Joint Knowledge and Power Management for Secure Semantic Communication Networks

Xuesong Liu, *Student Member, IEEE*, Yansong Liu, *Member, IEEE*, Haoyu Tang, *Member, IEEE*,
Fangzhou Zhao, *Student Member, IEEE*, Le Xia, *Member, IEEE*, and Yao Sun, *Senior Member, IEEE*

Abstract—Recently, semantic communication (SemCom) has shown its great superiorities in resource savings and information exchanges. However, while its unique background knowledge guarantees accurate semantic reasoning and recovery, semantic information security-related concerns are introduced at the same time. Since the potential eavesdroppers may have the same background knowledge to accurately decrypt the private semantic information transmitted between legal SemCom users, this makes the knowledge management in SemCom networks rather challenging in joint consideration with the power control. To this end, this paper focuses on jointly addressing three core issues of power allocation, knowledge base caching (KBC), and device-to-device (D2D) user pairing (DUP) in secure SemCom networks. We first develop a novel performance metric, namely semantic secrecy throughput (SST), to quantify the information security level that can be achieved at each pair of D2D SemCom users. Next, an SST maximization problem is formulated subject to secure SemCom-related delay and reliability constraints. Afterward, we propose a security-aware resource management solution using the Lagrange primal-dual method and a two-stage method. Simulation results demonstrate our proposed solution nearly doubles the SST performance and realizes less than half of the queuing delay performance compared to different benchmarks.

Index Terms—D2D semantic communication, semantic secrecy throughput, power control, knowledge base caching, user pairing.

I. INTRODUCTION

SEMANTIC communication (SemCom) has been recognized as a transformative paradigm that can greatly enhance information exchange efficiency and robustness by shifting the delivery focus from bits to semantics [1]. By employing advanced deep learning (DL) algorithms, a semantic encoder is deployed at the transmitter of the SemCom system to extract core semantic features and filter out extraneous content from the source information, thus reducing the number of bits required. Correspondingly, the receiver embeds a semantic decoder, which is jointly trained with the semantic encoder, to precisely recover the meaning from the received bits, even in the presence of bit errors caused by signal distortions under harsh channel conditions [2]. Most importantly, the semantic encoder and decoder should have the equivalent background knowledge for semantic reasoning and interpretation, and the higher the knowledge matches, the more accurate the semantics can be recovered.

The most common manner now to achieve SemCom is device-to-device (D2D) transmission [3], in which the delivery of semantic information is further improved, especially in scenarios with limited bandwidth, network congestion, or device arithmetic constraints [4]–[6]. Although D2D SemCom has great potential in improving communication efficiency and task execution, it is also facing new security and privacy challenges in coordination with the requirement of background knowledge matching among multiple SemCom users. In recent studies, a key concept of knowledge base (KB) is introduced in SemCom, which is deemed a small entity containing a specific domain of knowledge that can be cached in advance in local devices or shared during communication [7]. To be more concrete, the base station (BS) is a central knowledge scheduler responsible for distributing KBs to its serving users. However, eavesdroppers equipped with advanced coding models and the same KBs may exist in the cellular area, aiming to eavesdrop private information from D2D SemCom links. Especially if there exist some eavesdroppers adjacent to a D2D SemCom pair, the KBs shared between the two D2D users may be eavesdropped by malicious users and then the transmitted semantics can also be recovered by others, which may cause severe communication security issues. In this case, traditional information security performance metrics, e.g., secrecy bit throughput or secrecy outage probability [8], are no longer applicable to such a case, since these eavesdroppers concentrate more upon stealing semantic information rather than bit information. Accordingly, there is a pressing demand to explore the protection of semantic information for D2D SemCom underlying cellular networks.

In parallel, it is worth pointing out that different resource management solutions, particularly in knowledge and power allocation, can affect the overall semantic secrecy performance of D2D SemCom networks to a certain degree. For instance, in a scenario with multiple SemCom users (SUs) and eavesdroppers, if each SU is assigned with the suitable KBs matching its paired SU, the semantic information throughput that can be conveyed by the corresponding D2D link is obviously greater than that can be eavesdropped. This is due to the unique knowledge equivalence mechanism in SemCom, where the more the background knowledge of the transceiver matches, the more the semantic recovery rate can be accurate, i.e., more semantic information can be precisely obtained by the receiver.

Xuesong Liu, Fangzhou Zhao, Le Xia, and Yao Sun are with the James Watt School of Engineering, University of Glasgow, Glasgow G12 8QQ, UK (e-mail: {2928395L, 2430965z}@student.gla.ac.uk; xiale1995@outlook.com,Yao.Sun@glasgow.ac.uk;).

Yansong Liu (*Corresponding author: Yansong Liu.*) is with the School of Intelligent Engineering, Shandong Management University, China 250357 (e-mail: lys@sdmu.edu.cn).

Haoyu Tang is with the School of Software, Shandong University, China 250101 (e-mail: tanghao258@sdu.edu.cn).

Apart from this, the power control at each SU is paramount of importance, which has been a classical communication security challenge in cellular networks. Generally, given the fixed relative position of two SUs and one eavesdropper, if the distance between two paired SUs is closer than that between the transmitter SU and eavesdropper, the higher the transmit power, the more the bit secrecy throughput, and vice versa. Hence, both the power and KB management can lead to different semantic information secrecy performance of the SemCom network, which becomes rather critical and tricky, especially when jointly considering their optimality.

In fact, there has been some noteworthy related works paving the way for the development of SemCom. Benefit from advanced DL algorithms, the authors in [9] developed efficient deep multiple access in D2D scenarios to improve the performance of image transmission. [10] developed a generative AI-integrated end-to-end SemCom framework in a cloud-edge-mobile design for multimodal AIGC provisioning. In addition, a series of studies related to secure SemCom have also emerged. In [11], the authors proposed a privacy-preserving SemCom framework that leverages knowledge to protect sensitive information. Meanwhile, in [4], the authors pointed out the risk of eavesdropping on knowledge and identify resource allocation schemes in KBs as a viable future research direction. Some other works [8], [12], [13] also surveyed and discussed semantic information security in a comprehensive way, in particular comparing it to classical physical layer security. For the unique requirements of knowledge pairing in D2D SemCom, [14] proposed a reliable and low-latency knowledge matching method to solve the problem of semantic service provisioning among multiple vehicle-to-vehicle SemCom users. Besides, [15] developed a long-term robust resource allocation strategy to satisfy the trade-off between user satisfaction, queue stability and communication latency for D2D SemCom.

Nevertheless, to the best of our knowledge, none of these papers technically investigate the semantic information security-aware resource management problem for D2D SemCom. In this work, our main task is to seek optimal resource allocation in *secure SemCom network* (SSCN) to reach the best overall semantic network performance. Especially considering the unique knowledge matching demands between multiple SUs and the existence of eavesdroppers, devising the best resource management solution becomes quite indispensable, which can yield many benefits, such as further improving bandwidth utilization, and ensuring secure, efficient, and high-quality SemCom service provisioning. In view of these unique requirements, there exist three fundamental networking challenges as follows.

- *Challenge 1: How to find an appropriate metric to measure semantic network performance related to secure SemCom?* The longer it takes to allocate knowledge, the more susceptible transmitted information is to eavesdropping. Considering this, scheduling the delay a semantic packet spends in the queue buffer of each D2D SemCom link is of paramount importance. Moreover, traditional system metrics, e.g., bit throughput, is no longer applicable to measure the SSCN. Especially taking into account

the presence of eavesdroppers and semantic information importance in SemCom, how to define appropriate secure SemCom-related metrics should be the first challenge.

- *Challenge 2: How to accurately assign KBs to multiple SemCom users with different knowledge preferences?* In SemCom, the KB is the key to perform semantic inference and recovery. Considering that each SemCom user has its own knowledge preference and KB storage capacity, it is quite tricky for each user to realize the best caching policy for size-varying KBs. In particular, potential eavesdroppers pose continuous threats to semantic transmission. If being aware the knowledge distribution at eavesdroppers, how to achieve a secure knowledge base caching (KBC) model should be the second difficulty.

- *Challenge 3: How to determine the best D2D SemCom pairing strategy among all uses to maximize knowledge protection against eavesdropping?* Note that high semantic fidelity can be guaranteed based on equivalent knowledge matching conditions between two SemCom users. In line with different KBC situations, varying channel conditions, and the potential eavesdropping, it poses the third challenge to find the best D2D user pairing (DUP) for all SemCom users in order to realize the optimal semantic secrecy information delivery.

To address the above gaps that have never been jointly investigated in the current literature, in this paper, we jointly optimize KBC, DUP, and power allocation in the SSCN with the consideration of unique SemCom characteristics and semantic information security. Both theoretical analysis and numerical results validate the performance superiority of our proposed solution in terms of semantic secrecy throughput, queuing delay, and semantic knowledge satisfaction. In a nutshell, our main contributions are summarized in the following:

- We first identify two core resource management problems of KBC and DUP and construct their mathematical models. Specially, considering the potential eavesdropping center and its knowledge distribution conditions, we develop a novel metric, namely semantic secrecy throughput (SST), to measure the overall semantic information security for all SemCom users.

- By taking into account personal preference for different knowledge bases, we derive the average queuing latency and define the average semantic knowledge satisfaction for each potential D2D SemCom link in a mathematical manner. In this way, a joint SST-maximization problem is then formulated for power allocation, KBC, and DUP subject to several practical constraints.

- We propose an optimal security-aware resource management solution with polynomial-time complexity. Specifically, a Lagrange primal-dual method is first utilized to obtain the dual problem, which is able to be decomposed to multiple subproblems. Without loss of optimality, in each iteration, we leverage a two-stage method to separately solve these subproblems, where the first stage is to determine the KBC and power allocation strategies and the second stage is to finalize the DUP policy.

The remainder of this paper is organized as follows. Section

Fig. 1. The SSCN with KBC, DUP, and PC.

II first introduces the system model of SSCN and formulates a joint security-aware resource optimization problem. Then, we present our proposed solution and complexity analysis in Section III. Numerical results are demonstrated and discussed in Section IV, followed by the conclusions in Section V.

II. SYSTEM MODEL AND PROBLEM FORMULATION

In this section, the considered SSCN scenario is first elaborated along with the knowledge caching model, secrecy queuing model, and semantic secrecy throughput model. Then, the corresponding optimization problem is formulated.

A. SSCN Scenario

Consider a single-cell SSCN scenario as shown in Fig. 1, a total of M SUs are distributed within the coverage of one BS, and each SU $i \in \mathcal{M} = \{1, 2, \cdots, M\}$ is capable of providing D2D SemCom services to others. Without loss of generality, each SU is stipulated to be paired with only one (another) SU at a time for D2D SemCom. Herein, let $\beta_{i,j} \in \{0, 1\}$ denote the binary DUP indicator for a potential SU i-SU j ($j \in \mathcal{M}, j \neq i$) pair, where $\beta_{i,j} = 1$ means SU i is paired with SU j for D2D SemCom, otherwise $\beta_{i,j} = 1$. Besides, each D2D SemCom link has been pre-allocated an orthogonal subchannel with equal channel bandwidth W, and each SU i has a maximum allowable transmit power P_{max} for SemCom. Meanwhile, assume that there exists an eavesdropping center with a fixed spatial location, constantly attempting to eavesdrop on the conveyed semantic information from all D2D SemCom links and using the same subchannel and bandwidth resources as the correspondingly target links. Having these, let $G_{i,j}^D$ denote the channel power gain of the D2D link between the transmitter SU i and the receiver SU j, and let G_i^E denote the channel power gain of the eavesdropping link between the transmitter SU i and the eavesdropping center. Further denoting the transmit power of SU i as P_i, the signal-to-noise ratio (SNR) experienced by the

D2D link is $\gamma_{i,j}^D = P_i G_{i,j}^D / \zeta^2$, where ζ^2 is the noise power. Likewise, the SNR of the eavesdropping link regarding SU i is $\gamma_i^E = P_i G_i^E / \zeta^2$. Hence, the bit rates at the D2D link and the eavesdropping link should be $r_{i,j}^D = W \log_2 \left(1 + \gamma_{i,j}^D\right)$ and $r_i^E = W \log_2 \left(1 + \gamma_i^E\right)$, respectively.

