Cache-Enabled Physical Layer Security for Video Streaming in Backhaul-Limited Cellular Networks

Lin Xiang, *Student Member, IEEE,* Derrick Wing Kwan Ng, *Senior Member, IEEE,*
Robert Schober, *Fellow, IEEE,* and Vincent W.S. Wong, *Fellow, IEEE*

Abstract—In this paper, we propose a novel wireless caching scheme to enhance the physical layer security of video streaming in cellular networks with limited backhaul capacity. By proactively sharing video data across a subset of base stations (BSs) through both caching and backhaul loading, secure cooperative joint transmission of several BSs can be dynamically enabled in accordance with the cache status, the channel conditions, and the backhaul capacity. Assuming imperfect channel state information (CSI) at the transmitters, we formulate a two-stage non-convex mixed-integer robust optimization problem for minimizing the total transmit power while providing quality of service (QoS) and guaranteeing communication secrecy during video delivery, where the caching and the cooperative transmission policy are optimized in an offline video caching stage and an online video delivery stage, respectively. Although the formulated optimization problem turns out to be NP-hard, low-complexity polynomial-time algorithms, whose solutions are globally optimal under certain conditions, are proposed for cache training and video delivery control. Caching is shown to be beneficial as it reduces the data sharing overhead imposed on the capacity-constrained backhaul links, introduces additional secure degrees of freedom, and enables a power-efficient communication system design. Simulation results confirm that the proposed caching scheme achieves simultaneously a low secrecy outage probability and a high power efficiency. Furthermore, due to the proposed robust optimization, the performance loss caused by imperfect CSI knowledge can be significantly reduced when the cache capacity becomes large.

I. INTRODUCTION

THE rapidly growing video-on-demand (VoD) streaming traffic in cellular networks has introduced significant challenges for service providers as both the radio resources

L. Xiang and R. Schober are with the Institute for Digital Communications, Friedrich-Alexander University of Erlangen-Nuremberg, Erlangen 91058, Germany (Email: {lin.xiang, robert.schober}@fau.de).

D. W. K. Ng is with the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2052, Australia (Email: w.k.ng@unsw.edu.au).

V.W.S. Wong is with the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC V6T 1Z4, Canada (Email: vincentw@ece.ubc.ca).

in the radio access network (RAN) and the capacity of the backhaul links are limited [2], [3]. Wireless caching has been proposed to meet the stringent VoD streaming requirements in the fifth generation (5G) cellular networks [4]–[9]. Different from traditional wireless networking paradigms, wireless caching is a content-centric solution for VoD streaming and intelligently exploits the fact that the content requested by users is highly correlated [10]. By pre-storing the most popular files at base stations (BSs) and access points (APs), wireless caching enables quick access to these files via wireless networks and, consequently, improves the quality of service (QoS) in video streaming. Meanwhile, by reusing the cached content for transmission to multiple users, the backhaul traffic is significantly reduced [4]. Recently, caching has also been exploited as a physical layer mechanism to facilitate cooperative multiple-input multiple-output (MIMO) transmission [5]–[8] and cross-layer resource allocation [9]. These schemes effectively exploit the multiplexing and diversity gains introduced by caching for spectral efficiency enhancement in the RAN as well as energy savings in the entire cellular network. Therefore, wireless caching is an appealing option for supporting cellular VoD streaming while providing a capacity enhancement in both the RAN and the backhaul. More importantly, caching has introduced a new type of cellular resource, namely, the cache memory, and additional degrees of freedom to enhance the system performance. Recently, an advanced network architecture for exploiting distributed cache memories at BSs, referred as Fog-RAN, has been proposed for 5G in [11]–[14].

Meanwhile, due to the broadcast nature of wireless transmission, VoD streaming data is vulnerable to potential eavesdroppers such as non-paying subscribers and malicious attackers. Thus, secure video streaming schemes providing both video data protection and streaming QoS guarantees are needed in 5G cellular networks. However, secure data delivery in cache-enabled transmission was not considered until recently. The existing works [15]–[17] were motivated by the coded caching scheme proposed in [18]. Specifically, each user is equipped with a local cache to pre-store parts of a popular video content. By properly encoding the cached and the delivered content (e.g., via index coding), coded multicast delivery opportunities are enabled for serving

various user requests at high delivery rates [18]. In [15], a coded caching scheme was proposed to guarantee secure information delivery when eavesdroppers attempt to intercept the video data over a multicast link. To ensure communication secrecy, the cached and the delivered contents are encoded/encrypted using random secret keys and secure coded multicast delivery is enabled based on Shannon's one-time pad method. Coded caching was extended to device-to-device (D2D) cellular networks in [16], where a sophisticated key generation and encryption scheme was proposed. However, the encryption methods in [15], [16] can incur significant signaling overhead for sharing the secret keys; for example, one-time pad based methods typically require the size of the secret keys to be as large as the file size. In [17], a non-encryption based secure video delivery scheme, which prevents eavesdroppers from obtaining the number of coded packets required for successful video file recovery, was proposed for cache-enabled heterogeneous small cell networks. Nevertheless, an information-theoretic characterization of the secrecy considered in [17] is missing.

On the other hand, MIMO-based physical layer security (PLS) techniques have been proposed to guarantee information-theoretic secrecy in 5G cellular networks [19]–[21]. Different from one-time pad or encryption methods, PLS techniques opportunistically exploit the inherent properties of wireless channels to enhance communication secrecy without using secret keys. In an MIMO wiretap channel with full channel state information (CSI), information-theoretic studies have revealed that the maximal number of secure degrees of freedom (s.d.o.f.)[1] enabled by multiple antennas is given by $\min([N_t - N_e]^+, N_r)$ [19]–[22], where N_t, N_r, and N_e are the number of transmit, receive, and eavesdropping antennas, respectively, and $[x]^+ \triangleq \max(x, 0)$. Because of their superior performance and low overhead compared to encryption based methods, MIMO-based PLS techniques have been widely advocated for secure transmission in cellular networks. However, to the best of the authors' knowledge, despite the increasing interest in secure cache-enabled communication, the benefits of caching for PLS enhancement have not been investigated in the literature yet.

To fill this void, in this paper, we show that caching is an effective method to enhance the PLS of cellular VoD streaming. Specifically, assume that each BS is equipped with a cache. By caching the same video data at different BSs, several BSs can participate in cooperative joint secure transmission of the video data. Correspondingly, by exploiting the large transmit antenna array formed by the cooperating BSs, the s.d.o.f. can be significantly increased and the PLS can be enhanced [19]. Meanwhile, as caching

reduces the data sharing overhead typically needed for BS cooperation [23], the s.d.o.f. can be increased even if the cellular network has capacity-constrained backhaul links. Interestingly, since data caching takes place *a priori*, e.g., in the early mornings when cellular traffic is low, the overhead of data sharing and channel estimation incurred by the cache-enabled joint transmission is comparable to that of traditional coordinated beamforming [23, Section IV-A] adopted for intercell interference mitigation, particularly when the cache capacity is large.

In [1], we have investigated cache-enabled cooperative joint secure transmission assuming perfect knowledge of the channels between the BSs and the legitimate users as well as the channels between the BSs and the eavesdropper. Yet, in practical systems, the channels to passive eavesdroppers are not perfectly known since an eavesdropper can remain silent for long periods of time. This leads to a reduction of the s.d.o.f. for cooperative transmission [21], [22] and increases the likelihood of data leakage. In this paper, we extend [1] to mitigate the impact of imperfect CSI knowledge. To enhance communication secrecy, artificial noise (AN)-based jamming is applied to effectively interfere the eavesdropper's reception [24]–[26]. In particular, since the AN is generated randomly and locally at the BSs involving neither the cache nor the backhaul, cooperative jamming by all BSs is considered in this paper for achieving high power efficiency. Moreover, to prevent data leakage under imperfect CSI and as power efficiency is of paramount importance for the design of future communication systems, the joint optimization of cooperative transmission and AN-based jamming is formulated as a robust optimization problem for minimization of the total transmit power while providing QoS and guaranteeing communication secrecy during video delivery.

