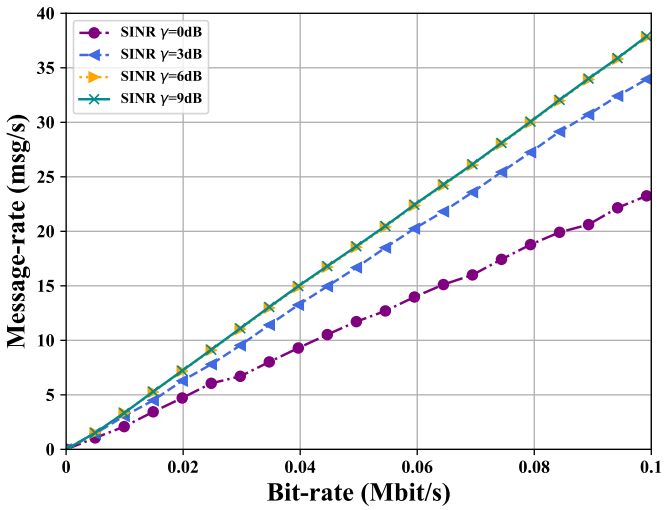

Figure 3.5: Demonstration of B2M transformation function under four SINRs in the PKM-based SC-Net.

According to (3.2), if the optimal semantic encoding strategy is guaranteed, it is believed that each MU can obtain a high message rate as the corresponding bit rate improves, which trend is related to its BLEU. To test this conjecture, we further draw the B2M transformation relationship (w.r.t. $S_{ij}^{P}(\cdot)$) under the same four SINRs in Fig. 3.5. Notably, the message rate (i.e., C_{ij}^{s}) obtained here is based on calculating the amount of messages correctly interpreted in a given time unit. As expected, we can see that the transformed message rate grows at a steady rate with increasing the bit rate, and the better the SINR, the higher the transformation rate of B2M.

The effectiveness of our UA and BA solution is demonstrated in the next two simulations, where the PKM-enabled BS means that each associated MU uses a well-trained Transformer decoding model under the perfectly matched training data. Fig. 3.6 first compares the proposed solution with the two baselines by evaluating the STM performance under varying numbers of MUs between 100 to 200. It is seen that the STM obtained by our solution always far outperforms the two baselines. Specifically, the proposed solution always maintains an average STM at 42.5 kmsg/s, which is around 6 kmsg/s higher than the max-SINR plus water-filling baseline and 13 kmsg/s than the max-SINR plus evenly-distributed

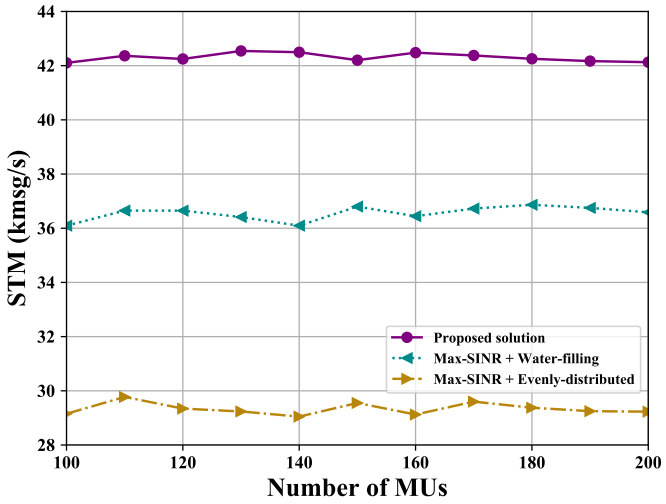

Figure 3.6: Comparison of the STM performance under different numbers of MUs in the PKM-based SC-Net.

baseline. Besides, the stable STM trend observed by all methods is because that the bandwidth budget of all BSs is reached in the BA phase, thus it is hard to improve the STM performance just by increasing the number of MUs.

Furthermore, similar comparisons are conducted with different numbers of PKM-enabled BSs between 10 to 20, as shown in Fig. 3.7. Consistent with the results in Fig. 3.6, the proposed solution gains an extra average STM around 6 kmsg/s compared with the max-SINR plus water-filling baseline, and around 12 kmsg/s compared with the max-SINR plus evenly-distributed baseline. In the meantime, we can see that the STM performance of our solution increases at the beginning, and then gradually tends to stabilize after exceeding 20 BSs. As there are more BSs that can provide MUs with PKM-based SemCom services, the MUs will correspondingly have more bandwidth resources available to achieve higher message rates. However, when the number of BSs surpasses a maximum threshold, the STM performance is believed to saturate and be even worsen, which is because of the severe channel interference incurred by the excess BSs.

3.5.2 Performance Evaluations in the IKM-based SC-Net

To evaluate the proposed two-stage solution in the IKM-based SC-Net, Fig. 3.8 first shows the comparisons of STM with different numbers of MUs, where three different semantic confidence levels of $\alpha = 55\%$, 75%, and 95% are taken into account. From this figure, we can see that the obtained STM increases with the number of MUs at the beginning, and soon remains stable after exceeding 130 MUs. This is because the bandwidth budget of some BSs starts to be reached after serving the high number of MUs, thereby the STM is inevitably stabilized, as

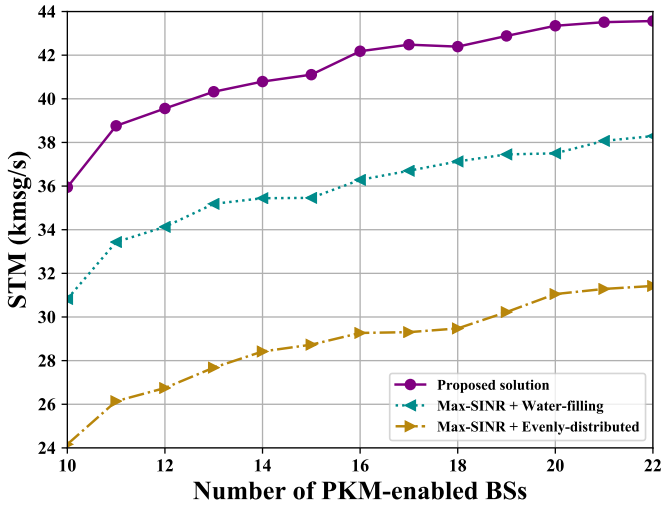

Figure 3.7: Comparison of the STM performance under different numbers of BSs in the PKM-based SC-Net.

mentioned earlier. Moreover, an upward trend of STM performance is observed along with the reduction of α, which can be explained from two perspectives. From the mathematical perspective, as introduced in (3.20), a higher value of α is equivalent to a lower achievable-bound on STM, which thus leads to the worse overall network performance. If we delve it in a semantical manner, due to the randomness of the knowledge matching degree in the IKM case, moderately increasing the preset semantic confidence can reduce the risk of getting below the expected message rate of each MU. Hence, some performance that compromises on STM are considered acceptable in alignment with the high preset semantic confidence level. Besides, it is seen in Fig. 3.8 that even at the highest required semantic confidence of $\alpha = 95\%$, the proposed solution can consistently outperform the two benchmarks.

A similar performance gain can be found in Fig. 3.9, where each solution is performed under two mean knowledge matching degrees of $\tau = 0.3$ and 0.7 w.r.t. $\boldsymbol{\beta}$ as in Definition 3. To be explicit, under each τ, our two-stage solution can always gain an extra STM performance around 2 kmsg/s to 6 kmsg/s with the increasing number of MUs when compared with the two baselines. In addition, the BA constraint of each method is always satisfied, hence, we can see the STM performance stabilizes from 130 MUs, which trend is consistent with that in the previous figure. As for the impact of different knowledge matching degrees, it always shows a reducing trend of STM as τ decreases. Since the higher τ represents the larger likelihood of having a good knowledge matching for SemCom, each MU can correspondingly obtain a higher accuracy of message interpretation, so that a better STM performance renders in the IKM-based SC-Net.

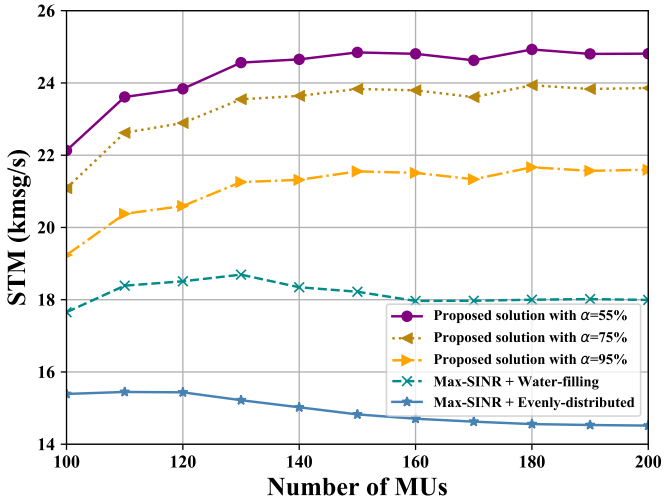

Figure 3.8: The STM performance against varying number of MUs under three semantic confidence levels in the IKM-based SC-Net.

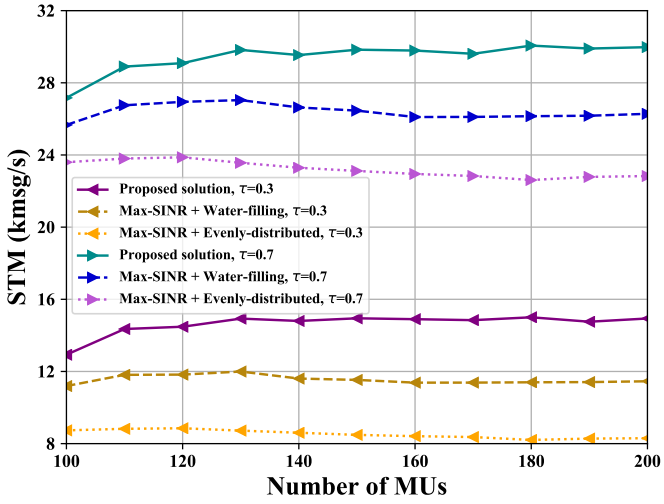

Figure 3.9: The STM performance against varying number of MUs under two average knowledge matching degrees in the IKM-based SC-Net.

Next, we evaluate the STM performance with different semantic confidence levels α and knowledge matching degrees τ in Fig. 3.10 and Fig. 3.11, respectively, under varying numbers of IKM-enabled BSs. Fig. 3.10 first presents the STM at different α, and the higher STM is seen again by the lower semantic confidence level, keeping the consistency with that in Fig. 3.8. Note that the lower semantic confidence level is generally inapplicable in practice, thus it may become tricky to consider in the IKM-based SC-Net how to strike a good balance between the preset risk level and the desired STM. As for the effect of τ, as expected that the higher knowledge matching degree still enables the better STM performance as shown in Fig. 3.11. Furthermore, we can always see a higher STM performance obtained by our solution when compared with the two baselines, and a slow growth trend of STM is also observed in all solutions as the number of BSs increases, which can be credit to more available bandwidth at MUs.

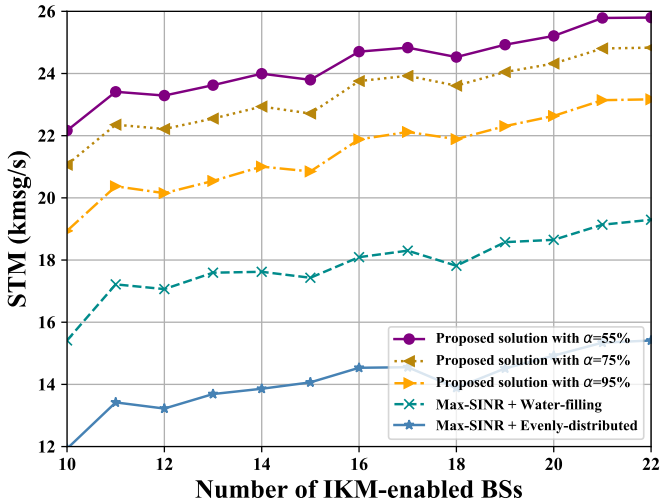

Figure 3.10: The STM performance against different numbers of BSs under three semantic confidence levels in the IKM-based SC-Net.

Finally, if we laterally compare the results in both PKM-based and IKM-based SC-Nets, it can be concluded that when the knowledge matching state of MUs changes from PKM to IKM, the penalty of STM performance is inevitable. Taking Fig. 3.6 and Fig. 3.8 as examples, in the same simulation settings, we observe an STM result around 42 kmsg/s by our solution in the PKM case, while only 25 kmsg/s STM is obtained in the IKM case. Due to the mismatching of partial KBs, the message generation and interpretation ability of IKM-based semantic coding models cannot be fully leveraged, which can incur a certain degree of semantic ambiguity compared with that in the PKM case. Therefore, it is of paramount importance to guarantee adequate knowledge matching degrees in SemCom to render a better network performance in the SC-Net.

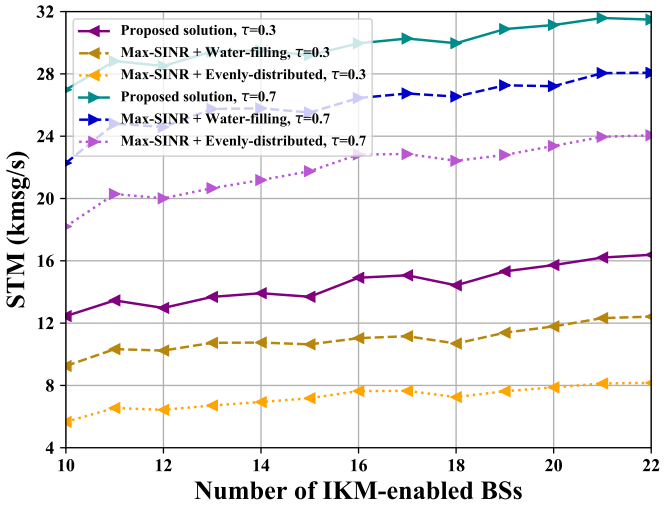

Figure 3.11: The STM performance against different numbers of BSs under two average knowledge matching degrees in the IKM-based SC-Net.

3.6 Conclusions

This chapter conducted a systematic study on SemCom from a networking perspective. Specifically, two typical scenarios of PKM-based and IKM-based SC-Nets were first identified, by which we presented their respective semantic channel models in combination with the existing works related to semantic information theory. After that, the concept of B2M transformation along with the new network performance metric STM were introduced in the two SC-Net scenarios, respectively. Then we formulated the joint optimization problem of UA and BA for each SC-Net scenario, followed by the corresponding solution proposed with the aim of STM maximization. Simulation results of both SC-Net scenarios demonstrated that our proposed solutions can always outperform two traditional benchmarks in terms of STM. The next chapter will investigate the energy efficiency problem for D2D SemCom in wireless cellular networks.

Chapter 4

Resource Allocation for D2D Semantic Communication Underlying Energy Efficiency-Driven Cellular Networks

4.1 Introduction

In this chapter, we explore the resource allocation problem in a D2D SemCom-enabled cellular network, where energy efficiency is specially considered as the target performance metric. As aforementioned, the energy efficiency metric has been widely adopted in many related works to provide a quantitative analysis of the power resource saving potential of a certain algorithm. In traditional BitCom networks, energy efficiency is typically defined as the ratio between the total data rate of all users and the total energy consumption (bits/Joule) [75, 78, 111]. In view of the unique characteristics of SemCom, we are encountering three fundamental networking challenges in the EE-SCN.

- *Challenge 1: How to measure the semantic-level performance for each SemCom user?* Different from the traditional communications, the semantic-level performance needs to be mathematically characterized for each SemCom user. In particular, mismatched background knowledge between the semantic encoder and decoder can cause a certain degree of semantic ambiguity as well as information distortion [7]. Hence, the first challenging problem lies in how to sketch a reasonable semantic performance based on

different knowledge-matching degrees.

- *Challenge 2: How to define the energy efficiency model in SemCom?* Due to the unique semantic coding process in SemCom, there may be additional energy consumption in SemCom-enabled mobile devices. Besides, the network performance per unit of consumed energy should be measured from a semantic-level perspective, which problem needs to be coupled with Challenge 1. As such, the second challenge becomes how to characterize the energy efficiency for each SemCom user appropriately.

- *Challenge 3: How to determine the best spectrum reusing pattern for each DUE and the power allocation scheme for all CUEs and DUEs?* Since the D2D SemCom method is employed in this work, with the target of maximizing energy efficiency, the unique knowledge-matching condition makes the spectrum reusing quite challenging in combination with the cellular SemCom method. Especially considering the power allocation issue, the joint solution should be tricky to determine.

To this end, this work proposed the optimal resource allocation solution by taking into account the knowledge-matching mechanism in the energy efficiency model. Numerical results demonstrate the superiority of our proposed solution in comparison with two different benchmarks. In a nutshell, the main contributions of this chapter are summarized as follows:

- We first define the performance metric of SemCom by employing the B2M function. Besides, the knowledge-matching state of each CUE and DUE is considered in the energy efficiency model, and thus the power consumption of each semantic data packet can be identified.

- We formulate an energy efficiency-maximization problem by jointly optimizing the power allocation and spectrum reusing indicators. The mathematical difficulties are then analyzed.

- We propose an optimal resource allocation solution for EE-SCNs. The primal fractional-form problem is transformed into the subtractive-form one, followed by utilizing a heuristic algorithm and a Hungarian algorithm.

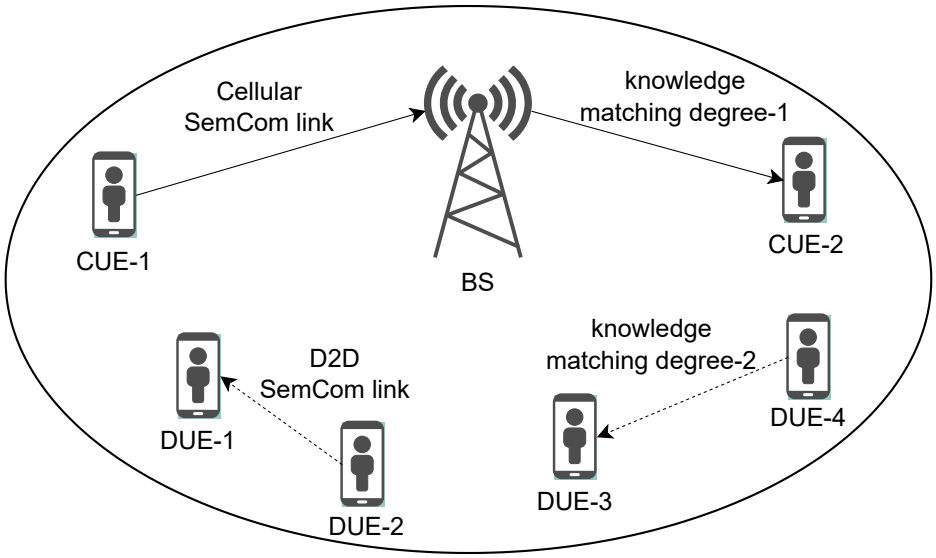

Figure 4.1: The overview of EE-SCN.

4.2 System Model

4.2.1 EE-SCN Scenario

Consider a single-cell EE-SCN scenario, where one BS is placed at the cell center to provide a total of M associated cellular users (CUEs) with wireless SemCom services, as shown in Fig. 4.1. Note that each CUE $i \in \mathcal{M} = \{1, 2, \cdots, M\}$ has been pre-allocated an orthogonal uplink subchannel with equal channel bandwidth W, while having a knowledge-matching degree τ_i^C $(0 \leqslant \tau_i^C \leqslant 1)$ with its communication counterpart. Meanwhile, a total of N $(N \leqslant M)$ pairs of D2D users (DUEs) coexist in the cell, and likewise, assume that each DUE $j \in \mathcal{N} = \{1, 2, \cdots, N\}$ has a knowledge-matching degree τ_j^D $(0 \leqslant \tau_j^D \leqslant 1)$ between its transmitter and receiver. For efficient spectrum utilization and interference management, each DUE can only reuse the subchannel of one CUE to execute the D2D SemCom services, and the subchannel of each CUE can be reused by at most one DUE. Here, let a binary variable $\alpha_{i,j} \in \{0, 1\}$ denote the spectrum reusing indicator, where $\alpha_{i,j} = 1$ represents that DUE j reuses the subchannel of CUE i and $\alpha_{i,j} = 0$ otherwise. Besides, the maximum transmit powers of CUEs and DUEs are denoted as P_{max}^C and P_{max}^D, respectively.

4.2.2 Channel Model and SemCom Model

For the data dissemination model in the EE-SCN, the channel power gain between each CUE i and the BS, the gain between DUE j and the BS, the gain between the transmitter and the receiver at DUE j, and the gain between each CUE i and

each DUE j are denoted as $G_{i,B}$, $G_{j,B}$, G_j^D, and $G_{i,j}$, respectively. If denoting the transmit power of each CUE i as P_i^C and the transmit power of each DUE j as P_j^D while combined with their potential spectrum reuse situations, the SINR of the uplink at CUE i is calculated by

$$\gamma_i^C = \frac{P_i^C G_{i,B}}{W\delta_0 + \sum_{j\in\mathcal{N}} \alpha_{i,j} P_j^D G_{j,B}}, \tag{4.1}$$

and the SINR of the link between the transceiver of DUE j is

$$\gamma_j^D = \frac{P_j^D G_j^D}{W\delta_0 + \sum_{i\in\mathcal{M}} \alpha_{i,j} P_i^C G_{i,j}}, \tag{4.2}$$

where δ_0 is noise power spectral density.

Hence, the bit throughput at CUE i is given by

$$r_i^C = W \log_2\left(1 + \gamma_i^C\right), \tag{4.3}$$

and the bit throughput at DUE j is

$$r_j^D = W \log_2\left(1 + \gamma_j^D\right). \tag{4.4}$$

Recall that the conveyed message itself becomes the sole focus of precise reception in SemCom rather than traditional transmitted bits in conventional bit communication, we again employ the performance metric developed in our previous work [21] to measure the message rate at each SemCom-enabled CUE and DUE via employing the B2M transformation function. To be specific, the B2M function is to output the semantic channel capacity (i.e., the achievable message rate in units of messages per unit time, *msg/s*) from input traditional Shannon channel capacity (i.e., the achievable bit rate in units of bits per unit time, *bit/s*) under the discrete memoryless channel. Especially, if the information source and destination have identical semantic reasoning capability, the B2M function can be approximated as linear [7, 21, 112]. For simplicity, each CUE and its communication counterpart are assumed to be equipped with the identical maturely-trained semantic coding models, and the same assumption also applies to the transceiver of each DUE. In this way, let $\Re_i^C(\cdot)$ denote the B2M function of each CUE i and let $\Re_j^D(\cdot)$ denote the B2M function of each DUE j, the message throughputs at CUE i and at DUE j are

$$q_i^C = \tau_i^C \Re_i^C\left(r_i^C\right), \tag{4.5}$$

and

$$q_j^D = \tau_j^D \Re_j^D \left(r_j^D\right). \tag{4.6}$$

Consequently, the overall message throughput of all SemCom-enabled uses in the EE-SCN should be

$$Q^{total} = \sum_{i\in\mathcal{M}} q_i^C + \sum_{j\in\mathcal{N}} q_j^D. \tag{4.7}$$

4.2.3 Energy Efficiency Model of SemCom Systems

In this work, we focus on the overall energy consumption including the contributions of the power amplifier and the unique semantic encoding circuit module at all SemCom-enabled CUEs and DUEs, which takes into account the knowledge-matching degree factor in the data packet unit. Specifically, most of the existing works consider the power dissipation in the power amplifier of each transmitter [34, 113, 114], which has been widely recognized as the noteworthy source of energy loss in the present wireless network. In line with this, we first define the power amplifier inefficiency coefficient as ξ ($\xi \geqslant 1$), which is a constant associated with the transmit power of each user. As such, the overall energy consumption (in *Watts*) for signal transmission in the EE-SCN is found by

$$E^{amp} = \xi \left(\sum_{i\in\mathcal{M}} P_i^C + \sum_{j\in\mathcal{N}} P_j^D \right). \tag{4.8}$$

In addition, different from the conventional model that pays attention to the circuit power consumption for processing the unit of bit data [114], we assume that such power loss in SemCom-enabled user equipments occurs pertinent to the knowledge-matching state of each semantic data packet. This assumption is justified since the content in each knowledge-mismatching packet necessarily requires more computing power and processing time for accurate contextual reasoning and interpretation than each knowledge-mismatching packet, due to the use of more sophisticated semantic-coding networks or the knowledge-sharing method, etc [8]. Therefore, for the semantic encoding module in each transmitter, its circuit power consumption for processing each knowledge-matching packet is assumed to be fixed and equal as P^{mat}, and for each knowledge-mismatching packet is P^{mis}, and we have $P^{mat} < P^{mis}$ in practice. If denoting the uniform size of all semantic data packets as L, according to the bit rate calculation in (4.3) and (4.4) combined with the knowledge-matching degree at each CUE and DUE, from a long-term perspective, the overall energy consumption for semantic encoding

becomes

$$E^{sc} = \sum_{i\in\mathcal{M}} \frac{r_i^C}{L}\left[\tau_i^C P^{mat} + (1-\tau_i^C)P^{mis}\right] + \sum_{j\in\mathcal{N}} \frac{r_j^D}{L}\left[\tau_j^D P^{mat} + (1-\tau_j^D)P^{mis}\right]. \tag{4.9}$$

Clearly, the total energy consumption becomes

$$E^{total} = E^{amp} + E^{sc}. \tag{4.10}$$

In view of the above, the energy efficiency of each SemCom system is defined as the total number of messages successfully conveyed to the receiver per Joule consumed energy, i.e., $\eta_{EE} = Q^{total}/E^{total}$.

4.2.4 Problem Formulation

For ease of illustration, we first define three variable sets $\boldsymbol{P^C} = \left\{P_i^C \mid i \in \mathcal{M}\right\}$, $\boldsymbol{P^D} = \left\{P_j^D \mid j \in \mathcal{N}\right\}$, and $\boldsymbol{\alpha} = \{\alpha_{i,j} \mid i \in \mathcal{M}, j \in \mathcal{N}\}$ that consist of all possible indicators pertinent to power allocation and spectrum reusing, respectively. Without loss of generality, the objective is to maximize the energy efficiency η_{EE} of EE-SCN by jointly optimizing $(\boldsymbol{P^C}, \boldsymbol{P^D}, \boldsymbol{\alpha})$, while subject to SemCom-relevant requirements alongside several practical system constraints. The problem is now formulated as follows:

$$\begin{aligned}
\mathbf{P0}: \max_{\boldsymbol{P^C},\boldsymbol{P^D},\boldsymbol{\alpha}} \quad & \eta_{EE} && (4.11)\\
\text{s.t.} \quad & q_i^C \geqslant q_{min}^C,\ \forall i \in \mathcal{M}, && (4.11a)\\
& q_j^D \geqslant q_{min}^D,\ \forall j \in \mathcal{N}, && (4.11b)\\
& 0 < P_i^C \leqslant P_{max}^C,\ \forall i \in \mathcal{M}, && (4.11c)\\
& 0 < P_j^D \leqslant P_{max}^D,\ \forall j \in \mathcal{N}, && (4.11d)\\
& \sum_{j\in\mathcal{N}} \alpha_{i,j} \leqslant 1,\ \forall i \in \mathcal{M}, && (4.11e)\\
& \sum_{i\in\mathcal{M}} \alpha_{i,j} = 1,\ \forall j \in \mathcal{N}, && (4.11f)\\
& \alpha_{i,j} \in \{0,1\},\ \forall (i,j) \in \mathcal{M}\times\mathcal{N}. && (4.11g)
\end{aligned}$$

Constraints (4.11a) and (4.11b) guarantee the minimum semantic-level performance that should be achieved at each CUE and DUE, respectively. Similarly, constraints (4.11c) and (4.11d) limit the maximum transmit power for each CUE and DUE, respectively. Then, constraint (4.11e) represents that the uplink subchannel of each CUE can be shared by at most one DUE, while constraint (4.11f)

stipulates that each DUE can only reuse one uplink subchannel of an existing CUE. Finally, constraint (4.11g) characterizes the binary properties of $\boldsymbol{\alpha}$.

Carefully examining **P0**, it can be observed that the optimization is rather challenging to be solved straightforwardly due to several intractable mathematical obstacles. First of all, **P0** involves both continuous and discrete variables, leading to an obvious NP-hard problem. Besides, the expression of the objective function η_{EE} is quite complicated alongside the constraints (4.11a) and (4.11b), which is nonconvex and thus generally requires a high-complexity solution procedure. Therefore, we propose an efficient power allocation and spectrum reusing strategy in the next section to reach the optimality of **P0**.

4.3 Optimal Resource Allocation for EE-SCNs

In this section, we illustrate how to design our optimal resource allocation solution to cope with the energy efficiency optimization problem in the EE-SCN. Specifically, the primal problem **P0** is first transformed, without loss of optimality, from its original fractional form into an equivalent subtractive form (referring to **P1**) by drawing on the Dinkelbach's method [115], while the convexity of **P1** is theoretically proved. Notice that **P1** should be solved in an iterative fashion [114], and in each iteration, we dedicatedly devise a three-stage method. In the first and second stages, **P1** is decomposed into $U = M \times N$ subproblems (referring to $\mathbf{P2}_{i,j}, \forall (i,j) \in \mathcal{M} \times \mathcal{N}$) and M subproblems (referring to $\mathbf{P3}_i, \forall i \in \mathcal{M}$), respectively. Among them, each $\mathbf{P2}_{i,j}$ corresponds to a potential spectrum reusing pair of CUE i and DUE j, and each $\mathbf{P3}_i$ corresponds to one CUE without spectrum reusing. As such, we aim to seek the optimal power allocation strategy with respect to (w.r.t.) $\mathbf{P2}_{i,j}$ for each potential CUE-DUE pair, and w.r.t. $\mathbf{P3}_i$ for each single CUE. After solving all these subproblems, the spectrum reusing policy w.r.t. $\boldsymbol{\alpha}$ is optimally finalized (referring to **P4**). In the end, we present the workflow of our solution along with its complexity analysis.

4.3.1 Fractional-to-Subtractive Problem Transformation

By observing the non-convex fractional-form objective function η_{EE} in **P0**, inspired by the Dinkelbach's method, we first transform η_{EE} into a subtractive-form function $F(\eta_{EE})$ w.r.t. η_{EE} to make **P0** tractable. The transformation process is established in accordance with the following proposition.

Proposition 2. P0 *must have the same optimal solution as*

$$\mathbf{P1}: \ F(\eta_{EE}) = \max_{\boldsymbol{P^C}, \boldsymbol{P^D}, \boldsymbol{\alpha}} \quad Q^{total} - \eta_{EE} \cdot E^{total} \tag{4.12}$$

$$\text{s.t.} \quad \textit{(4.11a)} - \textit{(4.11g)}, \tag{4.12a}$$

if and only if $F(\eta_{EE}) = 0$.

Proof. Please see Appendix B. □

From Proposition 2, our optimization goal becomes solving **P1** given any η_{EE} while requiring the iterative update for η_{EE} such that $F(\eta_{EE})$ approaches 0. Specifically, suppose in a certain iteration, say iteration t, η_{EE} is updated by

$$\eta_{EE}(t+1) = \frac{Q^{total}\left(\boldsymbol{P^{C*}}(t), \boldsymbol{P^{D*}}(t), \boldsymbol{\alpha^*}(t)\right)}{E^{total}\left(\boldsymbol{P^{C*}}(t), \boldsymbol{P^{D*}}(t), \boldsymbol{\alpha^*}(t)\right)}, \tag{4.13}$$

where $\left(\boldsymbol{P^{C*}}(t), \boldsymbol{P^{D*}}(t), \boldsymbol{\alpha^*}(t)\right)$ is the optimal solution to **P1** in iteration t. The above iterative update should be stopped when either reaching the maximum number of iterations or satisfying $F(\eta_{EE}(t)) < \epsilon$, where ϵ is a preset very small positive value. Notably, the convergence of $F(\eta_{EE}(t))$ can be guaranteed, the proof of which can refer to [114] and [115].

Given any η_{EE} in each iteration, we now concentrate upon how to reach the optimality of **P1**. However, solving such a problem is still tricky due to the mixed integer variables in its highly complex objective function (4.12). To this end, we propose a three-stage method to separately obtain the optimal power allocation indicators $\left(\boldsymbol{P^C}, \boldsymbol{P^D}\right)$ and the optimal spectrum reusing indicator $\boldsymbol{\alpha}$ with polynomial-time complexity.

4.3.2 Optimal Power Allocation for a Single CUE-DUE Pair

In the first stage, the power allocation scheme is considered for optimization at a specific pair of CUE i ($\forall i \in \mathcal{M}$) and DUE j ($\forall j \in \mathcal{N}$). As such, we construct $U = M \times N$ subproblems, each of which is denoted as $\mathbf{P2}_{i,j}$ and the objective is to maximize the energy efficiency of the single spectrum reusing pair. Herein, it is worth pointing out that the optimal solution to **P1** cannot be achieved by simply combining the obtained schemes of these $\mathbf{P2}_{i,j}$, but these power allocation schemes will be used to construct the subsequent spectrum reusing subproblem to finalize the joint optimal solution for **P1**. In particular, when DUE j reuses

the subchannel of CUE i (i.e., $\alpha_{i,j}=1$), given any η_{EE}, $\mathbf{P2}_{i,j}$ turns out to be

$$\mathbf{P2}_{i,j}: \max_{P_i^C, P_j^D} \quad \lambda_{i,j} \tag{4.14}$$

$$\text{s.t.} \quad \tau_i^C \Re_i^C\left(\overline{r_i^C}\right) \geqslant q_{min}^C, \tag{4.14a}$$

$$\tau_j^D \Re_j^D\left(\overline{r_j^D}\right) \geqslant q_{min}^D, \tag{4.14b}$$

$$0 \leqslant P_i^C \leqslant P_{max}^C, \tag{4.14c}$$

$$0 \leqslant P_j^D \leqslant P_{max}^D. \tag{4.14d}$$

The objective function $\lambda_{i,j}$ is presented by

$$\lambda_{i,j} = \tau_i^C \Re_i^C\left(\overline{r_i^C}\right) + \tau_j^D \Re_j^D\left(\overline{r_j^D}\right) - \eta_{EE}\left[\xi\left(P_i^C + P_j^D\right) + \frac{\overline{r_i^C}}{L/\left[\tau_i^C P^{mat} + (1-\tau_i^C)P^{mis}\right]} + \frac{\overline{r_j^D}}{L/\left[\tau_j^D P^{mat} + (1-\tau_j^D)P^{mis}\right]}\right], \tag{4.15}$$

while we define

$$\overline{r_i^C} = W \log_2\left(1 + \frac{P_i^C G_{i,B}}{W\delta_0 + P_j^D G_{j,B}}\right) \tag{4.16}$$

and

$$\overline{r_j^D} = W \log_2\left(1 + \frac{P_j^D G_j^D}{W\delta_0 + P_i^C G_{i,j}}\right) \tag{4.17}$$

for expression brevity.