Note that we allow SU i to provide D2D SemCom services to SU j only if the SNR $\gamma_{i,j}$ is above a specified threshold γ_0, hence the set of eligible communication neighbors of SU i is defined as $\mathcal{M}_i = \{j \mid j \in \mathcal{M}, j \neq i, \gamma_{i,j} \geqslant \gamma_0\}$. Moreover, let the BS act as a semantic service controller to efficient schedule and coordinate the whole secure D2D SemCom process based on the request and state information received from all participating SUs within its coverage.

B. Knowledge Base Caching Model

The acquisition of necessary background knowledge is known to be inevitable for accurate semantic inference and interpretation in SemCom. Note that each KB contains the background knowledge of only one particular application domain (e.g., music or sports), and thus different KBs are associated with different background knowledge, and holding some common KBs becomes the necessary condition to perform SemCom between two D2D SUs in accordance with the knowledge equivalence principle.[1] Especially considering the existence of the eavesdropping center, different knowledge base caching (KBC) schemes can significantly affect the overlapping degree of background knowledge between different nodes, since the transmitted semantic information can only be decoded under the equivalent background knowledge. In this work, suppose that there is a KB library $\mathcal{K}$ with a total of K differing KBs in the considered SSCN, and each KB $k \in \mathcal{K} = \{1, 2, \ldots, K\}$ is corresponding to a specific type of SemCom service. On this basis, all SUs should proactively download and cache their respective interested KBs from the BS to achieve D2D SemCom based on the desired semantic services. Note that each requires a unique storage size s_k, and each SU $i \in \mathcal{M}$ has a finite capacity C_i for its local KB storage. In this way, we define a binary KBC indicator as

$$\alpha_i^k = \begin{cases} 1, & \text{if KB } k \text{ is cached on SU } i; \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

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

Besides, it is noticed that different SUs may have different personal preferences for these KBs, thus resulting in the diversity of KB popularity. Naturally, the more popular the KBs, the higher the KBC probabilities. Herein, we assume that the KB popularity at each SU follows the Zipf distribution [17], [18].[2] Hence, the probability of SU i requesting its desired KB k-based SemCom services (generating the corresponding semantic data packets) is $p_i^k = \left(r_i^k\right)^{-\xi_i} / \sum_{e \in \mathcal{K}} e^{-\xi_i}, \forall(i, k) \in$

---

[1]The structure of a KB can roughly cover multiple computational ontologies, facts, rules and constraints associated to a specific domain [16]. In recent deep learning-driven semantic coding models, the KB is also treated as a training database serving a certain class of learning tasks [2].

[2]Note that other known probability distributions can also be adopted for KB popularity without changing the remaining modeling and solution.

$\mathcal{M} \times \mathcal{K}$, where ξ_i ($\xi_i \geqslant 0$) is the skewness of the Zipf distribution, and r_i^k is the popularity rank of KB k at SU i.[3] Based on p_i^k, we specially define a KBC-related metric η_i, namely *semantic knowledge satisfaction*, to measure the satisfaction degree of SU i caching its interested KBs, given as $\eta_i = \sum_{k \in \mathcal{K}} \alpha_i^k p_i^k$. It is further noticed that $\eta_i \geqslant \eta_0$, where η_0 is the unified minimum threshold that needs to be achieved at each SU.

As for the eavesdropping center, let us assume it has its own KB preference for illegal eavesdropping and decoding of their interested semantic information from these D2D SemCom links. Without loss of the generality, let r_E^k and ξ_E ($\xi_E \geqslant 0$) denote the popularity rank of KB k at the eavesdropping center and the skewness of its Zipf distribution, respectively. As such, the probability of KB k being held by the eavesdropping center can be expressed by $p_E^k = \left(r_E^k\right)^{-\xi_E} / \sum_{e \in \mathcal{K}} e^{-\xi_E}, \forall k \in \mathcal{K}$. In other words, if denoting α_E^k as the KBC indicator of the eavesdropping center, where $\alpha_E^k = 1$ indicates that KB k is cached and $\alpha_E^k = 0$ otherwise, we have $\Pr\left\{\alpha_E^k = 1\right\} = p_E^k$ and $\Pr\left\{\alpha_E^k = 0\right\} = 1 - p_E^k$.

C. Secrecy Delay Model

In secure communications, the longer the delay incurred in the communication process, the more likely it is that the transmitted information will be eavesdropped on [21]. Being aware of this, the knowledge matching based semantic packet queuing delay is jointly considered as a long-term metric in this work, which is to characterize the average sojourn time of semantic data packets in the receiver SU's queue buffer (following the first-come first-serve rule). This is mainly considered from the large-timescale perspective, as we only focus on the queuing latency under the steady-state of the semantic packet queuing system. Note that each semantic data packet is associated with a specific service type, i.e., a certain KB, while we assume that semantic data packets generated based on different KBs can co-exist in the queue and have independent average arrival rate and interpretation time. Besides, not all semantic data packets arriving at the receiver SU are always allowed to enter its queue, as some of them may mismatch the KBs currently held, rendering these packets uninterpretable [14]. As for these mismatched packets, we can choose the traditional bit transmission, where the additional delay introduced can be easily covered by applying SemCom due to the considerably saved bits and communication delay.

To preserve generality, we first suppose a Poisson data arrival process with average rate $\lambda_{i,j}^k = \lambda_{i,j}^D p_i^k$ for the D2D SemCom link between transmitter SU i and receiver SU j to account for semantic packet generation based on KB k, where $\lambda_{i,j}^D$ is the total arrival rate of all semantic packets from transmitter SU i to receiver SU j. Herein, if assuming that each semantic packet has an average size of L bits, given its achievable bit rate $r_{i,j}^D$ with variable P_i, we can approximately calculate $\lambda_{i,j}^D$ by $\lambda_{i,j}^D = r_{i,j}^D / L$. Combined with the KBC situation at SU i and SU j, the effective arrival rate of semantic

packets (i.e., semantic packets based on matched KB between them) in the queue is given as $\lambda_{i,j}^{eff} = \sum_{k \in \mathcal{K}} \alpha_i^k \alpha_j^k \lambda_{i,j}^k$. In parallel, let a random variable I_j^k denote the Markovian interpretation time [22] required by KB k-based packets at SU j with mean $1/\mu_j^k$, which is determined by the computing capability of the SU 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 SU should be treated as a general distribution [23]. If further taking into account the KB popularity, we can calculate the ratio of the amount of KB k-based packets to the total packets in the SU i-SU j pair's queue by $\epsilon_{i,j}^k = p_i^k / \sum_{f \in \mathcal{K}} \alpha_i^f \alpha_j^f p_i^f$. With the independence among packets based on different KBs, the interpretation time required by each packet in the queue is now expressed as $W_{i,j} = \sum_{k \in \mathcal{K}} \alpha_i^k \alpha_j^k \epsilon_{i,j}^k I_j^k$.

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 D2D pair in the SSCN 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* [24], the average queuing latency for the SU i-SU j pair, denoted as $\delta_{i,j}$, is determined as follows[4]

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

To calculate the close form of $\delta_{i,j}$, again leveraging the independence of I_j^k over k, we can obtain the expectation of the interpretation time for all semantic data packets by

$$\mathbb{E}\left[W_{i,j}\right] = \sum_{k \in \mathcal{K}} \alpha_i^k \alpha_j^k \epsilon_{i,j}^k \mathbb{E}\left[I_j^k\right] = \sum_{k \in \mathcal{K}} \frac{\alpha_i^k \alpha_j^k \epsilon_{i,j}^k}{\mu_j^k}, \quad (3)$$

and the variance of $W_{i,j}$ can be calculated by

$$
\begin{aligned}
Var\left(W_{i,j}\right) &= \sum_{k \in \mathcal{K}} \alpha_i^k \alpha_j^k \left(\epsilon_{i,j}^k\right)^2 Var\left[I_j^k\right] \\
&= \sum_{k \in \mathcal{K}} \alpha_i^k \alpha_j^k \left(\frac{\epsilon_{i,j}^k}{\mu_j^k}\right)^2.
\end{aligned}
\quad (4)
$$

By substituting (3) and (4) into (2), $\delta_{i,j}$ can be rewritten in (5), as shown at the bottom of the next page. Due to the security consideration, let δ_0 denote a unified secrecy latency threshold that should be satisfied at all associated D2D SemCom links.

D. Semantic Secrecy Throughput Measurement

Lately, a concept of *semantic triplet* is dedicatedly introduced in the realm of SemCom to represent the interpretable relationship between two specific semantic entities implied in source information [25]–[27], and its typical expression is *(Entity-A, Relationship, Entity-B)*, as depicted in the ex-emplification of Fig. 1. In the context of the above secrecy

---

[3]The KB popularity ranking of each SU can be analyzed and estimated based on its historical messaging records [14], [19], [20], which will not be discussed in this paper.

[4]In order to guarantee the steady-state of the queuing system, a condition of $\lambda_{i,j}^{eff} \mathbb{E}\left[W_{i,j}\right] < 1$ must be satisfied before proceeding [23]. 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.

queuing latency model, we assume that each semantic data packet contains exactly one semantic triple, which is corresponding to a certain KB. To be more specific, all source information of transmitters in SemCom is first encoded in the minimum unit of semantic triplets by semantic encoding models.[5] Afterward, each semantic triplet is encoded by each SU's channel encoder and then encapsulated into an L bits-size semantic data packet for wireless transmission [27]. As such, we clearly have the total number of semantic triplets transmitted from transmitter SU i to receiver SU j per second given by $\lambda_{i,j}^D$, as aforementioned, and likewise, obtain that for the eavesdropping link as $\lambda_i^E = r_i^E / L$.