We note that cache-enabled cooperative transmission for transmit power minimization and delivery time minimization has been investigated in [5]–[8] and [13], [14], respectively. However, these works did not consider PLS nor backhaul capacity constraints. Hence, the solutions proposed in [5]–[8], [13], [14] are not applicable to the considered problem, which motivates this work. The main contributions of this paper are as follows:

- We propose caching as a mechanism to enhance secure cellular video streaming with limited backhaul capacity. Thereby, the cache is instrumental for reducing the backhaul traffic and supporting secure data transmission via BS cooperation.
- Assuming imperfect CSI knowledge, we formulate a two-stage non-convex robust optimization problem for the minimization of the total BS transmit power while satisfying QoS and secrecy constraints for video delivery. Effective caching and delivery algorithms with

---

[1]Strictly positive s.d.o.f. indicate that the system's secrecy capacity can be scaled up by increasing the transmit power.

polynomial-time computational complexity are developed to solve the problem. In particular, we show that the proposed algorithms are asymptotically optimal when the cache capacity and the number of data sets adopted for cache training are sufficiently large.

- Simulation results show that the proposed schemes can efficiently utilize the cache capacity to enhance PLS and reduce the total BS transmit power. Moreover, the proposed robust algorithms ensure secure communication even for imperfect CSI knowledge.

The remainder of this paper is organized as follows. In Section II, we present the system model for cache enabled cooperative video delivery. The formulation and solution of the proposed two-stage robust control problem are provided in Sections III and IV, respectively. In Section V, the performance of the proposed robust algorithms is evaluated by simulation, and finally, Section VI concludes the paper.

Notation: Throughout this paper, $\mathbb{R}$ and $\mathbb{C}$ denote the sets of real and complex numbers, respectively; $\mathbb{C}^L$ denotes the set of $L \times L$ complex matrices; $\mathbf{I}_L$, $\mathbf{1}_L$, and $\mathbf{0}_L$ are the $L \times L$ identity, all-one, and zero matrices, respectively; $\mathrm{diag}(\mathbf{v})$ is a diagonal matrix with the diagonal elements given by the elements of vector $\mathbf{v}$; $(\cdot)^T$ and $(\cdot)^H$ are the transpose and complex conjugate transpose operators, respectively; $\mathrm{tr}(\cdot)$, $\mathrm{rank}(\cdot)$, $\det(\cdot)$, $\lambda_{\max}(\cdot)$, and $\|\cdot\|_F$ denote the trace, rank, determinant, maximum eigenvalue, and Frobenius norm of a square matrix, respectively; $\Pr(\cdot)$ and $\mathbb{E}(\cdot)$ denote the probability mass operator and the expectation operator, respectively; the circularly symmetric complex Gaussian distribution is denoted by $\mathcal{CN}(\boldsymbol{\mu}, \mathbf{C})$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$; $\sim$ stands for "distributed as"; $|\mathcal{X}|$ and $\mathrm{conv}(\mathcal{X})$ represent the cardinality and the convex hull of set $\mathcal{X}$, respectively; $\mathcal{X} \times \mathcal{Y}$ denotes the Cartesian product of sets $\mathcal{X}$ and $\mathcal{Y}$; $\mathbf{A} \succeq \mathbf{0}$ ($\mathbf{A} \succ \mathbf{0}$) indicates that matrix $\mathbf{A}$ is positive semidefinite (definite); $\nabla_{\mathbf{X}} f(\mathbf{X})$ denotes the complex-valued gradient of $f(\mathbf{X})$ with respect to matrix $\mathbf{X}$; finally, $\lfloor \cdot \rfloor$ denotes the rounding operator and $\binom{n}{k}$ is the binomial coefficient.

II. SYSTEM MODEL

We consider video streaming in the downlink of a multi-cell cellular network as shown in Fig. 1(a). A set of BSs, $\mathcal{M} = \{1, \ldots, M\}$, each equipped with N_t antennas, broadcast the video data to a set of single-antenna legitimate receivers (LRs), $\mathcal{K} = \{1, \ldots, K\}$. Since the broadcasted video data may be overheard by a passive eavesdropping receiver (ER), securing the video delivery is desirable to reduce the chance of potential information leakage. We assume that the ER is equipped with N_e antennas[2].

The video server located at the Internet edge owns a library of video files, $\mathcal{F} = \{1, \ldots, F\}$, which are intended

for delivery. The size of video file $f \in \mathcal{F}$ is V_f bits. The BSs are connected to the video server via dedicated "last-mile" wired backhaul links such as digital subscriber lines. We assume that the backhaul links are secure. However, since each backhaul is shared by different types of traffic (e.g., voice, data, multimedia, control signaling, etc.), the backhaul capacity available for supporting video streaming may be time-varying and limited. To reduce the backhaul traffic, a cache is deployed at each BS for pre-storing the video data.

The cache-enabled system is time-slotted and its operation is divided into two stages. In the first stage, a portion of the video files is cached at the BSs. To reduce the system overhead, the cache is updated when the network utilization is low, e.g., during early mornings. In the second stage, users request video files and in response, a subset of the BSs cooperate to address the requests. We assume that there is a central processor (CP) also located at the Internet edge, which is capable of performing computationally intensive signal processing tasks. The video caching and delivery control decisions are determined at the CP and conveyed to the BSs via the backhaul links. In this paper, we focus on static caching for notational convenience. Nevertheless, the proposed schemes can be extended to spatio-temporally dynamic caching by executing the corresponding algorithms multiple times across space and time. A list of key notations is shown in Table I.

A. Caching and Backhaul Loading in Two Stages

We assume hypertext transfer protocol (HTTP) based video streaming [27]. Thereby, file f is split/segmented into L subfiles/segments of equal sizes, i.e., V_f/L, and each subfile $(f,l) \in \mathcal{F} \times \mathcal{L}$ is delivered in one time slot[3], where $L \gg 1$ and $\mathcal{L} = \{1, \ldots, L\}$, cf. Fig. 1(b). Let binary variable $q_{f,l,m} \in \{0,1\}$ indicate the participation of BS $m \in \mathcal{M}$ in the cooperative transmission of subfile (f,l). The set of BSs cooperating for delivering subfile (f,l) is then defined as $\mathcal{M}_{f,l}^{\mathrm{Coop}} \triangleq \{m \in \mathcal{M} \mid q_{f,l,m} = 1\} \subseteq \mathcal{M}$. To facilitate cooperative BS transmission, the video data can be conveyed to the cooperating BSs in two manners: (1) caching the data ahead of time; (2) loading it via the backhaul links instantaneously during delivery. The data caching decisions are determined in the first stage based on the statistics or historical records of the user requests[4]. The cache status remains unchanged once it is updated according to the caching

---

[2]The ER may represent a set of distributed ERs with a total number of N_e antennas which are connected to perform joint eavesdropping.

[3]Herein, one time slot corresponds to a streaming session in HTTP video streaming. To facilitate a succinct formulation of the transmit power minimization problem for video file delivery, we neglect any inter-session delay caused by packet scheduling and user interruptions.

[4]Similar to file popularity estimation [4], the proposed cache training is performed based on profiles of user requests (and CSI) and does not require explicit *a priori* knowledge of the file popularity. Note that the proposed caching scheme is offline. Online cache training based on learning was studied in [28].