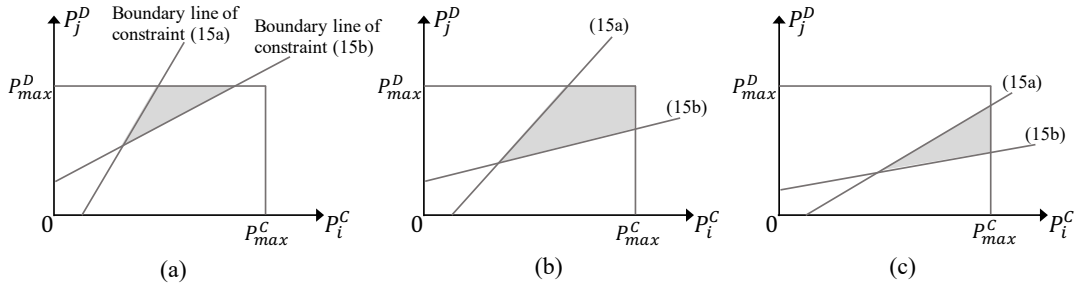

Figure 4.2: Three possible cases of the feasible power allocation region ψ for each pair of CUE i and DUE j w.r.t. $\mathbf{P2}_{i,j}$.

Intuitively, the more bits are transmitted, the more messages can be conveyed to the receiver, and thus, proceeding as [21], $\Re_i^C(\cdot)$ and $\Re_j^D(\cdot)$ are actually deemed two monotonically linearly increasing functions w.r.t. $\overline{r_i^C}$ and of $\overline{r_j^D}$, respectively. Combined with (4.16) and (4.17), it is seen that in the boundary case of either the constraint (4.14a) or (4.14b), P_j^D can be expressed as a linear function of P_i^C. Consequently, Fig. 4.2 depicts three possible cases for the closed feasible region ψ of the two-variable optimization $\mathbf{P2}_{i,j}$, according to different values of

channel gains, knowledge-matching degrees, the required minimum message rates and maximum transmit powers.

Having these and examining (4.15) again, $\lambda_{i,j}$ can be rephrased as a more concise form by combining all the constant coefficients of $\overline{r_i^C}$ and combining all the constant coefficients of $\overline{r_j^D}$, respectively. If denoting σ_i^C as the constant coefficient of $\overline{r_i^C}$ and σ_j^D as the constant coefficient of $\overline{r_j^D}$ after combining like terms, $\lambda_{i,j}$ can be rewritten as

$$\lambda_{i,j} = \sigma_i^C \overline{r_i^C} + \sigma_j^D \overline{r_j^D} - \eta_{EE}\xi\left(P_i^C + P_j^D\right). \tag{4.18}$$

Note that either σ_i^C or σ_j^D is possible to have both positive and negative, which can be easily obtained once the semantic coding models (w.r.t. $\Re_i^C(\cdot)$ or $\Re_j^D(\cdot)$) and other relevant system parameters (i.e., τ_i^C or τ_j^D, L, P^{mat}, and P^{mis}) are determined for each CUE i or DUE j.

Now, suppose that we first randomly generate a feasible solution in $\mathbf{P2}_{i,j}$'s feasible region of ψ, denoted as $\left(\widetilde{P_i^C}, \widetilde{P_j^D}\right)$, and substitute it into the first term $\sigma_i^C \overline{r_i^C}$, such that

$$\sigma_i^C \overline{r_i^C} = \sigma_i^C W \log_2\left(1 + \frac{\widetilde{P_i^C} G_{i,B}}{W\delta_0 + \widetilde{P_j^D} G_{j,B}}\right) = \lambda_0, \tag{4.19}$$

where λ_0 is the corresponding calculated value. Clearly, (4.19) can yield a line segment w.r.t. $\left(P_i^C, P_j^D\right) \in \psi$, given by

$$\begin{aligned} P_i^C &= \frac{G_{j,B}\left(2^{\lambda_0/\left(\sigma_i^C W\right)} - 1\right)}{G_{i,B}} P_j^D + \frac{W\delta_0\left(2^{\lambda_0/\left(\sigma_i^C W\right)} - 1\right)}{G_{i,B}} \\ &\triangleq k_0 P_j^D + b_0, \end{aligned} \tag{4.20}$$

where k_0 and b_0 are defined for expression brevity. Specifically, (4.20) inscribes the set of all $\left(P_i^C, P_j^D\right)$ points on a line segment through the feasible region ψ, and any point on this line segment leads to a fixed value λ_0 of the term $\sigma_i^C \overline{r_i^C}$.

Keeping this in mind, we then concentrate on the remaining terms of $\lambda_{i,j}$ by substituting (4.20) into (4.18), such that

$$\begin{aligned} &\sigma_j^D \overline{r_j^D} - \eta_{EE}\xi\left(P_i^C + P_j^D\right) \\ &= \sigma_j^D W \log_2\left(1 + \frac{G_j^D}{\left[\left(W\delta_0 + G_{i,j}b_0\right)/P_j^D\right] + G_{i,j}k_0}\right) - \eta_{EE}\xi\left[\left(k_0+1\right)P_j^D + b_0\right] \\ &\triangleq \widetilde{\lambda_1}\left(P_j^D\right), \end{aligned} \tag{4.21}$$

which is a form of one logarithmic function minus one linear function, and thus

$\widetilde{\lambda_1}\left(P_j^D\right)$ is an obviously convex or concave function only w.r.t. P_j^D. In other words, we must be able to find a maximum value of $\widetilde{\lambda_1}\left(P_j^D\right)$ over the point set given by the line segment (4.20).

Likewise, if we use an arbitrary feasible solution point in ψ but substitute it into the second term of $\sigma_j^D \overline{r_j^D}$ and then repeat the same process between (4.19) and (4.21), we can similarly obtain another convex/concave function, denoted as $\widetilde{\lambda_2}\left(P_i^C\right)$, to represent the remaining terms of $\sigma_i^C \overline{r_i^C} - \eta_{EE}\xi\left(P_i^C + P_j^D\right)$. Note that different from (4.20), in this process, another linear function w.r.t. P_i^C should be utilized to replace P_j^D, thereby $\widetilde{\lambda_2}\left(P_i^C\right)$ is a function only w.r.t. P_i^C.

In view of the above, the following proposition shows how $\widetilde{\lambda_1}\left(P_j^D\right)$ and $\widetilde{\lambda_2}\left(P_i^C\right)$ correlate to the optimality of $\mathbf{P2}_{i,j}$.

Proposition 3. *Given any ψ of $\mathbf{P2}_{i,j}$, let $\left(\overleftarrow{P_i^C}, \overleftarrow{P_j^D}\right)$ be the optimal point at any line segment w.r.t. $\widetilde{\lambda_1}\left(P_j^D\right)$ where $\overleftarrow{P_j^D} = \arg\max_{P_j^D \in \psi} \widetilde{\lambda_1}\left(P_j^D\right)$, and let $\left(\overrightarrow{P_i^C}, \overrightarrow{P_j^D}\right)$ be the optimal point at any line segment w.r.t. $\widetilde{\lambda_2}\left(P_i^C\right)$ where $\overrightarrow{P_i^C} = \arg\max_{P_i^C \in \psi} \widetilde{\lambda_2}\left(P_i^C\right)$. If denoting the optimal solution to $\mathbf{P2}_{i,j}$ as $\left(P_i^{C^*}, P_j^{D^*}\right)$, it must satisfy*

$$\left(P_i^{C^*}, P_j^{D^*}\right) \in \left\{\left(\overleftarrow{P_i^C}, \overleftarrow{P_j^D}\right) \mid \left(\overleftarrow{P_i^C}, \overleftarrow{P_j^D}\right) = \left(\overrightarrow{P_i^C}, \overrightarrow{P_j^D}\right) \in \psi\right\}. \tag{4.22}$$

Proof. Please see Appendix C. □

From Proposition 3, it is seen that $\mathbf{P2}_{i,j}$'s optimal solution $\left(P_i^{C^*}, P_j^{D^*}\right)$ must be the coincide point of two line segments in ψ for reaching the maxima of $\widetilde{\lambda_1}\left(P_j^D\right)$ and $\widetilde{\lambda_2}\left(P_i^C\right)$ at the same time. In line with this, we specially devise a heuristic searching algorithm to efficiently determine the optimal power allocation strategy for each possible CUE-DUE pair. In detail, our power allocation solution is illustrated as follows:

- *(1) Initial Feasible Solution Generation:* According to the feasible region ψ determined by all constraints of $\mathbf{P2}_{i,j}$, we need to first generate an initial feasible solution $\left(\widetilde{P_i^C}, \widetilde{P_j^D}\right)$ as the search starting point. For simplicity, $\left(\widetilde{P_i^C}, \widetilde{P_j^D}\right)$ can be set as one of the corners of ψ in Fig. 4.2.

- *(2) Optimal Line Point Searching for Maximum $\widetilde{\lambda_1}\left(P_j^D\right)$:* By executing the procedures in the context of (4.19)-(4.21), the close-form expression of $\widetilde{\lambda_1}\left(P_j^D\right)$ is easily obtained, where the value of P_j^D is constrained simultaneously by the line (4.20) and ψ. As a univariate convex optimization problem, its maximization can be efficiently realized via existing toolboxes such as CVXPY [102], thereby determining its optimal line point $\left(\overleftarrow{P_i^C}, \overleftarrow{P_j^D}\right)$.

- *(3) Optimal Line Point Searching for Maximum $\widetilde{\lambda_2}\left(P_i^C\right)$:* Based on the obtained $\left(\overleftarrow{P_i^C}, \overleftarrow{P_j^D}\right)$ from the previous line point searching, we substitute it into $\sigma_i^C \overline{r_i^C}$'s second term $\sigma_j^D \overline{r_j^D}$ to get the close-form expression of its corresponding line segment and $\widetilde{\lambda_2}\left(P_i^C\right)$. Similarly, its optimal line point $\left(\overrightarrow{P_i^C}, \overrightarrow{P_j^D}\right)$ can be obtained again using CVXPY.

- *(4) Searching Termination Check:* The above searching should be terminated once $\left(\overleftarrow{P_i^C}, \overleftarrow{P_j^D}\right) = \left(\overrightarrow{P_i^C}, \overrightarrow{P_j^D}\right)$, i.e., the optimal point of the previous line segment is also optimal for the current searching line segment. However, according to Proposition 3, such a coincide point may fall into a local optimum of $\mathbf{P2}_{i,j}$, hence we add it into an optimality list, denoted by $\mathcal{I}$, for record.

- *(5) Multiple Rounds of Searches:* To prevent the algorithm trapping into the local optimum, we set *step (1)-(4)* as one round of search and then repeat multiple rounds until reaching a preset maximum search round restriction, where $\left(\widetilde{P_i^C}, \widetilde{P_j^D}\right)$ is always changing. Consequently, the optimal power allocation strategy $\left(P_i^{C*}, P_j^{D*}\right)$ is finalized by iterating through all the points in $\mathcal{I}$.

4.3.3 Optimal Power Allocation for a single CUE without Spectrum Sharing

It is worth pointing out that due to the number of DUEs is less than the number of CUEs (i.e., $N \leqslant M$), there must be some CUEs' sub-channels that are not reused by any DUE. Accordingly, determining the optimal power allocation solution for each CUE i without spectrum sharing becomes necessary as well, and our second-stage problem $\mathbf{P3}_i$ ($\forall i \in \mathcal{M}$) is

$$\mathbf{P3}_i: \max_{P_i^C} \quad \check{\lambda}_i \tag{4.23}$$

$$\text{s.t.} \quad \tau_i^C \Re_i^C\left(\check{r}_i{}^C\right) \geqslant q_{min}^C, \tag{4.23a}$$

$$(4.14c), \tag{4.23b}$$

where $\check{r}_i{}^C = W \log_2\left(1 + \frac{P_i^C G_{i,B}}{W \delta_0}\right)$ and

$$\check{\lambda}_i = \tau_i^C \Re_i^C\left(\check{r}_i{}^C\right) - \frac{\eta_{EE} \check{r}_i{}^C}{L/\left[\tau_i^C P^{mat} + (1-\tau_i^C) P^{mis}\right]} - \eta_{EE} \xi P_i^C. \tag{4.24}$$

It is seen that $\mathbf{P3}_i$ is apparently a convex optimization problem as its objective function $\check{\lambda}_i$ is a linear combination of one logarithmic function and one linear function, the optimality of which can be easily obtained via CVXPY.

4.3.4 Spectrum Reusing Policy Optimization for EE-SCN

Given any η_{EE} in each iteration w.r.t. **P1**, our third-stage method is to leverage the obtained optimal power allocation strategies of $\mathbf{P2}_{i,j}$ and $\mathbf{P3}_i$ to finalize the spectrum reusing policy for all DUEs in the EE-SCN. Let $\lambda_{i,j}^*$ denote the maximum $\lambda_{i,j}$ at each possible spectrum reusing pair of CUE i and DUE j by solving each $\mathbf{P2}_{i,j}$ and $\check{\lambda}_i^*$ denote the maximum $\check{\lambda}_i$ at the single CUE i by solving each $\mathbf{P3}_i$. As such, the spectrum reusing problem is actually a variant problem about the weighted bipartite matching optimization, i.e.,

$$\mathbf{P4}: \max_{\boldsymbol{\alpha}} \quad \sum_{i\in\mathcal{M}}\sum_{j\in\mathcal{N}} \alpha_{i,j}\lambda_{i,j}^* + \sum_{i\in\mathcal{M}} \check{\lambda}_i^* \left(1 - \sum_{j\in\mathcal{N}} \alpha_{i,j}\right) \tag{4.25}$$

$$\text{s.t.} \quad (4.11\text{e}) - (4.11\text{g}). \tag{4.25a}$$

Combining the constraint (4.25a) with $N \leqslant M$, it is observed that **P4** is able to be decomposed into $\binom{M}{N}$ subproblems. Specifically, suppose that N out of M CUEs are arbitrarily selected to share their subchannels with N DUEs, where we have a total of $\binom{M}{N}$ different selection combinations. Hungarian method Clearly, each of them is a pure N-to-N bipartite matching problem, which can be efficiently solved in polynomial time by applying the Hungarian method [116]. As for the remaining $(M-N)$ CUEs without spectrum reusing, their respective optimality w.r.t. $\check{\lambda}_i^*$ is already known after confirming each combination. Therefore, the optimal spectrum reusing policy can be finalized by comparing the optimality among $\binom{M}{N}$ problems.

Regarding the computational complexity of the proposed solution, it is first seen that in each search round for solving each $\mathbf{P2}_{i,j}$, it takes several iterations (Lines 11-16) to determine one viable solution in $\mathcal{I}$, and in each iteration, the brute-force search needs to be executed once to obtain $\left(\overleftarrow{P_i^C}, \overleftarrow{P_j^D}\right)$ or $\left(\overrightarrow{P_i^C}, \overrightarrow{P_j^D}\right)$. Hence, if denoting its maximum number of iterations as H and the maximum number of staircase w.r.t. each brute-force search as $\tilde{H}$, then solving each $\mathbf{P2}_{i,j}$ would require complexity of $\mathcal{O}(\tilde{Q}H\tilde{H})$. Likewise, solving each $\mathbf{P3}_i$ only needs complexity of $\mathcal{O}(\tilde{H})$. Moreover, since the N-to-N bipartite matching problem w.r.t. **P4** can be solved by the Hungarian method with complexity of $\mathcal{O}(N^3)$, the complexity for solving **P4** becomes $\mathcal{O}(\binom{M}{N}N^3)$. Accordingly, the proposed Algorithm 1 has a polynomial-time overall complexity, given as $\mathcal{O}(QMN\tilde{Q}H\tilde{H}+$

Table 4.1: Simulation Parameters

Parameters	**Values**
System bandwidth	10 MHz
Number of CUEs (M)	50
Number of DUEs (N)	30
Maximum transmit power of each CUE (P_{max}^{C})	21 dBm [117]
Maximum transmit power of each DUE (P_{max}^{D})	21 dBm
Noise power spectral density (δ_0)	-174 dBm/Hz
Path loss model for cellular links	$128.1 + 37.6 \log_{10}(d \text{ [km]})$ dB
Path loss model for D2D links	$148 + 40 \log_{10}(d \text{ [km]})$ dB [118]
Semantic data packet size (L)	800 bits [119]

$Q\binom{M}{N}N^3$).

4.4 Numerical Results and Discussions

In this section, numerical evaluations are conducted to demonstrate the performance of our proposed power allocation and spectrum reusing solutions in the EE-SCN, where we employ Python 3.7-based PyCharm as the simulator platform and implement it in a workstation PC featuring the AMD Ryzen-9-7900X processor with 12 CPU cores and 128 GB RAM. In the basic system setup, we first model a single-cell circular area with a radius of 300 meters, in which multiple CUEs and DUEs are randomly dropped, and the distance between the transceiver of each DUE is randomly generated between 5 and 45 meters [120]. For the energy efficiency model, the class power amplifier is first preset for each user device with a fixed power efficiency of 35 percent [119], i.e., $\xi = 1/0.35 = 2.8571$. Moreover, following the research in [121], we assume that the circuit power consumptions for processing a knowledge-matching semantic packet P^{mat} and for a knowledge-mismatching semantic packet P^{mis} are 0.1 and 0.5 mW, respectively. For brevity, some other simulation parameters not mentioned in the context and their values are summarized in Table 6.1.

In SemCom-relevant settings, we simulate a general text transmission scenario to examine the proposed solution. Note that such performance test can also be accomplished with other content types like images or videos, and the reason we choose text is to leverage existing natural language processing models that have been well validated in SemCom-related works. Particularly, the Transformer in [1] is adopted as the unified semantic encoder for all SemCom links, and the PyTorch-based Adam optimizer is applied for model training with an initial learning rate of 0.001. Based on the public dataset extracted from the

proceedings of European Parliament [108], the expression of B2M function $\Re_i(\cdot)$ and $\Re_j(\cdot)$ at each SemCom link can be well approximated from extensive model tests [21]. Besides, the knowledge-matching degree τ_i^C for each CUE i and τ_j^D for each DUE j are randomly generated in the range of $0.4 \sim 1$, respectively. For the solution-related settings, the maximum number of iterations for updating η_{EE} w.r.t. **P1** is set as 100, and its convergence threshold ϵ is 0.001. In addition, the same minimum semantic throughput requirement is assumed for all CUEs and DUEs with $q_{min}^C = q_{min}^D = 30$ msg/s. It is worth mentioning that all the above parameter values are set by default unless otherwise specified, and all subsequent numerical results are obtained by averaging over a sufficiently large number of trials.

For comparison purposes, here we employ two resource allocation benchmarks in EE-SCNs: (I) Maximum power allocation plus random spectrum reusing [122], which means that each user is allocated with its maximum allowable transmit power while each DUE randomly reuses the subchannel of one CUE; (II) Random power allocation [123] plus distance-based spectrum reusing [124], where each user is allocated with the randomized transmit power while each DUE reuses the subchannel of the CUE furthest away from itself to reduce the interference impact as much as possible.

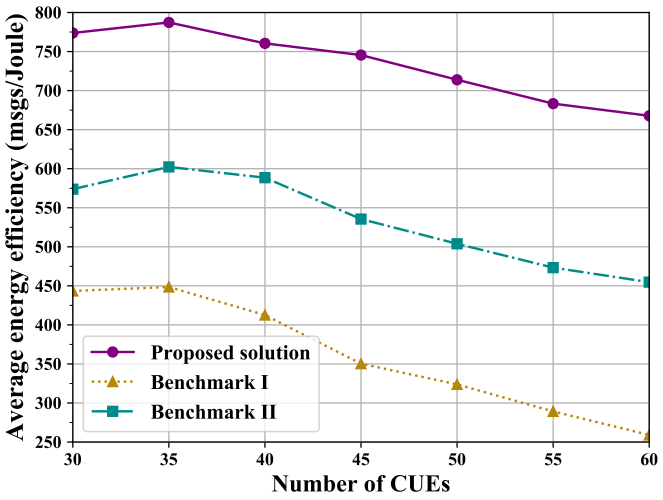

Figure 4.3: Average energy efficiency versus different numbers of CUEs.

Fig. 4.3 first shows the objective performance metric of η_{EE} obtained under different numbers of CUEs (i.e., M) between 30 and 60. Compared with the two benchmarks, it is seen that our proposed solution always guarantees a significant performance gain with the changes of M. For instance, when the number of CUEs is 35, an average energy efficiency of 786 msgs/Joule is observed by the proposed solution, which increases 47.7% performance compared to Benchmark I and 74.9%

compared to Benchmark II. Moreover, as the number of CUEs increases, the energy efficiency of the proposed solution rises at the beginning from 30 to 35, and then drops gradually. This is because before the point of 35, the number of DUEs is quite close to the number of CUEs, in which case each DUE should have more options to choose a better CEU for spectrum reusing as the number of CUEs grows and thus resulting a better energy efficiency. As the number of CUEs keeps increasing, due to the limited overall system bandwidth, the bandwidth allocated to each CUE is also fewer while the power consumption becomes greater, which obviously leads to a downward trend on the rendered x_{--} performance.

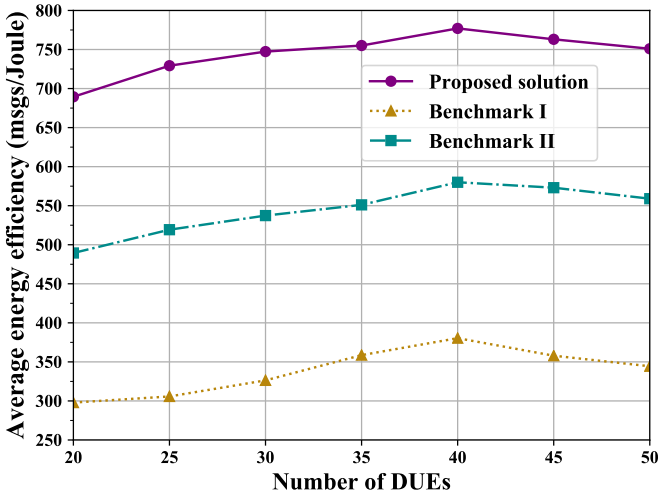

Figure 4.4: Average energy efficiency versus different numbers of DUEs.

Next, we compare the proposed solution with benchmarks under different number of DUEs (i.e., N) between 20 and 50. It is observed that the energy efficiency obtained by our solution still exceeds that of benchmarks at each point with a performance gain similar to the results in Fig. 4.3. However, the energy efficiency becomes higher with the number of DUEs at the beginning, and drops a little bit after reaching 40. The former phenomenon is because the performance gain on message throughput resulted from the increase of DUEs surpasses the impact of the power consumption increase. When the number of DUEs surpasses a maximum threshold, such performance increase eventually reaches the peak and is saturated and even worse, as significant interferences to CUEs' links will dominate the change in energy efficiency.

Apart from these, the impacts of the maximum allowable transmit power of CUEs ($i.e., P^C_{max}$) and DUEs (i.e., P^D_{max}) are tested in Fig. 4.5 and Fig. 4.6, respectively. In both figures, our proposed solution presents a better energy efficiency performance compared with the two benchmarks, which results are consistent with the previous two figures. Nevertheless, as the maximum transmit

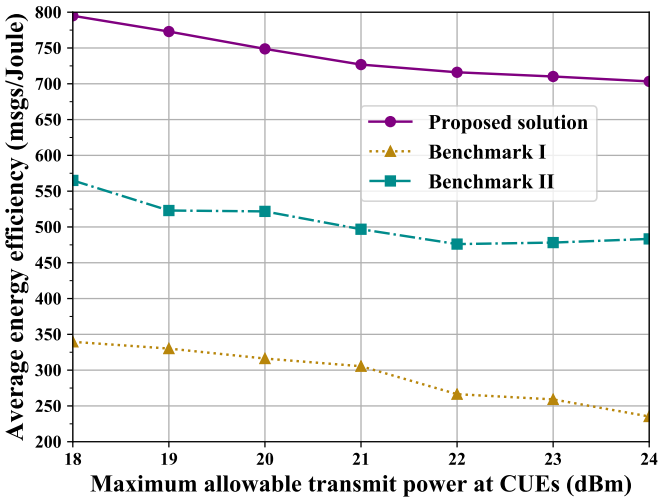

Figure 4.5: Average energy efficiency versus varying maximum transmit powers of CUEs.

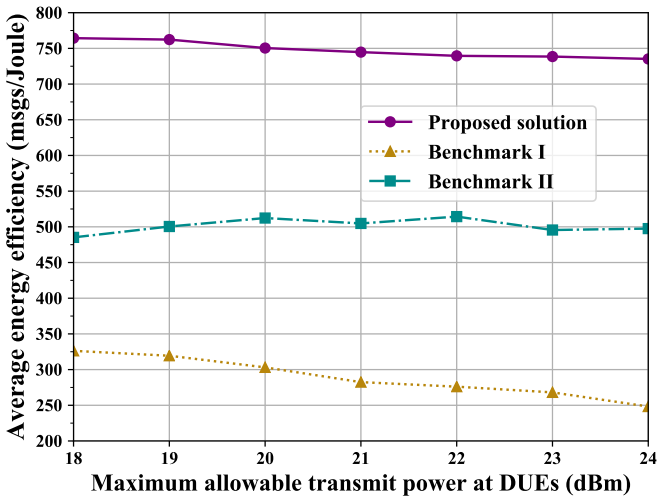

Figure 4.6: Average energy efficiency versus varying maximum transmit powers of DUEs.

power of either CUEs or DUEs rises, a downward trend is seen in both cases. In the first case, such a trend can be interpreted as that each CUE prefers to have a higher transmit power at the looser power constraint to ensure a better message throughput for these CUEs without spectrum reusing, with this comes more energy consumption at the power amplifier and semantic packet processing. Meanwhile, Fig. 4.6 depicts a slower downward trend compared with Fig. 4.5 in the performance presentation of our solution. This is because the interference resulted from the maximum allowable transmit power of DUEs can completely cover the power gain for the received signals of CUEs, thereby the changes in the optimal transmit power of each DUE should be small.

4.5 Conclusions

In this chapter, we jointly addressed the power allocation and spectrum reusing problems in the EE-SCNs, where two different groups of D2D SemCom users and cellular SemCom users coexist. First, the B2M transformation function was introduced to identify the message throughput obtained by each CUE and each DUE. Next, considering the knowledge-matching mechanism, the energy consumption model of SemCom is identified at a semantic data packet level. On this basis, a joint optimization problem was formulated with the aim of maximizing the average energy efficiency, followed by a corresponding optimal solution proposed. Among them, the primal fractional-form problem was first transformed into a subtractive-form one via Dinkelbach's method, and then we developed a heuristic algorithm and employed a Hungarian method to seek the optimal power allocation and spectrum reusing solutions, respectively. Numerical results show the performance superiority of our proposed solution in terms of energy efficiency compared with two different benchmarks. The next chapter will explore the resource allocation problem in hybrid semantic/bit communication scenario.

Chapter 5

Wireless Resource Optimization in Hybrid Semantic/Bit Communication Networks

5.1 Introduction

As aforementioned, recent advances in SemCom have attracted widespread attention, promising to significantly alleviate the scarcity of wireless resources in next-generation cellular networks. By leveraging cutting-edge DL algorithms, SemCom is capable of providing MUs with a variety of high-quality, large-capacity, and multimodal services, including typical multimedia content (e.g., text, image, and video streaming) and AIGC [14].

Nevertheless, there is still a missing investigation for a more practical yet novel next-generation cellular network paradigm, namely *hybrid semantic/bit communication networks* (HSB-Nets), where the two modes of SemCom and BitCom coexist. Note that SemCom typically requires more data processing time but produces higher semantic performance than BitCom at each transceiver, thereby determining an appropriate mode selection (MS) scheme for each MU becomes quite tricky. Most uniquely, the varying degrees of background knowledge matching among MUs can also affect the amount of allocated bandwidth in combination with different channel conditions. As such, if aiming at high semantic fidelity and low latency for a large-scale HSB-Net, we are encountering the following three fundamental challenges in resource management:

- *Challenge 1: How to unify performance metrics for both SemCom and BitCom in the HSB-Net?* Given the core mechanism of meaning delivery in SemCom, traditional metrics in BitCom, like bit rate or bit throughput, are

evidently no longer applicable to the SemCom links. Especially in such a hybrid scenario, it becomes necessary to align SemCom and BitCom to the same assessment basis to facilitate subsequent performance comparisons or overall network optimization, which raises the first nontrivial point.

- *Challenge 2: How to mathematically characterize the unique semantic-coding process in SemCom when combined with bit transmission?* Note that SemCom involves an extra semantic-coding process compared with BitCom before the bit data transmission at each link, which can be characterized from a packet-queuing perspective. In the semantic-coding process, due to diverse knowledge-matching degrees among different SemCom-enabled MUs, semantic data packet interpretation rates can vary [125], thereby resulting in distinct queuing delay and reliability performance. Combined with the subsequent indispensable packet-transmission queuing process, all of these constitute the second difficulty.

- *Challenge 3: How to determine the best communication mode for each MU with the joint consideration of UA and BA to optimize overall network performance?* Generally, each MU can select only one of the SemCom and BitCom modes at a time during the UA process, subject to its current knowledge-matching degree, channel condition, desired service quality, as well as latency and reliability budgets. Such a new MS problem, coupled with inherent practical constraints such as limited bandwidth resources and the single-base BS association requirement, poses the third challenge, i.e., seeking an optimal resource management strategy for the UA, MS, and BA to jointly optimize overall network performance in the HSB-Net.

In response to the challenges outlined above, in this chapter, we systematically investigate the UA, MS, and BA problems in the uplink of the HSB-Net and correspondingly propose an optimal strategy with the awareness of unique SemCom characteristics. Simulation results not only demonstrate the accuracy of our theoretical analysis for semantic data packet queuing, but also showcase the performance superiority of the proposed resource management solution in terms of realized message throughput compared with four benchmarks. Accordingly, our main contributions are summarized as follows:

- We unify the performance metrics for both SemCom and BitCom links by introducing the bit-rate-to-message-rate transformation mechanism to measure their respective achievable message throughputs. In this regard, the stochasticity of knowledge matching degree and channel state are particularly taking into account over different time slots. Correspondingly, we then

formulate an optimization problem to maximize the time-averaged overall message throughput of the HSB-Net by jointly correlating the UA, MS, and BA-related indicators. These first address the aforementioned *Challenge 1*.

- We specially model a two-stage tandem queue for each SemCom-enabled MU to capture the entire queuing process of its locally generated semantic packets, which fully incorporates the semantic coding and knowledge-matching characteristics with the traditional packet transmission. On this basis, the steady-state average packet loss ratio and queuing delay in both SemCom and BitCom cases are then mathematically derived to post the reliability and latency requirements in subsequent optimization. The contribution directly addresses *Challenge 2*.

- We theoretically prove the monotonicity of allocated bandwidth with respect to reliability and latency, and then develop an efficient resource management strategy to jointly solve the UA, MS, and BA problems with polynomial-time complexity. Specifically, the minimum bandwidth threshold is first fixed for each SemCom and BitCom link, following by a Lagrange primal-dual method and a preference list-based heuristic algorithm to finalize the UA and MS solutions. Afterward, the optimal BA strategy is further obtained by reallocating the remaining bandwidth of each BS to all its associated MUs. In this way, *Challenge 3* is finally well tackled.

5.2 System Model

5.2.1 HSB-Net Scenario

Consider an HSB-Net scenario as depicted in Fig. 5.1, the total of U MUs are distributed within the coverage of S BSs, where two communication modes of SemCom and BitCom are available for all MUs, while each MU can only select one mode and be associated with one BS at a time. Herein, let $x_{ij} \in \{0, 1\}$ denote the binary UA indicator, where $x_{ij} = 1$ means that MU $i \in \mathcal{U} = \{1, 2, \cdots, U\}$ is associated with BS $j \in \mathcal{J} = \{1, 2, \cdots, J\}$, and $x_{ij} = 0$ otherwise. Besides, we specially define the binary MS indicator as $y_{ij} \in \{0, 1\}$, where $y_{ij} = 1$ represents that the SemCom mode is selected for the link between MU i and BS j, and $y_{ij} = 0$ indicates that the BitCom mode is selected.[1] Meanwhile, the amount of bandwidth resource that BS j assigns to MU i is denoted as z_{ij}, while the total

[1]It is worth pointing out that y_{ij} is applicable to be an effective MS indicator only when $x_{ij} = 1, \forall (i, j) \in \mathcal{U} \times \mathcal{J}$.