Since each KB has its popularity rank at each SU as identified in the KBC model, it means that each semantic triplet also has the same preference level. Naturally, transmitting the higher-ranked semantic triplets contributes more valuable semantic information for each SemCom receiver. Inspired by this, we employ a performance metric called *semantic value* proposed in [27] to measure the semantic information importance of semantic triplets with different rankings at each SU. Proceeding as in [27], the calculation of semantic value for each KB k-based semantic triplet relies on SU i's Zipf distribution, which is exactly the numerator of p_i^k, given by $\left(r_i^k \right)^{-\xi_i}$. Note that the semantic value above is measured according to the transmitter's KB preference only, which is reasonable as the primary purpose of SemCom is generally determined by the sender, rather than the receiver. With these and further combining the effective semantic packet arrival of KB k from SU i to SU j given in $\lambda_{i,j}^{eff}$, i.e., $\alpha_i^k \alpha_j^k \lambda_{i,j}^k$, we can obtain the effective semantic value transmitted via the SU i-SU j D2D pair per second as

$$V_{i,j}^D = \sum_{k \in \mathcal{K}} \alpha_i^k \alpha_j^k \lambda_{i,j}^k \left(r_i^k \right)^{-\xi_i} = \frac{r_{i,j}^D}{L} \sum_{k \in \mathcal{K}} \alpha_i^k \alpha_j^k p_i^k \left(r_i^k \right)^{-\xi_i}. \tag{6}$$

Similarly, based on the semantic triplet transmission rate λ_i^E and the KBC situation α_E^k pertinent to the eavesdropping center, the average number of KB k-based semantic triplets that can be eavesdropped per second should be

$$V_i^E = \sum_{k \in \mathcal{K}} \lambda_i^E \alpha_i^k p_i^k \Pr \left\{ \alpha_E^k = 1 \right\} \left(r_i^k \right)^{-\xi_i}$$
$$= \frac{r_i^E}{L} \sum_{k \in \mathcal{K}} \alpha_i^k p_i^k p_E^k \left(r_i^k \right)^{-\xi_i}. \tag{7}$$

[5]This assumption is justified since core semantic information can be extracted by state-of-the-art DL models from multimedia services (e.g., text [28], image [29], and video [30]) to draw the semantic knowledge graph, which can be decomposed into multiple semantic triplets. In other words, it is reasonable that source information in any format can be conveyed in units of semantic triplets for SemCom service provisioning.

Therefore, drawing on the classical metric of secrecy throughput in traditional secure communications [8], [31], [32], we define here a new metric, namely *semantic secrecy throughput* (SST), by subtracting V_i^E in (7) from $V_{i,j}^D$ in (6) as

$$V_{i,j}^S = \left[V_{i,j}^D - V_i^E \right]^+$$
$$= \frac{1}{L} \left[\sum_{k \in \mathcal{K}} \alpha_i^k \left(r_i^k \right)^{-\xi_i} \left(r_{i,j}^D \alpha_j^k p_i^k - r_i^E p_i^k p_E^k \right) \right]^+, \tag{8}$$

where the operator $[\cdot]^+$ is to output the maximum value between its argument and zero. To provide an intuitive explanation, $V_{i,j}^S$ is the secure semantic value transmission rate at which the legitimate SU i-SU j pair is able to communicate safely without enough valuable semantic information being stolen by the eavesdropping center. If $V_{i,j}^D > V_i^E$, the D2D SemCom users can transmit semantic information to each other with a secure rate of $(V_{i,j}^D - V_i^E)$. Otherwise, any transmitted semantics at the SU i-SU j pair will be stolen by the eavesdroppers. If taking into account all DUP possibilities, the overall SST of SSCN should be the sum of $\beta_{i,j} V_{i,j}^S$ corresponding to all potential SU i-SU j cases, that is,

$$SST = \sum_{i \in \mathcal{M}} \sum_{j \in \mathcal{M}_i} \beta_{i,j} V_{i,j}^S$$
$$= \sum_{i \in \mathcal{M}} \sum_{j \in \mathcal{M}_i} \frac{\beta_{i,j}}{L} \left[\sum_{k \in \mathcal{K}} \alpha_i^k \left(r_i^k \right)^{-\xi_i} \left(r_{i,j}^D \alpha_j^k p_i^k - r_i^E p_i^k p_E^k \right) \right]^+. \tag{9}$$

E. Problem Formulation

For ease of illustration, we first define three variable sets $\boldsymbol{\alpha} = \left\{ \alpha_i^k \mid i \in \mathcal{M}, k \in \mathcal{K} \right\}$, $\boldsymbol{\beta} = \left\{ \beta_{i,j} \mid i \in \mathcal{M}, j \in \mathcal{M}_i \right\}$, and $\boldsymbol{P} = \{ P_i \mid i \in \mathcal{M} \}$ that consist of all possible indicators pertinent to KBC, DUP, and PC, respectively. Without loss of generality, the objective of our optimization is to maximize SST in (9) by jointly optimizing $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{P})$, while subject to SemCom-relevant latency and performance requirements alongside several practical system constraints. The problem

$$\delta_{i,j} = \frac{\left[\left(\sum_{k \in \mathcal{K}} \alpha_i^k \alpha_j^k \frac{p_i^k / \mu_j^k}{\sum_{f \in \mathcal{K}} \alpha_i^f \alpha_j^f p_i^f} \right)^2 + \sum_{k \in \mathcal{K}} \alpha_i^k \alpha_j^k \left(\frac{p_i^k / \mu_j^k}{\sum_{f \in \mathcal{K}} \alpha_i^f \alpha_j^f p_i^f} \right)^2 \right] \cdot \left(\sum_{k \in \mathcal{K}} \alpha_i^k \alpha_j^k \lambda_{i,j}^k \right)}{2 \left[1 - \left(\sum_{k \in \mathcal{K}} \alpha_i^k \alpha_j^k \lambda_{i,j}^k \right) \cdot \left(\sum_{k \in \mathcal{K}} \alpha_i^k \alpha_j^k \frac{p_i^k / \mu_j^k}{\sum_{f \in \mathcal{K}} \alpha_i^f \alpha_j^f p_i^f} \right) \right]}. \tag{5}$$

is specifically formulated as follows:

$$\textbf{P0}: \max_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{P}} \quad \sum_{i\in\mathcal{M}}\sum_{j\in\mathcal{M}_i} \beta_{i,j} V_{i,j}^S \tag{10}$$

$$\text{s.t.} \quad \sum_{k\in\mathcal{K}} \alpha_i^k \cdot s_k \leqslant C_i, \ \forall i\in\mathcal{M}, \tag{10a}$$

$$\eta_i \geqslant \eta_0, \ \forall i\in\mathcal{M}, \tag{10b}$$

$$\sum_{j\in\mathcal{M}_i} \beta_{i,j} = 1, \ \forall i\in\mathcal{M}, \tag{10c}$$

$$\beta_{i,j} = \beta_{j,i}, \ \forall (i,j)\in\mathcal{M}\times\mathcal{M}_i, \tag{10d}$$

$$\sum_{j\in\mathcal{M}_i} \beta_{i,j}\delta_{i,j} \leqslant \delta_0, \ \forall i\in\mathcal{M}, \tag{10e}$$

$$\sum_{j\in\mathcal{M}_i} \beta_{i,j} V_{i,j}^S \geqslant V_0, \ \forall i\in\mathcal{M}, \tag{10f}$$

$$\alpha_i^k \in \{0,1\}, \ \forall (i,k)\in\mathcal{M}\times\mathcal{K}, \tag{10g}$$

$$\beta_{i,j} \in \{0,1\}, \ \forall (i,j)\in\mathcal{M}\times\mathcal{M}_i \tag{10h}$$

$$0 \leqslant P_i \leqslant P_{max}, \ \forall i\in\mathcal{M}. \tag{10i}$$

Constraints (10a) guarantees that the total size of KBs cached at each SU cannot exceed its maximum storage capacity, while constraint (10b) represents the aforementioned semantic knowledge satisfaction requirement. Then, constraints (10c) and (10d) jointly model the single-D2D pairing limitation of SUs. Next, constraint (10e) indicates the maximum secrecy queuing latency threshold, while constraint (10f) is the minimum SST requirement at each D2D SemCom link indicating the reliability level of the system. Moreover, constraints (10g) and (10h) characterize the binary properties of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Finally, constraint (10i) determines the transmit power range for all SUs.

Carefully examining **P0**, it can be observed that the optimization is quite challenging to be solved straightforwardly due to several intractable mathematical obstacles. First of all, **P0** is an NP-hard optimization problem as demonstrated below. Consider a special case of **P0** where all $\boldsymbol{\beta}$- and $\boldsymbol{P}$-related constraints have been satisfied. In this case, constraints (10c)-(10d) and (10h)-(10i) can all be removed, and the primal problem degenerates into a variant 0-1 multi-knapsack problem that is known to be NP-hard [33], thereby **P0** is also NP-hard. Besides, **P0** involves both continuous and discrete variables, and its objective function (10) is quite complicated alongside constraints (10e) and (10f), 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 (10) and constraints (10e) and (10f). Therefore, a severe performance penalty and a high computational complexity will be incurred from the procedure of integer recovery due to the huge performance compromise on solving the nonconvex problem for relaxed variables [34]–[36]. In view of the above mathematical challenges, we propose an efficient resource allocation strategy in the next section to reach the optimality of **P0** and obtain the joint KBC, DUP, and PC solution.

III. PROPOSED SECURITY-AWARE RESOURCE ALLOCATION FOR SSCN

In this section, we illustrate how to achieve the optimal security-aware resource management in the SSCN. We first employ the Lagrange dual method to transform the primal problem **P0** to its dual optimization problem. Then, given dual variables in each iteration, the dual problem is decomposed into multiple subproblems, which are solved by the proposed two-stage method. Finally, the workflow of our solution and its complexity analysis are provided.

A. Primal-Dual Problem Transformation

To make **P0** tractable, we first incorporate constraints (10e) and (10f) into the objective function (10) by associating two Lagrange multipliers $\boldsymbol{\tau} = \{\tau_i \mid i\in\mathcal{M}\}$ and $\boldsymbol{\rho} = \{\rho_i \mid i\in\mathcal{M}\}$. As such, its Lagrange function is found by (11), as shown at the bottom of the next page, in which $\widetilde{L}_{\boldsymbol{\tau},\boldsymbol{\rho}}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{P})$ is defined for expression brevity. Then, the Lagrange dual problem of **P0** becomes

$$\textbf{D0}: \min_{\boldsymbol{\tau},\boldsymbol{\rho}} \quad D(\boldsymbol{\tau},\boldsymbol{\rho}) = g_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{P}}(\boldsymbol{\tau},\boldsymbol{\rho}) + \delta_0 \sum_{i\in\mathcal{M}} \tau_i - V_0 \sum_{i\in\mathcal{M}} \rho_i \tag{12}$$

$$\text{s.t.} \quad \tau_i \geqslant 0, \ \rho_i \geqslant 0, \ \forall i\in\mathcal{M}, \tag{12a}$$

where

$$g_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{P}}(\boldsymbol{\tau},\boldsymbol{\rho}) = \sup_{\boldsymbol{\alpha},\boldsymbol{\beta}} \widetilde{L}_{\boldsymbol{\tau},\boldsymbol{\rho}}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{P})$$
$$\text{s.t.} \quad (10a)-(10d),(10g)-(10i). \tag{13}$$

Notably, the optimality of the convex problem **D0** gives at least the best upper bound of **P0**, even if **P0** is nonconvex, according to the duality property [37]. Hence, our focus now naturally shifts to seeking the optimal solution to **D0**.