| $\mathcal{M}, \mathcal{K}, \mathcal{M}_{f,l}^{\text{Coop}}$	Sets of M BSs, K LRs, subset of cooperating BSs for delivery of subfile (f,l)
$\mathcal{F}, \mathcal{L}$	Sets of F video files and L subfiles per file
$\boldsymbol{\rho} \triangleq (k,f,l), \mathcal{S}$	Request of LR k for subfile (f,l) and set of user requests
$c_{f,l,m}, b_{f,l,m}, q_{f,l,m}$	Caching, backhaul loading, and cooperative delivery decisions for subfile (f,l) at BS m
$\mathbf{w}_{m,\boldsymbol{\rho}}, \mathbf{w}_{\boldsymbol{\rho}}$	Beamforming vectors for BS m and BS set $\mathcal{M}$
$\mathbf{v}, \mathbf{V}$	AN and its covariance matrix
$B_m^{\max}, C_m^{\max}$	Backhaul link capacity and cache size at BS m
$\Gamma_{\boldsymbol{\rho}}, R_{\boldsymbol{\rho}}, R_{\boldsymbol{\rho}}^{\text{sec}}$	SINR, achievable rate, and achievable secrecy rate at LR $\boldsymbol{\rho}$
$R_{\text{e},\boldsymbol{\rho}}$	Capacity of the ER for eavesdropping LR $\boldsymbol{\rho}$
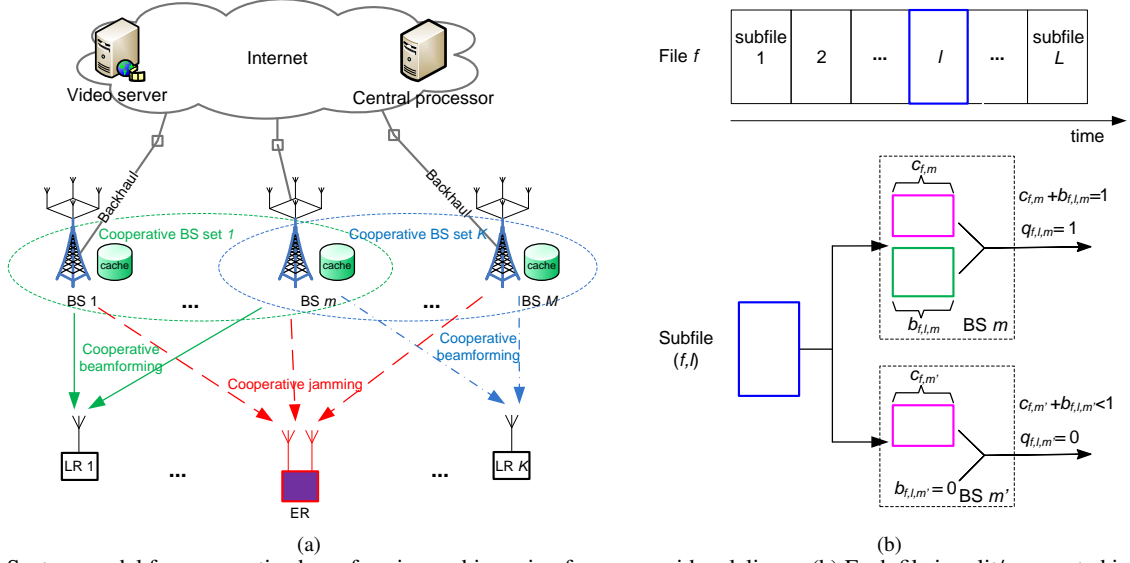
Fig. 1. (a) System model for cooperative beamforming and jamming for secure video delivery. (b) Each file is split/segmented into multiple subfiles/segments. Each subfile can be shared through both caching and backhaul loading to enable cooperative joint transmission.

decisions. Since the information of the user requests, the channel state, and the backhaul capacity are only known online at the time of request, joint optimization of backhaul loading and BS cooperative transmission is deferred to the second stage when this information is (partially) available. We note that two-stage control schemes have also been widely studied in the literature of stochastic control [29], [30] as they can incorporate online information for improved system performance. For a similar reason, two-stage control is a popular choice for operating cache-enabled cellular networks [5], [6], [9].

For the purpose of data sharing, we cache $c_{f,l,m} \in [0,1]$ portion and load via the backhaul $b_{f,l,m} \in [0,1]$ portion of subfile (f,l) at BS m, respectively. We assume that the same portion of each subfile of file f is cached at BS m, i.e., $c_{f,l,m} = c_{f,m}$, $\forall l \in \mathcal{L}$. Fig. 1(b) illustrates the caching and backhaul loading process for cooperative transmission of subfile (f,l). We can establish the following relation between caching, backhaul loading, and cooperation formation

$$b_{f,l,m} = (1 - c_{f,m})q_{f,l,m}. \tag{1}$$

That is, successful cooperative transmission of subfile (f,l) is possible, i.e., $q_{f,l,m} = 1$, only when its data symbols are fully available at BS m such that $b_{f,l,m} + c_{f,m} = 1$. Otherwise, (1) enforces $q_{f,l,m} = 0$ and $b_{f,l,m} = 0$.

The aggregate backhaul load of BS $m \in \mathcal{M}$ is given by

$$B_{m,l} = \sum_{f \in \mathcal{F}} b_{f,l,m} Q_f, \tag{2}$$

where the fixed parameter Q_f (in bps) represents the data rate required to load subfile (f,l) via the backhaul at BS m. We have $Q_f = V_f/(\tau L)$ or equivalently $\frac{c_{f,m}V_f}{L} + b_{f,l,m}Q_f\tau = \frac{V_f}{L}$, where τ denotes the duration of a time slot. Note that $B_{m,l} \leq \sum_{f \in \mathcal{F}} q_{f,l,m} Q_f$, in other words, caching reduces the backhaul capacity required for cooperative BS transmission.

B. Cooperative Beamforming and Jamming for Secure Video Delivery

When subfile (f,l) is available at BS subset $\mathcal{M}_{f,l}^{\text{Coop}}$, these BSs employ joint cooperative beamforming and jamming to deliver the subfile. Assume that an LR requests one (sub)file

at a time[5] [6]. We denote the request of LR k for subfile (f, l) by $\boldsymbol{\rho} \triangleq (k, f, l)$ and the set of user requests by $\mathcal{S} \subseteq \mathcal{K} \times \mathcal{F} \times \mathcal{L}$. Here, $\mathcal{S}$ is known at the beginning of the online delivery stage. With a slight abuse of notation, in the following, the LRs are also indexed by $\boldsymbol{\rho}$ when the requested (sub)files need to be specified. Note that there is a one-to-one correspondence between $\boldsymbol{\rho}$ and k.

The data symbols of subfile (f, l) for serving request $\boldsymbol{\rho}$ are denoted by $s_{\boldsymbol{\rho}} \in \mathbb{C}$ and are modeled as complex Gaussian random variables[6] with $s_{\boldsymbol{\rho}} \sim \mathcal{CN}(0, 1)$ [31, Ch. 5], [32, Ch. 9]. Let $\mathbf{w}_{m,\boldsymbol{\rho}} \in \mathbb{C}^{N_t \times 1}$ be the beamforming vector used at BS $m \in \mathcal{M}$ for sending symbol $s_{\boldsymbol{\rho}}$. As the cooperative BS set $\mathcal{M}_{f,l}^{\text{Coop}}$ dynamically changes with the cache status during backhaul loading, cf. (1), we impose

$$(1 - q_{f,l,m})\mathbf{w}_{m,\boldsymbol{\rho}} = \mathbf{0}, \quad \forall m \in \mathcal{M}, \ \forall \boldsymbol{\rho} \in \mathcal{S}, \qquad (3)$$

to adaptively adjust the beamforming vectors, $\mathbf{w}_{m,\boldsymbol{\rho}}$, according to the cooperation decisions, $q_{f,l,m}$. Based on (3), we have $\mathbf{w}_{m,\boldsymbol{\rho}} = \mathbf{0}, \forall m \notin \mathcal{M}_{f,l}^{\text{Coop}}$. Without loss of generality, let $\mathbf{w}_{\boldsymbol{\rho}} \triangleq [\mathbf{w}_{1,\boldsymbol{\rho}}^H, \ldots, \mathbf{w}_{M,\boldsymbol{\rho}}^H]^H$ $\in \mathbb{C}^{MN_t \times 1}$ be the joint beamforming vector for serving request $\boldsymbol{\rho}$. Furthermore, complex Gaussian distributed AN, $\mathbf{v} \in \mathbb{C}^{MN_t \times 1}$, is sent cooperatively by BS set $\mathcal{M}$ to proactively interfere the reception of the ER. We assume $\mathbf{v} \sim \mathcal{CN}(\mathbf{0}, \mathbf{V})$, where $\mathbf{V}$ is the covariance matrix of the artificial noise, i.e., $\mathbf{V} \triangleq \mathbb{E}\left[\mathbf{v}\mathbf{v}^H\right]$. The joint transmit signal of BS set $\mathcal{M}$, denoted as $\mathbf{x} \in \mathbb{C}^{MN_t \times 1}$, is thus given by

$$\mathbf{x} = \sum\nolimits_{\boldsymbol{\rho} \in \mathcal{S}} \mathbf{w}_{\boldsymbol{\rho}} s_{\boldsymbol{\rho}} + \mathbf{v}. \qquad (4)$$

Note that (3) and (4) describe a flexible BS cooperation topology, which allows any cooperative set $\mathcal{M}_{f,l}^{\text{Coop}} \subseteq \mathcal{M}$, provided that $\mathcal{M}_{f,l}^{\text{Coop}}$ ensures problem feasibility. For example, (3) and (4) correspond to joint transmission with full BS cooperation when $|\mathcal{M}_{f,l}^{\text{Coop}}| = M, \forall f$, and coordinated beamforming when $|\mathcal{M}_{f,l}^{\text{Coop}}| = 1, \forall f$, respectively.