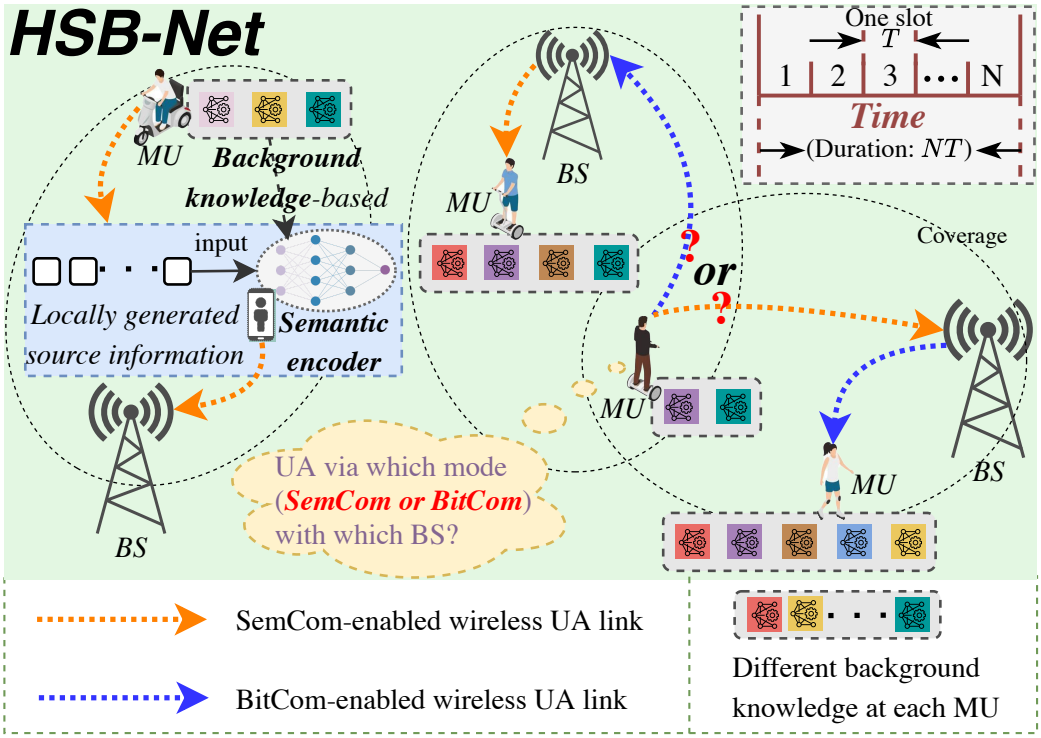

Figure 5.1: The HSB-Net scenario involving UA, MS, and BA in one time block.

bandwidth budget of BS j is denoted as Z_j. Moreover, time is equally partitioned into N consecutive time slots, each with the same duration length T.

5.2.2 Network Performance Metric

For the wireless propagation model, let $\gamma_{ij}(t)$ denote the SINR of the link between MU i and BS j at time slot t, $t = 1, 2, \cdots, N$. Note that $\gamma_{ij}(t)$ is assumed to be an independent and identically distributed (i.i.d.) random variable for different slots but remain constant during one slot [122, 126]. Since the conveyed message itself becomes the sole focus of precise reception in SemCom rather than traditional transmitted bits in BitCom, we proceed with the performance metric developed in our previous work [21] to measure the message rate for each SemCom-enabled MU via employing the B2M transformation function. To be specific, the B2M function is to output the semantic channel capacity (i.e., the achievable message rate in units of messages per unit time, *msg/s*) from input traditional Shannon channel capacity (i.e., the achievable bit rate in units of bits per unit time, *bit/s*) under the discrete memoryless channel.[2] If in an ideal condition, i.e., the information source and destination have identical semantic

[2]The semantic channel capacity with respect to B2M is derived from a semantic information-theoretical perspective, which is beyond the scope of this work and thus will not be discussed in-depth. More technical details can be found in [21] and [7].

reasoning capability and equivalent background knowledge, the B2M function can be approximated as linear. However, the B2M can also involve stochastic variables in the case of knowledge mismatch, resulting in the presentation of random message rates. Given this, let $\Re_{ij}(\cdot)$ denote the B2M function of the SemCom link between MU i and BS j, its instantaneous achievable message rate in time slot t should be

$$M_{ij}^{S}(t) = \beta_i(t)\Re_{ij}(z_{ij}\log_2\left(1+\gamma_{ij}\left(t\right)\right)). \tag{5.1}$$

Here, $\beta_i(t)$ represents the knowledge-matching degree between MU i and its communication counterpart at slot t, which is an i.i.d. random Gaussian variable ranging from 0 to 1 [21], having mean τ_i. To provide more details here, each message is first assumed to be associated with a specific SemCom service type based on Footnote **??**. Then, compared with the perfectly knowledge-matching case, only the messages related to the overlapped services can be effectively encoded/decoded in the knowledge-mismatching state in each slot, and $\beta_i(t)$ is the overlap proportion. Combined with the fact that the generation of source messages is generally a stochastic process [13], therefore, $\beta_i(t)$ is deemed as a random variable. In addition, other factors like the channel encoding scheme that may affect the message-rate measurement are assumed to be identical between different SemCom-enabled MUs for simplicity.c process [13], therefore, $\beta_i(t)$ is deemed as a random variable.

Likewise, for the BitCom link between MU i and BS j, considering it has an average B2M transformation ratio,[3] denoted by ρ_{ij}, to align with the semantic performance measurement of SemCom. In other words, we assume that each message in BitCom can be encoded into bits of fixed length on average [**?**], i.e., the reciprocal of ρ_{ij}, and thus its instantaneous achievable message rate in slot t is given by

$$M_{ij}^{B}(t) = \rho_{ij}z_{ij}\log_2\left(1+\gamma_{ij}\left(t\right)\right),\ 0<\rho_{ij}<1. \tag{5.2}$$

As such, if taking into account both SemCom (i.e., $y_{ij}=1$) and BitCom (i.e., $y_{ij}=0$) cases, we obtain the time-averaged message rate of each link as

$$M_{ij} = \frac{1}{N}\sum_{t=1}^{N}\left[y_{ij}M_{ij}^{S}(t) + (1-y_{ij})M_{ij}^{B}(t)\right]. \tag{5.3}$$

[3]This assumption is justified since the source-and-channel coding of BitCom for source information typically follows prescribed codebooks, and the variable length coding is adopted [4, 127]. Hence, based on each link's known channel state information, the proportion of messages that can be effectively decoded from a certain amount of transmitted bits can be averaged.

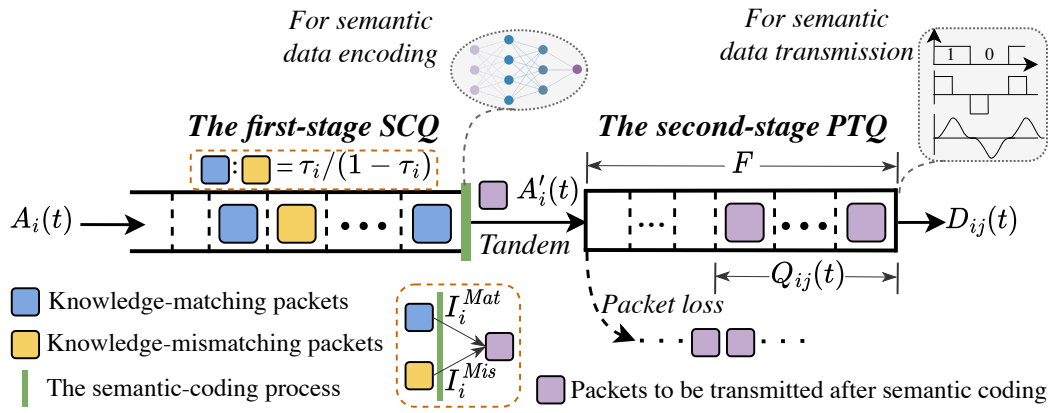

Figure 5.2: The two-stage tandem queue model at each SemCom-enabled MU.

5.2.3 Queuing Model

In this work, we focus on the differences in queuing models between SemCom and BitCom during data uplink transmission, where the queuing delay is employed as the latency metric to characterize the average sojourn time of a data packet in the queue buffer at each MU in the HSB-Net. Unlike existing studies that consider only packet queuing during the channel-coding process in BitCom [36, 128–130], the queuing delay at a SemCom-enabled MU should take into account the newly introduced semantic-coding process in combination with traditional packet transmission, as shown in Fig. 6.2. To this end, we first provide the following definition for clarity.

Definition 4. *A SemCom-enabled MU has a two-stage tandem queue,*[4] *named Semantic-Coding Queue (SCQ) and Packet-Transmission Queue (PTQ). As for a BitCom-enabled MU, only one PTQ is considered for its packet uplink transmission.*

To preserve the generality, the SCQ is assumed with infinite-size memory to handle all locally generated SemCom services, while the PTQ has a finite-size buffer that can accommodate up to F data packets to align with practical resource limitations and scheduling.[5] Moreover, the packets in both SCQ and PTQ are queued in a first-come-first-serve manner.

Based on the above, if a Poisson arrival process with average rate λ_i (in *packets/s*) of initial data packet generation is assumed for each MU i ($\forall i \in \mathcal{U}$), the number of arrival packets during slot t, denoted as $A_i(t)$, has the probability

[4] A two-stage tandem queue implies that the output of the first queue becomes the input of the second, and the packet processing in the two queues is independent of each other [36, 131].

[5] Note that the SCQ can also be modeled with a finite-size buffer whose queuing latency is derived similarly to that of the PTQ. Likewise, the above rationale applies to the PTQ as well.

mass function (PMF) as follows:

$$\Pr\left\{A_i\left(t\right)=k\right\}=\frac{\left(\lambda_i T\right)^k}{k!}\exp\left(-\lambda_i T\right),\ k=0,1,2,\cdots. \tag{5.4}$$

For the PTQ in both SemCom and BitCom cases, its packet departure rate depends on the number of packets sent out from MU i to BS j ($\forall j\in\mathcal{J}$) during slot t, denoted as $D_{ij}\left(t\right)$, which has the PMF as

$$\begin{aligned}\Pr\left\{D_{ij}\left(t\right)=k\right\}&=\Pr\left\{\left\lfloor\frac{Tz_{ij}\log_2\left(1+\gamma_{ij}\left(t\right)\right)}{L}\right\rfloor=k\right\}\\&=\Pr\left\{\gamma_{ij}\left(t\right)\leqslant 2^{\frac{(k+1)L}{Tz_{ij}}}-1\right\}-\Pr\left\{\gamma_{ij}\left(t\right)\leqslant 2^{\frac{kL}{Tz_{ij}}}-1\right\}.\end{aligned} \tag{5.5}$$

Here, $\lfloor\cdot\rfloor$ is the floor function that outputs the largest integer less than or equal to the input value, and all packets have the same size of L bits, $k=0,1,2,\cdots$. Clearly, given any reasonable probability distribution approximation of the SINR $\gamma_{ij}\left(t\right)$ (e.g., Gaussian distribution [126] or generalized Gamma distribution [132]), applying its cumulative distribution function (CDF) directly yields the close-form expression of (5.5). Besides, it is noteworthy that the obtained PMF of $D_{ij}\left(t\right)$ should be independent of time slot index t, as the randomness of each physical link's SINR is generally t-independent [122].

Next, we model the packet departure of SemCom-enabled SCQ and the packet arrival of SemCom-enabled PTQ, respectively. As mentioned earlier, each data packet generated at a SemCom-enabled sender MU requires a certain type of background knowledge, resulting in either a knowledge-matching or knowledge-mismatching state with its receiver. For illustration, let I_i^{Mat} denote the semantic-coding time required by a knowledge-matching packet with mean $1/\mu_i^{Mat}$ (in *s/packet*), and let I_i^{Mis} denote the semantic-coding time required by a knowledge-mismatching packet with mean $1/\mu_i^{Mis}$ ($\mu_i^{Mat}>\mu_i^{Mis}$ in practice[6]). Without loss of generality, I_i^{Mat} and I_i^{Mis} are assumed to be two exponential random variables independent of each other, which are determined by the specific semantic computing capability available at the MU i's terminal device. Having these, it is seen that the overall service time distribution of packets at SCQ should be treated as a general distribution [125]. Let us denote the average packet queuing latency of SCQ at each SemCom-enabled MU i by $\delta_i^{S_I}$, which will be analyzed in detail in the next section.

[6]Notice that the content in knowledge-mismatching packets necessarily requires more computational resources and processing time for accurate contextual reasoning and interpretation, due to the use of more sophisticated semantic-coding networks or the knowledge-sharing method, etc [8].

As for the number of packets arriving at the SemCom-enabled PTQ in slot t, denoted by $A'_i(t)$, it should exactly be the number of packets leaving its tandem SCQ in the same slot, according to the two-stage tandem structure in Definition 4. Meanwhile, it is not difficult to deduce that the knowledge-matching packets leaving the SCQ follow a Poisson process with mean μ_i^{Mat}, while the knowledge-mismatching packets leave as a Poisson process with mean μ_i^{Mis}. The former event occurs with probability τ_i and the latter happens with probability $(1-\tau_i)$. As such, $A'_i(t)$ should still satisfy the Poisson distribution with a PMF of

$$\Pr\left\{A'_i(t)=k\right\}=\frac{\left(\lambda'_i T\right)^k}{k!}\exp\left(-\lambda'_i T\right),\ k=0,1,2,\cdots, \tag{5.6}$$

where λ'_i is the average arrival rate (in *packets/s*), given as

$$\lambda'_i=\tau_i\mu_i^{Mat}+(1-\tau_i)\mu_i^{Mis}. \tag{5.7}$$

Considering the limited buffer size F of PTQ, we further assume that in any t, the packets to be transmitted leave the queue first and then the arriving packets enter it. Hence, the evolution of its queue length between two consecutive slots is

$$Q_{ij}\left(t+1\right)\triangleq\min\left\{\max\left\{Q_{ij}\left(t\right)-D_{ij}\left(t\right),0\right\}+A'_i(t),F\right\}, \tag{5.8}$$

where $Q_{ij}\left(t\right)$ denotes the queue length of PTQ for the link between MU i and BS j at slot t, $t=1,2,\cdots,N-1$.

Note that those packets arriving at a fully loaded PTQ in each slot will be blocked and dropped, which can affect achievable communication reliability and message rate performance. Accordingly, let θ_{ij}^S and θ_{ij}^B denote the average packet loss ratio of SemCom-enabled PTQ and BitCom-enabled PTQ, respectively, and each represents the proportion of packets failed to be delivered to all arriving packets. Likewise, let $\delta_{ij}^{S_2}$ and δ_{ij}^B denote the average packet queuing latency of SemCom-enabled PTQ and BitCom-enabled PTQ, respectively. Combined with the $\delta_i^{S_1}$ defined before, we obtain the overall average queuing latency of the link between SemCom-enabled MU i and BS j as $\delta_{ij}^S=\delta_i^{S_1}+\delta_{ij}^{S_2}$.

When considering both SemCom (i.e., $y_{ij}=1$) and BitCom (i.e., $y_{ij}=0$), the average queuing latency experienced by the link between any MU i and BS j should be

$$\delta_{ij}=y_{ij}\delta_{ij}^S+(1-y_{ij})\delta_{ij}^B. \tag{5.9}$$

Similarly, the average packet loss ratio that indicates the communication reliabil-

ity of the link is found by

$$\theta_{ij} = y_{ij}\theta_{ij}^{S} + (1 - y_{ij})\theta_{ij}^{B}. \tag{5.10}$$

In the subsequent section, we elaborate the derivations for the mathematical expressions of $\delta_{ij}^{S_1}$, $\delta_{ij}^{S_2}$, and θ_{ij}^{S}. Recalling the BitCom-enabled PTQ model and the SemCom-enabled PTQ model, it is seen that their sole distinction lies in their packet arrival processes, in which the former follows (5.4) and the latter follows (5.6). Therefore, δ_{ij}^{B} and θ_{ij}^{B} can be easily derived using the similar procedure as for $\delta_{ij}^{S_2}$ and θ_{ij}^{S}.

5.3 Queuing Analysis and Problem Formulation

5.3.1 Queuing Analysis for SCQ and PTQ

First for the SemCom-enabled SCQ, it should be noted that the average proportion of knowledge-matching packets to the total number of packets in the queue is exactly equal to the average knowledge-matching degree τ_i between MU i and its receiver.[7] Combined with the general distribution conclusion obtained earlier, the average semantic-coding time of a packet in the SCQ, denoted by I_i, becomes $I_i = \tau_i I_i^{Mat} + (1 - \tau_i) I_i^{Mis}$. Since I_i^{Mat} and I_i^{Mis} are independent of each other, we have its expectation as $\mathbb{E}[I_i] = \tau_i/\mu_i^{Mat} + (1 - \tau_i)/\mu_i^{Mis}$ and its variance as $\mathbb{V}(I_i) = \left(\tau_i/\mu_i^{Mat}\right)^2 + \left((1 - \tau_i)/\mu_i^{Mis}\right)^2$. Owing to the Markovian packet arrival and general-distribution packet departure, the SCQ follows an M/G/1 system, which has been widely used to capture data traffic in wireless networks [125]. In this case, we can directly apply the *Pollaczek-Khintchine formula* [133] to calculate the steady-state average packet queuing latency of SCQ $\delta_i^{S_1}$ by[8]

$$\begin{aligned}\delta_i^{S_1} &= \frac{\lambda_i\left(\mathbb{E}^2[I_i] + \mathbb{V}(I_i)\right)}{2\left(1 - \lambda_i\mathbb{E}[I_i]\right)} + \mathbb{E}[I_i] \\ &= \frac{\lambda_i\left[\tau_i(1-\tau_i)/\mu_i^{Mat}\mu_i^{Mis} + \left(\tau_i/\mu_i^{Mat}\right)^2 + \left((1-\tau_i)/\mu_i^{Mis}\right)^2\right]}{1 - \lambda_i\tau_i/\mu_i^{Mat} - \lambda_i(1-\tau_i)/\mu_i^{Mis}} + \frac{\tau_i}{\mu_i^{Mat}} + \frac{1-\tau_i}{\mu_i^{Mis}}.\end{aligned} \tag{5.11}$$

It is worth further noting that either I_i^{Mat} or I_i^{Mis} in (5.11) is independent of time slot index t, thus $\delta_i^{S_1}$ should be deemed a constant.

[7] This observation holds true when examined on a large timescale, and it assumes that each packet has the same probability of being generated locally.

[8] Applying the Pollaczek-Khintchine formula implies a prerequisite that $\lambda_i\mathbb{E}[I_i] < 1$ must be satisfied to guarantee a steady-state M/G/1 system [134]. Therefore, we consider that in the SCQ, the packet departure rate exceeds the packet arrival rate to make its queuing latency finite and solvable.

As for the SemCom-enabled PTQ, we at first introduce the following proposition to characterize its steady-state queue length $Q_{ij}(t) = 0, 1, 2, \cdots, k, \cdots, F$ in slot t.

Proposition 4. *For each $Q_{ij}(t)$ of PTQ, it must have a solvable and unique steady-state probability vector, denoted as $\boldsymbol{\alpha}_{ij} = \left[\alpha_{ij}^0, \alpha_{ij}^1, \cdots, \alpha_{ij}^F\right]^T$, where α_{ij}^k represents the steady-state probability of $Q_{ij}(t) = k$ when t tends to infinity.*

Proof. Please see Appendix D. □

From Proposition 4, the long-term average queue length of $Q_{ij}(t)$ can be obtained by computing its expectation, i.e., $\mathbb{E}\left[Q_{ij}(t)\right] = \sum_{k=0}^{F} k\alpha_{ij}^k$. Moreover, by combining $\boldsymbol{\alpha}_{ij}$ with the PMFs of PTQ's packet arrival as in (5.6) and packet departure as in (5.5), the average number of packets dropped at the steady-state PTQ during any slot t, denoted by G_{ij}, can be calculated by

$$\begin{aligned} G_{ij} = \sum_{l=1}^{F} \alpha_{ij}^l \Bigg[\sum_{k=0}^{l-1} \Pr\left\{D_{ij}=k\right\} \Bigg(\sum_{f=F-l+k}^{\infty} (f+l-k-F) \Pr\{A_i'=f\} \Bigg) \\ + \Bigg(\sum_{k=l}^{\infty} \Pr\left\{D_{ij}=k\right\} \Bigg) \Bigg(\sum_{f=F+1}^{\infty} (f-F) \Pr\{A_i'=f\} \Bigg) \Bigg] \\ + \alpha_{ij}^0 \sum_{k=F+1}^{\infty} (k-F) \Pr\left\{A_i'=k\right\}. \end{aligned} \tag{5.12}$$

As its average total packet arrival rate is λ_i', we have the steady-state average packet loss ratio of SemCom-enabled PTQ as follows:

$$\theta_{ij}^S = \frac{G_{ij}}{\lambda_i' T} = \frac{G_{ij}}{\tau_i \mu_i^{Mat} T + \mu_i^{Mis} T - \tau_i \mu_i^{Mis} T}. \tag{5.13}$$

Hence, the average effective packet arrival rate becomes $\lambda_i^{eff} = \left(1 - \theta_{ij}^S\right) \lambda_i' = \tau_i \mu_i^{Mat} + (1-\tau_i)\mu_i^{Mis} - G_{ij}/T$. As such, we can apply Little's law [135] to finalize the steady-state average queuing latency of SemCom-enabled PTQ as

$$\delta_{ij}^{S_2} = \frac{\mathbb{E}\left[Q_{ij}(t)\right]}{\lambda_i^{eff}} = \frac{\sum_{k=0}^{F} k\alpha_{ij}^k}{\tau_i \mu_i^{Mat} + (1-\tau_i)\mu_i^{Mis} - G_{ij}/T}. \tag{5.14}$$

Furthermore, to determine the expressions of BitCom-enabled average packet queuing latency δ_{ij}^B and the BitCom-enabled average packet loss ratio θ_{ij}^B, the same mathematical methods as the above can be employed, where only the PMF and mean of $A_i'(t)$ as in (5.6) in each relevant term need to be substituted with that of $A_i(t)$ as in (5.4). For brevity, the derivation details for δ_{ij}^B and θ_{ij}^B are omitted here.

5.3.2 Problem Formulation

For ease of illustration, we first define three variable sets $\boldsymbol{x} = \{x_{ij} \mid i \in \mathcal{U}, j \in \mathcal{J}\}$, $\boldsymbol{y} = \{y_{ij} \mid i \in \mathcal{U}, j \in \mathcal{J}\}$, and $\boldsymbol{z} = \{z_{ij} \mid i \in \mathcal{U}, j \in \mathcal{J}\}$ that consist of all possible indicators pertinent to UA, MS, and BA, respectively. Without loss of generality, the objective is to maximize the overall message throughput (i.e., the sum of the achievable message rates of all MUs) of the HSB-Net by jointly optimizing $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$, while subject to SemCom-relevant latency and reliability requirements alongside several practical system constraints. Notice that the message throughput performance M_{ij} in (5.3) is actually the ergodic capacity of each link over the timescale of a block when N is large enough, and thus can be computed through averaging the two time-dependent parameters $\gamma_{ij}(t)$ and $\beta_i(t)$ within it [136]. Accordingly, if denoting the long-term average of $M_{ij}^S(t)$ and $M_{ij}^B(t)$ as $\overline{M}_{ij}^S$ and $\overline{M}_{ij}^B$, respectively, when N tends to infinity in (5.3), our optimization objective becomes

$$\begin{aligned}\overline{M}_{ij} &= y_{ij}\overline{M}_{ij}^S + (1-y_{ij})\overline{M}_{ij}^B \\ &= y_{ij}\tau_i\Re_{ij}\big(z_{ij}\log_2\big(1+\overline{\gamma}_{ij}\big)\big) + \rho_{ij}z_{ij}(1-y_{ij})\log_2\big(1+\overline{\gamma}_{ij}\big),\end{aligned} \tag{5.15}$$

where $\overline{\gamma}_{ij}$ denotes the mean of $\gamma_{ij}(t)$ and τ_i is the mean of $\beta_i(t)$. Recalling the average queuing latency δ_{ij} as in (5.9) and the average packet loss ratio θ_{ij} as in (5.10), our joint optimization problem **P1** is now formulated as follows:

$$\mathbf{P1}: \max_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}} \quad \sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{J}} x_{ij}\overline{M}_{ij} \tag{5.16}$$

$$\text{s.t.} \quad \sum_{j\in\mathcal{J}} x_{ij} = 1,\ \forall i \in \mathcal{U}, \tag{5.16a}$$

$$\sum_{i\in\mathcal{U}} x_{ij}z_{ij} \leqslant Z_j,\ \forall j \in \mathcal{J}, \tag{5.16b}$$

$$x_{ij}\delta_{ij} \leqslant \delta_0,\ \forall (i,j) \in \mathcal{U}\times\mathcal{J}, \tag{5.16c}$$

$$x_{ij}\theta_{ij} \leqslant \theta_0,\ \forall (i,j) \in \mathcal{U}\times\mathcal{J}, \tag{5.16d}$$

$$\sum_{j\in\mathcal{J}} x_{ij}\overline{M}_{ij} \geqslant M_i^o,\ \forall i \in \mathcal{U}, \tag{5.16e}$$

$$x_{ij} \in \{0,1\},\ y_{ij} \in \{0,1\},\ \forall (i,j) \in \mathcal{U}\times\mathcal{J}. \tag{5.16f}$$

Constraints (5.16a) and (5.16b) mathematically model the single-BS constraint for UA and the maximum bandwidth resource constraint for BA, respectively. Constraints (5.16c) and (5.16d) ensure that the average queuing latency and the average packet loss ratio of the link between each MU and its associated BS cannot exceed their respective requirements δ_0 and θ_0. M_i^{th} in constraint (5.16e)

represents a minimum message rate threshold for each MU i's association link, while constraint (5.16f) characterizes the binary properties of both $\boldsymbol{x}$ and $\boldsymbol{y}$.

Carefully examining **P1**, it can be observed that the optimization is rather challenging due to several inevitable mathematical obstacles. First of all, **P1** is clearly a non-convex problem involving two complicated constraints (5.16c) and (5.16d), which leads to a high-complexity solution procedure. Another nontrivial point originates from the three different optimization variables, including two integer variables (i.e., $\boldsymbol{x}$ and $\boldsymbol{y}$) and one continuous variable (i.e., $\boldsymbol{z}$). In this respect, although we could first relax $\boldsymbol{x}$ and $\boldsymbol{y}$ to the continuous ones in a conventional manner, the problem after slack should still be non-convex and the subsequent integer recovery may lead to severe performance compromise [137]. In full view of the above difficulties, we propose an efficient solution in the next section to solve **P1** and obtain the joint optimal strategy for the UA, MS, and BA in the HSB-Net.

5.4 Optimal Resource Management in HSB-Net

To make P1 tractable, each z_{ij} $(\forall(i,j) \in \mathcal{U} \times \mathcal{J})$ is first fixed to two thresholds based on both the SemCom case and the BitCom case, respectively. Next, we determine the initial UA and MS strategies by employing a Lagrange primal-dual method, and then devise a preference list-based heuristic algorithm to reach the optimality. On this basis, the BA strategy is then optimally finalized by reallocating the bandwidth of each BS to all its associated MUs while accommodating their respective identified communication modes. Finally, we present the complexity analysis of the proposed solution.

5.4.1 Strategy Determination for UA and MS

There is a minimum bandwidth amount that BS j should allocate to each of its associated SemCom-enabled MU i and BitCom-enabled MU i to simultaneously meet the preset latency, reliability, and message throughout requirements. The feasibility behind this approach is established in accordance with the following proposition.

Proposition 5. *The steady-state average packet queuing latency δ_{ij} and average packet loss ratio θ_{ij} are monotonically non-increasing w.r.t. z_{ij} given any value of y_{ij}.*

Proof. Please see Appendix E. □

Proceeding as in [21], $\Re_{ij}(\cdot)$ is known to be a monotonically increasing function of z_{ij}, and thus $\overline{M}_{ij}$ should also monotonically increase w.r.t. z_{ij} in either the case of $y_{ij}=0$ or $y_{ij}=1$. Accordingly, we first consider the boundary situation of the inequality constraint (5.16e), i.e., $\overline{M}_{ij}=M_i^o$, the minimum z_{ij} required by the association link between MU i and BS j to perform SemCom (denoted by $z_{ij}^{S_M}$) and BitCom (denoted by $z_{ij}^{B_M}$), respectively, can be

$$z_{ij}^{S_M}=\frac{\Re_{ij}^{-1}\left(M_i^o/\tau_i\right)}{\log_2\left(1+\overline{\gamma}_{ij}\right)} \quad \text{and} \quad z_{ij}^{B_M}=\frac{M_i^o}{\rho_{ij}\log_2\left(1+\overline{\gamma}_{ij}\right)}, \tag{5.17}$$

where $\Re_{ij}^{-1}(\cdot)$ indicates the inverse function of $\Re_{ij}(\cdot)$ w.r.t. z_{ij}. Likewise, in the context of (5.9) and (5.10), we can also obtain the constraint (5.16c)-based minimum z_{ij} (denoted by $z_{ij}^{S_\delta}$ and $z_{ij}^{B_\delta}$) and constraint (5.16d)-based minimum z_{ij} (denoted by $z_{ij}^{S_\theta}$ and $z_{ij}^{B_\theta}$) in their respective inequality boundary situations. It is worth pointing out here that the feasible $z_{ij}^{S_\delta}$ solution may not exist if $\delta_i^{S_1}>\delta_0$, while $\delta_i^{S_2}$ cannot be negative. In such a case, we set $z_{ij}^{S_\delta}=+\infty$ to avoid the possibility of the MU selecting the SemCom mode in the subsequent solution.

Afterward, our aim is to find the optimal $\boldsymbol{x}^*=\left\{x_{ij}^* \mid i\in\mathcal{U}, j\in\mathcal{J}\right\}$ and $\boldsymbol{y}^*=\left\{y_{ij}^* \mid i\in\mathcal{U}, j\in\mathcal{J}\right\}$ by fixing each SemCom-associated z_{ij} term as $z_{ij}^{S_{th}}=\max\left\{z_{ij}^{S_M}, z_{ij}^{S_\delta}, z_{ij}^{S_\theta}\right\}$ and each BitCom-associated z_{ij} as $z_{ij}^{B_{th}}=\max\left\{z_{ij}^{B_M}, z_{ij}^{B_\delta}, z_{ij}^{B_\theta}\right\}$. As such, constraints (5.16c)-(5.16e) in the primal problem **P1** can be all removed, and then **P1** degenerates into

$$\mathbf{P1.1}: \max_{\boldsymbol{x},\boldsymbol{y}} \quad \sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{J}} x_{ij}\left[y_{ij}\overline{M}_{ij}^{S_{th}}+\left(1-y_{ij}\right)\overline{M}_{ij}^{B_{th}}\right] \tag{5.18}$$

$$\text{s.t.} \quad \sum_{i\in\mathcal{U}} x_{ij}\left[y_{ij}z_{ij}^{S_{th}}+\left(1-y_{ij}\right)z_{ij}^{B_{th}}\right]\leqslant Z_j,\ \forall j\in\mathcal{J}, \tag{5.18a}$$

$$(5.16a),\ (5.16f), \tag{5.18b}$$

where let $\overline{M}_{ij}^{S_{th}}=\tau_i\Re_{ij}\left(z_{ij}^{S_{th}}\log_2\left(1+\overline{\gamma}_{ij}\right)\right)$ and $\overline{M}_{ij}^{B_{th}}=\rho_{ij}z_{ij}^{B_{th}}\log_2\left(1+\overline{\gamma}_{ij}\right)$, both are regarded as known constants.

Regarding **P1.1**, we incorporate constraint (5.18a) into its objective function (5.18) by associating Lagrange multipliers $\boldsymbol{\eta}=\{\eta_j \mid j\in\mathcal{J}\}$. The associated

Lagrange function is

$$\begin{aligned}
&L\left(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\eta}\right)\\
=&\sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{J}}x_{ij}\left[y_{ij}\overline{M}_{ij}^{S_{th}}+\left(1-y_{ij}\right)\overline{M}_{ij}^{B_{th}}\right]\\
&\qquad+\sum_{j\in\mathcal{J}}\eta_j\left(Z_j-\sum_{i\in\mathcal{U}}x_{ij}\left[y_{ij}z_{ij}^{S_{th}}+\left(1-y_{ij}\right)z_{ij}^{B_{th}}\right]\right)\\
=&\sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{J}}\left[x_{ij}y_{ij}\left(\overline{M}_{ij}^{S_{th}}-\eta_j z_{ij}^{S_{th}}\right)+x_{ij}\left(1-y_{ij}\right)\left(\overline{M}_{ij}^{B_{th}}-\eta_j z_{ij}^{B_{th}}\right)\right]+\sum_{j\in\mathcal{J}}\eta_j Z_j\\
\triangleq&\ \widetilde{L}_{\boldsymbol{\eta}}\left(\boldsymbol{x},\boldsymbol{y}\right)+\sum_{j\in\mathcal{J}}\eta_j Z_j,
\end{aligned}\tag{5.19}$$

where $\widetilde{L}_{\boldsymbol{\eta}}\left(\boldsymbol{x},\boldsymbol{y}\right)$ is defined for expression brevity That way, the Lagrange dual problem of **P1.1** becomes

$$\mathbf{D1.1}:\ \min_{\boldsymbol{\eta}}\quad H\left(\boldsymbol{\eta}\right)=g_{\boldsymbol{x},\boldsymbol{y}}\left(\boldsymbol{\eta}\right)+\sum_{j\in\mathcal{J}}\eta_j Z_j \tag{5.20}$$

$$\text{s.t.}\quad \eta_j\geqslant 0,\ \forall j\in\mathcal{J}, \tag{5.20a}$$

where

$$\begin{aligned}
g_{\boldsymbol{x},\boldsymbol{y}}\left(\boldsymbol{\eta}\right)\ &=\ \sup_{\boldsymbol{x},\boldsymbol{y}}\ \widetilde{L}_{\boldsymbol{\eta}}\left(\boldsymbol{x},\boldsymbol{y}\right)\\
&\text{s.t.}\ \ (5.16\text{a}),\ (5.16\text{f}).
\end{aligned}\tag{5.21}$$

Notably, since (5.18) is convex and (5.18a) contains only linear and affine inequalities, according to the duality property [101], the above primal-dual transformation w.r.t. **D1.1** determines at least the best upper bound of **P1.1**. Hence, our focus now shifts to seeking $\boldsymbol{x}^*$ and $\boldsymbol{y}^*$ through solving problem (5.21) in an iterative fashion of updating $\boldsymbol{\eta}$ with a subgradient method [138].