Given the initial dual variable $\boldsymbol{\tau}$ and $\boldsymbol{\rho}$, we can solve problem (13) in the first place to find the optimal solution of $(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{P})$, the details of which will be presented in the subsequent subsections. After that, a subgradient method is employed for updating $\boldsymbol{\tau}$ and $\boldsymbol{\rho}$ to solve **D0** in an iterative fashion. Particularly, the partial derivatives with respect to (w.r.t.) $\boldsymbol{\tau}$ and $\boldsymbol{\rho}$ in $D(\boldsymbol{\tau},\boldsymbol{\rho})$ are set as the subgradient directions, respectively. Suppose in a certain iteration, say iteration t, each dual variable $\tau_i(t)$ $(i\in\mathcal{M})$ is updated by

$$\tau_i(t+1) = \left[\tau_i(t) - \nu_1(t)\cdot\left(\delta_0 - \sum_{j\in\mathcal{M}_i}\beta_{i,j}(t)\delta_{i,j}(t)\right)\right]^+, \tag{14}$$

and each $\rho_i(t)$ is updated by

$$\rho_i(t+1) = \left[\rho_i(t) + \nu_2(t)\cdot\left(V_0 - \sum_{j\in\mathcal{M}_i}\beta_{i,j}(t)V_{i,j}^S(t)\right)\right]^+. \tag{15}$$

$\nu_1(t)$ and $\nu_2(t)$ are the stepsizes w.r.t. the update of $\tau_i(t)$ and $\rho_i(t)$ in iteration t, respectively. Generally, the convergence of the subgradient descent method can be ensured with the proper stepsize preset in practice [38].

B. Dual Problem Decomposition for Optimality Guarantee

As discussed earlier, given $\boldsymbol{\tau}$ and $\boldsymbol{\rho}$ in each iteration, the optimal $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{P})$ need to be obtained by solving problem (13). However, solving such a problem is still challenging due to the inseparability of these variables $\widetilde{L}_{\boldsymbol{\tau}, \boldsymbol{\rho}}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{P})$. To this end, we propose a two-stage method to reach its optimality with a low computational complexity.

In the first stage, we focus on multiple independent KB caching subproblems, each corresponding to a potential D2D SU pair in the SSCN. Specifically, the performances of an SU i-SU j single pair (i.e., the sender SU i and the receiver SU j) and an SU j-SU i pair (i.e., the sender SU j and the receiver SU i) should be jointly considered, and for ease of distinction, we refer to the two case as the same *SU i, j pair*, $\forall (i, j) \in \mathcal{M} \times \mathcal{M}_i, j > i$. In other words, for each KBC subproblem, we have $\beta_{i,j} = \beta_{j,i} = 1$ in $\widetilde{L}_{\boldsymbol{\tau}, \boldsymbol{\rho}}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{P})$ corresponding to a given SU i, j pair, while all other SU pairs are not taken into account. Therefore, different KBC subproblems can be tackled independently, and in this way, let

$$\omega_{i,j} = \left[(1 + \rho_i) V_{i,j}^S - \tau_i \delta_{i,j} \right] + \left[(1 + \rho_j) V_{j,i}^S - \tau_j \delta_{j,i} \right]. \quad (16)$$

Clearly, we have $\omega_{i,j} = \omega_{j,i}$, thus only the case of $j > i$ needs to be considered for each potential SU i, j pair, and constraints (10c), (10d), and (10h) are satisfied at the same time.

Clearly, we have $U = \left(\sum_{i \in \mathcal{M}} |\mathcal{M}_i| \right) / 2$ subproblems in total, each of which is denoted as $\mathbf{P1}_{i,j}$, $\forall (i, j) \in \mathcal{M} \times \mathcal{M}_i, j > i$, to seek the optimal KBC sub-policy only for an individual SU i, j pair. Note that the optimal KBC solution to problem (13) 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 DUP and power allocation subproblems to finalize the joint optimal solution of $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{P})$ for (13). Given the dual variable $\boldsymbol{\tau}$ and $\boldsymbol{\rho}$ in each iteration, $\mathbf{P1}_{i,j}$ becomes

$$\mathbf{P1}_{i,j}: \max_{\{\alpha_i^k\}, \{\alpha_j^k\}, P_i, P_j} \omega_{i,j} \quad (17)$$

$$\text{s.t.} \quad \sum_{k \in \mathcal{K}} \alpha_i^k \cdot s_k \leqslant C_i, \quad (17a)$$

$$\sum_{k \in \mathcal{K}} \alpha_j^k \cdot s_k \leqslant C_j, \quad (17b)$$

$$\eta_i \geqslant \eta_0, \ \eta_j \geqslant \eta_0, \quad (17c)$$

$$\alpha_i^k \in \{0, 1\}, \alpha_j^k \in \{0, 1\}, \forall k \in \mathcal{K}, \quad (17d)$$

$$0 \leqslant P_i \leqslant P_{max}, \ 0 \leqslant P_j \leqslant P_{max}. \quad (17e)$$

If we can obtain the optimal KBC and power allocation sub-policies by solving $\mathbf{P1}_{i,j}$ for SU i (denoted as $\boldsymbol{\alpha}_{i_{(j)}}^*$ and $P_{i_{(j)}}^*$) and SU j (denoted as $\boldsymbol{\alpha}_{j_{(i)}}^*$ and $P_{j_{(i)}}^*$),[6] corresponding to the individual SU i, j pair. The following proposition explicitly shows how $\mathbf{P1}_{i,j}$ correlates to the problem (13).

Proposition 1. *Let* $\boldsymbol{\alpha}^* = [\boldsymbol{\alpha}_1^*, \boldsymbol{\alpha}_2^*, \cdots, \boldsymbol{\alpha}_M^*]^T$ *be the optimal KBC solution and* $\boldsymbol{P}^* = [P_1^*, P_2^*, \cdots, P_M^*]^T$ *be the optimal power allocation solution to the problem in (13) given dual variables* $\boldsymbol{\tau}$ *and* $\boldsymbol{\rho}$, *where* $\boldsymbol{\alpha}_i^*$ *and* P_i^* *represent the optimal KBC policy and power allocation policy of SU i, respectively. Then we have* $\forall i \in \mathcal{M}, \exists j \in \mathcal{M}_i$, *such that* $\boldsymbol{\alpha}_{i_{(j)}}^* = \boldsymbol{\alpha}_i^*$ *and* $P_{i_{(j)}}^* = P_i^*$.

Proof: Please see Appendix A. ∎

Based on Proposition 1, it is seen that the optimal KBC and power allocation policies of each SU can be obtained by solving a certain $\mathbf{P1}_{i,j}$. Moreover, due to the single-association requirement of D2D SemCom pairing, the optimal DUP strategy becomes the only key to finalize the optimal solution to (13). Hence, we first construct the optimal coefficient matrix for $\boldsymbol{\beta}$ in (13) to account for all DUP possibilities. By calculating the optimal $\omega_{i,j}$ (denoted as $\omega_{i,j}^*$, $\forall (i, j) \in \mathcal{M} \times \mathcal{M}_i$) in $\mathbf{P1}_{i,j}$, the optimal coefficient matrix is given by

$$\boldsymbol{\Omega} = \begin{bmatrix} +\infty & \omega_{1,2}^* & \omega_{1,3}^* & \cdots & \omega_{1,M}^* \\ \omega_{2,1}^* & +\infty & \omega_{2,3}^* & \cdots & \omega_{2,M}^* \\ \omega_{3,1}^* & \omega_{3,2}^* & +\infty & \cdots & \omega_{3,M}^* \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \omega_{M,1}^* & \omega_{M,2}^* & \omega_{M,3}^* & \cdots & +\infty \end{bmatrix}. \quad (18)$$

$\boldsymbol{\Omega}$ is an $M \times M$ symmetric matrix in which $\omega_{i,j}^* = \omega_{j,i}^*$, and all elements on its main diagonal are set to $+\infty$ to indicate $i \neq j$. In addition, some $\omega_{i,j}^*$s in $\boldsymbol{\Omega}$ may have a value of $+\infty$ when SU j is not the eligible neighbor of SU i, i.e., $j \notin \mathcal{M}_i$.

On this basis, we will focus on finding the optimal DUP strategy by constructing a related subproblem in the second stage. With the objective $\widetilde{L}_{\boldsymbol{\tau}, \boldsymbol{\rho}}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{P})$ and $\boldsymbol{\beta}$-related constraints in (13), the DUP subproblem is constructed as

$$\mathbf{P2}: \max_{\boldsymbol{\beta}} \frac{1}{2} \sum_{i \in \mathcal{M}} \sum_{j \in \mathcal{M}_i} \beta_{i,j} \omega_{i,j}^* \quad (19)$$

$$\text{s.t.} \quad (10c), (10d), (10h). \quad (19a)$$

Given any $\boldsymbol{\tau}$ and $\boldsymbol{\rho}$, the optimal DUP strategy $\boldsymbol{\beta}$ (denoted as $\boldsymbol{\beta}^* = \left[\beta_{1,j_1^*}, \beta_{2,j_2^*}, \cdots, \beta_{M,j_M^*} \right]^T$) can be straightforwardly

---

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

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{P}, \boldsymbol{\tau}, \boldsymbol{\rho}) = \sum_{i \in \mathcal{M}} \sum_{j \in \mathcal{M}_i} \beta_{i,j} V_{i,j}^S + \sum_{i \in \mathcal{M}} \tau_i \left(\delta_0 - \sum_{j \in \mathcal{M}_i} \beta_{i,j} \delta_{i,j} \right) + \sum_{i \in \mathcal{M}} \rho_i \left(\sum_{j \in \mathcal{M}_i} \beta_{i,j} V_{i,j}^S - V_0 \right)$$

$$= \sum_{i \in \mathcal{M}} \sum_{j \in \mathcal{M}_i} \beta_{i,j} \left[(1 + \rho_i) V_{i,j}^S - \tau_i \delta_{i,j} \right] + \delta_0 \sum_{i \in \mathcal{M}} \tau_i - V_0 \sum_{i \in \mathcal{M}} \rho_i \quad (11)$$

$$\triangleq \widetilde{L}_{\boldsymbol{\tau}, \boldsymbol{\rho}}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{P}) + \delta_0 \sum_{i \in \mathcal{M}} \tau_i - V_0 \sum_{i \in \mathcal{M}} \rho_i.$$

finalized by solving **P2**, where β_{i,j_i^*} ($\forall i \in \mathcal{M}$) means that SU j_i^* is the optimal D2D SemCom node for SU i, i.e., $\beta_{i,j_i^*} = 1$. Then, we feed the obtained $\boldsymbol{\beta}^*$ back to $\boldsymbol{\Omega}$ to further determine the optimal KBC and power allocation policies $\boldsymbol{\alpha}^*$ and $\boldsymbol{P}^*$ for all SUs. In other words, for any $i \in \mathcal{M}$, we have $\boldsymbol{\alpha}_i^* = \boldsymbol{\alpha}_{i_{(j^*)}}^*$ and $P_i^* = P_{i_{(j^*)}}^*$, which is established by giving the following proposition.

Proposition 2. *Given any dual variable $\boldsymbol{\tau}$ and $\boldsymbol{\rho}$, $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{P}^*)$ is exactly the optimal solution to problem (13).*

Proof: Please see Appendix B. ∎

Based on Proposition 2, we have proved that the proposed two-stage method is guaranteed to reach the optimality of problem (13) in each iteration,. Most importantly, the optimization to either $\mathbf{P1}_{i,j}$ or $\mathbf{P2}$ becomes quite tractable owing to the significantly reduced number of variables. In the next two subsections, we will elucidate our optimal solutions to $\mathbf{P1}_{i,j}$ and $\mathbf{P2}$, respectively.