C. Channel Model and Channel State Information

We assume a frequency flat fading channel model. The received signals at LR $\boldsymbol{\rho} \in \mathcal{S}$ and the ER, denoted by $y_{\boldsymbol{\rho}} \in \mathbb{C}$

and $\mathbf{y}_e \in \mathbb{C}^{N_e \times 1}$, respectively, are given by

$$y_{\boldsymbol{\rho}} = \mathbf{h}_{\boldsymbol{\rho}}^H \mathbf{x} + z_{\boldsymbol{\rho}} \qquad (5)$$
$$= \underbrace{\mathbf{h}_{\boldsymbol{\rho}}^H \mathbf{w}_{\boldsymbol{\rho}} s_{\boldsymbol{\rho}}}_{\text{desired signal}} + \underbrace{\sum\nolimits_{\boldsymbol{\rho}' \in \mathcal{S}, \boldsymbol{\rho}' \neq \boldsymbol{\rho}} \mathbf{h}_{\boldsymbol{\rho}}^H \mathbf{w}_{\boldsymbol{\rho}'} s_{\boldsymbol{\rho}'}}_{\text{multiuser interference}} + \underbrace{\mathbf{h}_{\boldsymbol{\rho}}^H \mathbf{v}}_{\text{injected AN}} + z_{\boldsymbol{\rho}},$$

and $\mathbf{y}_e = \mathbf{G}^H \mathbf{x} + \mathbf{z}_e$, where

$$\mathbf{h}_{\boldsymbol{\rho}} = [\mathbf{h}_{1,\boldsymbol{\rho}}^H, \ldots, \mathbf{h}_{M,\boldsymbol{\rho}}^H]^H \in \mathbb{C}^{MN_t \times 1} \quad \text{and} \qquad (6)$$
$$\mathbf{G} = [\mathbf{G}_1^H, \ldots, \mathbf{G}_M^H]^H \in \mathbb{C}^{MN_t \times N_e}$$

are the channel vectors/matrices from BS set $\mathcal{M}$ to LR $\boldsymbol{\rho}$ and the ER, respectively. $\mathbf{h}_{m,\boldsymbol{\rho}} \in \mathbb{C}^{N_t \times 1}$ and $\mathbf{G}_m \in \mathbb{C}^{N_t \times N_e}$ model the channels between BS $m \in \mathcal{M}$ and the corresponding LR/ER receivers; and $z_{\boldsymbol{\rho}} \sim \mathcal{CN}(0, \sigma^2)$ and $\mathbf{z}_e \sim \mathcal{CN}(\mathbf{0}, \sigma_e^2 \mathbf{I}_{N_e})$ are the zero-mean complex Gaussian noises at the LRs and the ER with variance σ^2 and covariance matrix $\sigma_e^2 \mathbf{I}_{N_e}$, respectively. Here, the channel vectors/matrices capture the joint effects of multipath fading and path loss; in addition, $z_{\boldsymbol{\rho}}$ and $\mathbf{z}_e$ represent the joint effects of thermal noise and possible out-of-system interference.

For channel estimation, we assume a time division duplex (TDD) system with slowly time-varying channels. At the beginning of each time slot, the LRs send orthogonal pilot sequences simultaneously in the uplink. Due to channel reciprocity, the LRs' downlink channels are estimated by measuring the pilot signals. We assume that the pilot sequences are long enough such that the CSI of the LRs can be reliably estimated at the CP[7]. In contrast, we consider imperfect CSI for the ER[8] since the ER does not directly interact with the BSs during channel estimation, which prevents a timely and accurate estimation of its CSI. Let $\widehat{\mathbf{G}} \in \mathbb{C}^{MN_t \times N_e}$ be the estimate of the ER channel matrix $\mathbf{G}$. We write $\mathbf{G} = \widehat{\mathbf{G}} + \Delta\mathbf{G}$, where $\Delta\mathbf{G}$ is the estimation error matrix and its value is unknown at the CP. To capture the effect of imperfect CSI, we assume that $\Delta\mathbf{G}$ lies in a continuous set of possible values given by

$$\mathcal{U}_e \triangleq \left\{ \Delta\mathbf{G} \in \mathbb{C}^{MN_t \times N_e} \mid \|\Delta\mathbf{G}\|_F^2 \leq \varepsilon_e^2 \right\}. \qquad (7)$$

Here, $\mathcal{U}_e$ is referred to as the uncertainty set associated with channel estimate $\widehat{\mathbf{G}}$ and $\varepsilon_e \geq 0$ is a measure for the accuracy of the channel estimate of the ER. In practice, the value of ε_e depends on the channel coherence time and the adopted channel estimation method. Note that the above imperfect

---

[5]If an LR requests several files simultaneously, orthogonal frequency multiple access (OFDMA) can be adopted to simultaneously deliver multiple video data streams to a single-antenna LR [31]. For a given subcarrier assignment, the model considered in this paper still applies at the expense of extra notations.

[6]Gaussian data symbols are capacity-achieving for the additive white Gaussian noise channel, and hence, are commonly assumed for analytical formulation of the achievable rate [31], [32]. On the other hand, if the modulation and coding schemes (MCS) employed in practical VoD systems are not capacity-achieving, the achievable rate expression can be amended by introducing an equivalent signal-to-noise ratio (SNR), $\frac{SNR}{\Delta}$, where $\Delta \geq 1$ defines the SNR gap between the channel capacity and the rate achievable with the adopted MCS [33].

[7]In practice, the CSI of the LRs can be imperfect if channel estimation errors occur, which degrades communication secrecy. The proposed robust optimization framework can be easily extended to tackle CSI imperfection at the LRs [25], [26]. However, as the imperfect CSI for the ER usually dominates other source of imperfection, we neglect the effect of imperfect CSI at the LRs to keep the paper readable. The results of this paper provide a performance upper bound for the case of imperfect CSI at LRs.

[8]Even though the ER is passive and silent, it may still be possible to estimate the CSI of the ER based on the local oscillator power that is unintentionally leaked from its radio frequency (RF) front end during eavesdropping [34].

CSI model is commonly adopted in the literature [25], [26], and the special case $\varepsilon_{\mathrm{e}} = 0$ corresponds to the optimistic scenario when $\mathbf{G}$ is perfectly known at the CP.

D. Achievable Secrecy Rate

The achievable rate at LR $\boldsymbol{\rho} \in \mathcal{S}$, denoted by $R_{\boldsymbol{\rho}}$, is given by

$$R_{\boldsymbol{\rho}} = \log\left(1 + \Gamma_{\boldsymbol{\rho}}\right), \tag{8}$$

$$\Gamma_{\boldsymbol{\rho}} = \frac{\frac{1}{\sigma^2}\left|\mathbf{h}_{\boldsymbol{\rho}}^H \mathbf{w}_{\boldsymbol{\rho}}\right|^2}{1 + \frac{1}{\sigma^2}\sum_{\boldsymbol{\rho}' \in \mathcal{S}, \boldsymbol{\rho}' \neq \boldsymbol{\rho}}\left|\mathbf{h}_{\boldsymbol{\rho}}^H \mathbf{w}_{\boldsymbol{\rho}'}\right|^2 + \frac{1}{\sigma^2}\mathbf{h}_{\boldsymbol{\rho}}^H \mathbf{V}\mathbf{h}_{\boldsymbol{\rho}}},$$

where $\Gamma_{\boldsymbol{\rho}}$ is the received signal-to-interference-plus-noise ratio (SINR) at LR $\boldsymbol{\rho}$. In (8), the term $\sum_{\boldsymbol{\rho}' \in \mathcal{S}, \boldsymbol{\rho}' \neq \boldsymbol{\rho}}\left|\mathbf{h}_{\boldsymbol{\rho}}^H \mathbf{w}_{\boldsymbol{\rho}'}\right|^2$ accounts for the interference power caused by multiuser transmission, cf. (5).