Before that, all cross terms of $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\widetilde{L}_{\boldsymbol{\eta}}\left(\boldsymbol{x},\boldsymbol{y}\right)$ need to be tackled for tractability, where $x_{ij}y_{ij}$ and $x_{ij}(1-y_{ij})$ are the only two ways of crossing. Combined with constraint (5.16f), we first extend the original BS indicator set $\mathcal{J}$ to $\mathcal{J}'=\{1,2,\cdots,J,J+1,J+2,\cdots,2J\}$ and define a new set of variables $\boldsymbol{\nu}=\{\nu_{ij'}\mid i\in\mathcal{U},j\in\mathcal{J}'\}$ w.r.t. (5.21), such that

$$\nu_{ij'}=\begin{cases}x_{ij'}y_{ij'}, & \text{if } j'\in\mathcal{J}=\{1,2,\cdots,J\};\\ x_{i(j'-J)}(1-y_{i(j'-J)}), & \text{if } j'\in\mathcal{J}'\backslash\mathcal{J}.\end{cases}\tag{5.22}$$

In parallel, we define another new set of constants $\boldsymbol{\xi}=\{\xi_{ij'}\mid i\in\mathcal{U},j\in\mathcal{J}'\}$ to

characterize the coefficient of each $\nu_{ij'}$, i.e.,

$$\xi_{ij'} = \begin{cases} \overline{M}_{ij'}^{S_{th}} - \eta_{j'} z_{ij'}^{S_{th}}, & \text{if } j' \in \mathcal{J}; \\ \overline{M}_{i(j'-J)}^{B_{th}} - \eta_{(j'-J)} z_{i(j'-J)}^{B_{th}}, & \text{if } j' \in \mathcal{J}' \backslash \mathcal{J}. \end{cases} \tag{5.23}$$

As such, given the initial dual variable $\boldsymbol{\eta}$, problem (5.21) should be straightforwardly converted to

$$\mathbf{P1.2}: \max_{\boldsymbol{\nu}} \quad \sum_{i \in \mathcal{U}} \sum_{j' \in \mathcal{J}'} \xi_{ij'} \nu_{ij'} \tag{5.24}$$

$$\text{s.t.} \quad \sum_{j' \in \mathcal{J}'} \nu_{ij'} = 1, \ \forall i \in \mathcal{U}, \tag{5.24a}$$

$$\nu_{ij'} \in \{0, 1\}, \ \forall (i, j') \in \mathcal{U} \times \mathcal{J}'. \tag{5.24b}$$

It is easily derived from **P1.2** that for any $i \in \mathcal{U}$, the optimal j' such that $\nu_{ij'} = 1$ is exactly the j' that enables the maximum $\xi_{ij'}$ compared with any other $j' \in \mathcal{J}'$. In other words, let $\widehat{j'} = \arg\max_{j' \in \mathcal{J}'} \xi_{ij'}, \forall i \in \mathcal{U}$, we can determine $\boldsymbol{x}^*$ and $\boldsymbol{y}^*$ for each MU i and BS j in the HSB-Net by

$$\begin{cases} x_{ij}^* = 1, \ y_{ij}^* = 1, & \text{if } \widehat{j'} \in \mathcal{J} \text{ and } j = \widehat{j'}; \\ x_{ij}^* = 1, \ y_{ij}^* = 0, & \text{if } \widehat{j'} \in \mathcal{J}' \backslash \mathcal{J} \text{ and } j = \widehat{j'} - J; \\ x_{ij}^* = 0, & \text{otherwise.} \end{cases} \tag{5.25}$$

Afterward, the partial derivatives w.r.t. $\boldsymbol{\eta}$ in the objective function $H(\boldsymbol{\eta})$ in **D1.1** are set as the subgradient direction in each update iteration. Now suppose that in a certain iteration, e.g., iteration l, in line with constraint (5.20a), each η_j ($j \in \mathcal{J}$) is updated as the following rule:

$$\eta_j(l+1) = \max\{\eta_j(l) - \epsilon(l) \cdot \nabla H(\eta_j), 0\}, \tag{5.26}$$

where

$$\nabla H(\eta_j) = Z_j - \sum_{i \in \mathcal{U}} x_{ij} \left[y_{ij} z_{ij}^{S_{th}} + (1 - y_{ij}) z_{ij}^{B_{th}} \right], \tag{5.27}$$

and $\epsilon(l)$ denotes the stepsize of the update in iteration l. In general, the convergence of the subgradient descent method can be guaranteed with a properly preset stepsize [139].

Nevertheless, it is worth noting that the above solutions cannot always directly reach the optimality of **P1.1**, as the BA constraint (5.18a) may be violated at some BSs within each iteration. In this case, here we additionally adopt a

preference list-based heuristic algorithm to project the solution obtained in each iteration back to the feasible set of (5.18a). To be concrete, $\boldsymbol{x}^*$ and $\boldsymbol{y}^*$ obtained by (5.22)-(5.27) are first leveraged to identify the index list of the BSs that violate (5.18a), denoted as $\widetilde{\mathcal{J}} = \left\{ j \mid j \in \mathcal{J}, \sum_{i\in\mathcal{U}} x_{ij}^* \left[y_{ij}^* z_{ij}^{S_{th}} + \left(1 - y_{ij}^*\right) z_{ij}^{B_{th}} \right] > Z_j \right\}$. Consider an arbitrary BS $\tilde{j} \in \widetilde{\mathcal{J}}$, let $\mathcal{U}_{\tilde{j}} = \{ i \mid i \in \mathcal{U}, x_{i\tilde{j}}^* = 1 \}$ store all its current associated MUs, the MU that consumes the largest amount of bandwidth among all MUs can be found by

$$\hat{i} = \arg\max_{i\in\mathcal{U}_{\tilde{j}}} \left[y_{i\tilde{j}}^* z_{i\tilde{j}}^{S_{th}} + \left(1 - y_{i\tilde{j}}^*\right) z_{i\tilde{j}}^{B_{th}} \right]. \tag{5.28}$$

Next, we assume that MU $\hat{i}$ has an initial variable set $\mathcal{J}_{\hat{i}}' = \mathcal{J}'$, which can be reckoned as its UA and MS preference list pertinent to optimizing **P1.2**. Since the solution $(x_{\hat{i}\tilde{j}}^*, y_{\hat{i}\tilde{j}}^*)$ is obviously inapplicable due to the insufficient bandwidth resources at BS $\tilde{j}$, let the corresponding index $j' = \tilde{j}$ in its SemCom case or $j' = \tilde{j} + J$ in its BitCom case be removed from MU $\hat{i}$'s preference list $\mathcal{J}_{\hat{i}}'$. That is,

$$\mathcal{J}_{\hat{i}}' = \begin{cases} \mathcal{J}_{\hat{i}}' \backslash \tilde{j}, & \text{if } y_{\hat{i}\tilde{j}}^* = 1; \\ \mathcal{J}_{\hat{i}}' \backslash \left(\tilde{j} + J\right), & \text{if } y_{\hat{i}\tilde{j}}^* = 0, \end{cases} \tag{5.29}$$

whereby its current optimal $\widehat{j'}$ becomes

$$\widehat{j'} = \arg\max_{j'\in\mathcal{J}_{\hat{i}}'} \xi_{\hat{i}j'}. \tag{5.30}$$

Calculating (5.25) again, we can update MU $\hat{i}$'s optimal UA and MS strategy $(x_{\hat{i}j}^*, y_{\hat{i}j}^*)$ over any BS $j \in \mathcal{J}$ as well as BS $\tilde{j}$'s UA list $\mathcal{U}_{\tilde{j}}$. After that, the satisfaction of constraint (5.18a) w.r.t. BS $\tilde{j}$ should be rechecked, and even if it is still in violation, we can repeat the operations between (5.28) to (5.30) until (5.18a) is eventually met at BS $\tilde{j}$. Likewise, the above procedure can be applied to any other BS $j \in \widetilde{\mathcal{J}}$ until the bandwidth constraints are fulfilled at all BSs after each iteration. In summary, by alternately updating $(\boldsymbol{x}, \boldsymbol{y})$ and $\boldsymbol{\eta}$ in combination with the proposed preference list-based heuristic algorithm, the UA and MS problems can be solved in the HSB-Net.

5.4.2 Optimal Solution for BA with Complexity Analysis

According to the obtained UA solution $\boldsymbol{x}^*$ and MS solution $\boldsymbol{y}^*$, we aim to reallocate all bandwidth resources of each BS j ($\forall j \in \mathcal{J}$) to all its associated MUs, thus a total of S BA subproblems w.r.t. **P1** are constructed. Based on Proposition 5

alongside the preset bandwidth threshold $z_{ij}^{S_{th}}$ and $z_{ij}^{B_{th}}$, each BA subproblem of BS j is formulated as follows:

$$\mathbf{P1.3}_j: \max_{\boldsymbol{z}} \quad \sum_{i\in\mathcal{U}_j^S} \overline{M}_{ij}^S + \sum_{i\in\mathcal{U}_j^B} \overline{M}_{ij}^B \tag{5.31}$$

$$\text{s.t.} \quad \sum_{i\in\mathcal{U}_j^S\cup\mathcal{U}_j^B} z_{ij} = Z_j, \tag{5.31a}$$

$$z_{ij} \geqslant z_{ij}^{S_{th}},\ \forall i \in \mathcal{U}_j^S, \tag{5.31b}$$

$$z_{ij} \geqslant z_{ij}^{B_{th}},\ \forall i \in \mathcal{U}_j^B, \tag{5.31c}$$

where $\mathcal{U}_j^S = \left\{ i \mid i \in \mathcal{U}, x_{ij}^* = 1, y_{ij}^* = 1 \right\}$ stands for the set of all SemCom-enabled MUs associated with BS j, and $\mathcal{U}_j^B = \left\{ i \mid i \in \mathcal{U}, x_{ij}^* = 1, y_{ij}^* = 0 \right\}$ represents the set of all BitCom-enabled MUs associated with BS j. Then given the convex property of $\Re_{ij}(\cdot)$ [21], we have the objective function and all constraints of each $\mathbf{P1.3}_j$ are convex, thereby efficient linear programming toolboxes such as CVPXY [102] can be directly applied to obtain the optimal BA solution for the HSB-Net.

In terms of the computational complexity of the proposed solution, determining the minimum z_{ij} allocated to each potential UA link first requires $\mathcal{O}(F^2)$ complexity to compute the one-step state transition probability matrix of its PTQ as given in the proof of Proposition 4, hence $\mathcal{O}(UJF^2)$ complexity is needed for obtaining $\mathbf{P1.1}$. Then, in each iteration of solving $\mathbf{D1.1}$, the complexity is $\mathcal{O}(UJ^2)$ for at most J violated BSs to find their respective largest bandwidth-consumed MUs in a group of UJ variables. As such, if let V denote the required number of iterations that leads to convergence of $\mathbf{D1.1}$, finalizing the UA and MS solutions needs a total of $\mathcal{O}(VUJ^2)$ complexity. Finally, since each $\mathbf{P1.3}_j$ can be solved by the linear programming method with complexity $\mathcal{O}(U^2)$ [140], the proposed wireless resource management solution has a polynomial-time overall complexity of $\mathcal{O}(UJ(F^2 + VJ + U))$.

5.5 Numerical Results and Discussions

In this section, numerical evaluations are conducted to demonstrate the performance of our proposed wireless resource management solution in the HSB-Net, where we employ Python 3.7-based PyCharm as the simulator platform and implement it in a workstation PC featuring the AMD Ryzen-9-7900X processor with 12 CPU cores and 128 GB RAM. To preserve generality, we first model a circular area with a radius of 300 meters, in which 200 MUs and 10 BSs are randomly dropped. Moreover, the SINR γ_{ij} follows a Gaussian distribution with standard

Table 5.1: Simulation Parameters

Parameters	**Values**
Bandwidth budget of each BS (Z_j)	15 MHz [141]
Transmit power of each MU	20 dBm
Noise power	-111.45 dBm [29]
Path loss model	$34 + 40 \log (d \text{ [m]})$
Time slot length (T)	1 ms
Number of bits in each packet (L)	800 bits
Packet buffer size of PTQ (F)	20
Maximum average packet queuing latency threshold (δ_0)	20 ms
Maximum average packet loss ratio threshold (θ_0)	0.01

deviation of 4 dB [126]. For brevity, some simulation parameters not mentioned in the context and their fixed values are summarized in Table 6.1.

In SemCom-relevant settings, we simulate a general text transmission scenario to examine the proposed solution. Note that such performance test can also be accomplished with other content types like images or videos, and the reason we choose text is to leverage existing natural language processing models that have been well validated in SemCom-related works. Particularly, the Transformer in [1] is adopted as the unified semantic encoder for all SemCom links, and the PyTorch-based Adam optimizer is applied for model training with an initial learning rate of 0.001. Based on the public dataset extracted from the proceedings of European Parliament [108], the expression of B2M function $\Re_{ij}(\cdot)$ at each SemCom link can be well approximated from extensive model tests [21].

As for the queuing modeling part, each MU's average knowledge-matching degree τ_i, minimum message rate threshold M_i^o, and BitCom-based B2M coefficient ρ_{ij} are randomly generated in the range of $0.6 \sim 1$, $50 \sim 100$, and $2 \times 10^{-5} \sim 2 \times 10^{-4}$, respectively. Besides, the average packet arrival rate λ_i is prescribed at 1000 packets/s for all MUs [142], while the average interpretation times of knowledge-matching and -mismatching packets in SCQ are considered as 8×10^{-4} and 1×10^{-3} s/packet, respectively. Furthermore, we set a dynamic stepsize of $\epsilon(l) = 1 \times 10^{-6}/l$ to update the Lagrange multipliers in (5.26), where the convergence of each trial can be always guaranteed. It is worth mentioning that all the above parameter values are set by default unless otherwise specified, and all subsequent numerical results are obtained by averaging over a sufficiently large number of trials.

For comparison purposes, here we employ four different resource management benchmarks in HSB-Nets by combining the max-SINR UA scheme (i.e., each MU is associated with the BS enabling the strongest SINR) with several differing MS

and BA schemes, respectively. To the best of our knowledge, no existing work has proposed any benchmark solutions dedicatedly for MS, and therefore, two heuristic schemes are developed as MS baselines: (MS-I) A *knowledge matching degree-based* method, where each MU selects the SemCom mode when its knowledge matching degree is above a preset threshold (e.g., a threshold of 0.8 has been used in our simulations), and otherwise selects the BitCom mode; (MS-II) A *SINR-based* method, where each MU selects the BitCom mode when its SINR is above a preset threshold (e.g., a threshold of 6 dB has been used in our simulations), and otherwise selects the SemCom mode.[9] In parallel, two typical BA schemes are adopted as baselines: (BA-I) The *water-filling* algorithm [109]; (BA-II) The *evenly-distributed* algorithm [30].

5.5.1 Queuing Model Validation

For starters, we simulate the entire packet queuing processes for SCQ and PTQ at a SemCom-enabled MU with a default average knowledge-matching degree $\tau_i = 1.0$ and a default SINR $\gamma_{ij} = 0$ dB to validate the analytical accuracy of the derived queuing model. In detail, the analysis results are based on the direct computation of average packet queuing latency and packet loss ratio as in (5.11)-(5.14). In contrast, the simulation results are calculated by generating various randomized processes (including Poisson packet arrival, general-distribution based SCQ-packet departure and SINR-stochasticity based PTQ-packet departure) and averaging over $10,000$ trials.

Figure 5.3 first depicts the average queuing latency $\delta_i^{S_1}$ at SCQ by increasing the initial packet arrival intensity λ_i from 750 to 1050 packets/s, where $\tau_i = 0.4$, 0.7, and 1.0 are all taking into account. It is seen that the analytical curve basically agrees with the simulated one as λ_i grows, and the higher the τ_i, the closer the two latency curves in values. This can be explained by that the lower τ_i indicates the worse semantic inference capability for packet departure at SCQ, resulting in more uncertainty, i.e., higher fluctuation, on each randomly generated semantic-coding time. However, in our queuing analysis, the semantic-coding times of all knowledge-mismatching packets are simply approximated to have the same rate of $1/\mu_i^{Mis}$, which ignores the discrepancy between different knowledge-matching degrees, and thus rendering the numerical bias between simulated and analytical results in the lower τ_i region. Besides, the average queuing latency

[9]The MS-I scheme is intuitive since the higher the knowledge matching degree, the better the semantic-related performance [21]. As for the MS-II scheme, this is because SemCom shows more powerful anti-noise capability in the low-SINR region [1], while BitCom ensures higher content transmission accuracy in the high-SINR region [3], thereby MS-II should be applicable.

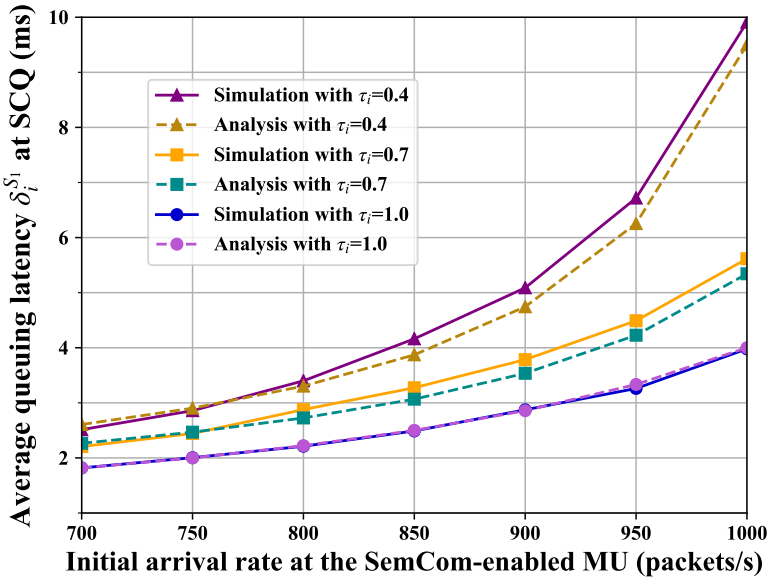

Figure 5.3: Simulated and analytical results w.r.t. average queuing latency $\delta_i^{S_1}$ at SCQ for varying packet arrival rates and average knowledge-matching degrees.

increases with the packet arrival rate in each case, which trend is obvious as the semantic-c

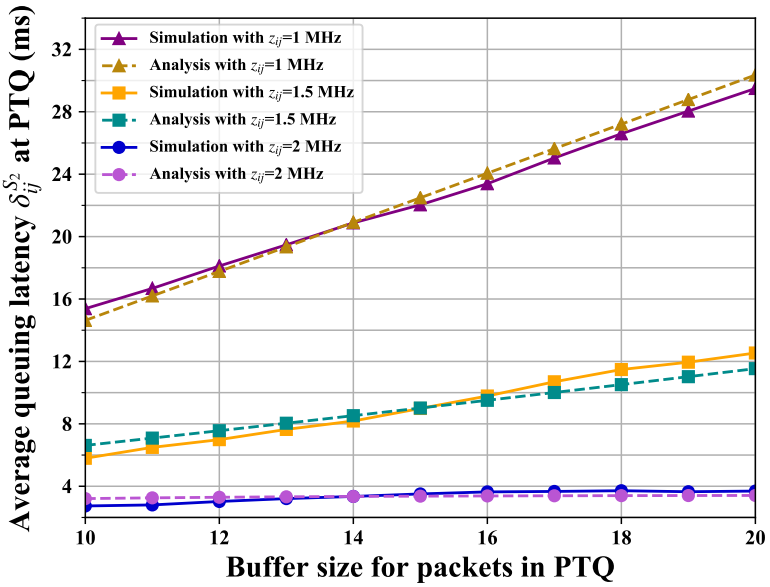

Figure 5.4: Simulated and analytical results w.r.t. average queuing latency $\delta_{ij}^{S_2}$ at PTQ for varying packet buffer sizes and allocated bandwidth resources.

Next, we compare the simulated and analytical results of PTQ in terms of its average queuing latency and packet loss ratio in Fig. 5.4 and Fig. 5.5, respectively, which are basically consistent in both cases. By varying PTQ's buffer size F from 10 to 22, Fig. 5.4 shows a moderate increasing trend in average queuing latency $\delta_{ij}^{S_2}$ with different allocated bandwidth $z_{ij} = 1$, 1.5, and 2 MHz. This is logical

since the buffer with a larger size is more likely to hold a long queue length, resulting in more average waiting time per packet according to (5.14). Moreover, it can be observed that the less the bandwidth assigned to the MU, the higher the $\delta_{ij}^{S_2}$ while the steeper the upward trend. Herein, the former phenomenon is reasonable due to the low packet departure rate as in (5.5). The latter is because that as the given z_{ij} grows, the rapid departure of packets gradually dominates the queuing process of PTQ, thereby the small changes in the buffer size could only have

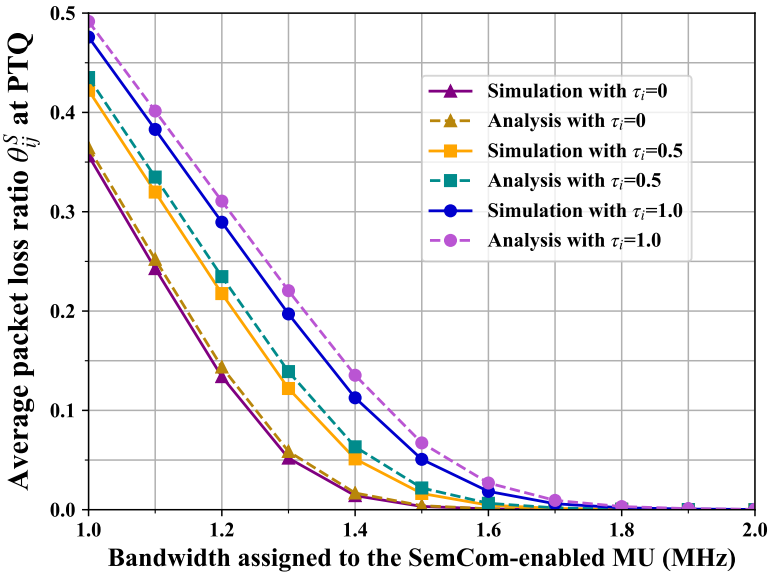

Figure 5.5: Simulated and analytical results w.r.t. average packet loss ratio θ_{ij}^{S} at PTQ for varying bandwidth resources and knowledge-matching degrees.

Meanwhile, Fig. 5.5 presents the average packet loss ratio θ_{ij}^{S} at PTQ versus the allocated bandwidth z_{ij} under three average knowledge-matching degrees of $\tau_i = 0$, 0.5, and 1.0, where the simulated results can always fit the analytical ones well. Specifically, the obtained θ_{ij}^{S} first decreases with z_{ij}, and then converges close to 0 when z_{ij} exceeds around 1.8 MHz. The rationale behind this is similar to Fig. 5.4, i.e., the packets arriving at the PTQ with a higher departure rate are less likely to be blocked. Furthermore, combined (5.7) with the setting of $\mu_i^{Mat} > \mu_i^{Mis}$, it is seen that the higher the τ_i, the higher the packet arrival rate of PTQ, and thus the greater the likelihood that its buffer tends to be full. Notably, the average packet loss ratio of PTQ and the overall queuing latency of both SCQ and PTQ should be taken into account together to meet the preset delay and reliability requirements. For instance, θ_{ij}^{S} can reach the threshold of $\theta_0 = 0.01$ by assigning 1.55 MHz bandwidth to the same MU with $\tau_i = 0.5$. However, even the default $\lambda_i = 1000$ packets/s will cause the queuing delay of 9.1 ms at SCQ and 11.5 ms at PTQ (i.e., the total of 20.6 ms) in the same case, which has exceeded

the threshold $\delta_0 = 20$ ms.

5.5.2 Performance of the Proposed Solution

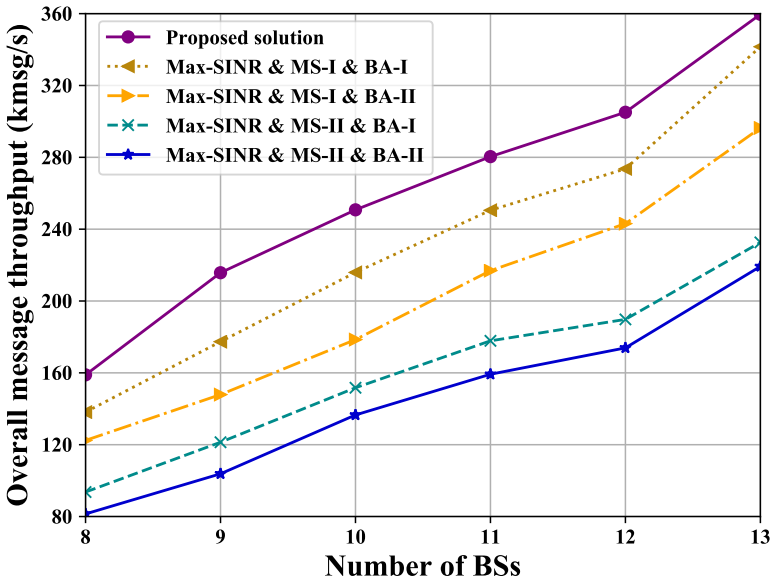

Figure 5.6: Time-averaged overall message throughput (*kmsg/s*) versus different numbers of BSs in the HSB-Net.

To validate our proposed resource management solution, we test the overall message throughput of HSB-Net under different numbers of BSs and MUs in Fig. 5.6 and Fig. 5.7, respectively, in comparison with the four benchmarks. As first elucidated in Fig. 5.6, by varying the number of BSs from 8 to 13, the message throughput performance of the proposed solution gradually increases from around 160 to 360 kmsg/s (1 kmsg/s = 1000 msg/s), and consistently outperforms these benchmarks. For example, a performance gain of the proposed solution is about 29.9 kmsg/s compared with the benchmark of Max-SINR plus MS-I plus BA-I and 102.6 kmsg/s compared with the benchmark of Max-SINR plus MS-II plus BA-I when 11 BSs are located in the HSB-Net. Here, such an uptrend is apparent since more BSs represent that more bandwidth resources are available for MUs to achieve higher message rates. Particularly in such an uplink scenario of HSB-Net, the increase in the number of BSs does not have any impact on channel interference, and hence a stable growth is observed.

By contrast, Fig. 5.7 demonstrates a downward trend of message throughput performance when rising the number of MUs from 140 to 240. To be concrete, the overall network performance is already saturated at the very beginning in holding 140 MUs and then deteriorates with the addition of MUs, as the effect of severe channel interference from excessive MUs starts to dominate the more availability

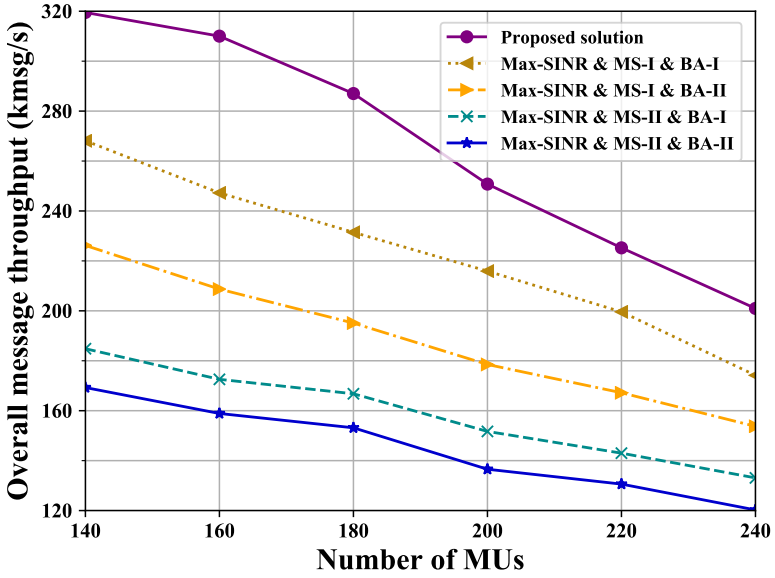

Figure 5.7: Time-averaged overall message throughput (*kmsg/s*) versus different numbers of MUs in the HSB-Net.

of resources. In the meantime, it can be seen that our solution still surpasses all the four benchmarks with a significant performance gain. For instance, with 160 MUs in the HSB-Net, the proposed solution realizes a message throughput of about 310 kmsg/s, i.e., 1.5 times that of the Max-SINR plus MS-I plus BA-II scheme ar

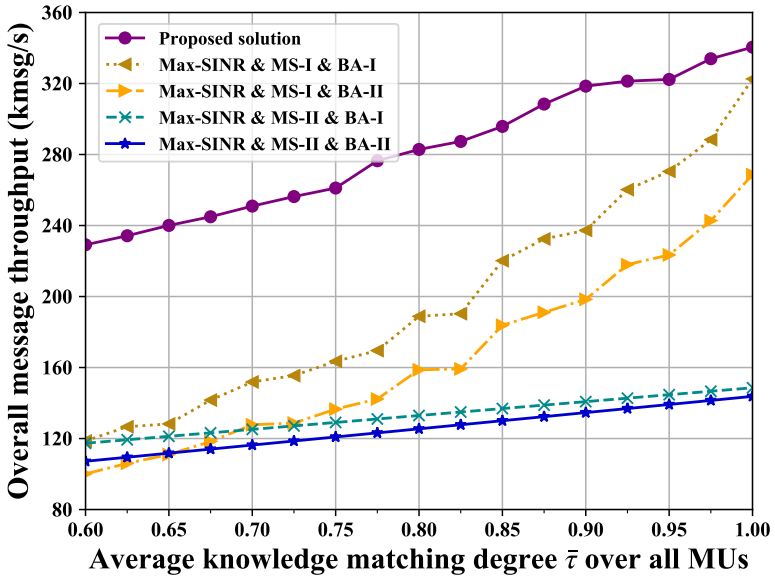

Figure 5.8: Time-averaged overall message throughput (*kmsg/s*) versus different average knowledge-matching degrees over all 200 MUs in the HSB-Net.

In addition, we compare the message throughput performance with varying overall average knowledge-matching degree $\bar{\tau} = \frac{1}{U}\sum_{i\in\mathcal{U}}\tau_i$ as shown in Fig. 5.8.

Again, our solution still outperforms these benchmarks with the considerable performance gain, especially in the low $\bar{\tau}$ region. Besides, a growing message throughput is observed by all solutions as $\bar{\tau}$ increases, and our solution and the MS-I scheme are more affected by changes in $\bar{\tau}$ compared to the MS-II. The former trend is intuitive since the larger $\bar{\tau}$ means that there is a greater likelihood for the HSB-Net having MUs with the high B2M transformation rates. The latter is first due to the message-throughput-priority design in our objective function (5.16), and therefore, our solution is more likely to generate more SemCom-enabled MUs with larger $\bar{\tau}$. Likewise, more SemCom-enabled MUs can exist in the same case according to the prescribed MS-I scheme, while the number of SemCom-enabled MUs is only affected by SINR in MS-II, and thus keeps stable irrespective of the change in $\bar{\tau}$.

Finally, the CDFs of the message rate M_{ij} rendered at all links are plotted in Fig. 5.9. Although most MUs in our solution only get the lower message rates compared with these benchmarks, this is reasonable since our optimization to **P1** focuses on the maximization for overall message throughput of all MUs in the HSB-Net. Hence, it can be interpreted as that the proposed solution choose to sacrifice user semantic fairness in favor of devoting more bandwidth resources to a smaller number of MUs with better average knowledge-matching degrees, B2M transformation, and SINRs.

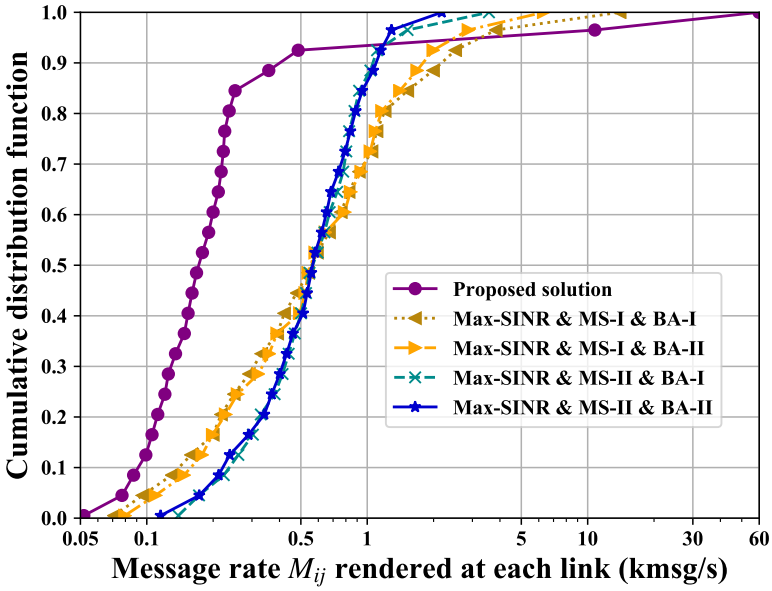

Figure 5.9: The CDF of the message rate M_{ij} obtained by the associated link.

5.6 Conclusions

In this chapter, we investigated the wireless resource management problem in a novel yet practical network scenario, i.e., HSB-Net, where SemCom and BitCom modes are available for selection by all MUs. To measure SemCom and BitCom with the same performance metric, a B2M transformation function was first introduced to identify the message throughput of each associated link. Then, considering the unique semantic coding and knowledge matching mechanisms in SemCom, we modelled a two-stage tandem queuing system for the transmission of semantic packets, followed by the theoretical derivation for average packet loss ratio and queuing latency. On this basis, a joint optimization problem was formulated to maximize the overall message throughput of HSB-Net. Afterward, a Lagrange primal-dual method was employed and a preference list-based heuristic algorithm was developed to seek the optimal UA, MS, and BA solutions with the low computational complexity. Numerical results finally validated the accuracy of our queuing analysis and the performance superiority of our proposed solution in terms of overall message throughput compared with four different benchmarks. The next chapter will study how to apply SemCom to V2V networks for efficient semantic service provisioning.