C. Near-Optimal KBC and Power Allocation for A D2D Pair

Carefully examining $\mathbf{P1}_{i,j}$, it is seen that the two binary variables $\boldsymbol{\alpha}_i$ and $\boldsymbol{\alpha}_j$ are the main tricky points during the solving process. Suppose that we fix $\boldsymbol{\alpha}_i$ and $\boldsymbol{\alpha}_j$ to any feasible solutions in the first place, only constraint (17e) remains and then we just need to concentrate upon its objective function $\omega_{i,j}$ w.r.t. variable P_i and P_j. Note that $\delta_{i,j}$ and $\delta_{j,i}$ in $\omega_{i,j}$ are monotonically increasing functions on P_i and P_j, respectively, which fact is quite obvious by observing (5). Moreover, when $\boldsymbol{\alpha}_i$ and $\boldsymbol{\alpha}_j$ are fixed to be known, $V_{i,j}^S$ can be rephrased to a more precise form, that is

$$V_{i,j}^S = \left[\theta_{i,j}^D \log_2 \left(1 + \frac{P_i G_{i,j}^D}{\delta^2} \right) - \theta_i^E \log_2 \left(1 + \frac{P_i G_i^E}{\delta^2} \right) \right]^+, \quad (20)$$

where $\theta_{i,j}^D$ and θ_i^E are constant parameters obtained by substituting the fixed $\boldsymbol{\alpha}_i$ and $\boldsymbol{\alpha}_j$ into (6) and (7), respectively. From (20), we can easily calculate $V_{i,j}^S$'s first order derivative w.r.t. P_i and obtain its monotonicity between 0 and P_{max}. Likewise, the monotonicity of P_j in $V_{j,i}^S$ can be easily determined at the same time. As such, since $V_{i,j}^S$ and $V_{j,i}^S$ are convex functions while $\delta_{i,j}$ $\delta_{j,i}$ are linear functions, their linear combination $\omega_{i,j}$ must also be convex given any $\boldsymbol{\rho}$ and $\boldsymbol{\tau}$. Therefore, if $\boldsymbol{\alpha}_i$ and $\boldsymbol{\alpha}_j$ can be fixed first, we can directly apply efficient linear programming toolboxes like CVXPY [39] to find the optimum P_i and P_j for the best $\omega_{i,j}$. As such, we propose a heuristic KBC search algorithm by drawing on tabu search [40] to efficiently determine the near-optimal solution for each $\mathbf{P1}_{i,j}$. In detail, the search process is illustrated as follows:

- *Initial KBC Solution Generation:* In the beginning, we first determine a feasible solution of $\boldsymbol{\alpha}_i$ and $\boldsymbol{\alpha}_j$ (jointly denoted as a $2K$-dimensional vector $\boldsymbol{\alpha}_{i,j}^I = [\boldsymbol{\alpha}_i^I, \boldsymbol{\alpha}_j^I]$) as the search starting point, where $\boldsymbol{\alpha}_i^I$ and $\boldsymbol{\alpha}_j^I$ are two K-dimensional vectors with all elements being 0 initialized to represent the KBC sub-policies at SU i and SU j, respectively. Considering constraint (17c), let $\hat{\mathcal{K}}$ and $\check{\mathcal{K}}$ denote two variable sets to record η_i-relevant information, where $\hat{\mathcal{K}} = \mathcal{K}$ and $\check{\mathcal{K}} = \emptyset$ are initialized. Having these,

we determine a certain KB k_0 that leads to the highest sum of KB preferences of SU i and SU j by

$$k_0 = \arg\max_{k \in \hat{\mathcal{K}}} \left(p_i^n + p_j^n \right), \quad (21)$$

such as

$$\alpha_i^{k_0} = 1 \quad \text{and} \quad \alpha_j^{k_0} = 1. \quad (22)$$

Then, let $\hat{\mathcal{K}} = \hat{\mathcal{K}} \backslash n_0$ and $\check{\mathcal{K}} = \check{\mathcal{K}} \cup \{n_0\}$, and repeat the two procedures in (21) and (22) until both SUs meet constraint (17c). However, it should note that constraint (17a) or (17b) may be violated during the above process. Once that happens, the corresponding KBC indicator of the KB with the maximum size in $\check{\mathcal{K}}$, i.e.,

$$k_1 = \arg\max_{k \in \check{\mathcal{K}}} s_k, \quad (23)$$

should be reset by

$$\alpha_i^{k_1} = 0 \quad \text{and} \quad \alpha_j^{k_1} = 0. \quad (24)$$

As a result, we can finally generate an initial feasible solution of $\boldsymbol{\alpha}_{i,j}^I$ to problem $\mathbf{P1}_{i,j}$. Meanwhile, as mentioned earlier, the optimal P_i^I and P_j^I corresponding to $\boldsymbol{\alpha}_{i,j}^I$ can be obtained.

- *Neighboring Space Exploration:* Our search algorithm is started from all solutions neighboring to the obtained $\boldsymbol{\alpha}_{i,j}^I$, and let $\mathcal{H}(\cdot)$ denote the neighboring solution set of its input solution. Next, we need to find the optimal solution within $\mathcal{H}(\boldsymbol{\alpha}_{i,j}^I)$ that can yield the minimum value of $\omega_{i,j}$, and then replacing the previous solution $\boldsymbol{\alpha}_{i,j}^I$ as well as P_i^I and P_j^I with the current solution, denoted by $(\boldsymbol{\alpha}_{i,j}^C, P_i^C, P_j^C)$. By repeating the above search process in an iteration fashion, we can obtain multiple different current solution in different search iterations. Besides, let $\mathcal{I}$ denote an optimal solution list with a given maximum length, and once we have a new better $\boldsymbol{\alpha}_{i,j}^C$ from searching $\mathcal{H}(\boldsymbol{\alpha}_{i,j}^C)$, the solution of $(\boldsymbol{\alpha}_{i,j}^C, P_i^C, P_j^C)$ should be added into $\mathcal{I}$ if it is not already in $\mathcal{I}$. The existence of $\mathcal{I}$ is to prevent each search from looping back to previously visited solution spaces, i.e., avoiding trapping into the local optimum. Technically, $\mathcal{H}(\boldsymbol{\alpha}_{i,j}^C)$ is given as

$$\mathcal{H}(\boldsymbol{\alpha}_{i,j}^C) = \left\{ \boldsymbol{\alpha}_{i,j} \colon \|\boldsymbol{\alpha}_{i,j} - \boldsymbol{\alpha}_{i,j}^C\| \leqslant \sigma, \\ \boldsymbol{\alpha}_{i,j} \notin \mathcal{I}, \boldsymbol{\alpha}_{i,j} \in \psi \right\}, \quad (25)$$

where σ is the maximum neighboring length, and ψ is $\mathbf{P1}_{i,j}$'s entire feasible solution space.

- *Optimal Solution Update and Termination Check:* Let $(\boldsymbol{\alpha}_{i,j}^*, P_i^*, P_j^*)$ denote a variable vector to record the best solution obtained so far. Specifically, if we find the current $(\boldsymbol{\alpha}_{i,j}^C, P_i^C, P_j^C)$ that is better than $(\boldsymbol{\alpha}_{i,j}^*, P_i^*, P_j^*)$ in any iteration, this should not be added into $\mathcal{I}$ but let

$$\boldsymbol{\alpha}_{i,j}^* = \boldsymbol{\alpha}_{i,j}^C, \quad P_i^* = P_i^C, \quad \text{and} \quad P_j^* = P_j^C. \quad (26)$$

Moreover, a search termination criterion is checked by either reaching a maximum number of iterations or a performance growth threshold in terms of $\omega_{i,j}$.

D. Optimal DUP Scheme

After solving each $\mathbf{P1}_{i,j}$, the optimal coefficient matrix $\boldsymbol{\Omega}$ as well as $\mathbf{P2}$ can be determined. Clearly, its objective function and constraints (10c) and (10d) are all linear, the only challenge to solve $\mathbf{P2}$ is the 0-1 constraint in (10h). Herein, we relax $\boldsymbol{\beta}$ into the continuous variable, denoted as $\boldsymbol{\beta}^R$, between 0 and 1, which makes $\mathbf{P2}$ a linear programming problem that is able to be efficiently tackled by CVXPY. Afterward, we need to recover $\boldsymbol{\beta}^R$ to its binary property, and for this, we propose a heuristic DUP strategy as follows. First, we determine one single SU i'-SU j' pair by

$$\beta^*_{i',j'} = \beta^*_{j',i'} = 1 \tag{27}$$

if

$$\left(i', j'\right) = \arg \max_{i \in \mathcal{M}, j \in \mathcal{M}_i, j > i} \beta^R_{i,j}. \tag{28}$$

For the remaining $\beta^*_{i,j}$ w.r.t. SU i' and SU j', we have

$$\begin{cases} \beta^*_{i',j} = \beta^*_{j,i'} = 0, & \forall j \in \mathcal{M}_{i'}, j \neq j' \\ \beta^*_{j',i} = \beta^*_{i,j'} = 0, & \forall i \in \mathcal{M}_{j'}, i \neq i' \end{cases}. \tag{29}$$

Then we let $\mathcal{M} = \mathcal{M} \setminus \{i,j\}$, and repeat the above processes until finalizing the optimal DUP solutions for all SU pairs. It is observed that the number of variables is only $\left(\sum_{i \in \mathcal{M}} |\mathcal{M}_i|\right)/2$ in solving $\mathbf{P2}$, which is a fairly acceptable problem scale in practice, and the performance compromise is believed to be small.

E. Algorithm Analysis

To better demonstrate the full picture of the proposed solution, we summarize the relevant technical points and enclose them in the following Algorithm 1.

Algorithm 1 The Proposed Resource Management for SSCN

Input: *The system parameters M, K, W, L, ζ, P_{max}, γ_0, η_0, δ_0, V_0, and the parameters of the eavesdropping center and all SUs $G^D_{i,j}$, G^E_i, s_k, C_i, r^k_i, r^k_E, ξ_i, ξ_E, μ^k_j*

Output: *The optimal KBC strategy $\boldsymbol{\alpha}^*$, optimal DUP strategy $\boldsymbol{\beta}^*$, and optimal power allocation strategy $\boldsymbol{P}^*$*