To guarantee secure VoD streaming, the proposed secure delivery scheme is designed to avoid VoD data leakage even under worst-case conditions. Specifically, we assume that the ER can eavesdrop the information intended for LR $\boldsymbol{\rho}$ after canceling the interference caused by all other LRs. This is possible if the ER adopts advanced receiver structures such as successive interference cancellation decoders [31, Ch. 8.3]. Thus, the achievable secrecy rate for LR $\boldsymbol{\rho}$ is modeled as [19], [25],

$$R_{\boldsymbol{\rho}}^{\mathrm{sec}} = \left[R_{\boldsymbol{\rho}} - R_{\mathrm{e},\boldsymbol{\rho}}\right]^+, \ \boldsymbol{\rho} \in \mathcal{S}, \tag{9}$$

where $R_{\mathrm{e},\boldsymbol{\rho}}$ denotes the capacity of the ER for decoding subfile (f, l) of LR $\boldsymbol{\rho}$ and is given by [31, Chapter 8]

$$R_{\mathrm{e},\boldsymbol{\rho}} = \log\det\left(\mathbf{I}_{N_{\mathrm{e}}} + \frac{1}{\sigma_{\mathrm{e}}^2}\mathbf{G}\mathbf{Z}_{\mathrm{e},\boldsymbol{\rho}}^{-1}\mathbf{G}^H \mathbf{w}_{\boldsymbol{\rho}} \mathbf{w}_{\boldsymbol{\rho}}^H\right), \ \boldsymbol{\rho} \in \mathcal{S}, \tag{10}$$

with $\mathbf{Z}_{\mathrm{e},\boldsymbol{\rho}} = \mathbf{I}_{N_{\mathrm{e}}} + \frac{1}{\sigma_{\mathrm{e}}^2}\mathbf{G}^H \mathbf{V}\mathbf{G}$.

According to (9), for power-efficient secure video delivery, the cooperative beamforming and jamming has to ensure a certain rate difference between $R_{\boldsymbol{\rho}}$ and $R_{\mathrm{e},\boldsymbol{\rho}}$ while consuming as little transmit power as possible. This means that multiuser interference should be reduced through cooperative beamforming across a virtual transmit antenna array whose size, according to (1) and (3), depends on the caching decisions. At the same time, the covariance matrix $\mathbf{V}$ for cooperative jamming has to be properly designed to degrade the capacity of the ER channel, cf. (10), without causing interference to the LRs, cf. (8). Therefore, caching, cooperative beamforming, and AN-based jamming have to be judiciously and jointly optimized to reap the benefits of cache-enabled secrecy.

III. ROBUST TWO-STAGE PROBLEM FORMULATION

In this section, we formulate a two-stage robust optimization problem for minimization of the total BS transmit power while taking into account the secrecy/QoS constraints of the LRs and the imperfect CSI knowledge regarding the ER. In the first stage, the cached video data is optimized *offline*.

In the second stage, the cooperative transmission strategy is optimized *online* for a given cache and backhaul status. For ease of discussion, we first present the formulation of the second-stage problem.

A. Second-Stage Online Delivery Control

The BS cooperation formation policy $\mathbf{D}_{\mathrm{II},1} \triangleq [q_{f,l,m}, b_{f,l,m}]$ and the cooperative transmission policy $\mathbf{D}_{\mathrm{II},2} \triangleq [\mathbf{w}_{\boldsymbol{\rho}}, \mathbf{V}]$ (including beamforming and jamming) are optimized in the second stage. For this purpose, we assume that the set of user requests $\mathcal{S}$ is given and the cache status $\{c_{f,m}\}$ has already been determined in the first stage. Let $\mathbf{D}_{\mathrm{II}} \triangleq [\mathbf{D}_{\mathrm{II},1}, \mathbf{D}_{\mathrm{II},2}]$ be the optimization space of the second (delivery) stage. The considered optimization problem is formulated in (11) at the top of the next page, where B_m^{max} is the backhaul capacity available for video sharing at BS m, P_m^{max} is the maximum transmit power at BS m, and $\boldsymbol{\Lambda}_m$ is an $MN_{\mathrm{t}} \times MN_{\mathrm{t}}$ diagonal matrix defined as $\boldsymbol{\Lambda}_m = \mathrm{diag}(\mathbf{0}_{(m-1)N_{\mathrm{t}} \times 1}^T, \mathbf{1}_{N_{\mathrm{t}} \times 1}^T, \mathbf{0}_{(M-m)N_{\mathrm{t}} \times 1}^T)$, such that $\mathrm{tr}\left(\mathbf{w}_{m,\boldsymbol{\rho}} \mathbf{w}_{m,\boldsymbol{\rho}}^H\right) = \mathrm{tr}\left(\boldsymbol{\Lambda}_m \mathbf{w}_{\boldsymbol{\rho}} \mathbf{w}_{\boldsymbol{\rho}}^H\right)$ holds. By constraints C1, C2, and C4, BS m is only allowed to cooperate in transmitting subfile (f, l) if $q_{f,l,m} = 1$ or $b_{f,l,m} + c_{f,l,m} = 1$; otherwise, if $q_{f,l,m} = 0$ and $b_{f,l,m} + c_{f,l,m} < 1$, BS m cannot cooperate as $\mathbf{w}_{\boldsymbol{\rho}} = \mathbf{0}$. Constraints C3 and C5 limit the maximum backhaul rate and the maximum transmit power of each BS, respectively. C6 guarantees a minimum required video delivery rate, $R_{\boldsymbol{\rho}}^{\mathrm{req}}$, to provide video streaming QoS for LR $\boldsymbol{\rho}$. C7 ensures that the capacity of the ER is kept below a maximum tolerable secrecy threshold $R_{\mathrm{e},\boldsymbol{\rho}}^{\mathrm{tol}}$ for video data protection. Due to the imperfect CSI of the ER, the capacity of the ER is constrained for all possible estimation error matrices within the uncertainty set, i.e., for any $\Delta\mathbf{G} \in \mathcal{U}_{\mathrm{e}}$, in C7, to provide robustness in secure communication. C6 and C7 together guarantee a minimum achievable secrecy rate of $R_{\boldsymbol{\rho}}^{\mathrm{sec}} = [R_{\boldsymbol{\rho}}^{\mathrm{req}} - R_{\mathrm{e},\boldsymbol{\rho}}^{\mathrm{tol}}]^+$ for LR $\boldsymbol{\rho}$, provided that problem R0 is feasible.

Remark 1. If, instead of (7), the probability distribution of the channel estimation error matrix is known, e.g., if $\Delta\mathbf{G}$ is complex-valued Gaussian distributed, the robust design in R0 is still applicable and can ensure a probabilistic secrecy outage constraint, e.g., $\mathrm{Pr}(R_{\boldsymbol{\rho}}^{\mathrm{sec}} < [R_{\boldsymbol{\rho}}^{\mathrm{req}} - R_{\mathrm{e},\boldsymbol{\rho}}^{\mathrm{tol}}]^+) \leq \alpha$, by following a similar approach as in [25]. Meanwhile, if perfect CSI knowledge of the ER is available, i.e., $\varepsilon_{\mathrm{e}} = 0$ in C7, the optimal solution of R0 provides a performance upper bound for the case of imperfect CSI. However, the methods for solving R0 for $\varepsilon_{\mathrm{e}} = 0$ and $\varepsilon_{\mathrm{e}} > 0$ differ slightly in how C7 is dealt with, cf. Section IV.