Chapter 6

xURLLC-Aware Service Provisioning in Vehicular Networks: A Semantic Communication Perspective

6.1 Introduction

Nowadays, the scarcity of available communication resources, such as bandwidth and energy, is envisioned to exacerbate to unprecedented levels and to be the most challenging problem in the near future, especially considering traditional communications-based massive vehicular networks. Fortunately, SemCom beyond the conventional Shannon paradigm has recently been recognized as a promising remedy for communication resource savings and transmission reliability promotion [1, 3, 5–8, 25, 143], which, therefore, inspires us to investigate the potential of exploiting SemCom to perform efficient service provisioning in vehicular networks.

Apart from many superiorities, it is worth pointing out that equivalent background knowledge should be of paramount importance to eliminate semantic ambiguity, which has led to a key concept of KB in the realm of SemCom [3, 6, 8, 25, 143]. Specifically, a single KB is deemed a small information entity that contains background knowledge corresponding to only one particular application domain [6]. Since different KBs are associated with different background knowledge, holding some common KBs becomes the necessary condition to perform SemCom between two vehicles in accordance with the knowledge equivalence principle. Moreover, through employing differing KBs, semantic information related to different application domains can be accurately exchanged among vehicles,

thereby efficiently achieving service provisioning in a way of SemCom.

Nevertheless, to the best of our knowledge, none of the existing work has ever explored the potential of applying SemCom to V2V networks in alignment with stringent latency and reliability requirements, which should be rather challenging as explained below. In full view of the novel paradigm of *SemCom-enabled vehicular networks* (SCVNs), the task lies in seeking the optimal solution to efficiently provide all participating vehicle users (VUEs) with diverse SemCom-empowered services. However, it is noticed that enabling the next-generation ultra-reliable and low-latency communication (xURLLC) remains indispensable, especially when pursuing adequate semantic fidelity for large-scale V2V communications. Uniquely, the reliability requirement originates from the aforementioned strict knowledge equivalence condition, while the latency requirement is related to varying processing efficiencies of semantic interpretation models. In summary, we are encountering three fundamental networking challenges in SCVN.

- *Challenge 1: How to measure performance in terms of reliability and latency when introducing SemCom into vehicular networks?* Notice that data packets transmitted in traditional V2V communications generally consider only one type of queuing process, in which different packets have the same distributions of arrival and interpretation [122]. However, semantic data packets in SCVN related to various SemCom-empowered services may result in different queuing and processing delays due to the different semantic interpretation models equipped on VUEs. Hence, it is not trivial to accurately measure and assume prior information about the latency performance in SCVN. Besides, given the core mechanism of semantic delivery, it should be more reasonable to characterize the reliability performance from the knowledge equivalence perspective between any two associated VUEs for V2V communications. All of the above constitutes the first and the main challenge.

- *Challenge 2: How to construct appropriate KBs at each VUE for better SemCom-empowered service provisioning?* Considering varying practical KB sizes, personal preferences on different SemCom services, and the limited vehicular storage capacities, there is a pressing need to devise an optimal *knowledge base construction* (KBC) policy that is not only proactive but also collaborative for all VUEs to construct their respective appropriate KBs for better service provisioning. Notably, this challenge is inevitable in xURLLC-aware SCVN, since the remote KB access approach (i.e., the approach that each VUE remotely accesses its required KBs via RSUs, Cloud

servers or core networks) can incur intolerable communication overhead and transmission latency. Therefore, the local and distributed KBC approach should be more applicable for each VUE to well perform SemCom.

- *Challenge 3: How to select the best vehicle node for each VUE from multiple candidate neighbors to optimize service provisioning related overall network performance?* As mentioned earlier, selecting vehicle pairs for realizing service provisioning is very much distinct in SCVN due to the knowledge equivalence condition. Combined with different KBs constructed at numerous VUEs and unstable wireless link quality, it can be challenging to well solve the service provisioning-driven vehicle pairing problem, namely *vehicle service pairing* (VSP).

In line with the above, it is particularly worthwhile to note that challenges 2 and 3 are closely coupled, which makes it indispensable to jointly seek the optimal KBC and VSP policy for all VUEs to meet the xURLLC requirements. Moreover, efficient SemCom-empowered service provisioning is excepted after addressing the three challenges to yield a bunch of benefits in SCVN, such as improving V2V information interaction efficiency, reducing data traffic congestion, and ensuring high-quality vehicular services.

In this chapter, we propose a novel SemCom-empowered Service Supplying Solution (S^4) in SCVN with the awareness of meeting the xURLLC requirements. Both theoretical analysis and numerical results demonstrate the performance superiority of S^4 in terms of average queuing latency, semantic data packet throughput, user knowledge matching degree, and user knowledge preference satisfaction compared with two different benchmarks. In a nutshell, our main contributions are summarized as follows:

- We identify two fundamental yet unique problems KBC and VSP in SCVN by fully incorporating SemCom-related characteristics with vehicular network scenarios. In particular, individual VUE preference for different KBs is considered in the KBC, while the VSP of two adjacent VUEs takes into account the strict matching requirement between their respective constructed KBs.

- We theoretically derive the KB matching based queuing latency for a VUE pair in SCVN. Then, through carefully analyzing the unique queuing features of received semantic data packets, a joint latency-minimization problem is mathematically formulated subject to several KBC and VSP-related reliability constraints and other practical system limitations.

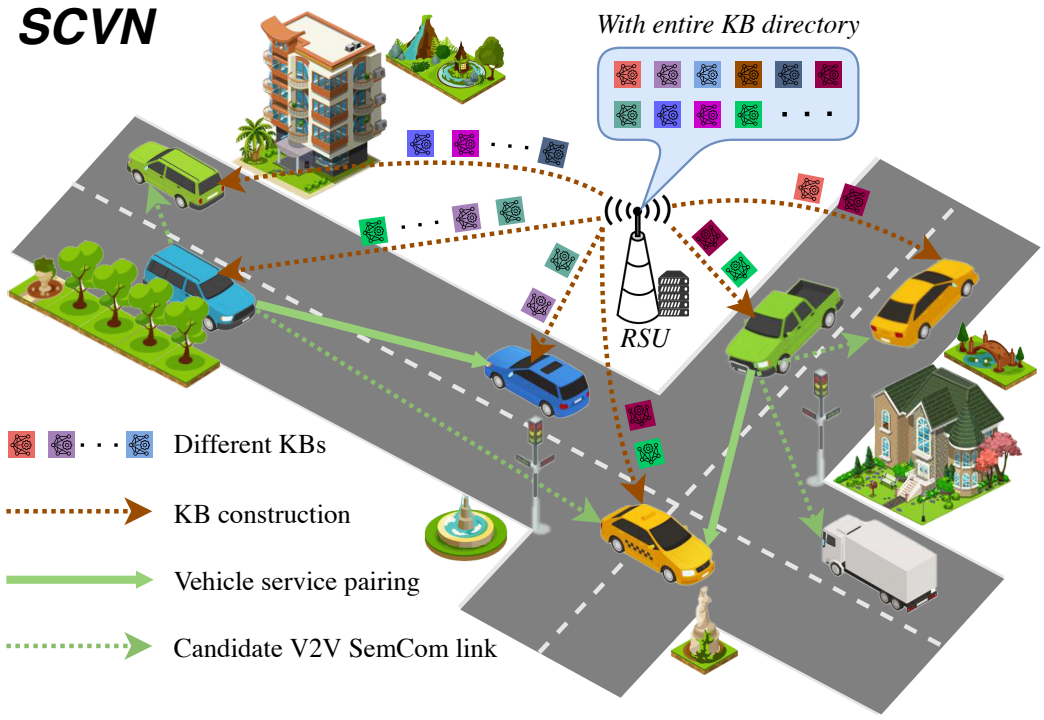

Figure 6.1: The SCVN scenario with KBC and VSP.

- We develop an efficient solution named S^4 to tackle the above optimization problem, and its optimality is theoretically proved by two propositions. Specifically, a primal-dual problem transformation method is first exploited in S^4 to obtain the corresponding Lagrange dual problem, followed by a two-stage method dedicated to solving multiple subproblems with a low computational complexity. Given the dual variables in each iteration, the first stage is to obtain the optimal KBC sub-policy for each potential VUE pair, whereby the second stage is able to finalize the optimal solutions of KBC and VSP for all VUEs in SCVN.

6.2 System Model

In this section, the considered SCVN scenario is first elaborated along with the knowledge storage model and vehicle pairing model. Then, the knowledge matching based queuing latency for semantic packets is derived.

6.2.1 SCVN Scenario

Consider an SCVN scenario as shown in Fig. 6.1, the total of V VUEs are distributed within the coverage of a single roadside unit (RSU), and each VUE $i \in \mathcal{V} = \{1, 2, \ldots, V\}$ is capable of providing SemCom-empowered services to

others. For the wireless propagation model, let $\gamma_{i,j}$ denote the SINR experienced by the V2V link between VUE i and VUE j ($j \neq i$). Essentially, we allow VUE i to communicate with VUE j if their SINR value $\gamma_{i,j}$ is above a prescribed threshold γ_0. In this manner, the set of communication neighbors of VUE i is defined as $\mathcal{V}_i = \{j | j \in \mathcal{V}, j \neq i, \gamma_{i,j} \geqslant \gamma_0\}$, $\forall i \in \mathcal{V}$. Moreover, it is known that the RSU has powerful communication, computing, and storage capabilities and can provide stable communication coverage [40]. Hence, in this work, let the RSU act as a semantic service controller in the SCVN to efficiently schedule and coordinate the whole SemCom-empowered service provisioning process based on the request and state information received from all participating VUEs within its coverage.

6.2.2 Vehicular Knowledge Storage Model

Due to the unique mechanism of semantic interpretation, the acquisition of necessary background knowledge is inevitable for all SemCom-enabled transceivers. In this work, assuming that all VUEs are able to proactively download and construct their respective required KBs from the RSU, where each VUE i has a finite capacity C_i for its local KB storage.

Meanwhile, suppose that there is a KB library $\mathcal{K}$ with a total of N differing KBs in the considered SCVN, and each requires a unique storage size s_n, $n \in \mathcal{K} = \{1, 2, \ldots, N\}$. Furthermore, we define a binary KBC indicator as

$$\alpha_i^n = \begin{cases} 1, & \text{if KB } n \text{ is constructed at VUE } i; \\ 0, & \text{otherwise.} \end{cases} \tag{6.1}$$

It is worth mentioning that the same KB cannot be constructed repeatedly at one VUE for reducing redundancy and for promoting the storage efficiency.

Besides, it is noticed that different VUEs may have different preferences for these KBs corresponding to their required services, thus resulting in the diversity of KB popularity. Naturally, the more popular the KBs, the higher the KBC probabilities. Therefore, without loss of generality, we assume that the KB popularity at each VUE follows Zipf distribution [144].[1] Hence, the probability of VUE i requesting its desired KB n-based services (generating the corresponding semantic data packets) is $p_i^n = (r_i^n)^{-\xi_i} / \sum_{e \in \mathcal{K}} e^{-\xi_i}, \forall (i, n) \in \mathcal{V} \times \mathcal{K}$, where ξ_i ($\xi_i \geqslant 0$) is the skewness of the Zipf distribution, and r_i^n is the popularity rank of KB n at VUE i.[2] Based on p_i^n, we specially develop a KBC-related metric η_i,

[1] Other known probability distributions can also be adopted without changing the remaining modeling and solution.

[2] The KB popularity rank of each VUE can be estimated based on its historical messaging records, which will not be discussed in this work.

namely *knowledge preference satisfaction*, to measure the satisfaction degree of VUE i constructing its interested KBs as

$$\eta_i = \sum_{n \in \mathcal{K}} \alpha_i^n p_i^n. \tag{6.2}$$

It is further required that $\eta_i \geqslant \eta_0$, where η_0 is the unified minimum threshold that needs to be achieved at each VUE.

6.2.3 Vehicle Pairing Model for SemCom

Apart from equipping with suitable KBs, it is also essential for each VUE to select an appropriate VUE from its neighbors for SemCom-empowered service provisioning. It is worthwhile to re-emphasize that the necessary condition for performing SemCom is that the two VUEs (transmitter and receiver) hold common KBs. Moreover, the single association is required for all VUEs in the SCVN for practical purposes, i.e., each VUE can be paired with only one (another) VUE at a time.

Let $\beta_{i \looparrowright j}$ denote the binary VSP indicator for a VUE i-VUE j pair (suppose that VUE i is the sender and VUE j is the receiver), where

$$\beta_{i \looparrowright j} = \begin{cases} 1, & \text{if VUE } i \text{ is associated with VUE } j; \\ 0, & \text{otherwise.} \end{cases} \tag{6.3}$$

Note that the presented communication performance (such as latency, reliability, and throughput) should be different when swapping the roles of sender and receiver in the same VUE pair, since the KBs utilized for SemCom are determined by the sender's preference. For this reason, we use the notation $\looparrowright$ here as an auxiliary illustration to specify the roles of the sender and receiver in each VUE pair.

6.2.4 Knowledge Matching Based Queuing Model

As depicted in Fig. 6.2, the knowledge matching based semantic packet queuing delay is employed as the latency metric of SemCom, to characterize the average sojourn time of semantic data packets in the receiver VUE's queue buffer (following the first-come first-serve rule). For better illustration, three major differences between SemCom-based and traditional communication-based queuing models are listed below: 1) Each semantic data packet is associated with a specific service type, i.e., a certain KB; 2) Semantic data packets generated based

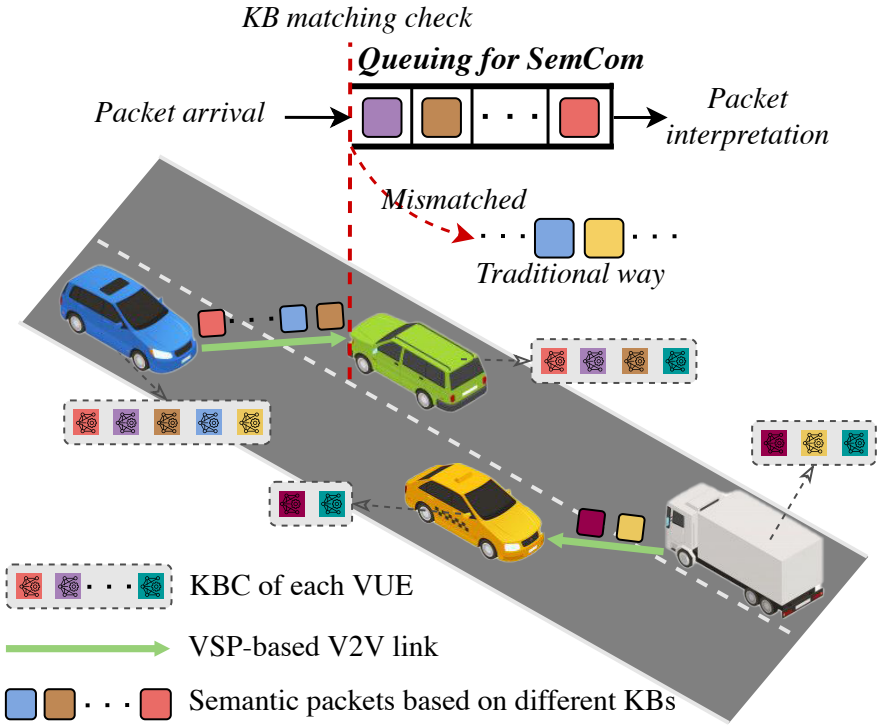

Figure 6.2: The knowledge matching based queuing model for semantic data packets transmitted between VUEs in the SCVN.

on different KBs can co-exist in the queue, and have independent average arrival rate and interpretation time; 3) Not all semantic data packets arriving at the receiver VUE are always allowed to enter its queue, as some of them may mismatch the KBs currently held, rendering these packets uninterpretable. To avoid pointless queuing, these mismatched packets may have to choose traditional communication channels for information transfer, and will not be counted in the arrival process of the queue.

To preserve generality, we first suppose a Poisson data arrival process with average rate $\lambda_i^n = \lambda_i p_i^n$ for a sender VUE i to account for its local semantic packet generation based on KB n, where λ_i is the total arrival rate of all semantic packets at VUE i. In line with this, we can obtain the overall arrival rate of semantic packets from sender VUE i to receiver VUE j as $\sum_{n\in\mathcal{K}} \alpha_i^n \lambda_i^n$, and the effective arrival rate of semantic packets (i.e., these KB-matched semantic packets) in the queue is given as $\sum_{n\in\mathcal{K}} \alpha_i^n \alpha_j^n \lambda_i^n$, thereby the arrival rate of mismatched semantic packets should be the value of the former minus the latter, that is, $\sum_{n\in\mathcal{K}} \alpha_i^n \left(1-\alpha_j^n\right) \lambda_i^n$. Herein, denoting the ratio of the arrival rate of mismatched semantic packets to the arrival rate of all received semantic packets at VUE i-VUE j pair as $\theta_{i\nrightarrow j}$,

namely *knowledge mismatch degree*, which is explicitly calculated by

$$\theta_{i\looparrowright j} = \frac{\sum_{n\in\mathcal{K}} \alpha_i^n \left(1-\alpha_j^n\right)\lambda_i^n}{\sum_{n\in\mathcal{K}} \alpha_i^n \lambda_i^n}. \tag{6.4}$$

In parallel, let a random variable I_j^n denote the Markovian interpretation time [145] required by KB n-based packets at VUE j with mean $1/\mu_j^n$, which is determined by the computing capability of the vehicle and the type of the desired KB. However, since multiple packets based on different KBs are allowed to queue at the same time, it is seen that the interpretation time distribution for a receiver VUE should be treated as a general distribution [134]. If further taking into account the KB popularity, we can calculate the ratio of the amount of KB n-based packets to the total packets in the VUE i-VUE j pair's queue by $\epsilon_{i\looparrowright j}^n = p_i^n / \sum_{f\in\mathcal{K}} \alpha_i^f \alpha_j^f p_i^f$. With the independence among packets based on different KBs, the interpretation time required by each packet in the queue is now expressed as $W_{i\looparrowright j} = \sum_{n\in\mathcal{K}} \alpha_i^n \alpha_j^n \epsilon_{i\looparrowright j}^n I_j^n$.

Since the Markovian arrival process leads to the correlated packet arrivals while the service pattern of packets obeys a general distribution, the queue of each VUE pair can be modeled as an M/G/1 system, which has been widely used to model data traffic in wireless networks. According to the *Pollaczek-Khintchine formula* [133], the average queuing latency for the VUE i-VUE j pair, denoted as $\delta_{i\looparrowright j}$, is determined as follows[3]

$$\delta_{i\looparrowright j} = \frac{\lambda_{i\looparrowright j}^{eff} \cdot \left(\mathbb{E}^2\left[W_{i\looparrowright j}\right] + Var\left(W_{i\looparrowright j}\right)\right)}{2\left(1-\lambda_{i\looparrowright j}^{eff}\cdot\mathbb{E}\left[W_{i\looparrowright j}\right]\right)}. \tag{6.5}$$

On this basis, again leveraging the independence of I_j^n over n, we can then obtain the expectation of the interpretation time for all semantic data packets in the queue by

$$\mathbb{E}\left[W_{i\looparrowright j}\right] = \sum_{n\in\mathcal{K}} \alpha_i^n \alpha_j^n \epsilon_{i\looparrowright j}^n \mathbb{E}\left[I_j^n\right] = \sum_{n\in\mathcal{K}} \frac{\alpha_i^n \alpha_j^n \epsilon_{i\looparrowright j}^n}{\mu_j^n}, \tag{6.6}$$

and the variance of $W_{i\looparrowright j}$ is given by

$$Var\left(W_{i\looparrowright j}\right) = \sum_{n\in\mathcal{K}} \alpha_i^n \alpha_j^n \left(\epsilon_{i\looparrowright j}^n\right)^2 Var\left[I_j^n\right] = \sum_{n\in\mathcal{K}} \alpha_i^n \alpha_j^n \left(\frac{\epsilon_{i\looparrowright j}^n}{\mu_j^n}\right)^2. \tag{6.7}$$

[3]In order to guarantee the steady-state of the queuing system, a condition of $\lambda_{i\looparrowright j}^{eff}\mathbb{E}\left[W_{i\looparrowright j}\right] < 1$ must be satisfied before proceeding [134]. In this work, we assume that the packet interpretation rate is larger than the packet arrival rate to make the queuing latency finite and thus solvable.

By substituting (6.6) and (6.7) into (6.5), $\delta_{i\looparrowright j}$ can be rewritten as

$$\delta_{i\looparrowright j}=\frac{\left[\left(\sum_{n\in\mathcal{K}}\alpha_i^n\alpha_j^n\frac{p_i^n/\mu_j^n}{\sum_{f\in\mathcal{K}}\alpha_i^f\alpha_j^f p_i^f}\right)^2+\sum_{n\in\mathcal{K}}\alpha_i^n\alpha_j^n\left(\frac{p_i^n/\mu_j^n}{\sum_{f\in\mathcal{K}}\alpha_i^f\alpha_j^f p_i^f}\right)^2\right]\left(\sum_{n\in\mathcal{K}}\alpha_i^n\alpha_j^n\lambda_i^n\right)}{2\left[1-\left(\sum_{n\in\mathcal{K}}\alpha_i^n\alpha_j^n\lambda_i^n\right)\cdot\left(\sum_{n\in\mathcal{K}}\alpha_i^n\alpha_j^n\frac{p_i^n/\mu_j^n}{\sum_{f\in\mathcal{K}}\alpha_i^f\alpha_j^f p_i^f}\right)\right]}. \tag{6.8}$$

6.3 Problem Formulation

In line with the xURLLC requirements, it is of paramount importance to achieve the optimality of the overall queuing delay in the SCVN, while being subject to several SemCom-relevant reliability requirements as well as practical system constraints. To that end, we identify and formulate a latency-minimization problem in a joint optimization manner of the KBC indicator α_i^n and the VSP indicator $\beta_{i\looparrowright j}$. For ease of illustration, we define a matrix $\boldsymbol{\alpha}=\{\alpha_i^n|i\in\mathcal{V},n\in\mathcal{K}\}$ and a matrix $\boldsymbol{\beta}=\{\beta_{i\looparrowright j}|i\in\mathcal{V},j\in\mathcal{V}_i\}$ consisting of all binary variables of KBC and VSP, respectively. On the basis of these, our joint optimization problem is formulated as follows:

$$\mathbf{P0}:\ \min_{\boldsymbol{\alpha},\boldsymbol{\beta}}\quad\sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{V}_i}\beta_{i\looparrowright j}\delta_{i\looparrowright j} \tag{6.9}$$

$$\text{s.t.}\quad\sum_{n\in\mathcal{K}}\alpha_i^n\cdot s_n\leqslant C_i,\ \forall i\in\mathcal{V}, \tag{6.9a}$$

$$\eta_i\geqslant\eta_0,\ \forall i\in\mathcal{V}, \tag{6.9b}$$

$$\sum_{j\in\mathcal{V}_i}\beta_{i\looparrowright j}=1,\ \forall i\in\mathcal{V}, \tag{6.9c}$$

$$\beta_{i\looparrowright j}=\beta_{j\looparrowright i},\ \forall(i,j)\in\mathcal{V}\times\mathcal{V}_i, \tag{6.9d}$$

$$\sum_{j\in\mathcal{V}_i}\beta_{i\looparrowright j}\theta_{i\looparrowright j}\leqslant\theta_0,\ \forall i\in\mathcal{V}, \tag{6.9e}$$

$$\alpha_i^n\in\{0,1\},\ \forall(i,n)\in\mathcal{V}\times\mathcal{K}, \tag{6.9f}$$

$$\beta_{i\looparrowright j}\in\{0,1\},\ \forall(i,j)\in\mathcal{V}\times\mathcal{V}_i. \tag{6.9g}$$

Constraint (6.9a) ensures that the total size of KBs constructed at each vehicle cannot exceed its maximum storage capacity, while constraint (6.9b) corresponds to the aforementioned knowledge preference satisfaction requirement for each VUE. Constraints (6.9c) and (6.9d) mathematically model the single-association requirement of VUEs. Constraint (6.9e) represents that the knowledge mismatch degree of each VUE pair should not be over the threshold θ_0, which guarantees

sufficiently high reliability of semantic information delivery. Constraints (6.9f) and (6.9g) characterize the binary properties of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively.

Carefully examining **P0**, it is seen that addressing this problem is rather challenging due to several inevitable mathematical obstacles. First of all, **P0** is an NP-hard optimization problem as demonstrated below. Consider a special case of **P0** where all $\boldsymbol{\beta}$-related constraints are satisfied. In this case, constraints (6.9c)-(6.9e) and (6.9g) can all be removed, and the primal problem degenerates into a classical 0-1 multi-knapsack problem that is known to be NP-hard [146]. Hence, **P0** is also NP-hard. Another nontrivial point originates from the complicated objective function, which prevents us from using the conventional two-step solution (i.e., relaxation and recovery) to approach optimality. In more detail, the problem after relaxing $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ should still be a nonconvex optimization problem owing to the non-convexity preserved in the objective function (6.9) and constraint (6.9e). Therefore, a severe performance penalty will be incurred from the procedure of integer recovery due to the huge performance compromise on solving the nonconvex problem for relaxed variables [137, 147, 148]. In view of the above mathematical challenges, we propose an efficient solution S^4 in the subsequent section to solve **P0** and obtain the joint optimal KBC and VSP solution.

6.4 Proposed S^4 Solution

In this section, we illustrate how to design our proposed solution S^4 to cope with the SemCom-empowered service provisioning problem **P0** in vehicular networks. As depicted in Fig. 6.3, a Lagrange dual method is first leveraged to eliminate the cross-term constraints in **P0** with a corresponding dual optimization problem transformed (referring to **D0** in Section IV.A). Then given the dual variable in each iteration, we dedicatedly develop a two-stage method to determine $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ for the dual problem, where the optimality will be theoretically proved in Section IV.B. Specifically, in the first stage, we subtly construct U ($U = \left(\sum_{i\in\mathcal{V}} |\mathcal{V}_i|\right)/2$) subproblems (referring to $\mathbf{P1}_{i,j}, \forall\,(i,j) \in \mathcal{V} \times \mathcal{V}_i, j > i$), each of which aims to independently seek the optimal KBC sub-policy (with respect to only α_i^n and α_j^n) for each individual VUE pair (as detailed in Section IV.C). After solving all the U subproblems, the optimal coefficient matrix $\boldsymbol{\Omega}$ is obtained for all potential VUE pairs in the SCVN, by which we further construct a new subproblem (referring to **P2**) in the second stage to find the optimal VSP strategy for $\boldsymbol{\beta}$ (as detailed in Section IV.D). In the end, we present the workflow of S^4 along with its complexity analysis in Section IV.E.

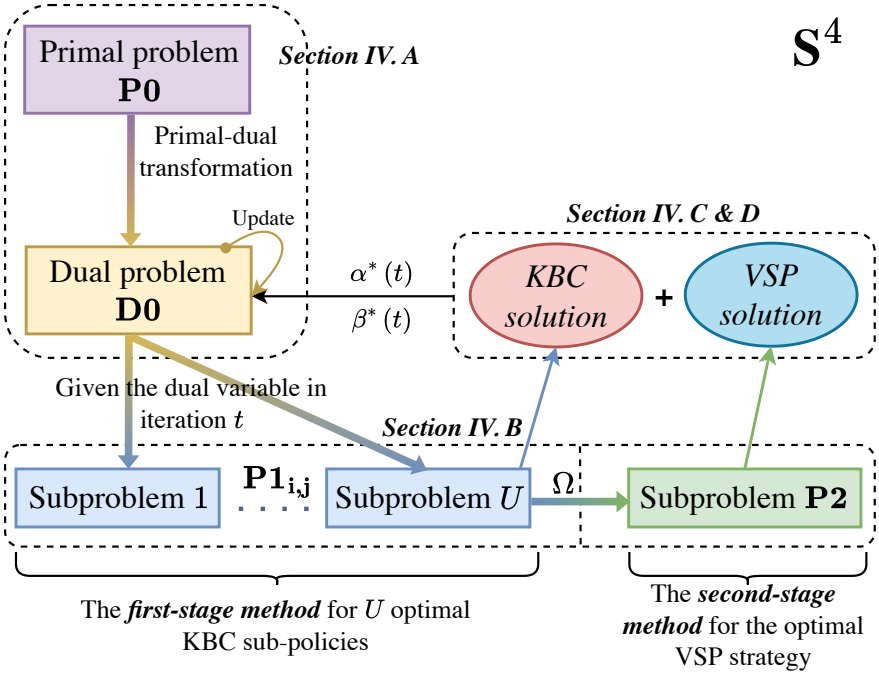

Figure 6.3: Illustration of the proposed solution S^4.

6.4.1 Primal-Dual Problem Transformation

We first incorporate constraint (6.9e) into the objective function (6.9) by associating a Lagrange multiplier $\boldsymbol{\tau} = \{\tau_i | i \in \mathcal{V}\}$. That way, the associated Lagrange function is obtained by

$$\begin{aligned} L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}) &= \sum_{i\in\mathcal{V}} \sum_{j\in\mathcal{V}_i} \beta_{i\looparrowright j} \delta_{i\looparrowright j} + \sum_{i\in\mathcal{V}} \tau_i \left(\sum_{j\in\mathcal{V}_i} \beta_{i\looparrowright j} \theta_{i\looparrowright j} - \theta_0 \right) \\ &= \sum_{i\in\mathcal{V}} \sum_{j\in\mathcal{V}_i} \beta_{i\looparrowright j} \left(\delta_{i\looparrowright j} + \tau_i \theta_{i\looparrowright j} \right) - \theta_0 \sum_{i\in\mathcal{V}} \tau_i \\ &\triangleq \widetilde{L}_{\boldsymbol{\tau}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) - \theta_0 \sum_{i\in\mathcal{V}} \tau_i, \end{aligned} \tag{6.10}$$

where $\widetilde{L}_{\boldsymbol{\tau}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is defined for expression brevity. Then, the Lagrange dual problem of **P0** should be formulated as

$$\mathbf{D0}: \quad \max_{\boldsymbol{\tau}} \quad D(\boldsymbol{\tau}) = g_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{\tau}) - \theta_0 \sum_{i\in\mathcal{V}} \tau_i \tag{6.11}$$

$$\text{s.t.} \quad \tau_i \geqslant 0, \ \forall i \in \mathcal{V}, \tag{6.11a}$$

where we have

$$\begin{aligned} g_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{\tau}) \ = \ & \inf_{\boldsymbol{\alpha},\boldsymbol{\beta}} \ \widetilde{L}_{\boldsymbol{\tau}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \\ \text{s.t.} \ \ & (6.9\text{a}) - (6.9\text{d}), (6.9\text{f}), (6.9\text{g}). \end{aligned} \tag{6.12}$$

Notably, the optimality of the convex problem **D0** gives at least the best lower bound of **P0**, even if **P0** is nonconvex, according to the duality property [101]. Hence, our focus now naturally shifts to seeking the optimal solution to **D0**.

Given the initial dual variable $\boldsymbol{\tau}$, we can solve problem (6.12) in the first place to find the optimal solution $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, the details of which will be presented in the next subsection. After that, a subgradient method is employed for updating $\boldsymbol{\tau}$ to solve **D0** in an iterative fashion, as shown in Fig. 6.3. Specifically, the partial derivatives with respect to $\boldsymbol{\tau}$ in the objective function $D(\boldsymbol{\tau})$ are set as the subgradient direction in each iteration. Now suppose that in a certain iteration, say iteration t, each dual variable $\tau_i(t)$ $(i \in \mathcal{V})$ is updated as

$$\tau_i(t+1) = \left[\tau_i(t) - \nu(t) \cdot \left(\theta_0 - \sum_{j \in \mathcal{V}_i} \beta_{i \looparrowright j}(t)\, \theta_{i \looparrowright j}(t)\right)\right]^+. \tag{6.13}$$

The operator $[\cdot]^+$ here is to output the maximum value between its argument and zero, ensuring that $\boldsymbol{\tau}$ must be non-negative as constrained in (6.11a). $\nu(t)$ is the stepsize in iteration t and generally, the convergence of the subgradient descent method can be ensured with the proper stepsize [138].

6.4.2 Two-Stage Method Based on KBC and VSP

As discussed before, for a given $\boldsymbol{\tau}$ in each iteration, the optimal $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ need to be determined by solving problem (6.12). However, solving such a problem is still rather tricky due to the mathematical inseparability of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in the highly complex objective function $\widetilde{L}_{\boldsymbol{\tau}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$. To this end, we propose a two-stage method to obtain the exactly optimal $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ with a low computational complexity.

In the first stage, we focus on multiple independent KB construction subproblems, each corresponding to a potential VUE pair in the SCVN. In particular, here the performances of the VUE i-VUE j pair (i.e., the sender VUE i and the receiver VUE j) and the VUE j-VUE i pair (i.e., the sender VUE j and the receiver VUE i) need to be considered together, and for ease of distinction, we refer to the two as a *VUE i, j pair*, $\forall (i,j) \in \mathcal{V} \times \mathcal{V}_i, j > i$. In other words, for any KBC subproblem, we have $\beta_{i \looparrowright j} = \beta_{j \looparrowright i} = 1$ in $\widetilde{L}_{\boldsymbol{\tau}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ corresponding to a given VUE i, j pair, while all other VUE pairs are not considered. Therefore, different KBC subproblems can be solved independently, and in this way, let

$$\omega_{i,j} = (\delta_{i \looparrowright j} + \tau_i \theta_{i \looparrowright j}) + (\delta_{j \looparrowright i} + \tau_j \theta_{j \looparrowright i}). \tag{6.14}$$

Obviously, we have $\omega_{i,j} = \omega_{j,i}$, thus only one case of $j > i$ needs to be investigated

for each potential VUE i, j pair.