1: *Initialize dual problem iteration index $t \leftarrow 1$, $\tau_i(1)$, $\rho_i(1)$, $\nu_1(1)$, and $\nu_2(1)$ to proper positive values for $\mathbf{D0}$*
2: *Set the maximum number of iterations Q*
3: **while** $t \leqslant Q$ **do**
4:   **for** $i \leftarrow 1$ *to* M **do**
5:     **for** $j \leftarrow 1$ *to* M **do**
6:       **if** $j \in \mathcal{M}_i$ *and* $j > i$ **then**
7:         *Find $\left(\boldsymbol{\alpha}^I_{i,j}, P^I_i, P^I_j\right)$ in the context of (21)-(24)*
8:         *Initialize KBC search iteration index as $t' \leftarrow 1$, $\mathcal{I}(1) \leftarrow \emptyset$, $\boldsymbol{\alpha}^C_{i,j}(1) \leftarrow \boldsymbol{\alpha}^*_{i,j} \leftarrow \boldsymbol{\alpha}^I_{i,j}$, $P^C_i(1) \leftarrow P^*_i \leftarrow P^I_i$, and $P^C_j(1) \leftarrow P^*_j \leftarrow P^I_j$*
9:         *Set a suitable maximum neighboring length σ and maximum number of iterations Q' for $\mathbf{P1}_{i,j}$*
10:         **while** $t' \leqslant Q'$ **do**
11:           *Determine $\mathcal{H}\left(\boldsymbol{\alpha}^C_{i,j}(t')\right)$ by (25)*
12:           *Find the best KBC solution in $\mathcal{H}\left(\boldsymbol{\alpha}^C_{i,j}(t')\right)$ and then obtain the corresponding optimal power allocation solution by CVXPY*
13:           *Assign them to $\boldsymbol{\alpha}^C_{i,j}(t'+1)$, $P^C_i(t'+1)$, and $P^C_j(t'+1)$, respectively*
14:           **if** $\left(\boldsymbol{\alpha}^C_{i,j}(t'+1), P^C_i(t'+1), P^C_j(t'+1)\right)$ *is better than* $\left(\boldsymbol{\alpha}^*_{i,j}, P^*_i, P^*_j\right)$ **then**
15:             *Update $\left(\boldsymbol{\alpha}^*_{i,j}, P^*_i, P^*_j\right) \leftarrow \left(\boldsymbol{\alpha}^C_{i,j}(t'+1), P^C_i(t'+1), P^C_j(t'+1)\right)$*
16:             *Keep $\mathcal{I}(t'+1) \leftarrow \mathcal{I}(t')$*
17:           **else**
18:             *Update $\mathcal{I}(t'+1) \leftarrow \mathcal{I}(t') \cup \left\{\left(\boldsymbol{\alpha}^C_{i,j}(t'+1), P^C_i(t'+1), P^C_j(t'+1)\right)\right\}$*
19:           **end if**
20:           *Update $t' \leftarrow t'+1$*
21:         **end while**
22:         *Calculate $\omega^*_{i,j}$ by substituting $\left(\boldsymbol{\alpha}^*_{i,j}, P^*_i, P^*_j\right)$ into (16)*
23:         $\omega^*_{j,i} \leftarrow \omega^*_{i,j}$
24:       **end if**
25:     **end for**
26:   **end for**
27:   *Generate $\boldsymbol{\Omega}(t)$ according to (18)*
28:   *Solve $\mathbf{P2}$ by CVXPY and obtain $\boldsymbol{\beta}^R(t)$*
29:   *Finalize $\boldsymbol{\beta}^*(t)$ by (27)-(29)*
30:   *Finalize $\boldsymbol{\alpha}^*(t)$ and $\boldsymbol{P}^*(t)$ by feeding $\boldsymbol{\beta}^*(t)$ back to $\boldsymbol{\Omega}(t)$*
31:   *Update $\tau_i(t+1)$ and $\rho_i(t+1)$ by (14) and (15), respectively*
32:   *Update $\nu_1(t+1)$ and $\nu_2(t+1)$ under a given rule*
33:   $t \leftarrow t+1$
34: **end while**
35: **return** $\left(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{P}^*\right) \leftarrow \left(\boldsymbol{\alpha}^*(Q), \boldsymbol{\beta}^*(Q), \boldsymbol{P}^*(Q)\right)$

Regarding the computational complexity of Algorithm 1, it is first observed that for each $\mathbf{P1}_{i,j}$, we need to obtain all feasible KBC solutions within $\mathcal{H}\left(\boldsymbol{\alpha}^C_{i,j}\right)$ in any of its iterations. Note that σ is a very small parameter compared with N in (25), the complexity of determining the neighborhood solution space is approximately $\mathcal{O}\left(\binom{2K}{\sigma}\right) = \mathcal{O}(K^\sigma)$. With the maximum number of search iteration Q', solving each $\mathbf{P1}_{i,j}$ needs complexity $\mathcal{O}(Q'K^\sigma)$. Besides, the CVXPY toolbox employed for the relaxed $\mathbf{P2}$ requires $\mathcal{O}(M^4)$ complexity [41] to solve a group of $\left(\sum_{i \in \mathcal{M}} |\mathcal{M}_i|\right)$ DUP variables. Since there is a total of $\left(\left(\sum_{i \in \mathcal{M}} |\mathcal{M}_i|\right)/2\right)$ subproblems $\mathbf{P1}_{i,j}$ and one subproblem $\mathbf{P2}$ need to be tackled in each iteration of $\mathbf{D0}$, its corresponding complexity is $\mathcal{O}\left(M^2 Q'K^\sigma + M^4\right)$. Further combining the maximum number of dual problem iteration Q, the proposed Algorithm 1 has a polynomial-time overall complexity of $\mathcal{O}\left(QM^2\left(Q'K^\sigma + M^2\right)\right)$.

IV. NUMERICAL RESULTS AND DISCUSSIONS

In this section, numerical evaluations are conducted to demonstrate the performance of our proposed resource allocation solution in the SSCN, 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

TABLE I
SIMULATION PARAMETERS

Parameters	Values
Subchannel bandwidth (W)	0.1 MHz
Number of SUs (M)	100
Number of KBs (K)	12
Size of KB k (s_k)	$1 \sim 5$ units (randomly) [14]
KB storage capacity of SUs (C_i)	24 units
Skewness of the Zipf distribution w.r.t. each SU's KB preference (ξ_i)	1.2
Average interpretation time of KB n-based semantic packets ($1/\mu_i^n$)	$5 \times 10^{-3} \sim 1 \times 10^{-2}$ s/packet (randomly) [14]
Maximum transmit power of each SU (P_{max})	21 dBm [42]
Noise power (ζ^2)	-111.45 dBm [43]
Path loss model	$34 + 40 \log (d \text{ [m]})$ [44]
Average number of bits required for encoding one semantic triplet (L)	800 bits [45]
Minimum semantic knowledge satisfaction threshold (η_0)	0.5 [18]
Maximum queuing delay threshold (δ_0)	5 ms [46]
Minimum SST threshold (V_0)	50 [47]

Fig. 2. Average SST vs. different numbers of SUs.

with a radius of 300 meters, in which multiple SUs and one eavesdropping center are randomly dropped, while multiple KBs are preset to provide SUs with a variety of distinct SemCom services. Correspondingly, we set a uniform KB storage capacity for all SUs, and each of them has a randomly generated storage size. For brevity, other relevant simulation parameters not mentioned in the context along with their values are summarized in Table I.

Besides, the preference rankings for all KBs (i.e., r_i^n) at the eavesdropping center and each SU are generated independently and randomly, where their Zipf distributions are assumed to have the same skewness 1.2. The average interpretation time for semantic packets based on each KB n is considered to be the same for all SUs. Here, we randomly generate $1/\mu_i^n$ in a range of $5 \times 10^{-3} \sim 1 \times 10^{-2}$ s/packet w.r.t. each KB n, as the average interpretation time of different KBs-based packets is different from each other. Further, the minimum knowledge preference satisfaction threshold η_0, the maximum queuing delay threshold δ_0, and the maximum SST threshold V_0 are prescribed as 0.5, 5 ms, and 50, respectively. Notably, 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 different benchmark schemes of SemCom-empowered service provisioning herein [14], [48], [49]: 1) Random Power allocation with Distance-first pairing (RPD) strategy, which assumes each SU has a random assigned transmit power while choosing its nearest SU for DUP; 2) Maximum Power allocation with Knowledge-first pairing (MPK), in which each SU is assigned with a transmit power of P_{max} while selecting its neighboring SU with the highest KB matching degree for DUP. In addition,
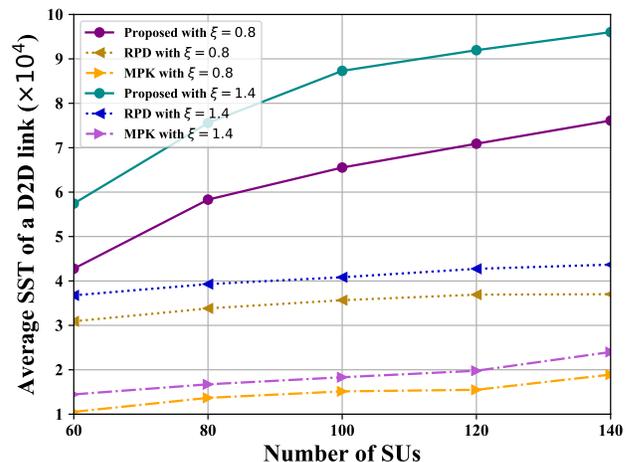
a personal preference-first KBC policy is considered for RPD and MPK, which allows each SU to cache KBs with the highest preferences until η_0 is satisfied, and then randomly cache the remaining KBs until reaching maximum capacity.

We first verify the average SST performance and the average queuing latency performance obtained by each D2D SemCom link under varying numbers of SUs in Fig. 2 and Fig. 3, respectively, in which two different skewness values of the Zipf distribution of $\xi = 0.8$ and $\xi = 1.4$ are taken into account. In comparison with the two benchmark schemes, our solution always renders an obvious SST performance gain at any number of SUs, while reaching a positive SST performance representing that the semantic information security has been adequately guaranteed in the SSCN. For example, in Fig. 2, our solution reaches a high SST performance of around 7.7×10^4 at 140 SUs with $\xi = 0.8$, which is much higher than around 3.8×10^4 SST of the RPD and around 1.9×10^4 SST of the MPK. In addition, in Fig. 3, a lower queuing latency is seen by our solution, which is less than half of that obtained by the comparison schemes. Note that as the increase of the number of SUs, both the SST performance and queuing latency show a grow trend. The former is because that the more SUs mean the more neighbors for each SU, and thus each SU should have a higher chance to find a better SemCom counterpart for D2D pairing and to achieve a better SST performance. However, such an increase will be eventually stabilized since the bandwidth budget is already fixed. Likewise, for the latter queuing delay trend, when each SU has a better semantic value rate, the semantic data packet arrival rate becomes correspondingly higher, thereby leading to a higher queuing delay. Furthermore, it is observed that the higher skewness of the Zipf distribution is with the better SST and a worse queuing delay. The rationale behind the former phenomenon is that the higher skewness makes each SU easier meet the semantic satisfaction as it only needs to cache the KBs with high probabilities, and the semantic triplets generated based on these KBs precisely have the high semantic value. Meanwhile, such a situation happens to cause that each SU is more difficult to find a high knowledge-matching D2D pair due to the higher deviation, hence the queuing delay
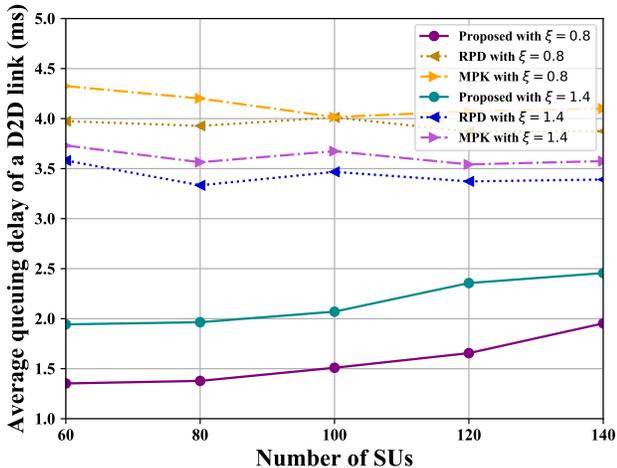
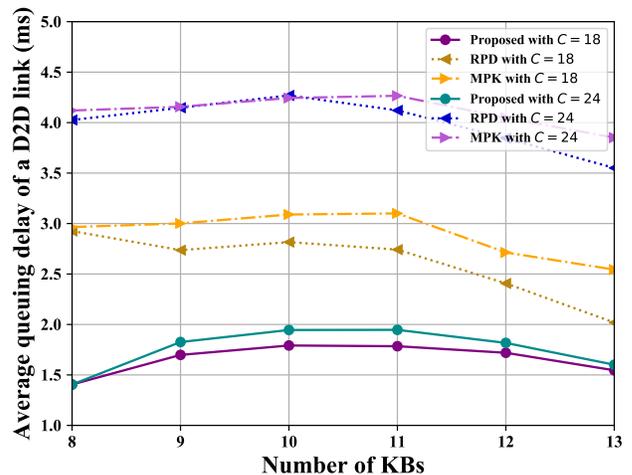Fig. 3. Average queuing delay vs. different numbers of SUs.