B. First-Stage Offline Cache Training

A historical data driven approach [9], [29] is adopted for optimizing the offline caching in the first stage. Assume that Ω sets of typical scenario data are available for training

$$\text{R0: minimize} \quad f_{\mathrm{II}} \triangleq \sum_{\boldsymbol{\rho} \in \mathcal{S}} \mathrm{tr}(\mathbf{w}_{\boldsymbol{\rho}} \mathbf{w}_{\boldsymbol{\rho}}^H + \mathbf{V}) \tag{11}$$

$$\text{subject to} \quad \text{C1: } b_{f,l,m} = (1 - c_{f,m}) q_{f,l,m}, \ (f,l) \in \mathcal{F} \times \mathcal{L}, \ m \in \mathcal{M},$$

$$\text{C2: } b_{f,l,m} \in [0,1], \ q_{f,l,m} \in \{0,1\}, \ (f,l) \in \mathcal{F} \times \mathcal{L}, \ m \in \mathcal{M},$$

$$\text{C3: } \sum_{f \in \mathcal{F}} b_{f,l,m} Q_f \le B_m^{\max}, \ \forall l, \forall m, \quad \text{C4: } (1 - q_{f,l,m}) \mathbf{w}_{\boldsymbol{\rho}} = \mathbf{0}, \ \forall m, \forall \boldsymbol{\rho},$$

$$\text{C5: } \mathrm{tr}\left(\boldsymbol{\Lambda}_m \left(\sum_{\boldsymbol{\rho} \in \mathcal{S}} \mathbf{w}_{\boldsymbol{\rho}} \mathbf{w}_{\boldsymbol{\rho}}^H + \mathbf{V} \right) \right) \le P_m^{\max}, \quad m \in \mathcal{M},$$

$$\text{C6: } R_{\boldsymbol{\rho}} \ge R_{\boldsymbol{\rho}}^{\mathrm{req}}, \ \boldsymbol{\rho} \in \mathcal{S}, \qquad \text{C7: } \max_{\Delta \mathbf{G} \in \mathcal{U}_e} R_{e,\boldsymbol{\rho}} \le R_{e,\boldsymbol{\rho}}^{\mathrm{tol}}, \ \boldsymbol{\rho} \in \mathcal{S},$$

$$\text{variables} \quad \mathbf{D}_{\mathrm{II}} = [q_{f,l,m}, b_{f,l,m}, \mathbf{w}_{\boldsymbol{\rho}}, \mathbf{V}],$$

---

the cache, e.g., obtained from past system records. Each set of scenario data consists of the user requests, CSI, and the available backhaul capacities at a particular time instant. The scenario data is indexed by $\omega \in \{1, \ldots, \Omega\}$. Let $\mathbf{C}_{\mathrm{I}} \triangleq [c_{f,m}, \mathbf{D}_{\mathrm{I},\omega}]$ be the first-stage (caching) optimization space, where $\mathbf{D}_{\mathrm{I},\omega} \triangleq [q_{f,l,m,\omega}, b_{f,l,m,\omega}, \mathbf{w}_{\boldsymbol{\rho},\omega}, \mathbf{V}_\omega]$ denotes the auxiliary delivery decisions for scenario ω during training. We define the feasible delivery set for scenario ω by $\mathcal{D}_{\mathrm{I},\omega} \triangleq \{\mathbf{D}_{\mathrm{I},\omega} \mid \text{C1, C2, C4–C7}\}$, where C1, C2, and C4–C7 need to be reformulated with an augmented system state space. For example, C1 and C2 are rewritten as

$$\text{C1: } b_{f,l,m,\omega} = (1 - c_{f,m}) q_{f,l,m,\omega}, \tag{12}$$

$$\text{C2: } c_{f,m}, b_{f,l,m,\omega} \in [0,1], \ q_{f,l,m,\omega} \in \{0,1\},$$

and C4–C7 are similarly formulated.

The objective of the first-stage problem is the minimization of the average transmit power for the considered scenarios, i.e.,

$$\text{Q0: minimize} \ \frac{1}{\Omega} \sum_{\omega=1}^{\Omega} f_{\mathrm{I},\omega} \tag{13}$$

$$\text{subject to } \overline{\text{C3}}: \frac{1}{\Omega}\sum_{\omega=1}^{\Omega}\sum_{f \in \mathcal{F}} b_{f,l,m,\omega} Q_f \le \frac{1}{\Omega}\sum_{\omega=1}^{\Omega} B_{m,\omega}^{\max}, \forall m,$$

$$\text{C8: } \sum_{f \in \mathcal{F}} c_{f,m} V_f \le C_m^{\max}, \ m \in \mathcal{M},$$

$$\text{C9: } \mathbf{D}_{\mathrm{I},\omega} \in \mathcal{D}_{\mathrm{I},\omega}, \ \omega \in \{1, \ldots, \Omega\},$$

$$\text{variables } \mathbf{C}_{\mathrm{I}} = [c_{f,m}, q_{f,l,m,\omega}, b_{f,l,m,\omega}, \mathbf{w}_{\boldsymbol{\rho},\omega}, \mathbf{V}_\omega],$$

where $f_{\mathrm{I},\omega} \triangleq \sum_{\boldsymbol{\rho} \in \mathcal{S}} \mathrm{tr}(\mathbf{w}_{\boldsymbol{\rho},\omega} \mathbf{w}_{\boldsymbol{\rho},\omega}^H + \mathbf{V}_\omega)$ is the instantaneous transmit power for scenario ω and $C_m^{\max}$ is the cache capacity at BS m. $\overline{\text{C3}}$ is an average backhaul capacity constraint and C8 is the cache capacity constraint. Considering $\overline{\text{C3}}$ and C8, the number of cooperating BSs during online VoD streaming depends largely on the values of $C_m^{\max}$ and $B_{m,\omega}^{\max}$. Note that $\overline{\text{C3}}$ is a relaxation of the per-scenario backhaul capacity constraint $\sum_{f \in \mathcal{F}} b_{f,l,m,\omega} Q_f \le B_{m,\omega}^{\max}$, $\omega \in \{1, \ldots, \Omega\}$. In the caching phase, when the actual backhaul capacity at the time of delivery is uncertain, $\overline{\text{C3}}$ avoids

the conservative use of the backhaul links, considering that the actual cooperative transmission decisions are deferred to the second stage when the available backhaul capacity is known. Furthermore, $\overline{\text{C3}}$ facilitates the design of low-complexity asymptotically optimal cache training as will be revealed in Section IV-C.

Problems R0 and Q0 are non-convex mixed-integer non-linear programs (MINLPs)[9] due to non-convex constraints C6, C7, and the binary optimization variables $q_{f,l,m} \in \{0,1\}$ and $q_{f,l,m,\omega} \in \{0,1\}$. Moreover, problems R0 and Q0 involve bilinear constraints C1 and C4. Even setting aside secrecy, the formulated two-stage control problem is more practical compared to the problems considered in [6]–[8], as it accounts for binary cooperation formation decisions, capacity-constrained backhaul links, and imperfect CSI. However, MINLPs are generally NP-hard and there are no known polynomial-time algorithms to solve them optimally [35]. In Section IV, to strike a balance between computational complexity and optimality, we present two effective polynomial-time suboptimal algorithms for solving problems R0 and Q0, respectively. We further show that the proposed algorithms are asymptotically optimal when the cache capacity and the number of scenarios considered are sufficiently large, respectively.

IV. SOLUTION OF TWO-STAGE PROBLEM

In this section, the solutions of problems R0 and Q0 are presented. We first tackle problem R0. The approach employed for solving problem R0 is then extended to solve problem Q0.

A. Optimal Solution of Problem R0 in Large Cache Capacity Regime

We first discuss special conditions under which problem R0 is solvable in polynomial time. The corresponding results provide the basis for solving problem R0 in the general case.

---

[9]For a non-convex MINLP, even if the integer constraints are relaxed to convex constraints, the problem remains non-convex [35].