In this context, we now construct $U = \left(\sum_{i\in\mathcal{V}} |\mathcal{V}_i|\right)/2$ subproblems, each of which is denoted as $\mathbf{P1}_{i,j}$ to seek the optimal KBC sub-policy only for an individual VUE i, j pair. Herein, it is worth pointing out that the optimal KBC solution to problem (6.12) cannot be achieved by simply combining the obtained sub-policies of these $\mathbf{P1}_{i,j}$, but these sub-policies will be used to construct the subsequent VSP subproblem to finalize the joint optimal solution of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ for (6.12). Given the dual variable $\boldsymbol{\tau}$ in each iteration,[4] $\mathbf{P1}_{i,j}$ becomes

$$\mathbf{P1}_{i,j}: \min_{\{\alpha_i^n\},\{\alpha_j^n\}} \quad \omega_{i,j} \tag{6.15}$$

$$\text{s.t.} \quad \sum_{n\in\mathcal{K}} \alpha_i^n \cdot s_n \leqslant C_i, \tag{6.15a}$$

$$\sum_{n\in\mathcal{K}} \alpha_j^n \cdot s_n \leqslant C_j, \tag{6.15b}$$

$$\eta_i \geqslant \eta_0,\ \eta_j \geqslant \eta_0, \tag{6.15c}$$

$$\alpha_i^n \in \{0,1\}, \alpha_j^n \in \{0,1\},\ \forall n \in \mathcal{K}. \tag{6.15d}$$

By solving $\mathbf{P1}_{i,j}$,[5] we can obtain the optimal KBC sub-policies for VUE i (denoted as $\boldsymbol{\alpha}^*_{i_{(j)}}$) and VUE j (denoted as $\boldsymbol{\alpha}^*_{j_{(i)}}$),[6] corresponding to the individual VUE i, j pair. The following proposition explicitly shows how the sub-policy of $\mathbf{P1}_{i,j}$ correlates to the solution of problem (6.12).

Proposition 6. *Let $\boldsymbol{\alpha}^* = [\boldsymbol{\alpha}_1^*, \boldsymbol{\alpha}_2^*, \cdots, \boldsymbol{\alpha}_V^*]^T$ be the optimal KBC solution to the problem in (6.12) given the dual variable $\boldsymbol{\tau}$, where $\boldsymbol{\alpha}_i^*$ represents the optimal KBC policy of VUE i. Then we have $\forall i \in \mathcal{V}$, $\exists j \in \mathcal{V}_i$, such that $\boldsymbol{\alpha}^*_{i_{(j)}} = \boldsymbol{\alpha}_i^*$.*

Proof. Please see Appendix F. □

From Proposition 6, it is observed that the optimal KBC policy of each VUE can be found by solving a certain $\mathbf{P1}_{i,j}$. Hence, considering the single-association requirement of V2V pairing, the optimal VSP strategy becomes the only key to finalize the optimal solution to (6.12). To achieve this, we first obtain the optimal coefficient matrix for $\boldsymbol{\beta}$ in (6.12) to account for all VSP possibilities. By calculating optimum $\omega_{i,j}$ (denoted as $\omega^*_{i,j}, \forall (i,j) \in \mathcal{V} \times \mathcal{V}_i$) in $\mathbf{P1}_{i,j}$, the optimal

[4]For simplicity, we omit $\boldsymbol{\tau}$ from all notations associated with $\mathbf{P1}_{i,j}$ and $\mathbf{P2}$ in this chapter.

[5]The solution details of $\mathbf{P1}_{i,j}$ as well as $\mathbf{P2}$ will be introduced in the subsequent Subsection C and D, respectively.

[6]For auxiliary illustration, we use $(\cdot)$ in the subscript to specify the VUE pair attribute (relation) for each VUE's KBC sub-policy obtained from $\mathbf{P1}_{i,j}$.

coefficient matrix is formed as

$$\mathbf{\Omega} = \begin{bmatrix} +\infty & \omega_{1,2}^* & \omega_{1,3}^* & \cdots & \omega_{1,V}^* \\ \omega_{2,1}^* & +\infty & \omega_{2,3}^* & \cdots & \omega_{2,V}^* \\ \omega_{3,1}^* & \omega_{3,2}^* & +\infty & \cdots & \omega_{3,V}^* \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \omega_{V,1}^* & \omega_{V,2}^* & \omega_{V,3}^* & \cdots & +\infty \end{bmatrix}. \tag{6.16}$$

$\mathbf{\Omega}$ is a $V \times V$ symmetric matrix where $\omega_{i,j}^* = \omega_{j,i}^*$, and all elements on its main diagonal are set to $+\infty$ to indicate the fact that a VUE cannot communicate with itself, i.e., $j \neq i$. Besides, note that some $\omega_{i,j}^*$s in $\mathbf{\Omega}$ also have a value $+\infty$ if VUE j is not the direct neighbor of VUE i, i.e., $j \notin \mathcal{V}_i$.

Next, we concentrate upon the optimal vehicle service pairing strategy by constructing a new subproblem in the second stage. In line with the objective $\widetilde{L}_{\boldsymbol{\tau}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $\boldsymbol{\beta}$-related constraints in (6.12), the VSP subproblem is written as

$$\mathbf{P2}: \min_{\boldsymbol{\beta}} \quad \frac{1}{2} \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{V}_i} \beta_{i \curvearrowright j} \omega_{i,j}^* \tag{6.17}$$

$$\text{s.t.} \quad (6.9\text{c}), (6.9\text{d}), (6.9\text{g}). \tag{6.17a}$$

Given any $\boldsymbol{\tau}$, the optimal $\boldsymbol{\beta}$ (denoted as $\boldsymbol{\beta}^* = \left[\beta_{1 \curvearrowright j_1^*}, \beta_{2 \curvearrowright j_2^*}, \cdots, \beta_{V \curvearrowright j_V^*}\right]^T$) can be directly finalized by solving **P2**, where $\beta_{i \curvearrowright j_i^*}$ $(\forall i \in \mathcal{V})$ indicates that VUE j_i^* is the optimal SemCom node for VUE i, i.e., $\beta_{i \curvearrowright j_i^*} = 1$. Afterward, we feed back the obtained $\boldsymbol{\beta}^*$ to $\mathbf{\Omega}$ to further finalize the optimal KBC policy $\boldsymbol{\alpha}^*$ for all VUEs. In the context of the solution to $\mathbf{P1}_{i,j}$, the approach to finalize $\boldsymbol{\alpha}^*$ can be stated more precisely as: for any $i \in \mathcal{V}$, we have $\boldsymbol{\alpha}_i^* = \boldsymbol{\alpha}_{i_{(j_i^*)}}^*$. The rationale behind this is established in accordance with the following proposition.

Proposition 7. *Given any dual variable $\boldsymbol{\tau}$, $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ is exactly the optimal solution to the problem in (6.12).*

Proof. Please see Appendix G. □

From Proposition 7, it is seen that for problem (6.12) in each iteration, the proposed two-stage method is ensured to find the optimal solution. Apart from this, the optimization difficulty of each subproblem, either $\mathbf{P1}_{i,j}$ or **P2**, is considered to be greatly decreased due to the reduced number of the optimization variables. In what follows, we will present our optimal solutions to $\mathbf{P1}_{i,j}$ and **P2**, respectively.

6.4.3 Near-Optimal Solution for KBC

Carefully examining $\mathbf{P1}_{i,j}, \forall\,(i,j) \in \mathcal{V} \times \mathcal{V}_i, j > i$, it can be observed that $\delta_{i\nrightarrow j}$ mainly makes the objective function $\omega_{i,j}$ still highly complex and nonconvex with two binary variables $\boldsymbol{\alpha}_i$ and $\boldsymbol{\alpha}_j$. To that end, a modified metaheuristic algorithm based on tabu search (TS) is employed here to efficiently determine a near-optimal KB construction policy for two VUEs in the VUE i,j pair with the consideration of SemCom features. In detail, the KBC solution is illustrated as follows:

- *Initial Feasible Solution Generation:* As the iterative search algorithm, an initial feasible solution (denoted as a $2N$-dimensional vector $\boldsymbol{\alpha}_{i,j}^I$) is needed as the search starting point [149]. To speed up convergence and enhance optimization performance, we heuristically adopt a KB preference and KB matching-aware approach to generate $\boldsymbol{\alpha}_{i,j}^I$ for a better initial solution performance. More concretely, first let two N-dimensional vectors $\boldsymbol{\alpha}_i^I$ and $\boldsymbol{\alpha}_j^I$ denote the KBC solutions of VUE i and VUE j, respectively, with all elements being 0 for initialization. Meanwhile, suppose there are two variable sets, denoted as $\hat{\mathcal{K}}$ and $\check{\mathcal{K}}$, to record KB-relevant information, where $\hat{\mathcal{K}} = \mathcal{K}$ and $\check{\mathcal{K}} = \emptyset$ are initialized. With these, we attempt to find a KB n_0 with the highest sum of KB preferences of the two VUEs by

$$n_0 = \arg\max_{n \in \hat{\mathcal{K}}} \left(p_i^n + p_j^n\right). \tag{6.18}$$

Then, in order to meet the knowledge mismatch requirement, the values of both $\boldsymbol{\alpha}_i^I$ and $\boldsymbol{\alpha}_j^I$ are updated as

$$\alpha_i^{n_0} = 1 \quad \text{and} \quad \alpha_j^{n_0} = 1. \tag{6.19}$$

Next, we let $\hat{\mathcal{K}} = \hat{\mathcal{K}} \backslash n_0$ and $\check{\mathcal{K}} = \check{\mathcal{K}} \cup \{n_0\}$, and then repeat the two procedures in (6.18) and (6.19) until both VUEs satisfy the minimum knowledge preference satisfaction requirement in constraint (6.15c). However, notice that constraint (6.15a) or (6.15b) may be violated during the above process, in which case we need to find the maximum-size KB in $\check{\mathcal{K}}$ by

$$n_1 = \arg\max_{n \in \check{\mathcal{K}}} s_n, \tag{6.20}$$

and then reset the corresponding KBC indicators in $\boldsymbol{\alpha}_i^I$ and $\boldsymbol{\alpha}_j^I$ to 0, i.e.,

$$\alpha_i^{n_1} = 0 \quad \text{and} \quad \alpha_j^{n_1} = 0. \tag{6.21}$$

As a result, we obtain an initial feasible solution

$$\boldsymbol{\alpha}_{i,j}^{I} = \left[\boldsymbol{\alpha}_{i}^{I}, \boldsymbol{\alpha}_{j}^{I}\right]. \tag{6.22}$$

- *Neighboring Solution Searching:* Let a $2N$-dimensional vector $\boldsymbol{\alpha}_{i,j}^{C}$ store the current solution in each search iteration, and $\mathcal{H}\left(\boldsymbol{\alpha}_{i,j}^{C}\right)$ denote its neighboring solution set, which should not include solutions that are already recorded in a tabu list, denoted as a set $\mathcal{I}$ (which will be explained later). Naturally, our $\mathcal{H}\left(\boldsymbol{\alpha}_{i,j}^{C}\right)$ is defined as

$$\mathcal{H}(\boldsymbol{\alpha}_{i,j}^{C}) = \left\{\boldsymbol{\alpha}_{i,j} \colon \|\boldsymbol{\alpha}_{i,j} - \boldsymbol{\alpha}_{i,j}^{C}\| \leqslant \sigma, \boldsymbol{\alpha}_{i,j} \notin \mathcal{I}, \boldsymbol{\alpha}_{i,j} \in \psi\right\}, \tag{6.23}$$

  where σ is the size of the maximum neighborhood space, and ψ represents the feasible solution space of $\mathbf{P1}_{i,j}$. Based on the above definitions, we are able to find the best solution within the current $\mathcal{H}\left(\boldsymbol{\alpha}_{i,j}^{C}\right)$ that can yield the minimum value of $\omega_{i,j}$ in an iterative fashion.

- *Tabu List Update:* The tabu list $\mathcal{I}$ is a special memory mechanism for preventing subsequent searches from looping back to previously visited solutions so as to avoid trapping into the local optimum [150]. In this way, whenever there is a newly obtained $\boldsymbol{\alpha}_{i,j}^{C}$, it should be added into $\mathcal{I}$ (cannot exceed its given maximum length [149]). Besides, let $\boldsymbol{\alpha}_{i,j}^{*}$ denote another vector dedicated to storing the best solution obtained so far. Particularly, once $\boldsymbol{\alpha}_{i,j}^{C}$ is found to be better than $\boldsymbol{\alpha}_{i,j}^{*}$ in any iteration, it will not be added into $\mathcal{I}$ but we will have

$$\boldsymbol{\alpha}_{i,j}^{*} = \boldsymbol{\alpha}_{i,j}^{C}. \tag{6.24}$$

- *Algorithm Termination Check:* Before commencing a new iteration, an algorithm termination criterion needs to be checked, which can be either a maximum iteration restriction or a performance improvement threshold of $\boldsymbol{\alpha}_{i,j}^{*}$ under a certain number of consecutive iterations.

6.4.4 Optimal Solution for VSP

After the KBC sub-policy $\boldsymbol{\alpha}_{i,j}^{*}$ is found for each potential VUE i, j pair, we can determine the optimal coefficient matrix $\boldsymbol{\Omega}$ to solve the VSP subproblem **P2**. Since its objective function and two equality constraints (6.9c) and (6.9d) are all linear, the only challenge is the 0-1 constraint in (6.9g).

With regards to this, we first relax $\boldsymbol{\beta}$ into a continuous variable between 0 and 1 to make **P2** a linear programming problem, which can be efficiently solved with toolboxes such as CVXPY [102]. Then the obtained continuous solution, denoted as $\boldsymbol{\beta}^R$, needs to be restored to the binary state under the original constraints. Here, we heuristically finalize the optimal VSP strategy $\boldsymbol{\beta}^*$ by

$$\beta^*_{i' \curvearrowright j'} = \beta^*_{j' \curvearrowright i'} = 1, \tag{6.25}$$

if and only if

$$\left(i^{'}, j^{'}\right) = \arg \max_{i \in \mathcal{V}, j \in \mathcal{V}_i, j > i} \beta^R_{i \curvearrowright j}. \tag{6.26}$$

Meanwhile, for the remaining $\beta^*_{i \curvearrowright j}$ with respect to VUE $i^{'}$ and VUE $j^{'}$, we naturally have

$$\begin{cases} \beta^*_{i' \curvearrowright j} = \beta^*_{j \curvearrowright i'} = 0, & \forall j \in \mathcal{V}_{i'}, j \neq j^{'} \\ \beta^*_{j' \curvearrowright i} = \beta^*_{i \curvearrowright j'} = 0, & \forall i \in \mathcal{V}_{j'}, i \neq i^{'} \end{cases}. \tag{6.27}$$

Then we let $\mathcal{V} = \mathcal{V} \backslash \{i, j\}$, and repeat the above progresses until determining the optimal VSP solution for all VUEs. It can be seen that the number of variables is actually only $\left(\sum_{i \in \mathcal{V}} |\mathcal{V}_i|\right)/2$ when solving **P2**, which is a fairly acceptable problem scale in practice. Hence, the performance compromise of our proposed heuristic VSP solution is believed to be small.

6.4.5 Workflow of S^4 and Complexity Analysis

In order to further demonstrate the full picture of the proposed solution S^4, we summarize the relevant technical points and present them in the following Algorithm 1.

Algorithm 1 The Proposed Solution S^4

Input: The SCVN parameters s_n, C_i, r_i^n, ξ_i, λ_i, μ_i^n, η_0, θ_0

Output: The optimal KBC policy $\boldsymbol{\alpha}^*$ and the optimal VSP strategy $\boldsymbol{\beta}^*$ for each VUE $i, \forall i \in \mathcal{V}$

Initialize $t = 0$, $\tau_i(0)$ and $\nu(0)$ to proper positive values;
Set the maximum number of iterations M for **D0**;
while $t < M$ **do**
  for $i = 1$ to V **do**
    for each $j \in \mathcal{V}_i$ **do**
      if $j > i$ **then**
        Determine the initial solution $\boldsymbol{\alpha}^I_{i,j}$ by (6.18)-(6.22);
        Initialize the TS iteration as $\tilde{t} = 0$, the Tabu list
          $\mathcal{I}(0) = \emptyset$, and $\boldsymbol{\alpha}^C_{i,j}(0) = \boldsymbol{\alpha}^*_{i,j}(0) = \boldsymbol{\alpha}^I_{i,j}$;

Set the neighborhood size σ and the maximum
number of iterations Z for solving each $\mathbf{P1}_{i,j}$;
while $\tilde{t} < Z$ **do**
Determine $\mathcal{H}\left(\boldsymbol{\alpha}_{i,j}^{C}\left(\tilde{t}\right)\right)$ by (6.23);
Find the best feasible solution in $\mathcal{H}\left(\boldsymbol{\alpha}_{i,j}^{C}\left(\tilde{t}\right)\right)$
and assign it to $\boldsymbol{\alpha}_{i,j}^{C}(\tilde{t}+1)$;
if $\boldsymbol{\alpha}_{i,j}^{C}(\tilde{t}+1)$ is better than $\boldsymbol{\alpha}_{i,j}^{*}(\tilde{t})$ **then**
Update $\boldsymbol{\alpha}_{i,j}^{*}(\tilde{t}+1) = \boldsymbol{\alpha}_{i,j}^{C}(\tilde{t}+1)$;
Keep $\mathcal{I}\left(\tilde{t}+1\right) = \mathcal{I}\left(\tilde{t}\right)$;
else
Keep $\boldsymbol{\alpha}_{i,j}^{*}(\tilde{t}+1) = \boldsymbol{\alpha}_{i,j}^{*}(\tilde{t})$;
Update $\mathcal{I}\left(\tilde{t}+1\right) = \mathcal{I}\left(\tilde{t}\right) \cup \left\{\boldsymbol{\alpha}_{i,j}^{C}(\tilde{t}+1)\right\}$;
end if
Update $\tilde{t} = \tilde{t} + 1$;
end while
Calculate $\omega_{i,j}^{*}$ by substituting $\boldsymbol{\alpha}_{i,j}^{*}(Z)$ into (6.14);
Assign $\omega_{j,i}^{*} = \omega_{i,j}^{*}$;
end if
end for
end for
Renew the optimal coefficient matrix $\boldsymbol{\Omega}(t)$ by (6.16);
Solve the relaxed **P2** by CVXPY and obtain $\boldsymbol{\beta}^{R}(t)$;
Finalize $\boldsymbol{\beta}^{*}(t)$ by (6.25)-(6.27);
Finalize $\boldsymbol{\alpha}^{*}(t)$ by feeding $\boldsymbol{\beta}^{*}(t)$ back into $\boldsymbol{\Omega}(t)$;
Update $\tau_i(t+1)$ by (6.13);
Update $\nu(t+1)$ under a given rule;
Update $t = t + 1$;
end while

In line with this algorithm, the main flow of S^4 working in the practical SCVN is demonstrated as follows:

- *Network Initialization:* In the initial phase, each VUE i ($\forall i \in \mathcal{V}$) generates a Lagrange multiplier parameter τ_i and records all KBC-related status information, including its available KB storage (i.e., C_i), preferences for different KBs (i.e., $r_i^n, \forall n \in \mathcal{K}$, and ξ_i), and average local arrival rate as well as interpretation time for semantic data packets (i.e., λ_i and $1/\mu_i^n$). Then, all VUEs need to upload the above parameters to the RSU for subsequent implementation.

- *Optimal Policy Determination for KBC and VSP:* In this phase, the RSU first measures the SINR between VUEs to determine each VUE i's communication neighbors, i.e., $\mathcal{V}_i$. Having these, the RSU is capable of computing the current optimal KBC sub-policy for each individual VUE i, j pair ($j \in \mathcal{V}_i$) (referring to steps 4-29 in Algorithm 1). Afterward, the current optimal VSP and KBC policies can be jointly obtained in steps 30-33. Nevertheless, the Lagrange multiplier of each VUE is required to be updated with a given rule (steps 34-36), thus the RSU should repeat the procedures in steps 4-33 until satisfying the algorithm termination criterion, so as to finalize the optimal KBC and VSP policies $\boldsymbol{\alpha}^*$ and $\boldsymbol{\beta}^*$.

- *SemCom-Empowered Service Provisioning:* Once each VUE receives the feedback information, it can request and download KBs from the RSU (according to $\boldsymbol{\alpha}^*$), and then pair with one of its neighbors (according to $\boldsymbol{\beta}^*$) to provide corresponding services for each other.

Herein, it is worth pointing out that the above workflow of S^4 is executed in a periodic timeline, and all decisions to update relevant parameters should be made at the end of each period. Besides, it can be observed that there are only four rounds of signaling interactions to implement a completed and successive KBC and VSP process, including vehicular information collection, optimal KBC and VSP policy assignment, KB downloading, and vehicle pairing. For each signaling interplay, only a few bits are needed to complete the functional confirmation work, hence the overall signaling overhead should be an apparently tolerable level in practice.

In terms of the computational complexity of S^4, it is first seen that for a single $\mathbf{P1}_{i,j}$, each feasible KBC solution within $\mathcal{H}\left(\boldsymbol{\alpha}_{i,j}^C\right)$ needs to be computed once in any of its iterations. Combining that σ is given as a small parameter compared to N in (6.23), the complexity in each iteration can be estimated by $\mathcal{O}\left(\binom{2N}{\sigma}\right) = \mathcal{O}\left(N^\sigma\right)$. If the maximum number of its iterations is assumed to be Z, then solving each $\mathbf{P1}_{i,j}$ would require complexity $\mathcal{O}\left(ZN^\sigma\right)$. In addition, the linear programming method is utilized for the relaxed $\mathbf{P2}$, where $\mathcal{O}\left(V^4\right)$ complexity is needed [140] to solve a group of $\left(\sum_{i\in\mathcal{V}}|\mathcal{V}_i|\right)$ VSP variables. Moreover, note that in any iteration of $\mathbf{D0}$, a total of $\left(\left(\sum_{i\in\mathcal{V}}|\mathcal{V}_i|\right)/2\right)$ subproblems $\mathbf{P1}_{i,j}$ and one subproblem $\mathbf{P2}$ need to be solved simultaneously, thereby its corresponding complexity is $\mathcal{O}\left(V^2ZN^\sigma + V^4\right)$. If denoting the maximum number of iterations that can make $\mathbf{D0}$ to converge as M, the proposed S^4 would have a polynomial-time overall complexity, given as $\mathcal{O}\left(MV^2\left(ZN^\sigma + V^2\right)\right)$.

Table 6.1: Simulation Parameters

Parameters	**Values**
Cell radius of the RSU	500 m
VUE drop model	Spatial Poisson process [151]
Number of lanes	3 in each direction (6 in total)
Lane width	4 m
Absolute velocity of VUEs	70 km/h [136]
Density of VUEs	Average inter-vehicle distance is 2.5 sec $\times$ absolute velocity of VUEs [151]
Transmit power of VUEs	20 dBm [152]
Noise power	-114 dBm
Path loss model	$128.1 + 37.6 \log(d \text{ [km]})$ [130]
Channel fading model	Log-normal shadowing distribution with standard deviation of 8 dB + Rayleigh fast fading [130]

6.5 Numerical Results and Discussions

In this section, numerical evaluations are conducted to demonstrate the performance of the proposed solution S^4 in SCVNs, where we employ Python 3.7-based PyCharm as the simulator platform and implement it in a computer with six CPU cores and Inter Core i7 processor. To preserve generality, we model a multi-lane freeway passing through a single cell with the RSU at its center, and 60 VUEs are dropped on the lanes according to the spatial Poisson process [151]. For brevity, some simulation parameters not mentioned in the context as well as their corresponding values can be found in Table 6.1 [130, 136, 151, 152]. As for the settings relevant to SemCom, a total of 12 different KBs are preset to provide VUEs with a variety of distinct services, and each of them has a storage size randomly distributed from 1 to 5 units. Correspondingly, we set a uniform KB storage capacity of 24 units for all VUEs. Besides, each VUE's preference ranking for all KBs (i.e., r_i^n) is generated independently and randomly, where their respective Zipf distributions are assumed to have the same skewness 1.0. Likewise, either the average arrival rate of total semantic data packets or the average interpretation time for packets based on the same KB n, is considered to be the same for all VUEs. Here, we fix the average total arrival rate λ_i at 100 packets/s [142]

and randomly generate the value of $1/\mu_i^n$ in a range of $5 \times 10^{-3} \sim 1 \times 10^{-2}$ s/packet with respect to different KB n, as the average interpretation time of different KBs-based packets is different from each other, which is to guarantee the steady-state of the queuing system at each VUE pair, as has mentioned in Footnote 4. Further, the minimum knowledge preference satisfaction threshold η_0 and the maximum knowledge mismatch degree threshold θ_0 are prescribed as 0.5 and 0.1, respectively. Notably, all these parameter values are set by default unless otherwise specified, and all subsequent numerical results are obtained by averaging over a sufficiently large number of trials.

For comparison purposes, we utilize two different benchmarks of SemCom-empowered service provisioning herein: 1) Distance-first pairing (DFP) strategy which assumes each VUE to choose its nearest unpaired VUE for V2V pairing; 2) Knowledge-first pairing (KFP) in which each VUE selects its neighboring unpaired VUE with the highest KB matching degree for V2V pairing. In the meantime, a personal preference-first KBC policy is considered for both benchmarks, which allows each VUE to construct KBs with the highest preferences until η_0 is satisfied, and then randomly select these unconstructed KBs until reaching respective maximum capacity.

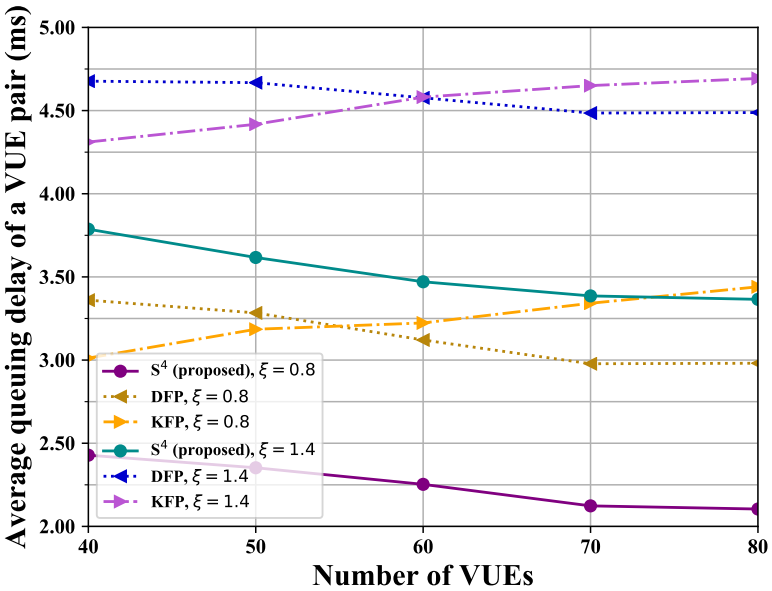

Figure 6.4: Average queuing latency of a VUE pair vs. varying numbers of VUEs.

Fig. 6.4 first depicts the average queuing latency performance of a VUE pair against varying numbers of VUEs, where two different KB preference skewness $\xi = 0.8$ and $\xi = 1.4$ are considered. In this figure, the latency of S^4 declines at the beginning with the number of VUEs, then remains stable beyond 70 VUEs, and it can always outperform both benchmarks with an average latency reduction of

around 1 ms at any ξ. The rationale behind this trend is that the more neighbors each VUE can have, the better chance of achieving the low queuing latency for each VUE pair, which will be eventually stabilized when reaching the respective best achievable latency with a fixed bandwidth budget. Moreover, it is observed that a larger ξ causes a higher latency penalty, since the vast majority of VUEs' KBC is concentrated on a small number of KBs when ξ increases. Clearly, a larger ξ will make each participant more difficult to find the best VUE with the low latency under the given knowledge mismatching requirement θ_0, thus resulting in a degraded performance.

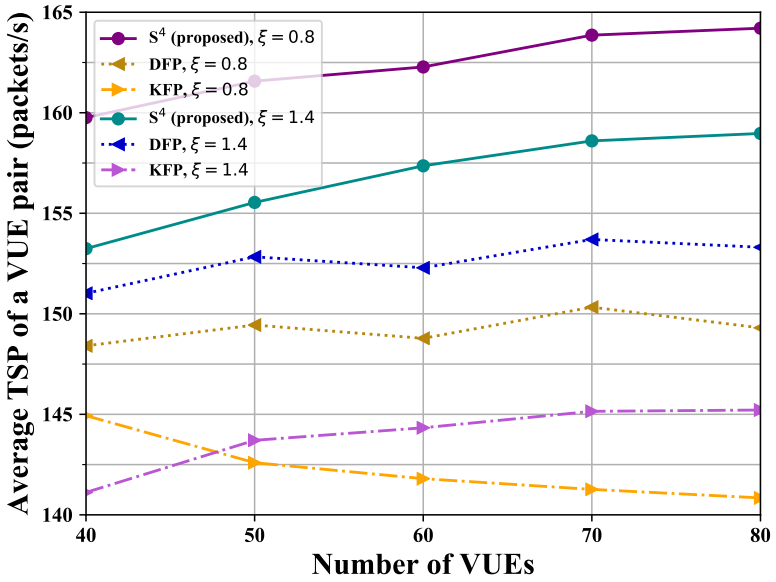

Figure 6.5: Average TSP of a VUE pair vs. varying numbers of VUEs.

The above analysis also applies to Fig. 6.5, which compares all the three methods under the same settings as Fig. 6.4 to demonstrate the performance of the average throughput in semantic packets (TSP). Specifically, the TSP represents the total number of semantic packets that can be interpreted by a VUE pair per second, whose value is determined based on $\mathbb{E}[W_{i \leftrightarrows j}]$ in (6.6). Likewise, a higher TSP is obtained as the number of VUEs increases, and our S^4 is still far better than the two benchmarks at any point, e.g., with an average performance gain of 14 packets/s compared with DFP and 20 packets/s with KFP at $\xi = 0.8$. Again, we see a better TSP when the KB popularity is diluted by a smaller ξ.

Next, we explore the impact of varying number of KBs on the average queuing latency of a VUE pair with different VUE capacities $C = 18$ and $C = 24$, as demonstrated in Fig. 6.6. It can be found that the latency drops fast at the beginning, and then rises slightly after exceeding 10 KBs, whereas the performance of our S^4 still surpasses the benchmarks. This trend is attributed to the fact that

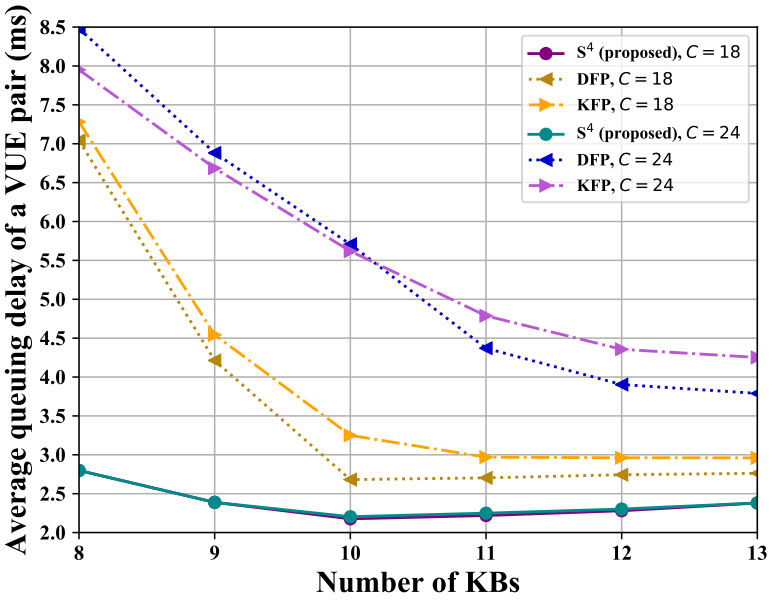

Figure 6.6: Average queuing latency of a VUE pair with varying numbers of KBs.

more KBs imply less discrepancy in VUEs' preferences for different KBs given the fixed ξ, thereby at first leading to the higher probability for two paired VUEs constructing the KBs with high interpretation rates so as to render a lower delay. However, such performance gains will be saturated and even worsen when these KBs with low interpretation rates become inevitably dominant in order to meet the minimum knowledge preference satisfaction threshold η_0. Besides, it is seen that different VUE capacities have little effect on the latency of S^4, although the larger capacity can construct more KBs. This is due to the latency-minimization objective we particularly focus on in the delay-sensitive SCVN, and only the KBs with low interpretation time should be selected. Afterward, we draw the TSP performance against different numbers of KBs with $\eta_0 = 0.4$ and $\eta_0 = 0.6$, as shown in Fig. 6.7. The similar trend to Fig. 6.6 is observed here as well, i.e., the TSP of S^4 rises at the beginning and then falls after 10 KBs, and a much higher TSP is provided compared with the benchmarks. Meanwhile, it is noticed that a lower η_0 brings a better TSP, as fewer KBs need to be constructed to guarantee the high average interpretation rates in the queue, as mentioned earlier.

In addition, we validate the average knowledge preference satisfaction $\bar{\eta} = \frac{1}{V}\sum_{i\in\mathcal{V}}\eta_i$ reached at each VUE with varying numbers of KBs as shown in Fig. 6.8, where $\xi = 0.8$, $\xi = 1.4$, $\theta_0 = 0.1$, and $\theta_0 = 0.2$ are taken into account. As the number of KBs increases, a lower $\bar{\eta}$ is obtained, which is to prevent these unnecessary KBs from being constructed while satisfying η_0 to the greatest extent. For the two curves with different ξ, referring to the analysis of Fig. 6.4, a higher ξ indicates a more concentrated KB preference, which means some extra KBs need

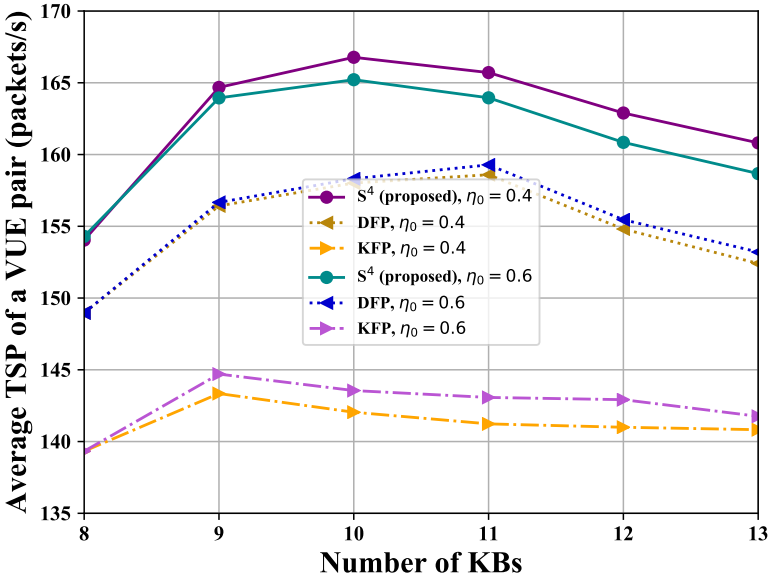

Figure 6.7: Average TSP of a VUE pair with varying numbers of KBs.

to be constructed to meet the maximum θ_0 requirement. Because of this, we also see a lower $\bar{\eta}$ at a higher θ_0, since a more tolerable knowledge mismatch degree is more likely to avoid the unnecessary KBC.