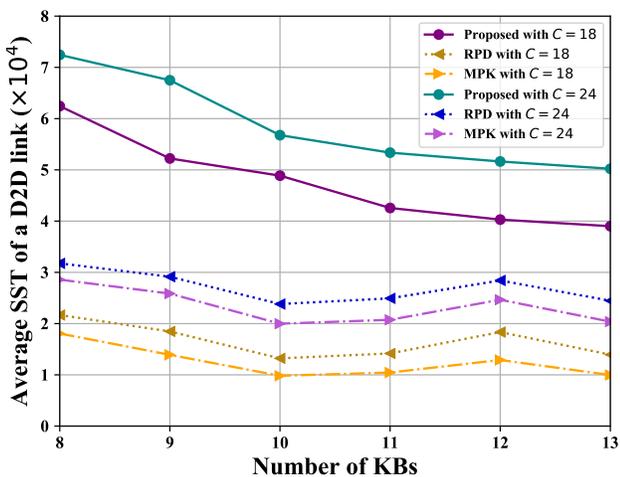Fig. 5. Average queuing delay vs. different numbers of KBs.

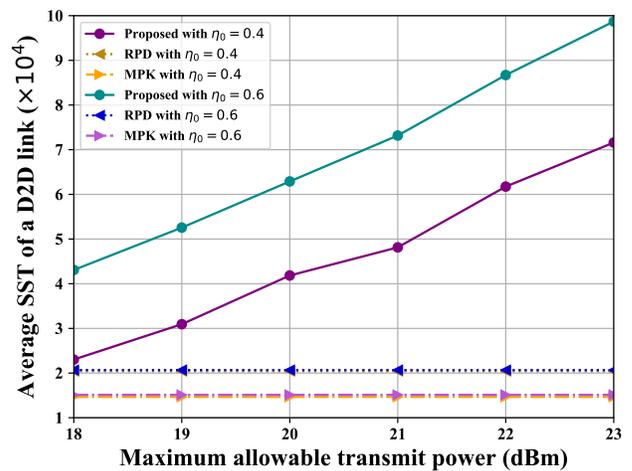Fig. 4. Average SST vs. different numbers of KBs.

Fig. 6. Average queuing delay vs. different maximum allowable transmit powers.

becomes higher as well compared with the low skewness.

Fig. 4 and 5 compare the average SST and queuing delay performance of each D2D SemCom link with the benchmarks under varying numbers of KBs, respectively, with two different KB storage capacities of $C = 18$ and $C = 24$. In both figures, it is seen that our proposed solution far superior to either the RPD or the MPK at each point. For instance, when the number of KBs is set to 10 and $C = 24$, an average SST performance of 5.8×10^4 is observed by the proposed solution in Fig. 4, which is around 1.9 times higher than that of the MPK and around 1.4 times higher than that of the RPD. Specially, in Fig. 4, our solution shows a downtrend of SST with the increase of the number of KBs. This is because that the more the number of KBs, the less the discrepancy of KB preference between SUs and the eavesdropping center, resulting in the higher probability for the eavesdropping center having the same KBs as SUs, which inevitably causes a lower semantic secrecy information rate. Besides, a larger KB capacity C has a better SST, which can be interpreted as the more KBs are cached at each SU, the more semantic value is transmitted.

In Fig. 5, an increase trend of the average queuing delay performance is first seen by the proposed solution with the number of KBs, and then decline slowly after 11 KBs. This is attribute to the fact that at the very beginning, the more KBs indicates a higher effective packet arrival rate, which can lead to a grow trend of queuing delay. However, such performance saturation will get better as a higher probability for two SUs caching the same KBs with high interpretation rates, which dominates the overall queuing delay performance change. Moreover, there is only a small queuing delay gap between the cases with different KB storage capacities. This is because that our solution always gives high caching priority to the KBs with lower interpretation times, and once the semantic knowledge satisfaction threshold or the queuing delay threshold is met, the remaining capacity will not be in use.

In addition, Fig. 6 and Fig. 7 validate the average queuing delay and SST performance under varying maximum allowable transmit powers, respectively, under two semantic knowledge satisfaction thresholds of $\eta_0 = 0.4$ and $\eta_0 = 0.6$. Consistent with the previous results, our solution still outperforms the two benchmarks with a significant performance gain. Specifically, at the point of $P_{max} = 21$ dBm and $\eta_0 = 0.6$, the proposed solution reaches an average SST performance of
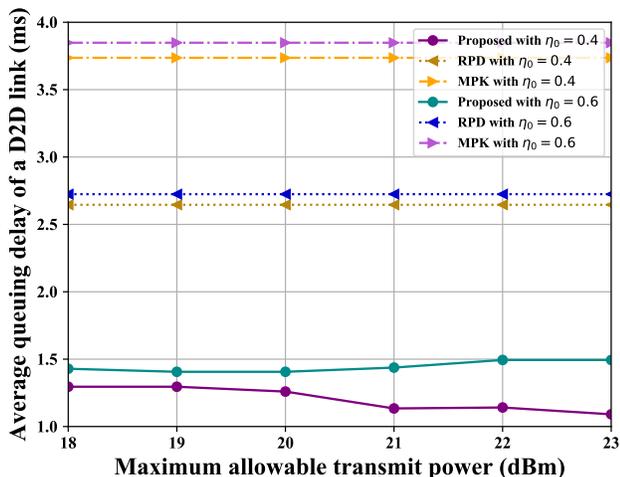
Fig. 7. Average queuing delay vs. different maximum allowable transmit powers.

around 73000 in Fig. 6, which is 53000 higher than that of RPD, and obtains an average queuing delay of around 1.4 ms in Fig. 7, which is 2.3 ms lower than that of MPK. Apart from this, note first that in Fig. 6, the SST performance increases as P_{max}. This is because that based on our efficient DUP solution, most of the SUs can always find a best SU for D2D SemCom pairing to guarantee a positive SST, in which case each SU is able to be assigned with a highest transmit power to improve the semantic value transmission rate as much as possible. Also, the higher η_0 leads to the better SST, which trend is quite obvious due to the looser KBC constraints.

Finally, as in Fig. 7, it is observed that the average queuing latency performance obtained by all resource allocation schemes shows a similar stable trend as the growth of P_{max}. This phenomenon can be understood by the fact that we just set a queuing latency limitation in the optimization problem, and our main focus is on the maximization of SST rather than minimizing the queuing delay. Hence, the corresponding solution here is just to satisfy the delay constraint, resulting in the stable trend. Nevertheless, our proposed solution still has a better delay performance compared with the benchmarks, as aforementioned. In the meantime, it is noticed that the lower the η_0, the lower the average queuing delay. This is because that the lower semantic knowledge satisfaction threshold clearly makes each SU cache less KBs, thus the knowledge overlapping rate between the two paring SUs is more likely to be smaller. Consequently, the effective semantic packet arrival rate at a lower η_0 should be smaller than that at a higher η_0, rendering a better queuing latency performance.

V. CONCLUSIONS

This paper has explored the semantic information security-aware resource allocation in the SSCN. We have first identified two key problems of KBC and DUP with the joint consideration of power allocation. Then, the knowledge matching based queuing delay has been derived and a novel performance metric of SST has been developed. On this basis, we have formulated a joint SST maximization problem, followed by a

corresponding solution proposed. Our solution first leveraged primal-dual transformation and decomposed the dual problem into multiple subproblems, which can be separately solved by our devised two-stage method. Compared with the two benchmarks, simulation results have shown superiorities in SST, average queuing delay, and semantic knowledge satisfaction.

We hope this paper can provide some valuable insights for follow-up research on secure SemCom. While this work establishes fundamental principles for secure resource allocation in D2D SemCom scenarios, real-world deployment in 5G/6G networks would require addressing additional challenges. For instance, in cellular-connected D2D scenarios, our KBC-DUP strategies may be capable of being implemented at edge servers to enable semantic-aware user pairing. However, practical constraints such as dynamic KBC across mobile users, semantic protocol standardization gaps, and latency overheads for SST calculation in multi-hop network environments need dedicated investigation, which will be systematically explored through testbed validation in our subsequent work.

APPENDIX A
PROOF OF PROPOSITION 1

Given the optimal KBC solution $\boldsymbol{\alpha}^*$ and optimal power allocation solution $\boldsymbol{P}^*$, let $\boldsymbol{\beta}^* = \left[\beta_{1,j_1^*}, \beta_{2,j_2^*}, \cdots, \beta_{M,j_M^*}\right]^T$ be the corresponding optimal DUP solution to the problem in (13) under the same dual variable $\boldsymbol{\tau}$ and $\boldsymbol{\rho}$, where β_{i,j_i^*} ($\forall i \in \mathcal{M}$) indicates that SU j_i^* is the optimal D2D SemCom counterpart for SU i, i.e., $\beta_{i,j_i^*} = 1$.

According to $\omega_{i,j}$ defined in (16), the objective function $\widetilde{L}_{\boldsymbol{\tau},\boldsymbol{\rho}}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{P})$ in (13) can be rewritten as

$$\widetilde{L}_{\boldsymbol{\tau},\boldsymbol{\rho}}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{P}) = \frac{1}{2}\sum_{i\in\mathcal{M}}\sum_{j\in\mathcal{M}_i}\beta_{i,j}\omega_{i,j} = \sum_{i\in\mathcal{M}}\sum_{j\in\mathcal{M}_i,j>i}\beta_{i,j}\omega_{i,j}, \tag{30}$$

then we substitute $\boldsymbol{\beta}^*$ into (30) and yield

$$\widetilde{L}_{\boldsymbol{\tau},\boldsymbol{\rho}}(\boldsymbol{\alpha},\boldsymbol{\beta}^*,\boldsymbol{P}) = \frac{1}{2}\sum_{i\in\mathcal{M}}\omega_{i,j_i^*} = \sum_{i\in\mathcal{M},j_i^*>i}\omega_{i,j_i^*}, \tag{31}$$

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

Clearly, if $\boldsymbol{\alpha}^*$ and $\boldsymbol{P}^*$ are further substituted into (31), we can directly reach the optimality of the problem (13). Since different SU i,j_i^* pairs should be independent of each other in practical systems, it means that different terms of ω_{i,j_i^*} are independent of each other as well in $\widetilde{L}_{\boldsymbol{\tau},\boldsymbol{\rho}}(\boldsymbol{\alpha},\boldsymbol{\beta}^*,\boldsymbol{P})$. Therefore, we are able to conclude that reaching the optimality of $\widetilde{L}_{\boldsymbol{\tau},\boldsymbol{\rho}}(\boldsymbol{\alpha},\boldsymbol{\beta}^*,\boldsymbol{P})$ is equivalent to reaching the optimality of each ω_{i,j_i^*}, where the optimality can be reached when $\boldsymbol{\alpha} = \boldsymbol{\alpha}^*$ and $\boldsymbol{P} = \boldsymbol{P}^*$ are met at the same time. As such, $\boldsymbol{\alpha}_i^*$ and P_i^* must be the optimal KBC and power allocation solutions in terms of the problem to maximize ω_{i,j_i^*}. Further noticing that $\omega_{i,j}$ is the objective of $\mathbf{P1}_{i,j}, \forall (i,j) \in \mathcal{M}\times\mathcal{M}_i, j>i$, where $\boldsymbol{\alpha}_{i_{(j)}}^*$ and $P_{i_{(j)}}^*$ are the corresponding optimal solutions, hence the equality $\boldsymbol{\alpha}_{i_{(j)}}^* = \boldsymbol{\alpha}_i^*$ and $P_{i_{(j)}}^* = P_i^*$ hold when $j = j_i^*$. This completes the proof.