Let $\mathcal{F}(\mathcal{S}) \triangleq \{f \mid (k,f,l) \in \mathcal{S}\}$ be the set of files requested by the LRs in $\mathcal{S}$, where $\mathcal{F}(\mathcal{S}) \subseteq \mathcal{F}$. We define $F(\mathcal{S}) \triangleq |\mathcal{F}(\mathcal{S})|$, which satisfies $F(\mathcal{S}) \leq \min\{|\mathcal{S}|, F\}$. For a given cache status $\{c_{f,m}\}$, C1 is an affine equality constraint which enables the elimination of the backhaul loading decision variables $b_{f,l,m}$. As a result, problem R0 is reformulated as

R0: minimize f_{II} (14)

$\quad$ subject to C4, C5, C6, C7, $\mathbf{V} \succeq \mathbf{0}$,

$$\overline{\mathrm{C2}}: q_{f,l,m} \in \{0,1\}, (f,l) \in \mathcal{F} \times \mathcal{L}, m \in \mathcal{M},$$

$$\widetilde{\mathrm{C3}}: \sum_{f \in \mathcal{F}(\mathcal{S})} q_{f,l,m} Q_{f,m} \leq B_m^{\max}, \forall l, \forall m,$$

$\quad$ variables $\mathbf{D}_{\mathrm{II}} = [q_{f,l,m}, b_{f,l,m}, \mathbf{w}_{\boldsymbol{\rho}}, \mathbf{V}]$,

where $Q_{f,m} \triangleq (1 - c_{f,m})Q_f$ is the "effective" data rate required for loading subfile (f,l) into BS m via the backhaul link and hence, constraints $\widetilde{\mathrm{C3}}$ and C3 are equivalent. We have the following lemma for the BS cooperation formation in (14).

Lemma 1 (Monotonicity of R0). Let f_{II}^1 and f_{II}^2 denote the optimal objective values for given cooperation sets $\mathcal{M}_{f,l}^{\mathrm{Coop},1}$ and $\mathcal{M}_{f,l}^{\mathrm{Coop},2}$, respectively. If $\mathcal{M}_{f,l}^{\mathrm{Coop},1} \subseteq \mathcal{M}_{f,l}^{\mathrm{Coop},2}, \forall (f,l) \in \mathcal{F}(\mathcal{S}) \times \mathcal{L}$, then $f_{\mathrm{II}}^1 \geq f_{\mathrm{II}}^2$ holds.

Proof: Assume that $\mathbf{V}$ is given. If $\mathcal{M}_{f,l}^{\mathrm{Coop},i}$ is adopted for solving problem R0, let $\mathcal{D}_{\boldsymbol{\rho}}^i$ and $\mathcal{D}_{m,\boldsymbol{\rho}}^i$ be the resulting feasible sets of $\mathbf{w}_{\boldsymbol{\rho}}$ and $\mathbf{w}_{m,\boldsymbol{\rho}}$, respectively, where $\mathcal{D}_{\boldsymbol{\rho}}^i = \prod_{m=1}^M \mathcal{D}_{m,\boldsymbol{\rho}}^i, i = 1,2$. Considering constraint C4, we have $\mathbf{0} \in \mathcal{D}_{m,\boldsymbol{\rho}}^i$ if $m \in \mathcal{M}$, and $\mathcal{D}_{m,\boldsymbol{\rho}}^i = \{\mathbf{0}\}$ if $m \notin \mathcal{M}_{f,l}^{\mathrm{Coop},i}$. Besides, $\mathcal{D}_{m,\boldsymbol{\rho}}^1 = \mathcal{D}_{m,\boldsymbol{\rho}}^2$ if $m \in \mathcal{M}_{f,l}^{\mathrm{Coop},1}$ and $m \in \mathcal{M}_{f,l}^{\mathrm{Coop},2}$. Thus, $\mathcal{D}_{\boldsymbol{\rho}}^1 \subseteq \mathcal{D}_{\boldsymbol{\rho}}^2$ holds if $\mathcal{M}_{f,l}^{\mathrm{Coop},1} \subseteq \mathcal{M}_{f,l}^{\mathrm{Coop},2}$. We have $f_{\mathrm{II}}^1 \geq f_{\mathrm{II}}^2$ for any given $\mathbf{V}$, since the objective function of R0 for given $\mathbf{V}$ is only a function of $\mathbf{w}_{\boldsymbol{\rho}}$. This completes the proof. ∎

In systems with large cache capacity, the backhaul capacity constraints C3/$\widetilde{\mathrm{C3}}$ can be removed and there is no loss of optimality as the cache can effectively offload the backhaul traffic. Besides, fully cooperative transmission with cooperation set $\mathcal{M}_{f,l}^{\mathrm{F-Coop}} = \mathcal{M}$ is feasible and, based on Lemma 1, optimal considering $\mathcal{M}_{f,l}^{\mathrm{Coop}} \subseteq \mathcal{M}_{f,l}^{\mathrm{F-Coop}}, \forall \mathcal{M}_{f,l}^{\mathrm{Coop}}$.

Furthermore, the non-convex problem R0 becomes polynomial time solvable. This is in fact due to the more general result that, whenever the cooperation sets are fixed (i.e., the cooperation formation decision $\mathbf{D}_{\mathrm{II},1} = [q_{f,l,m}, b_{f,l,m}]$ is known *a priori*), the resulting problem, denoted by R0($\mathbf{D}_{\mathrm{II},2}$),

R0($\mathbf{D}_{\mathrm{II},2}$): minimize $\quad f_{\mathrm{II}}$ (15)

$\quad\quad\quad$ subject to $\quad$ C4, C5, C6, C7, $\mathbf{V} \succeq \mathbf{0}$,

$\quad\quad\quad$ variables $\quad \mathbf{D}_{\mathrm{II},2} = [\mathbf{w}_{\boldsymbol{\rho}}, \mathbf{V}]$,

is polynomial time solvable. In particular, despite the seem-ingly non-convex constraints C4, C6, and C7 included in R0($\mathbf{D}_{\mathrm{II},2}$), we will reveal the hidden convexity of R0($\mathbf{D}_{\mathrm{II},2}$) and show that R0($\mathbf{D}_{\mathrm{II},2}$) can be solved optimally and efficiently via an equivalent convex problem. For this purpose, C4, C6, and C7 are first transformed into convex forms below.

We first apply the big-M reformulation [35] of bilinear constraint C4, which gives

$$\overline{\mathrm{C4}}: \mathrm{tr}\left(\boldsymbol{\Lambda}_m \mathbf{w}_{\boldsymbol{\rho}} \mathbf{w}_{\boldsymbol{\rho}}^H\right) \leq q_{f,l,m} P_m^{\max}, m \in \mathcal{M}, \boldsymbol{\rho} \in \mathcal{S}. \quad (16)$$

We note that $\overline{\mathrm{C4}}$ is equivalent to C4, as both constraints enforce $\mathbf{w}_{m,\boldsymbol{\rho}} = \mathbf{0}$ whenever BS $m \notin \mathcal{M}_{f,l}^{\mathrm{Coop}}$ cannot participate in the cooperative transmission of subfile (f,l). For example, if $q_{f,l,m} = 0$, we have $\mathrm{tr}\left(\boldsymbol{\Lambda}_m \mathbf{w}_{\boldsymbol{\rho}} \mathbf{w}_{\boldsymbol{\rho}}^H\right) = \|\mathbf{w}_{m,\boldsymbol{\rho}}\|_2^2 = 0$, which results in $\mathbf{w}_{m,\boldsymbol{\rho}} = \mathbf{0}$; on the other hand, if $q_{f,l,m} = 1$, $\overline{\mathrm{C4}}$ is inactive due to C5. Moreover, for a given BS cooperation decision $q_{f,l,m}$, $\overline{\mathrm{C4}}$ reduces to a convex quadratic inequality constraint.