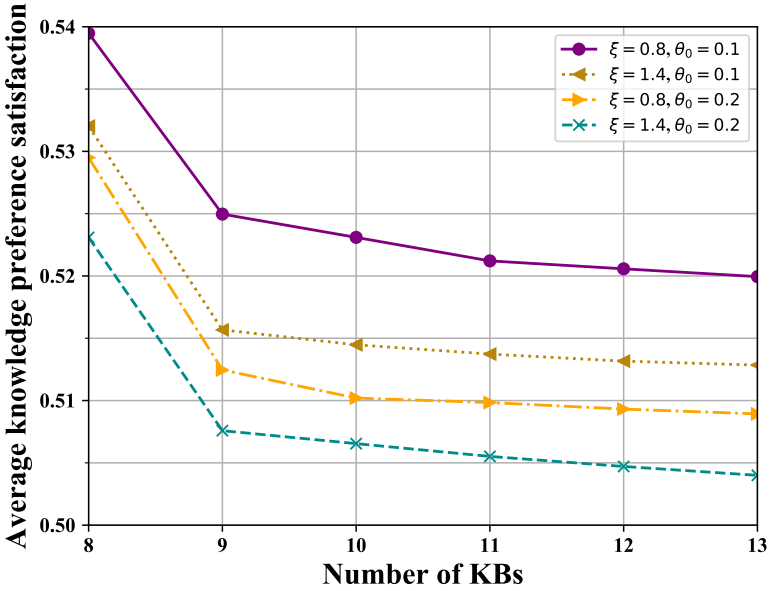

Figure 6.8: Average knowledge preference satisfaction reached at a VUE with varying numbers of KBs.

Fig. 6.9 presents the effect of varying ξ on the average queuing delay compared between different η_0 and θ_0. As expected, the latency of all three methods increases with ξ, and such an upward trend is consistent with the previous results. Specially, it is observed that the curve of $\eta_0 = 0.4$ and $\theta_0 = 0.2$ brings the best latency performance. This can be understood as that either the lower satisfaction

threshold or the higher mismatch tolerance can avoid the construction of unnecessary KBs with low interpretation rates, as discussed before. Naturally, the better latency performance is obtained when the constraints become less stringent.

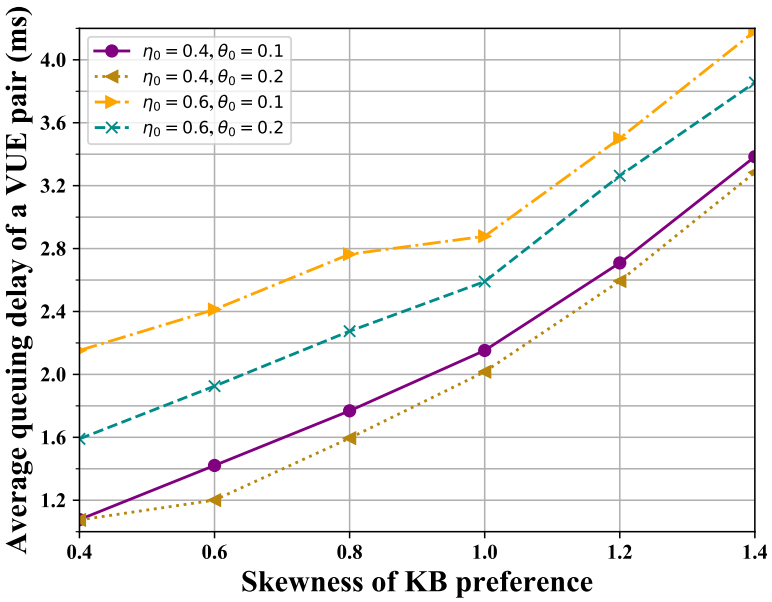

Figure 6.9: Average queuing latency of a VUE pair with varying skewness of VUEs' KB preferences.

Finally, we plot the average knowledge matching degree, defined as $\bar{\rho} = \frac{1}{V}\sum_{i\in\mathcal{V}, j\in\mathcal{V}_i}(1-\theta_{i\hookrightarrow j})$, with two different ξ in Fig. 6.10. It can be seen that under the threshold of $\theta_0 = 0.1$, the proposed solution S^4 is always above a 97.5% match degree, which is much higher than that of benchmarks. Furthermore, a higher ξ causes the slight drop of $\bar{\rho}$, which exactly proves the conclusion of Fig. 6.4, i.e., a more concentrated KB preference makes it more difficult for VUEs to find a highly matching neighbor in the VSP phase, especially for the one that can bring lower latency and meet all constraints at the same time.

6.6 Conclusions

In this chapter, we proposed a novel solution S^4 to tackle the SemCom-empowered service provisioning problem in the SCVN. To align with the stringent xURLLC requirements, the KB matching based queuing latency expression of semantic data packets was first derived, and then we identified and formulated the fundamental problem of KBC and VSP to minimize the queuing latency for all VUE pairs. After the primal-dual problem transformation, a two-stage method was developed specifically to jointly solve multiple subproblems related to KBC and VSP with low computational complexity, and the solution optimality has been

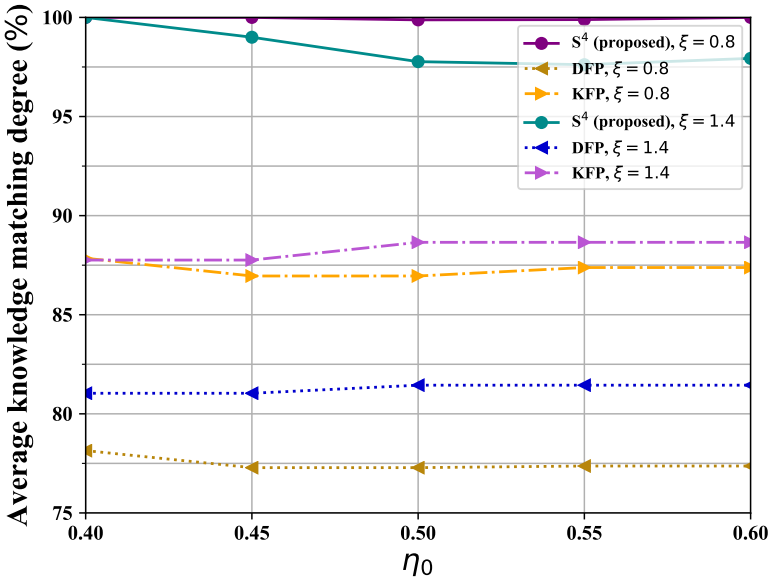

Figure 6.10: Average knowledge matching degree of a VUE pair vs. different knowledge preference satisfaction requirements.

theoretically proved. Numerical results verified the sufficient performance superiority of S^4 in terms of both latency and reliability by comparing it with two different benchmarks.

Chapter 7

Conclusions and Future Trends

In full view of the promising application prospects of SemCom in next-generation wireless communication networks, a profound cognition of the pertinent resource management becomes particularly meaningful and indispensable. In this chapter, the main work and contributions of this thesis are drawn, together with a brief discussion on the future trends of wireless resource management in intelligent SemCom-enabled networks.

7.1 Conclusions

This thesis separately discussed the resource management optimization in four differing SemCom cellular network architectures, including the SC-Net, the EE-SCN, the HSB-Net, and the SCVN. Extensive numerical simulations were conducted for each network scenario and the results consistently proved the superiority and reliability of our proposed solutions in terms of diverse performance metrics (such as STM, queuing latency, packet drop ratio, energy efficiency, TSP, and knowledge preference satisfaction) compared with differing benchmarks. The core contributions of this thesis can be summarized as follows.

The first work regarding the SC-Net conducted a systematic study on SemCom from a networking perspective. Specifically, the concept of B2M transformation was introduced to measure the overall network performance metric of STM in the SC-Net. Being aware of different knowledge-matching states of each MU, the SC-Net was specially categorized into two different types of PKM-based and IKM-based SC-Nets. In each SC-Net case, we formulated a corresponding STM-maximization problem jointly considering the UA and BA factors. Then, we proposed the optimal resource management solution for each SC-Net scenario, where a primal-dual decomposition method with a Lagrange-multiplier method was employed for the PKM-based one, and a chance-constrained model with an

interior-point method plus a heuristic algorithm was developed for the IKM-based one.

In the second work, we studied the optimal spectrum reusing pattern for each D2D SemCom user and the the optimal power allocation scheme for each SemCom user in a single-cell EE-SCN. Among them, the power consumption model was defined by considering different knowledge-matching states of all users, and thus determining the energy efficiency expression. By leveraging a factional-to-subtractive transformation method and a two-stage method, the corresponding energy efficiency-optimization problem is solved.

Next, the resource management in the novel yet practical network scenario of HSB-Net was investigated, in which two modes of SemCom and BitCom coexist. A two-stage tandem queuing system was dedicatedly modeled to identify the transmission process of semantic packets, followed by the theoretical derivation for average packet loss ratio and queuing latency. Having these as the SemCom-related constraints, a joint STM-maximization problem was formulated, followed by a Lagrange primal-dual method and a preference list-based heuristic algorithm proposed as the UA, MS, and BA solutions with low computational complexity.

Finally, for the last SCVN scenario, we proposed a novel solution S^4 to address the SemCom-empowered service provisioning problem that involves the KBC and VSP. With the awareness of xURLLC requirements, the KB matching based queuing latency of semantic data packets was first derived, and a corresponding latency-minimization problem was formulated. To cope with this problem, we utilized a primal-dual problem transformation and a two-stage method to specifically solve multiple subproblems related to KBC and VSP with low computational complexity, where the solution optimality has been theoretically proved.

7.2 Future Trends

The next-generation SemCom cellular networks promise to provide significant improvements for resource utilization and information interaction. In spite of many superiorities, the proposed resource management schemes in this thesis still impose some associated and nontrivial challenges that should be discussed before unlocking the full potential of SemCom networks. For instance, it turns out that KB matching should be a crucial factor in affecting the quality of SemCom service provisioning and the STM performance of SC-Net, therefore, a pressing need for effective KB matching algorithms inevitably arises in future research venues. Besides, this work provides foundations to properly generalize resource management solutions to other complicated SC-Net scenarios, e.g., PKM-IKM-coexistence net-

works. Moreover, some insights obtained from this work should inspire new resource management strategies in alignment with different new SemCom-relevant objectives, such as the accuracy of message interpretation, latency of end-to-end links, and user fairness in a semantical sense. Particularly, if the knowledge-matching condition is unknown in practice, how to accurately measure it as well as the semantic performance should be an interesting research direction. Apart from that, the connection between SemCom and energy efficiency still remains on the packet level, which can be extended to a more general and reasonable aspect such as the semantic level. Meanwhile, the performance metric of SemCom still follows the previous work, i.e., using STM, which may have other more suitable choices, such as semantic similarity or semantic value. Note that there may exist other forms of semantic-coding-related energy consumption, which is worth exploring in future works. In addition, other relevant networking issues in the HSB-Net, such as communication mode switching or semantic fairness-driven power or resource block allocation, inevitably arise, which can treat this work as the fundamental theoretical framework for reference. Since this work is limited to long-term network optimization under known knowledge-matching degrees, the further problem about instantaneous decision-making for MS and BA in the unaware background knowledge condition could be future research. Similarly, the semantic performance metric can be considered to be replaced with other metrics to enhance the generality of this work. Furthermore, in terms of the scenario of applying SemCom in vehicular networks, other advanced networking issues, such as semantic-aware resource allocation or semantic transceiver design, can treat our fourth work as the fundamental theoretical framework for reference. Especially considering the practical cases, the mobility of VUEs should be further taken into account while executing the resource scheduling. Since this work is limited to determining optimal instantaneous KBC and VSP policies for VUEs with known KB popularity, the further problem in an expanded SCVN scenario of considering high user mobility and unaware user preferences will be investigated in our future research. One of the other future trends in resource management in SemCom-enabled cellular networks can be how to realize the best tradeoff between computational and communication resources at each mobile device, which problem seems to be tricky due to the sophisticated semantic coding mechanism and limited resources. Further considering the priority of different semantics at different MUs, research regarding the age of semantic information in practical SemCom systems may be another interesting future trend.

Appendix A

Proof of Proposition 1

Consider a downlink case between MU i and BS j, first suppose that there is a set of source information modeled as a stochastic process $\{W_m\}$ $(m = 1, 2, \ldots, M)$ at the BS j side, where each W_m is independent of each other [13]. Besides, let K_m denote the KB required by source information W_m for SemCom, and let $\mathcal{K}_{ij}$ denote the set of KBs matched between MU i and BS j. Hence, we can define a KB matching indicator for source information as follows:

$$Z_m = \begin{cases} 1, & \text{if } K_m \in \mathcal{K}_{ij} \\ 0, & \text{otherwise} \end{cases}. \tag{A.1}$$

Further given the knowledge matching degree τ_{ij} of the link between MU i and BS j, thus the probability of successful matching of W_m, i.e., the probability of $Z_m = 1$, becomes τ_{ij}. Moreover, based on the same link, its different source information $\{W_m\}$ should have the same probability of successful matching, which is absolutely irrelevant to the knowledge matching situations of other links. As such, $\{Z_m\}$ obeys the identical binomial distribution w.r.t. τ_{ij}, such that

$$\begin{cases} \Pr(Z_m = 1) = \tau_{ij} \\ \Pr(Z_m = 0) = 1 - \tau_{ij} \end{cases}, \tag{A.2}$$

where $\Pr(\cdot)$ is the probability measure.

With these, the random knowledge matching coefficient β_{ij} can be now expressed as the mean of the sum of $\{Z_m\}$ from a statistical average point of view as M approaches infinity, that is,

$$\beta_{ij} = \lim_{M \to +\infty} \frac{1}{M} \sum_{m=1}^{M} Z_m. \tag{A.3}$$

Based on (A.2) and (A.3), the classical central limit theorem [153] can be directly applied to determine the distribution of β_{ij}, i.e., $\beta_{ij} \sim \mathcal{N}\left(\tau_{ij}, \tau_{ij}\left(1-\tau_{ij}\right)\right)$.

Appendix B

Proof of Proposition 2

First let $\left(\boldsymbol{P^{C*}}, \boldsymbol{P^{D*}}, \boldsymbol{\alpha^*}\right)$ and $\left(\widehat{\boldsymbol{P^C}}, \widehat{\boldsymbol{P^D}}, \widehat{\boldsymbol{\alpha}}\right)$ denote the optimal solution and an arbitrary feasible solution to **P0**, respectively. Clearly, we have

$$\eta_{EE}^{*} = \frac{Q^{total}\left(\boldsymbol{P^{C*}}, \boldsymbol{P^{D*}}, \boldsymbol{\alpha^*}\right)}{E^{total}\left(\boldsymbol{P^{C*}}, \boldsymbol{P^{D*}}, \boldsymbol{\alpha^*}\right)} \geqslant \frac{Q^{total}\left(\widehat{\boldsymbol{P^C}}, \widehat{\boldsymbol{P^D}}, \widehat{\boldsymbol{\alpha}}\right)}{E^{total}\left(\widehat{\boldsymbol{P^C}}, \widehat{\boldsymbol{P^D}}, \widehat{\boldsymbol{\alpha}}\right)}. \tag{B.1}$$

It is not difficult to find that $E^{total} > 0$ holds in any solution, the following conclusions can be easily derived from (B.1), i.e.,

$$Q^{total}\left(\boldsymbol{P^{C*}}, \boldsymbol{P^{D*}}, \boldsymbol{\alpha^*}\right) - \eta_{EE}^{*} E^{total}\left(\boldsymbol{P^{C*}}, \boldsymbol{P^{D*}}, \boldsymbol{\alpha^*}\right) = 0, \tag{B.2}$$

and

$$Q^{total}\left(\widehat{\boldsymbol{P^C}}, \widehat{\boldsymbol{P^D}}, \widehat{\boldsymbol{\alpha}}\right) - \eta_{EE}^{*} E^{total}\left(\widehat{\boldsymbol{P^C}}, \widehat{\boldsymbol{P^D}}, \widehat{\boldsymbol{\alpha}}\right) \leqslant 0. \tag{B.3}$$

Combined with the fact that **P0** and **P1** have the same feasible region as they have the identical constraints (4.11a)-(4.11g), it is obvious seen from (B.2) and (B.3) that $\left(\boldsymbol{P^{C*}}, \boldsymbol{P^{D*}}, \boldsymbol{\alpha^*}\right)$ must also be the optimal solution to **P1** if $F\left(\eta_{EE}^{*}\right) = 0$.

Meanwhile, considering the other two remaining cases of $F\left(\eta_{EE}\right) > 0$ and $F\left(\eta_{EE}\right) < 0$ in **P1**, and let $\left(\overline{\boldsymbol{P^{C*}}}, \overline{\boldsymbol{P^{D*}}}, \overline{\boldsymbol{\alpha^*}}\right)$ and $\left(\widetilde{\boldsymbol{P^{C*}}}, \widetilde{\boldsymbol{P^{D*}}}, \widetilde{\boldsymbol{\alpha^*}}\right)$ be their optimal solutions to **P1** respectively. Then we have

$$Q^{total}\left(\overline{\boldsymbol{P^{C*}}}, \overline{\boldsymbol{P^{D*}}}, \overline{\boldsymbol{\alpha^*}}\right) - \eta_{EE} E^{total}\left(\overline{\boldsymbol{P^{C*}}}, \overline{\boldsymbol{P^{D*}}}, \overline{\boldsymbol{\alpha^*}}\right) > 0, \tag{B.4}$$

and

$$Q^{total}\left(\widetilde{\boldsymbol{P^{C*}}}, \widetilde{\boldsymbol{P^{D*}}}, \widetilde{\boldsymbol{\alpha^*}}\right) - \eta_{EE} E^{total}\left(\widetilde{\boldsymbol{P^{C*}}}, \widetilde{\boldsymbol{P^{D*}}}, \widetilde{\boldsymbol{\alpha^*}}\right) < 0. \tag{B.5}$$

Again leveraging $E^{total} > 0$, given any η_{EE}, (B.4) yields

$$\frac{Q^{total}}{E^{total}} = \eta_{EE} < \frac{Q^{total}\left(\overline{\boldsymbol{P^{C*}}}, \overline{\boldsymbol{P^{D*}}}, \overline{\boldsymbol{\alpha^{*}}}\right)}{E^{total}\left(\overline{\boldsymbol{P^{C*}}}, \overline{\boldsymbol{P^{D*}}}, \overline{\boldsymbol{\alpha^{*}}}\right)}, \tag{B.6}$$

and likewise, (B.5) yields

$$\frac{Q^{total}}{E^{total}} = \eta_{EE} > \frac{Q^{total}\left(\widetilde{\boldsymbol{P^{C*}}}, \widetilde{\boldsymbol{P^{D*}}}, \widetilde{\boldsymbol{\alpha^{*}}}\right)}{E^{total}\left(\widetilde{\boldsymbol{P^{C*}}}, \widetilde{\boldsymbol{P^{D*}}}, \widetilde{\boldsymbol{\alpha^{*}}}\right)}. \tag{B.7}$$

From (B.6) and (B.7), it can be concluded that for any feasible solution to **P0**, it must not be the optimal solution to **P1** if $F\left(\eta_{EE}^{*}\right) \neq 0$. This completes the proof.

Appendix C

Proof of Proposition 3

This proposition can be proved by using contradiction. According to all constraints of $\mathbf{P2}_{i,j}$, if the optimization problem is supposed to be solvable, the optimal power allocation solution $\left(P_i^{C^*}, P_j^{D^*}\right)$ must fall into the non-empty ψ. Here, we first assume that $\left(P_i^{C^*}, P_j^{D^*}\right)$ is not the coincide point of the two line segments while making $\widetilde{\lambda_1}\left(P_j^D\right)$ and $\widetilde{\lambda_2}\left(P_i^C\right)$ simultaneously reach their respective maxima, i.e.,

$$\left(P_i^{C^*}, P_j^{D^*}\right) \notin \left\{\left(\overleftarrow{P_i^C}, \overleftarrow{P_j^D}\right) | \left(\overleftarrow{P_i^C}, \overleftarrow{P_j^D}\right) = \left(\overrightarrow{P_i^C}, \overrightarrow{P_j^D}\right) \in \psi\right\}. \tag{C.1}$$

This means $\left(P_i^{C^*}, P_j^{D^*}\right)$ should not be the optimal point for at least one line segment w.r.t. $\widetilde{\lambda_1}\left(P_j^D\right)$ or w.r.t. $\widetilde{\lambda_2}\left(P_i^C\right)$. For illustration, let $\tilde{l}$ denote the line segment through $\left(P_i^{C^*}, P_j^{D^*}\right)$.

However, it is also noticed that ψ must be a closed and bounded region in one of the three cases in Fig. 4.2. Due to the convex property of $\widetilde{\lambda_1}\left(P_j^D\right)$ or $\widetilde{\lambda_2}\left(P_i^C\right)$, there must be another point $\left(\overline{P_i^C}, \overline{P_j^D}\right)$ on the same line segment $\tilde{l}$, leading to a larger value $\widetilde{\lambda_1}\left(P_j^D\right)$ or w.r.t. $\widetilde{\lambda_2}\left(P_i^C\right)$. That is, one of the following cases must exist:

$$\widetilde{\lambda_1}\left(\overline{P_j^D}\right) > \widetilde{\lambda_1}\left(P_j^{D^*}\right) \quad \text{or} \quad \widetilde{\lambda_2}\left(\overline{P_i^C}\right) > \widetilde{\lambda_2}\left(P_i^{C^*}\right). \tag{C.2}$$

Meanwhile, since $\left(P_i^{C^*}, P_j^{D^*}\right)$ and $\left(\overline{P_i^C}, \overline{P_j^D}\right)$ are on the same line segment, this leads to another fact that $\sigma_i^C \overline{r_i^C}$ (in the left case of (C.2)) or $\sigma_j^D \overline{r_j^D}$ (in the right case of (C.2)) gets the same value at the two points. Combined with (4.18), clearly, either of the cases contradicts the assumption that $\left(P_i^{C^*}, P_j^{D^*}\right)$ is the optimal solution in ψ. This ends the proof.

Appendix D

Proof of Proposition 4

It is first found from (5.8) that given any queue length a at slot t (i.e., $Q_{ij}(t) = a$, $0 \leqslant a \leqslant F$), $Q_{ij}(t+1)$ is determined only by $A'_i(t)$ and $D_{ij}(t)$. Apparently, the stochastic $Q_{ij}(t)$ across all slots forms a discrete-time Markov process, herein we define $\omega_{ij}^{a \looparrowright b}(t) = \Pr\{Q_{ij}(t+1) = b \mid Q_{ij}(t) = a\}$ as its one-step state transition probability from state a at slot t to state b at slot $t+1$, $0 \leqslant b \leqslant F$. Since the PMFs of both $A'_i(t)$ and $D_{ij}(t)$ are independent of t, as mentioned earlier, we re-denote them by A'_i and D_{ij} for brevity, respectively. As such, $\omega_{ij}^{a \looparrowright b}(t)$ can be expressed as $\omega_{ij}^{a \looparrowright b}$ as well.

Having these, we have the one-step state transition probability matrix of SemCom-enabled PTQ as

$$\mathbf{\Omega}_{ij} = \begin{bmatrix} \omega_{ij}^{0 \looparrowright 0} & \omega_{ij}^{0 \looparrowright 1} & \cdots & \omega_{ij}^{0 \looparrowright F} \\ \omega_{ij}^{1 \looparrowright 0} & \omega_{ij}^{1 \looparrowright 1} & \cdots & \omega_{ij}^{1 \looparrowright F} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{ij}^{F \looparrowright 0} & \omega_{ij}^{F \looparrowright 1} & \cdots & \omega_{ij}^{F \looparrowright F} \end{bmatrix}, \tag{D.1}$$

where each $\omega_{ij}^{a\hookrightarrow b}$ can be explicitly calculated by

$$\omega_{ij}^{a\hookrightarrow b}=\begin{cases}\Pr\{A_i'=b\}, & \text{if } a=0, 1\leqslant b\leqslant F-1;\\ \sum_{k=F}^{\infty}\Pr\{A_i'=k\}, & \text{if } a=0, b=F;\\ \Pr\{A_i'=0\}\sum_{k=a}^{\infty}\Pr\{D_{ij}=k\}, & \text{if } 0\leqslant a\leqslant F,\ b=0;\\ \Pr\{A_i'=b\}\sum_{k=a}^{\infty}\Pr\{D_{ij}=k\}+\sum_{l=0}^{b-1}\Pr\{A_i'=l\}\Pr\{D_{ij}=a-b+l\}, & \\ \qquad\qquad \text{if } 1\leqslant a\leqslant F, 1\leqslant b\leqslant a, b\neq F;\\ \Pr\{A_i'=b\}\sum_{k=a}^{\infty}\Pr\{D_{ij}=k\}+\sum_{l=0}^{a-1}\Pr\{D_{ij}=l\}\Pr\{A_i'=b-a+l\}, & \\ \qquad\qquad \text{if } 1\leqslant a<b\leqslant F-1;\\ \sum_{k=b}^{\infty}\Pr\{A_i'=k\}\sum_{l=a}^{\infty}\Pr\{D_{ij}=l\} & \\ \quad+\sum_{l=0}^{a-1}\left(\Pr\{D_{ij}=l\}\sum_{k=b-a+l}^{\infty}\Pr\{A_i'=k\}\right), & \\ \qquad\qquad \text{if } 1\leqslant a\leqslant F,\ b=F.\end{cases}\tag{D.2}$$

Further noticing that for the queue state transited from $Q_{ij}(t)=0$ to $Q_{ij}(t+1)=0$, the transition probability is

$$\begin{aligned}\Pr\{Q_{ij}(t+1)=0\mid Q_{ij}(t)=0\}&=\Pr\{A_i'=0\}\\&=\exp\left(-\tau_i\mu_i^{Mat}T-(1-\tau_i)\mu_i^{Mis}T\right)>0,\end{aligned}\tag{D.3}$$

which proves that $Q_{ij}(t)=0$ is aperiodic. Besides, combined with a fact that each $\omega_{ij}^{a\hookrightarrow b}$ is time independent and each $Q_{ij}(t)$ has a finite state space, $\{Q_{ij}(t)\mid t=1,2,\cdots,N\}$ is time-homogeneous, irreducible, and aperiodic. Therefore, according to [154], there must be a unique steady-state probability vector $\boldsymbol{\alpha}_{ij}=\left[\alpha_{ij}^0,\alpha_{ij}^1,\cdots,\alpha_{ij}^F\right]^T$, which can be obtained by simultaneously solving

$$\boldsymbol{\Omega}_{ij}^T\boldsymbol{\alpha}_{ij}=\boldsymbol{\alpha}_{ij}\qquad\text{and}\qquad\sum_{k=0}^{F}\alpha_{ij}^k=1.\tag{D.4}$$

This completes the proof.

Appendix E

Proof of Proposition 5

It is worth noting in the first place that here we only show the proof in the SemCom-enabled queuing model case of $y_{ij} = 1$ for exemplification (given any pair of MU i and BS j), since the proof in the BitCom case can be similarly derived based on their analogous modeling for PTQ. Notice that $\delta_{ij}^{S_1}$ is a known constant as in (5.11), $\delta_{ij}^{S_2}$ in (5.14) and θ_{ij}^{S} in (5.13) become the only two factors that z_{ij} can influence. Further combining that A_i' is irrelevant with z_{ij} as in (5.6), let us at first present a lemma of how z_{ij} relates the distribution of D_{ij}.

Lemma 1: The CDF of D_{ij} decreases as z_{ij} increases.

To prove Lemma 1, we first derive the CDF of D_{ij} from its PMF given in (5.5) as follows:

$$\begin{aligned}
\Pr\{D_{ij} \leqslant k\} &= \sum_{f=0}^{k} \Pr\{D_{ij} = f\} \\
&= \Pr\left\{\gamma_{ij} \leqslant 2^{\frac{(k+1)L}{Tz_{ij}}} - 1\right\} - \Pr\left\{\gamma_{ij} \leqslant 2^{\frac{kL}{Tz_{ij}}} - 1\right\} \\
&\quad + \Pr\left\{\gamma_{ij} \leqslant 2^{\frac{kL}{Tz_{ij}}} - 1\right\} - \Pr\left\{\gamma_{ij} \leqslant 2^{\frac{(k-1)L}{Tz_{ij}}} - 1\right\} \\
&\quad + \Pr\left\{\gamma_{ij} \leqslant 2^{\frac{(k-1)L}{Tz_{ij}}} - 1\right\} - \Pr\left\{\gamma_{ij} \leqslant 2^{\frac{(k-2)L}{Tz_{ij}}} - 1\right\} \\
&\quad + \cdots\cdots + \Pr\left\{\gamma_{ij} \leqslant 2^{\frac{L}{Tz_{ij}}} - 1\right\} - \Pr\{\gamma_{ij} \leqslant 0\} \\
&= \Pr\left\{\gamma_{ij} \leqslant 2^{\frac{(k+1)L}{Tz_{ij}}} - 1\right\} - \Pr\{\gamma_{ij} \leqslant 0\}, \ k = 0, 1, 2, \cdots,
\end{aligned} \tag{E.1}$$

where slot index t is omitted from all notations associated with the SINR γ_{ij} for brevity due to its independence w.r.t. t as aforementioned. Given arbitrary known CDF of γ_{ij}, which is independent with z_{ij}, we clearly have that $\Pr\{D_{ij} \leqslant k\}$ is a monotonically decreasing function of z_{ij}. This also implies that $\Pr\{D_{ij} \geqslant k\}$ monotonically increases w.r.t. z_{ij}.

Now, let us consider two complementary queuing state subspaces of queue length Q_{ij}, denoted by $\overleftarrow{\mathcal{F}_c} = \{0, 1, 2, \cdots, c\}$ and $\overrightarrow{\mathcal{F}_c} = \{c+1, c+2, \cdots, F\}$, $c = 0, 1, 2, \cdots, F-1$. Given any current state c, it can only transit to either a smaller state in $\overleftarrow{\mathcal{F}_c}$ or a larger state in $\overrightarrow{\mathcal{F}_c}$ in the next step, and the probabilities of the two transition cases occurring sum to 1. According to the one-step transition probability $\omega_{ij}^{a \looparrowright b}$ expressed in (D.2), the probability of state c transiting to any state in $\overleftarrow{\mathcal{F}_c}$ should be computed by

$$\begin{aligned}
&\omega_{ij}^{c \looparrowright 0} + \omega_{ij}^{c \looparrowright 1} + \cdots\cdots + \omega_{ij}^{c \looparrowright c} \\
=\ & \Pr\{A_i' = 0\} + \Pr\{A_i' = 1\} \sum_{l=1}^{\infty} \Pr\{D_{ij} = l\} \\
& + \Pr\{A_i' = 2\} \sum_{l=2}^{\infty} \Pr\{D_{ij} = l\} + \cdots\cdots + \Pr\{A_i' = c\} \sum_{l=c}^{\infty} \Pr\{D_{ij} = l\} \\
=\ & \Pr\{A_i' = 0\} + \sum_{k=1}^{c} \left(\Pr\{A_i' = k\} \sum_{l=k}^{\infty} \Pr\{D_{ij} = l\} \right) \\
=\ & \Pr\{A_i' = 0\} + \sum_{k=1}^{c} \left(\Pr\{A_i' = k\} \Pr\{D_{ij} \geqslant k\} \right).
\end{aligned} \tag{E.2}$$

According to Lemma 1, (E.2) is clearly a monotonically increasing function of z_{ij} due to its $\Pr\{D_{ij} \geqslant k\}$ term. In other words, we have that the probability of any fixed state c transiting to a state in $\overrightarrow{\mathcal{F}_c}$ monotonically decreases w.r.t. z_{ij}. Further combined with the obtained steady-state probability vector $\boldsymbol{\alpha}_{ij}$, if denoting the cumulative distribution of the queuing system staying in the state space $\overleftarrow{\mathcal{F}_c}$ as $W_{ij}^{(c)} = \sum_{l=0}^{c} \alpha_{ij}^{l}$, $W_{ij}^{(c)}$ is increasing w.r.t. z_{ij} for any c as well.

By leveraging a fact that $W_{ij}^{(F)} = 1$, let us first rephrase the numerator term of $\delta_{ij}^{S_2}$ in (5.14) as follows:

$$\begin{aligned}
\mathbb{E}\left[Q_{ij}(t)\right] &= \sum_{k=1}^{F} \alpha_{ij}^{k} + \sum_{k=2}^{F} \alpha_{ij}^{k} + \cdots + \sum_{k=F-1}^{F} \alpha_{ij}^{k} + \alpha_{ij}^{F} \\
&= \left(1 - W_{ij}^{(0)}\right) + \left(1 - W_{ij}^{(1)}\right) + \cdots + \left(1 - W_{ij}^{(F-1)}\right),
\end{aligned} \tag{E.3}$$

whereby the conclusion that $\mathbb{E}\left[Q_{ij}(t)\right]$ is monotonically decreasing w.r.t. z_{ij} holds.