APPENDIX B
PROOF OF PROPOSITION 2

If $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{P}^*)$ is not the optimal solution to (13), there must be an optimal solution, denoted as $\bar{\boldsymbol{\alpha}} = [\bar{\boldsymbol{\alpha}}_1, \bar{\boldsymbol{\alpha}}_2, \cdots, \bar{\boldsymbol{\alpha}}_M]^T$, $\bar{\boldsymbol{\beta}} = \left[\beta_{1,\bar{j}_1}, \beta_{2,\bar{j}_2}, \cdots, \beta_{M,\bar{j}_M} \right]^T$, and $\bar{\boldsymbol{P}} = \left[\bar{P}_1, \bar{P}_2, \cdots, \bar{P}_M \right]^T$, such that

$$\widetilde{L}_{\boldsymbol{\tau},\boldsymbol{\rho}}(\bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}}, \bar{\boldsymbol{P}}) > \widetilde{L}_{\boldsymbol{\tau},\boldsymbol{\rho}}(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{P}^*). \tag{32}$$

On one hand, as $\boldsymbol{\beta}^*$ is the optimal solution to **P2**, for $\bar{\boldsymbol{\beta}} \neq \boldsymbol{\beta}^*$, we have $\widetilde{L}_{\boldsymbol{\tau},\boldsymbol{\rho}}(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{P}^*) > \widetilde{L}_{\boldsymbol{\tau},\boldsymbol{\rho}}(\boldsymbol{\alpha}^*, \bar{\boldsymbol{\beta}}, \boldsymbol{P}^*)$. On the other hand, for the DUP case of $\boldsymbol{\beta} = \bar{\boldsymbol{\beta}}$, the optimal KBC and power allocation policies have been determined by solving each $\mathbf{P1}_{i,\bar{j}_i}$, i.e., $(\boldsymbol{\alpha}^*, \boldsymbol{P}^*)$. In other words, we have

$$\widetilde{L}_{\boldsymbol{\tau},\boldsymbol{\rho}}(\bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}}, \bar{\boldsymbol{P}}) < \widetilde{L}_{\boldsymbol{\tau},\boldsymbol{\rho}}(\boldsymbol{\alpha}^*, \bar{\boldsymbol{\beta}}, \boldsymbol{P}^*) < \widetilde{L}_{\boldsymbol{\tau},\boldsymbol{\rho}}(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{P}^*), \tag{33}$$

which leads to a contradiction between (32) and (33). Accordingly, the assumption of existing another optimal solution of $(\bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}}, \bar{\boldsymbol{P}})$ cannot hold, and thus $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \boldsymbol{P}^*)$ is exactly the optimal solution to problem (13).

REFERENCES

[1] 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.

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

[3] Q. Ye, J. Li, K. Qu, W. Zhuang, X. S. Shen, and X. Li, "End-to-end quality of service in 5G networks: Examining the effectiveness of a network slicing framework," *IEEE Vehicular Technology Magazine*, vol. 13, no. 2, pp. 65–74, 2018.

[4] D. Won, G. Woraphonbenjakul, A. B. Wondmagegn, A. T. Tran, D. Lee, D. S. Lakew, and S. Cho, "Resource management, security, and privacy issues in semantic communications: A survey," *IEEE Communications Surveys & Tutorials*, 2024.

[5] V. Niazmand and Q. Ye, "Joint task offloading, DNN pruning, and computing resource allocation for fault detection with dynamic constraints in industrial IoT," *IEEE Transactions on Cognitive Communications and Networking*, 2025.

[6] W. Zhuang, Q. Ye, F. Lyu, N. Cheng, and J. Ren, "SDN/NFV-empowered future IoV with enhanced communication, computing, and caching," *Proceedings of the IEEE*, vol. 108, no. 2, pp. 274–291, 2019.

[7] Y. Sun, L. Zhang, L. Guo, J. Li, D. Niyato, and Y. Fang, "S-RAN: Semantic-aware radio access networks," *arXiv preprint arXiv:2407.11161*, 2024.

[8] H. Du, J. Wang, D. Niyato, J. Kang, Z. Xiong, M. Guizani, and D. I. Kim, "Rethinking wireless communication security in semantic Internet of Things," *IEEE Wireless Communications*, vol. 30, no. 3, pp. 36–43, 2023.

[9] W. Zhang, K. Bai, S. Zeadally, H. Zhang, H. Shao, H. Ma, and V. C. Leung, "DeepMA: End-to-end deep multiple access for wireless image transmission in semantic communication," *IEEE Transactions on Cognitive Communications and Networking*, 2023.

[10] L. Xia, Y. Sun, C. Liang, L. Zhang, M. A. Imran, and D. Niyato, "Generative AI for semantic communication: Architecture, challenges, and outlook," *IEEE Wireless Communications*, 2024.

[11] X. Liu, Y. Sun, R. Cheng, L. Xia, H. Abumarshoud, L. Zhang, and M. A. Imran, "Knowledge-assisted privacy preserving in semantic communication," *IEEE Wireless Communications*, 2024.

[12] M. Shen, J. Wang, H. Du, D. Niyato, X. Tang, J. Kang, Y. Ding, and L. Zhu, "Secure semantic communications: Challenges, approaches, and opportunities," *IEEE Network*, vol. 38, no. 4, pp. 197–206, 2023.

[13] Z. Yang, M. Chen, G. Li, Y. Yang, and Z. Zhang, "Secure semantic communications: Fundamentals and challenges," *IEEE network*, 2024.

[14] 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.

[15] 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.

[16] M. Chein and M.-L. Mugnier, *Graph-based Knowledge Representation: Computational Foundations of Conceptual Graphs*. Springer Science & Business Media, 2008.

[17] S. T. Piantadosi, "Zipf's word frequency law in natural language: A critical review and future directions," *Psychonomic Bulletin & Review*, vol. 21, pp. 1112–1130, 2014.

[18] L. Xia, Y. Sun, D. Niyato, K. Ma, J. Kang, and M. A. Imran, "Knowledge base aware semantic communication in vehicular networks," in *ICC 2023-IEEE International Conference on Communications*. IEEE, 2023, pp. 3989–3994.

[19] N. B. Hassine, D. Marinca, P. Minet, and D. Barth, "Popularity prediction in content delivery networks," in *2015 IEEE 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*. IEEE, 2015, pp. 2083–2088.

[20] G. Li, J. Wu, J. Li, K. Wang, and T. Ye, "Service popularity-based smart resources partitioning for fog computing-enabled industrial Internet of Things," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 10, pp. 4702–4711, 2018.

[21] C. Li, C. She, N. Yang, and T. Q. Quek, "Secure transmission rate of short packets with queueing delay requirement," *IEEE Transactions on Wireless Communications*, vol. 21, no. 1, pp. 203–218, 2021.

[22] 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.

[23] S. M. Ross, *Introduction to Probability Models*. Academic Press, 2014.

[24] F. Pollaczek, "Über eine Aufgabe der Wahrscheinlichkeitstheorie. I," *Mathematische Zeitschrift*, vol. 32, no. 1, pp. 64–100, 1930.

[25] S. Yang, J. Tian, H. Zhang, J. Yan, H. He, Y. Jin *et al.*, "TransMS: Knowledge graph embedding for complex relations by multidirectional semantics," in *IJCAI*, 2019, pp. 1935–1942.

[26] X. Chen, S. Jia, and Y. Xiang, "A review: Knowledge reasoning over knowledge graph," *Expert Systems with Applications*, vol. 141, p. 112948, 2020.

[27] S. Gao, X. Qin, L. Chen, Y. Chen, K. Han, and P. Zhang, "Importance of semantic information based on semantic value," *IEEE Transactions on Communications*, 2024.

[28] 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.

[29] 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.

[30] 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.

[31] D. Qiao, M. C. Gursoy, and S. Velipasalar, "Secure wireless communication and optimal power control under statistical queueing constraints," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 3, pp. 628–639, 2011.

[32] D. W. K. Ng, E. S. Lo, and R. Schober, "Multiobjective resource allocation for secure communication in cognitive radio networks with wireless information and power transfer," *IEEE transactions on vehicular technology*, vol. 65, no. 5, pp. 3166–3184, 2015.

[33] H. Kellerer, U. Pferschy, and D. Pisinger, "Multidimensional knapsack problems," in *Knapsack Problems*. Springer, 2004, pp. 235–283.

[34] H. Tuy, T. Hoang, T. Hoang, V.-n. Mathématicien, T. Hoang, and V. Mathematician, *Convex Analysis and Global Optimization*. Springer, 1998.

[35] C. H. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*. Courier Corporation, 1998.

[36] 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.

[37] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[38] S. Boyd, L. Xiao, and A. Mutapcic, "Subgradient methods," *Lecture Notes of EE392o, Stanford University, Autumn Quarter*, vol. 2004, pp. 2004–2005, 2003.

[39] 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.

[40] F. Glover, "Tabu search-part I," *ORSA Journal on Computing*, vol. 1, no. 3, pp. 190–206, 1989.

[41] 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.

[42] 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.

[43] 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.

[44] Q. Ye, W. Zhuang, S. Zhang, A.-L. Jin, X. Shen, and X. Li, "Dynamic radio resource slicing for a two-tier heterogeneous wireless network," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 10, pp. 9896–9910, 2018.

[45] L. Xia, Y. Sun, D. Niyato, L. Zhang, and M. A. Imran, "Wireless resource optimization in hybrid semantic/bit communication networks," *IEEE Transactions on Communications*, 2024.

[46] G. Carlucci, L. De Cicco, and S. Mascolo, "Controlling queuing delays for real-time communication: The interplay of E2E and AQM algorithms," *ACM SIGCOMM Computer Communication Review*, vol. 46, no. 3, pp. 1–7, 2018.

[47] L. Xia, Y. Sun, H. Sun, R. Q. Hu, D. Niyato, and M. A. Imran, "Joint power and spectrum orchestration for D2D semantic communication underlying energy-efficient cellular networks," *arXiv preprint arXiv:2501.18350*, 2025.

[48] 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.

[49] 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.