Regarding θ_{ij}^{S} in (5.12), which is also served as the key term in the denominator of $\delta_{ij}^{S_2}$, we restructure the formula by highlighting all its implicit terms that

transform to $\Pr\{D_{ij} \leqslant k\}$ and $W_{ij}^{(c)}$, and obtain

$$\begin{aligned} G_{ij} = \sum_{f=1}^{F} \Pr\{A_i' = f\} \left[\sum_{k=0}^{f-1} \Pr\{D_{ij} \leqslant k\} \left(1 - W_{ij}^{(F-f+k)}\right) \right] \\ + \sum_{f=F+1}^{\infty} \Pr\{A_i' = f\} \left[(f-F) + \sum_{k=0}^{F-1} \Pr\{D_{ij} \leqslant k\} \left(1 - W_{ij}^{(k)}\right) \right]. \end{aligned} \tag{E.4}$$

Again employing Lemma 1, we have that G_{ij} monotonically decreases w.r.t. z_{ij}, thereby θ_{ij}^{S} and $\delta_{ij}^{S_2}$ should have the same decreasing property. Finally, note that $\delta_{ij}^{S} = \delta_{ij}^{S_1} + \delta_{ij}^{S_2} \geqslant \delta_{ij}^{S_1} > 0$ always holds in practice, δ_{ij}^{S} must be monotonically non-increasing w.r.t. z_{ij}, which completes the proof.

Appendix F

Proof of Proposition 6

Given the optimal KBC solution $\boldsymbol{\alpha}^*$, let $\boldsymbol{\beta}^* = \left[\beta_{1\looparrowright j_1^*}, \beta_{2\looparrowright j_2^*}, \cdots, \beta_{V\looparrowright j_V^*}\right]^T$ be the corresponding optimal VSP solution to the problem in (6.12) under the same dual variable $\boldsymbol{\tau}$, where $\beta_{i\looparrowright j_i^*}$ $(\forall i \in \mathcal{V})$ indicates that VUE j_i^* is the optimal SemCom node for VUE i, i.e., $\beta_{i\looparrowright j_i^*} = 1$.

From $\omega_{i,j}$ defined in (6.14), the objective function $\widetilde{L}_{\boldsymbol{\tau}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ in (6.12) can be rewritten as

$$\widetilde{L}_{\boldsymbol{\tau}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2} \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{V}_i} \beta_{i\looparrowright j} \omega_{i,j} = \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{V}_i, j > i} \beta_{i\looparrowright j} \omega_{i,j}, \tag{F.1}$$

then we substitute $\boldsymbol{\beta}^*$ into (F.1) and yield

$$\widetilde{L}_{\boldsymbol{\tau}}(\boldsymbol{\alpha}, \boldsymbol{\beta}^*) = \frac{1}{2} \sum_{i \in \mathcal{V}} \omega_{i,j_i^*} = \sum_{i \in \mathcal{V}, i < j_i^*} \omega_{i,j_i^*}, \tag{F.2}$$

where ω_{i,j_i^*} is the term only related to VUE i, j_i^* pair.

Undoubtedly, if $\boldsymbol{\alpha}^*$ is further substituted into (F.2), we can straightforwardly reach the optimality of the problem in (6.12). Since different VUE i, j_i^* pairs are independent of each other, it means that different terms related to ω_{i,j_i^*} are independent of each other as well in $\widetilde{L}_{\boldsymbol{\tau}}(\boldsymbol{\alpha}, \boldsymbol{\beta}^*)$. Therefore, we can directly draw an important conclusion that achieving the optimality of $\widetilde{L}_{\boldsymbol{\tau}}(\boldsymbol{\alpha}, \boldsymbol{\beta}^*)$ is equivalent to achieving the optimality of each ω_{i,j_i^*}, where the optimality can be reached when $\boldsymbol{\alpha} = \boldsymbol{\alpha}^*$.

In view of the above, we know that $\boldsymbol{\alpha}_i^*$ must be the optimal solution of $\omega_{i,j_i^*}, \forall i \in \mathcal{V}$. Further combined with another fact that $\omega_{i,j}$ is the objective of $\mathbf{P1}_{i,j}, \forall (i,j) \in \mathcal{V} \times \mathcal{V}_i, j > i$ where $\boldsymbol{\alpha}_{i_{(j)}}^*$ is the corresponding optimal solution, we ensure that the equality $\boldsymbol{\alpha}_{i_{(j)}}^* = \boldsymbol{\alpha}_i^*$ holds when $j = j_i^*$.

Appendix G

Proof of Proposition 7

Suppose that $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ is not optimal for the problem in (6.12), which means there must exist another solution, denoted as $\bar{\boldsymbol{\alpha}} = [\bar{\boldsymbol{\alpha}}_1, \bar{\boldsymbol{\alpha}}_2, \cdots, \bar{\boldsymbol{\alpha}}_V]^T$ and $\bar{\boldsymbol{\beta}} = \left[\beta_{1\bar{\jmath}_1}, \beta_{2\bar{\jmath}_2}, \cdots, \beta_{V\bar{\jmath}_V}\right]^T$, such that

$$\widetilde{L}_{\boldsymbol{\tau}}(\bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}}) < \widetilde{L}_{\boldsymbol{\tau}}(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*). \tag{G.1}$$

On the one hand, since $\boldsymbol{\beta}^*$ is the optimal solution to **P2**, for $\bar{\boldsymbol{\beta}} \neq \boldsymbol{\beta}^*$, we have $\widetilde{L}_{\boldsymbol{\tau}}(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) < \widetilde{L}_{\boldsymbol{\tau}}(\boldsymbol{\alpha}^*, \bar{\boldsymbol{\beta}})$. On the other hand, directly applying the conclusions in Proposition 1, it is seen that $\forall i \in \mathcal{V}$, we have $\boldsymbol{\alpha}^*_{i_{(j)}} = \bar{\boldsymbol{\alpha}}_i$ when $j = \bar{\jmath}_i$. Combined with the previous assumption that $\bar{\boldsymbol{\beta}}$ is the optimal VSP solution to problem (6.12), $\omega^*_{i,\bar{\jmath}_i} = \bar{\omega}_{i,\bar{\jmath}_i}$ holds such that

$$\widetilde{L}_{\boldsymbol{\tau}}(\bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}}) = \widetilde{L}_{\boldsymbol{\tau}}(\boldsymbol{\alpha}^*, \bar{\boldsymbol{\beta}}) > \widetilde{L}_{\boldsymbol{\tau}}(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*). \tag{G.2}$$

However, there is a contradiction between (G.1) and (G.2). Consequently, the assumption cannot hold, which means that $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ is exactly the optimal solution to problem (6.12).

Bibliography

[1] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2663–2675, 2021.

[2] H. Zhang, H. Wang, Y. Li, K. Long, and A. Nallanathan, "DRL-driven dynamic resource allocation for task-oriented semantic communication," *IEEE Transactions on Communications*, vol. 71, no. 7, pp. 3992–4004, 2023.

[3] L. Xia, Y. Sun, C. Liang, D. Feng, R. Cheng, Y. Yang, and M. A. Imran, "WiserVR: Semantic communication enabled wireless virtual reality delivery," *IEEE Wireless Communications*, vol. 30, no. 2, pp. 32–39, 2023.

[4] L. Xia, Y. Sun, C. Liang, L. Zhang, M. A. Imran, and D. Niyato, "Generative AI for semantic communication: Architecture, challenges, and outlook," *arXiv preprint arXiv:2308.15483*, 2023.

[5] W. Weaver, "Recent contributions to the mathematical theory of communication," *ETC: A Review of General Semantics*, pp. 261–281, 1953.

[6] E. C. Strinati and S. Barbarossa, "6G networks: Beyond Shannon towards semantic and goal-oriented communications," *Computer Networks*, vol. 190, p. 107930, 2021.

[7] J. Bao, P. Basu, M. Dean, C. Partridge, A. Swami, W. Leland, and J. A. Hendler, "Towards a theory of semantic communication," in *2011 IEEE Network Science Workshop*. IEEE, 2011, pp. 110–117.

[8] X. Luo, H.-H. Chen, and Q. Guo, "Semantic communications: Overview, open issues, and future research directions," *IEEE Wireless Communications*, pp. 1–10, 2022.

[9] H. Xie and Z. Qin, "A lite distributed semantic communication system for Internet of Things," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 1, pp. 142–153, 2020.

[10] Z. Weng and Z. Qin, "Semantic communication systems for speech transmission," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2434–2444, 2021.

[11] P. Basu, J. Bao, M. Dean, and J. Hendler, "Preserving quality of information by using semantic relationships," *Pervasive and Mobile Computing*, vol. 11, pp. 188–202, 2014.

[12] R. Carnap and Y. Bar-Hillel, "An outline of a theory of semantic information," 1952.

[13] J. Liu, S. Shao, W. Zhang, and H. V. Poor, "An indirect rate-distortion characterization for semantic sources: General model and the case of Gaussian observation," *IEEE Transactions on Communications*, vol. 70, no. 9, pp. 5946–5959, 2022.

[14] P. Zhang, W. Xu, H. Gao, K. Niu, X. Xu, X. Qin, C. Yuan, Z. Qin, H. Zhao, J. Wei *et al.*, "Toward wisdom-evolutionary and primitive-concise 6G: A new paradigm of semantic communication networks," *Engineering*, vol. 8, pp. 60–73, 2022.

[15] D. Huang, X. Tao, F. Gao, and J. Lu, "Deep learning-based image semantic coding for semantic communications," in *2021 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2021, pp. 1–6.

[16] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[17] K. Davaslioglu and E. Ayanoglu, "Quantifying potential energy efficiency gain in green cellular wireless networks," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 4, pp. 2065–2091, 2014.

[18] Z. Cai, J. Hao, P. Tan, S. Sun, and P. Chin, "Efficient encoding of IEEE 802.11 n LDPC codes," *Electronics Letters*, vol. 42, no. 25, p. 1, 2006.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[20] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[21] L. Xia, Y. Sun, D. Niyato, X. Li, and M. A. Imran, "Joint user association and bandwidth allocation in semantic communication networks," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 2, pp. 2699–2711, 2024.

[22] J. Cao, X. Weng, R. Khirodkar, J. Pang, and K. Kitani, "Observation-centric SORT: Rethinking SORT for robust multi-object tracking," *arXiv preprint arXiv:2203.14360*, 2022.

[23] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A survey on knowledge graphs: Representation, acquisition, and applications," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 2, pp. 494–514, 2022.

[24] L. Lu, R. Wu, H. Lin, J. Lu, and J. Jia, "Video frame interpolation with transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3532–3542.

[25] G. Shi, Y. Xiao, Y. Li, and X. Xie, "From semantic communication to semantic-aware networking: Model, architecture, and open problems," *IEEE Communications Magazine*, vol. 59, no. 8, pp. 44–50, 2021.

[26] M. Chein and M.-L. Mugnier, *Graph-based knowledge representation: Computational foundations of conceptual graphs*. Springer Science & Business Media, 2008.

[27] T. RAN, "Scenarios and requirements for small cell enhancements for E-UTRA and E-UTRAN (Release 12), 3GPP," TR 36, Tech. Rep.

[28] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE communications surveys & tutorials*, vol. 18, no. 3, pp. 1617–1655, 2016.

[29] H. Boostanimehr and V. K. Bhargava, "Unified and distributed QoS-driven cell association algorithms in heterogeneous networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 3, pp. 1650–1662, 2014.

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

[31] D. Liu, L. Wang, Y. Chen, M. Elkashlan, K.-K. Wong, R. Schober, and L. Hanzo, "User association in 5G networks: A survey and an outlook,"

IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1018–1044, 2016.

[32] Y. Teng, M. Liu, F. R. Yu, V. C. Leung, M. Song, and Y. Zhang, "Resource allocation for ultra-dense networks: A survey, some research issues and challenges," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2134–2168, 2018.

[33] Y. Chen, J. Li, W. Chen, Z. Lin, and B. Vucetic, "Joint user association and resource allocation in the downlink of heterogeneous networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 7, pp. 5701–5706, 2015.

[34] Y. Jiang, Y. Zou, H. Guo, T. A. Tsiftsis, M. R. Bhatnagar, R. C. de Lamare, and Y.-D. Yao, "Joint power and bandwidth allocation for energy-efficient heterogeneous cellular networks," *IEEE Transactions on Communications*, vol. 67, no. 9, pp. 6168–6178, 2019.

[35] N. Wang, E. Hossain, and V. K. Bhargava, "Joint downlink cell association and bandwidth allocation for wireless backhauling in two-tier HetNets with large-scale antenna arrays," *IEEE Transactions on Wireless Communications*, vol. 15, no. 5, pp. 3251–3268, 2016.

[36] Y. Wang, X. Tao, Y. T. Hou, and P. Zhang, "Effective capacity-based resource allocation in mobile edge computing with two-stage tandem queues," *IEEE Transactions on Communications*, vol. 67, no. 9, pp. 6221–6233, 2019.

[37] H. Zhang, H. Wang, Y. Li, K. Long, and A. Nallanathan, "DRL-driven dynamic resource allocation for task-oriented semantic communication," *IEEE Transactions on Communications*, 2023.

[38] N. Zhao, Y.-C. Liang, D. Niyato, Y. Pei, M. Wu, and Y. Jiang, "Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 11, pp. 5141–5152, 2019.

[39] Y. Xu, G. Gui, H. Gacanin, and F. Adachi, "A survey on resource allocation for 5G heterogeneous networks: Current research, future trends and challenges," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 2, pp. 668–695, 2021.

[40] J. Wang, J. Liu, and N. Kato, "Networking and communications in autonomous driving: A survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1243–1274, 2018.

[41] C. She, C. Sun, Z. Gu, Y. Li, C. Yang, H. V. Poor, and B. Vucetic, "A tutorial on ultrareliable and low-latency communications in 6G: Integrating domain knowledge into deep learning," *Proceedings of the IEEE*, vol. 109, no. 3, pp. 204–246, 2021.

[42] U. Eco, *A theory of semiotics*. Indiana University Press, 1979, vol. 217.

[43] C. W. Morris, "Foundations of the Theory of Signs," in *International Encyclopedia of Unified Science*. Chicago University Press, 1938, pp. 1–59.

[44] J. Barwise and J. Perry, "Situations and attitudes," *The Journal of Philosophy*, vol. 78, no. 11, pp. 668–691, 1981.

[45] L. Floridi, "Outline of a theory of strongly semantic information," *Minds and machines*, vol. 14, pp. 197–221, 2004.

[46] ——, "Philosophical conceptions of information," in *Formal theories of information: From Shannon to semantic information theory and general concepts of information*. Springer, 2009, pp. 13–53.

[47] S. D'Alfonso, "On quantifying semantic information," *Information*, vol. 2, no. 1, pp. 61–101, 2011.

[48] Y. Zhong, "A theory of semantic information," *China communications*, vol. 14, no. 1, pp. 1–17, 2017.

[49] A. Kolchinsky and D. H. Wolpert, "Semantic information, autonomous agency and non-equilibrium statistical physics," *Interface focus*, vol. 8, no. 6, p. 20180041, 2018.

[50] M. Kountouris and N. Pappas, "Semantics-empowered communication for networked intelligent systems," *IEEE Communications Magazine*, vol. 59, no. 6, pp. 96–102, 2021.

[51] A. Rényi, "On measures of entropy and information," in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, vol. 4. University of California Press, 1961, pp. 547–562.

[52] A. Jiang, Y. Li, and J. Bruck, "Error correction through language processing," in *2015 IEEE Information Theory Workshop (ITW)*. IEEE, 2015, pp. 1–5.

[53] N. Farsad, M. Rao, and A. Goldsmith, "Deep learning for joint source-channel coding of text," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2326–2330.

[54] E. Bourtsoulatze, D. B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 3, pp. 567–579, 2019.

[55] M. Rao, N. Farsad, and A. Goldsmith, "Variable length joint source-channel coding of text using deep neural networks," in *2018 IEEE 19th international workshop on signal processing advances in wireless communications (SPAWC)*. IEEE, 2018, pp. 1–5.

[56] B. Güler, A. Yener, and A. Swami, "The semantic communication game," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 4, pp. 787–802, 2018.

[57] M. Jankowski, D. Gündüz, and K. Mikolajczyk, "Joint device-edge inference over wireless links with pruning," in *2020 IEEE 21st international workshop on signal processing advances in wireless communications (SPAWC)*. IEEE, 2020, pp. 1–5.

[58] Z. Yang, M. Chen, Z. Zhang, and C. Huang, "Energy efficient semantic communication over wireless networks with rate splitting," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 5, pp. 1484–1495, 2023.

[59] R. Kaewpuang, M. Xu, W. Y. B. Lim, D. Niyato, H. Yu, J. Kang, and X. Shen, "Cooperative resource management in quantum key distribution (QKD) networks for semantic communication," *IEEE Internet of Things Journal*, vol. 11, no. 3, pp. 4454–4469, 2024.

[60] L. Yan, Z. Qin, R. Zhang, Y. Li, and G. Y. Li, "Resource allocation for text semantic communications," *IEEE Wireless Communications Letters*, 2022.

[61] D. Lopez-Perez, X. Chu, and J. Zhang, "Dynamic downlink frequency and power allocation in OFDMA cellular networks," *IEEE Transactions on Communications*, vol. 60, no. 10, pp. 2904–2914, 2012.

[62] L. P. Qian, Y. J. A. Zhang, Y. Wu, and J. Chen, "Joint base station association and power control via Benders' decomposition," *IEEE Transactions on Wireless Communications*, vol. 12, no. 4, pp. 1651–1665, 2013.

[63] Q. Li, R. Q. Hu, Y. Qian, and G. Wu, "Intracell cooperation and resource allocation in a heterogeneous network with relays," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 4, pp. 1770–1784, 2012.

[64] J. Ghimire and C. Rosenberg, "Resource allocation, transmission coordination and user association in heterogeneous networks: A flow-based unified approach," *IEEE Transactions on Wireless Communications*, vol. 12, no. 3, pp. 1340–1351, 2013.

[65] D. Fooladivanda and C. Rosenberg, "Joint resource allocation and user association for heterogeneous wireless cellular networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 1, pp. 248–257, 2012.

[66] S. Corroy, L. Falconetti, and R. Mathar, "Dynamic cell association for downlink sum rate maximization in multi-cell heterogeneous networks," in *2012 IEEE international conference on communications (ICC)*. IEEE, 2012, pp. 2457–2461.

[67] D. Fooladivanda, A. Al Daoud, and C. Rosenberg, "Joint channel allocation and user association for heterogeneous wireless cellular networks," in *2011 IEEE 22nd International Symposium on Personal, Indoor and Mobile Radio Communications*. IEEE, 2011, pp. 384–390.

[68] K. Shen and W. Yu, "Distributed pricing-based user association for downlink heterogeneous cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1100–1113, 2014.

[69] R. Madan, J. Borran, A. Sampath, N. Bhushan, A. Khandekar, and T. Ji, "Cell association and interference coordination in heterogeneous LTE-A cellular networks," *IEEE Journal on selected areas in communications*, vol. 28, no. 9, pp. 1479–1489, 2010.

[70] R. Sun, M. Hong, and Z.-Q. Luo, "Joint downlink base station association and power control for max-min fairness: Computation and complexity," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 6, pp. 1040–1054, 2015.

[71] M. Hong and Z.-Q. Luo, "Distributed linear precoder optimization and base station selection for an uplink heterogeneous network," *IEEE transactions on signal processing*, vol. 61, no. 12, pp. 3214–3228, 2013.

[72] M. Feng, T. Jiang, D. Chen, and S. Mao, "Cooperative small cell networks: High capacity for hotspots with interference mitigation," *IEEE Wireless Communications*, vol. 21, no. 6, pp. 108–116, 2014.

[73] Ł. Budzisz, F. Ganji, G. Rizzo, M. A. Marsan, M. Meo, Y. Zhang, G. Koutitas, L. Tassiulas, S. Lambert, B. Lannoo *et al.*, "Dynamic resource provisioning for energy efficiency in wireless access networks: A survey and an outlook," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 4, pp. 2259–2285, 2014.

[74] J. Wu, Y. Zhang, M. Zukerman, and E. K.-N. Yung, "Energy-efficient base-stations sleep-mode techniques in green cellular networks: A survey," *IEEE communications surveys & tutorials*, vol. 17, no. 2, pp. 803–826, 2015.

[75] A. Mesodiakaki, F. Adelantado, L. Alonso, and C. Verikoukis, "Energy-efficient context-aware user association for outdoor small cell heterogeneous networks," in *2014 IEEE International Conference on Communications (ICC)*. IEEE, 2014, pp. 1614–1619.

[76] H. Zhu, S. Wang, and D. Chen, "Energy-efficient user association for heterogenous cloud cellular networks," in *2012 IEEE Globecom Workshops*. IEEE, 2012, pp. 273–278.

[77] A. M. Geoffrion, "Generalized benders decomposition," *Journal of optimization theory and applications*, vol. 10, pp. 237–260, 1972.

[78] L. Su, C. Yang, Z. Xu, and A. F. Molisch, "Energy-efficient downlink transmission with base station closing in small cell networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 4784–4788.

[79] E. Chavarria-Reyes, I. F. Akyildiz, and E. Fadel, "Energy consumption analysis and minimization in multi-layer heterogeneous wireless systems," *IEEE Transactions on mobile computing*, vol. 14, no. 12, pp. 2474–2487, 2015.

[80] Y. S. Soh, T. Q. Quek, M. Kountouris, and H. Shin, "Energy efficient heterogeneous cellular networks," *IEEE Journal on selected areas in communications*, vol. 31, no. 5, pp. 840–850, 2013.

[81] L. Xia, Y. Sun, D. Niyato, L. Zhang, L. Zhang, and M. A. Imran, "Hybrid semantic/bit communication based networking problem optimization," *arXiv preprint arXiv:2408.07820*, 2024.

[82] M. I. Ashraf, M. Bennis, C. Perfecto, and W. Saad, "Dynamic proximity-aware resource allocation in vehicle-to-vehicle (V2V) communications," in *2016 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2016, pp. 1–6.

[83] W. Zhao and S. Wang, "Resource sharing scheme for device-to-device communication underlaying cellular networks," *IEEE transactions on communications*, vol. 63, no. 12, pp. 4838–4848, 2015.

[84] M. I. Ashraf, C.-F. Liu, M. Bennis, and W. Saad, "Towards low-latency and ultra-reliable vehicle-to-vehicle communication," in *2017 European Conference on Networks and Communications (EuCNC)*. IEEE, 2017, pp. 1–5.

[85] C.-F. Liu and M. Bennis, "Ultra-reliable and low-latency vehicular transmission: An extreme value theory approach," *IEEE Communications Letters*, vol. 22, no. 6, pp. 1292–1295, 2018.

[86] S. Samarakoon, M. Bennis, W. Saad, and M. Debbah, "Federated learning for ultra-reliable low-latency V2V communications," in *2018 IEEE global communications conference (GLOBECOM)*. IEEE, 2018, pp. 1–7.

[87] M. K. Abdel-Aziz, S. Samarakoon, C.-F. Liu, M. Bennis, and W. Saad, "Optimized age of information tail for ultra-reliable low-latency communications in vehicular networks," *IEEE Transactions on Communications*, vol. 68, no. 3, pp. 1911–1924, 2019.

[88] H. S. Dhillon, R. K. Ganti, F. Baccelli, and J. G. Andrews, "Modeling and analysis of K-tier downlink heterogeneous cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 550–560, 2012.

[89] H. S. Dhillon, R. K. Ganti, and J. G. Andrews, "Load-aware modeling and analysis of heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 4, pp. 1666–1677, 2013.

[90] W. C. Cheung, T. Q. Quek, and M. Kountouris, "Throughput optimization, spectrum allocation, and access control in two-tier femtocell networks," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 561–574, 2012.

[91] Y. Lin, W. Bao, W. Yu, and B. Liang, "Optimizing user association and spectrum allocation in HetNets: A utility perspective," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 6, pp. 1025–1039, 2015.

[92] Y. Chen, "Convolutional neural network for sentence classification," Master's thesis, University of Waterloo, 2015.

[93] O. Vinyals and Q. Le, "A neural conversational model," *arXiv preprint arXiv:1506.05869*, 2015.

[94] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[95] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[96] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[97] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[98] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[99] Y. Guo, Y. Liu, T. Georgiou, and M. S. Lew, "A review of semantic segmentation using deep neural networks," *International journal of multimedia information retrieval*, vol. 7, pp. 87–93, 2018.

[100] S. H. Low and D. E. Lapsley, "Optimization flow control. I. Basic algorithm and convergence," *IEEE/ACM Transactions on Networking*, vol. 7, no. 6, pp. 861–874, 1999.

[101] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[102] S. Diamond and S. Boyd, "CVXPY: A Python-embedded modeling language for convex optimization," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2909–2913, 2016.

[103] S. Kataoka, "A stochastic programming model," *Econometrica: Journal of the Econometric Society*, pp. 181–196, 1963.

[104] A. Charnes and W. W. Cooper, "Chance-constrained programming," *Management Science*, vol. 6, no. 1, pp. 73–79, 1959.

[105] A. Prékopa, *Stochastic Programming*. Springer Science & Business Media, 2013, vol. 324.

[106] F. A. Potra and S. J. Wright, "Interior-point methods," *Journal of Computational and Applied Mathematics*, vol. 124, no. 1-2, pp. 281–302, 2000.

[107] A. V. Fiacco and G. P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. SIAM, 1990.

[108] P. Koehn *et al.*, "Europarl: A parallel corpus for statistical machine translation," in *MT Summit*, vol. 5. Citeseer, 2005, pp. 79–86.

[109] P. He, L. Zhao, S. Zhou, and Z. Niu, "Water-filling: A geometric approach and its application to solve generalized radio resource allocation problems," *IEEE transactions on Wireless Communications*, vol. 12, no. 7, pp. 3637–3647, 2013.

[110] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[111] D. Liu, Y. Chen, K. K. Chai, and T. Zhang, "Joint uplink and downlink user association for energy-efficient HetNets using Nash bargaining solution," in *2014 IEEE 79th Vehicular Technology Conference (VTC Spring)*. IEEE, 2014, pp. 1–5.

[112] L. Xia, Y. Sun, D. Niyato, L. Zhang, and M. A. Imran, "Wireless resource optimization in hybrid semantic/bit communication networks," *arXiv preprint arXiv:2404.04162*, 2024.

[113] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. A. Imran, D. Sabella, M. J. Gonzalez, O. Blume *et al.*, "How much energy is needed to run a wireless network?" *IEEE wireless communications*, vol. 18, no. 5, pp. 40–49, 2011.

[114] S. Guo, Y. Shi, Y. Yang, and B. Xiao, "Energy efficiency maximization in mobile wireless energy harvesting sensor networks," *IEEE Transactions on Mobile Computing*, vol. 17, no. 7, pp. 1524–1537, 2017.

[115] W. Dinkelbach, "On nonlinear fractional programming," *Management science*, vol. 13, no. 7, pp. 492–498, 1967.

[116] D. B. West *et al.*, *Introduction to graph theory*. Prentice hall Upper Saddle River, 2001, vol. 2.

[117] P. Pawar and A. Trivedi, "Joint uplink-downlink resource allocation for D2D underlaying cellular network," *IEEE Transactions on Communications*, vol. 69, no. 12, pp. 8352–8362, 2021.

[118] H. Esmat, M. M. Elmesalawy, and I. Ibrahim, "Uplink resource allocation and power control for D2D communications underlaying multi-cell mobile networks," *AEU-International Journal of Electronics and Communications*, vol. 93, pp. 163–171, 2018.

[119] Q. Wang, M. Hempstead, and W. Yang, "A realistic power consumption model for wireless sensor network devices," in *2006 3rd Annual IEEE Communications Society on Sensor and ad-hoc Communications and Networks*, vol. 1. IEEE, 2006, pp. 286–295.

[120] C. Kai, L. Xu, J. Zhang, and M. Peng, "Joint uplink and downlink resource allocation for D2D communication underlying cellular networks," in *2018 10th International Conference on Wireless Communications and Signal Processing (WCSP)*. IEEE, 2018, pp. 1–6.

[121] D. Halperin, B. Greenstein, A. Sheth, and D. Wetherall, "Demystifying 802.11n power consumption," in *Proceedings of the 2010 International Conference on Power Aware Computing and Systems*. USENIX Association, 2010, p. 1.

[122] C. Guo, L. Liang, and G. Y. Li, "Resource allocation for vehicular communications with low latency and high reliability," *IEEE Transactions on Wireless Communications*, vol. 18, no. 8, pp. 3887–3902, 2019.

[123] T.-S. Kim and S.-L. Kim, "Random power control in wireless ad hoc networks," *IEEE Communications Letters*, vol. 9, no. 12, pp. 1046–1048, 2005.

[124] D. D. Ningombam and S. Shin, "Non-orthogonal resource sharing optimization for D2D communication in LTE-A cellular networks: A fractional frequency reuse-based approach," *Electronics*, vol. 7, no. 10, p. 238, 2018.

[125] L. Xia, Y. Sun, D. Niyato, D. Feng, L. Feng, and M. A. Imran, "xURLLC-aware service provisioning in vehicular networks: A semantic communication perspective," *IEEE Transactions on Wireless Communications*, vol. 23, no. 5, pp. 4475–4488, 2024.

[126] A. L. Moustakas and P. Kazakopoulos, "SINR statistics of correlated MIMO linear receivers," *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6490–6500, 2013.

[127] R. C. Merkle, "Secure communications over insecure channels," *Communications of the ACM*, vol. 21, no. 4, pp. 294–299, 1978.

[128] F. Meshkati, H. V. Poor, S. C. Schwartz, and R. V. Balan, "Energy-efficient resource allocation in wireless networks with quality-of-service constraints," *IEEE Transactions on Communications*, vol. 57, no. 11, pp. 3406–3414, 2009.

[129] L. Xu and W. Zhuang, "Energy-efficient cross-layer resource allocation for heterogeneous wireless access," *IEEE Transactions on Wireless Communications*, vol. 17, no. 7, pp. 4819–4829, 2018.

[130] G. Ding, J. Yuan, G. Yu, and Y. Jiang, "Two-timescale resource management for ultrareliable and low-latency vehicular communications," *IEEE Transactions on Communications*, vol. 70, no. 5, pp. 3282–3294, 2022.

[131] C.-H. Wu, M. E. Lewis, and M. Veatch, "Dynamic allocation of reconfigurable resources in a two-stage tandem queueing system with reliability considerations," *IEEE Transactions on Automatic Control*, vol. 51, no. 2, pp. 309–314, 2006.

[132] P. Li, D. Paul, R. Narasimhan, and J. Cioffi, "On the distribution of SINR for the MMSE MIMO receiver and performance analysis," *IEEE Transactions on Information Theory*, vol. 52, no. 1, pp. 271–286, 2005.

[133] F. Pollaczek, "Über eine Aufgabe der Wahrscheinlichkeitstheorie. I," *Mathematische Zeitschrift*, vol. 32, no. 1, pp. 64–100, 1930.

[134] S. M. Ross, *Introduction to Probability Models*. Academic Press, 2014.

[135] J. D. Little and S. C. Graves, "Little's law," *Building Intuition: Insights from Basic Operations Management Models and Principles*, pp. 81–100, 2008.

[136] L. Liang, G. Y. Li, and W. Xu, "Resource allocation for D2D-enabled vehicular communications," *IEEE Transactions on Communications*, vol. 65, no. 7, pp. 3186–3197, 2017.

[137] S. Burer and A. N. Letchford, "Non-convex mixed-integer nonlinear programming: A survey," *Surveys in Operations Research and Management Science*, vol. 17, no. 2, pp. 97–106, 2012.

[138] S. Boyd, L. Xiao, and A. Mutapcic, "Subgradient methods," *Lecture Notes of EE392o, Stanford University, Autumn Quarter*, vol. 2004, pp. 2004–2005, 2003.

[139] W. Jiang, G. Feng, and S. Qin, "Optimal cooperative content caching and delivery policy for heterogeneous cellular networks," *IEEE Transactions on Mobile Computing*, vol. 16, no. 5, pp. 1382–1393, 2016.

[140] Y. T. Lee, Z. Song, and Q. Zhang, "Solving empirical risk minimization in the current matrix multiplication time," in *Conference on Learning Theory*. PMLR, 2019, pp. 2140–2157.

[141] L. Xu, A. Nallanathan, J. Yang, and W. Liao, "Power and bandwidth allocation for cognitive heterogeneous multi-homing networks," *IEEE Transactions on Communications*, vol. 66, no. 1, pp. 394–403, 2017.

[142] Y. Song, K. W. Sung, and Y. Han, "Impact of packet arrivals on Wi-Fi and cellular system sharing unlicensed spectrum," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 12, pp. 10 204–10 208, 2016.

[143] L. Xia, Y. Sun, X. Li, G. Feng, and M. A. Imran, "Wireless resource management in intelligent semantic communication networks," in *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2022, pp. 1–6.

[144] S. T. Piantadosi, "Zipf's word frequency law in natural language: A critical review and future directions," *Psychonomic Bulletin & Review*, vol. 21, no. 5, pp. 1112–1130, 2014.

[145] G. Lavee, E. Rivlin, and M. Rudzsky, "Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 39, no. 5, pp. 489–504, 2009.

[146] H. Kellerer, U. Pferschy, and D. Pisinger, "Multidimensional knapsack problems," in *Knapsack Problems*. Springer, 2004, pp. 235–283.

[147] H. Tuy, T. Hoang, T. Hoang, V.-n. Mathématicien, T. Hoang, and V. Mathematician, *Convex Analysis and Global Optimization*. Springer, 1998.

[148] C. H. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*. Courier Corporation, 1998.

[149] S. Salhi, "Defining tabu list size and aspiration criterion within tabu search methods," *Computers & Operations Research*, vol. 29, no. 1, pp. 67–86, 2002.

[150] F. Glover, "Tabu search-part I," *ORSA Journal on Computing*, vol. 1, no. 3, pp. 190–206, 1989.

[151] *3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Study on LTE-Based V2X Services; (Release 14)*, document TR 36.885, V2.0.0, 3GPP, Jun. 2016.

[152] R. Zhang, R. Lu, X. Cheng, N. Wang, and L. Yang, "A UAV-enabled data dissemination protocol with proactive caching and file sharing in V2X networks," *IEEE Transactions on Communications*, vol. 69, no. 6, pp. 3930–3942, 2021.

[153] H. Fischer, *A History of the Central Limit Theorem: From Classical to Modern Probability Theory*. Springer, 2011.

[154] G. Bolch, S. Greiner, H. De Meer, and K. S. Trivedi, *Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications*. John Wiley & Sons, 2006.