Next, let $\mathbf{W}_{\boldsymbol{\rho}} \triangleq \mathbf{w}_{\boldsymbol{\rho}} \mathbf{w}_{\boldsymbol{\rho}}^H \succeq \mathbf{0}$ and $\mathbf{H}_{\boldsymbol{\rho}} \triangleq \mathbf{h}_{\boldsymbol{\rho}} \mathbf{h}_{\boldsymbol{\rho}}^H$. The QoS constraint C6 can be transformed into affine inequality constraints,

$$\mathrm{C6} \iff \Gamma_{\boldsymbol{\rho}} \geq \kappa_{\boldsymbol{\rho}}^{\mathrm{req}} \quad (17)$$

$$\iff \overline{\mathrm{C6}}: \frac{1}{\kappa_{\boldsymbol{\rho}}^{\mathrm{req}}} \mathrm{tr}\left(\mathbf{W}_{\boldsymbol{\rho}} \mathbf{H}_{\boldsymbol{\rho}}\right) \geq \sigma^2 + \sum_{\boldsymbol{\rho}' \neq \boldsymbol{\rho}} \mathrm{tr}\left(\mathbf{W}_{\boldsymbol{\rho}'} \mathbf{H}_{\boldsymbol{\rho}}\right), \forall \boldsymbol{\rho},$$

where $\kappa_{\boldsymbol{\rho}}^{\mathrm{req}} \triangleq 2^{R^{\mathrm{req}}} - 1$. C6 and $\overline{\mathrm{C6}}$ are equivalent if and only if the following constraint holds

$$\mathrm{C10}: \mathbf{W}_{\boldsymbol{\rho}} \succeq \mathbf{0} \text{ and } \mathrm{rank}(\mathbf{W}_{\boldsymbol{\rho}}) \leq 1. \quad (18)$$

Finally, Proposition 1 is applied to transform C7.

Proposition 1. Assume that C10 holds. If $\varepsilon_{\mathrm{e}} > 0$, the secrecy constraint C7 can be equivalently transformed into a convex linear matrix inequality (LMI) as

$$\widetilde{\mathrm{C7}}: \mathbf{U}_{\mathrm{e}}^H \left(\mathbf{W}_{\boldsymbol{\rho}} - \frac{\kappa_{\boldsymbol{\rho}}^{\mathrm{tol}}}{\sigma_{\mathrm{e}}^2} \mathbf{V}\right) \mathbf{U}_{\mathrm{e}} \preceq \quad (19)$$

$$\mathrm{diag}\left((\kappa_{\boldsymbol{\rho}}^{\mathrm{tol}} - \delta_{\mathrm{e}})\mathbf{1}_{1 \times N_{\mathrm{e}}}, \frac{\delta_{\mathrm{e}}}{\varepsilon_{\mathrm{e}}^2}\mathbf{1}_{1 \times MN_{\mathrm{t}}}\right), \delta_{\mathrm{e}} \geq 0,$$

where $\mathbf{U}_{\mathrm{e}} = [\widehat{\mathbf{G}}, \mathbf{I}_{MN_{\mathrm{t}}}]$ and $\kappa_{\boldsymbol{\rho}}^{\mathrm{tol}} \triangleq \sigma_{\mathrm{e}}^2(2^{R_{\mathrm{e},\boldsymbol{\rho}}^{\mathrm{tol}}} - 1)$. Moreover, if $\varepsilon_{\mathrm{e}} = 0$, C7 is equivalent to

$$\overline{\mathrm{C7}}: \mathbf{G}^H \mathbf{W}_{\boldsymbol{\rho}} \mathbf{G} \preceq \sigma_{\mathrm{e}}^2 \kappa_{\boldsymbol{\rho}}^{\mathrm{tol}} \mathbf{Z}_{\mathrm{e},\boldsymbol{\rho}}, \quad (20)$$

where $\mathbf{Z}_{\mathrm{e},\boldsymbol{\rho}}$ is defined after (10).

Proof: Please refer to Appendix A. ∎

Now, by applying the above transformations and relaxing the rank constraint $\mathrm{rank}(\mathbf{W}_{\boldsymbol{\rho}}) \leq 1$, i.e., removing it from C10, we obtain the following convex semidefinite program (SDP),

R1: minimize $\text{tr}\big(\sum_{\boldsymbol{\rho}\in\mathcal{S}}\mathbf{W}_{\boldsymbol{\rho}}+\mathbf{V}\big)$ (21)

subject to $\overline{\text{C4}}$: $\text{tr}\left(\boldsymbol{\Lambda}_m\mathbf{W}_{\boldsymbol{\rho}}\right)\leq q_{f,l,m}P_m^{\max},\ \forall m,\forall\boldsymbol{\rho}$,

$\overline{\text{C5}}$: $\text{tr}\big(\boldsymbol{\Lambda}_m\big(\sum_{\boldsymbol{\rho}\in\mathcal{S}}\mathbf{W}_{\boldsymbol{\rho}}+\mathbf{V}\big)\big)\leq P_m^{\max},\ \forall m,$

$\overline{\text{C6}},\ \widetilde{\text{C7}}/\overline{\text{C7}},\ \overline{\text{C10}}$: $\mathbf{W}_{\boldsymbol{\rho}}\succeq\mathbf{0},\mathbf{V}\succeq\mathbf{0},\ \forall\boldsymbol{\rho}$,

variables $\mathbf{W}_{\boldsymbol{\rho}},\mathbf{V},\delta_{\text{e}}$,

where $\widetilde{\text{C7}}$ and $\overline{\text{C7}}$ are applied when $\varepsilon_{\text{e}}>0$ and $\varepsilon_{\text{e}}=0$, respectively. Generally, problem R1 achieves a lower bound on the optimal value of R0($\mathbf{D}_{\text{II},2}$). If the solution of R1 further satisfies $\text{rank}(\mathbf{W}_{\boldsymbol{\rho}})\leq 1$, $\boldsymbol{\rho}\in\mathcal{S}$, then the optimal solution of problem R0($\mathbf{D}_{\text{II},2}$) is readily available based on $\mathbf{W}_{\boldsymbol{\rho}}=\mathbf{w}_{\boldsymbol{\rho}}\mathbf{w}_{\boldsymbol{\rho}}^H$, i.e., the relaxation is tight. For the problem at hand, however, the relaxation is always tight, which is established in the following theorem.

Theorem 1 (Rank of Optimal $\mathbf{W}_{\boldsymbol{\rho}}$). Problems R0($\mathbf{D}_{\text{II},2}$) and R1 are equivalent in the sense that both problems have the same optimal value; in particular, the optimal solution $\mathbf{W}_{\boldsymbol{\rho}}^*$ of R1 satisfies $\text{rank}(\mathbf{W}_{\boldsymbol{\rho}}^*)\leq 1$, $\boldsymbol{\rho}\in\mathcal{S}$, and the optimal beamforming vector $\mathbf{w}_{\boldsymbol{\rho}}^*$ of R0($\mathbf{D}_{\text{II},2}$) is given by the principal eigenvector of $\mathbf{W}_{\boldsymbol{\rho}}^*$. This result holds for $\varepsilon_{\text{e}}\geq 0$.

Proof: Please refer to Appendix B. ∎

Based on Theorem 1, problem R0($\mathbf{D}_{\text{II},2}$) can be efficiently solved using standard convex optimization algorithms. For example, the interior-point method [36], [37], which is implemented in existing numerical convex program solvers such as CVX [38], is applicable. The resulting computational complexity of solving problem R0($\mathbf{D}_{\text{II},2}$) with respect to the number of LRs, K, the number of eavesdropping antennas, N_{e}, and the total number of transmit antennas, MN_{t}, is given by [39, Theorem 3.12]

$$\Theta^{\text{sdp}}=\mathcal{O}\left(\underbrace{\sqrt{\Phi}\log\left(\epsilon^{-1}\right)}_{\text{Number of iterations}}\underbrace{\left(\Xi\Phi^3+\Xi^2\Phi^2+\Xi^3\right)}_{\text{Complexity per iteration}}\right),\quad(22)$$

with $\Phi\triangleq MN_{\text{t}}(K+1)$ and $\Xi\triangleq MK+M+2K$. Herein, $\epsilon>0$ is the desired solution accuracy specified for the adopted numerical solver and $\mathcal{O}(\cdot)$ denotes the big-O notation. Based on (22), the algorithm for solving problem R0($\mathbf{D}_{\text{II},2}$) has a polynomial-time computational complexity, which is desirable for real time implementation [38].

B. General Case: Greedy Iterative Solution of R0

For limited cache capacity, problem R0 is NP-hard due to the non-convex constraints C2 and C3. The optimal solution has to be determined by enumerating all possible cooperation sets satisfying the backhaul capacity constraint C3. For this purpose, enumeration methods such as exhaustive search and branch-and-bound [35] are applicable. However, although the remaining cooperative beamforming problem, i.e., problem R0($\mathbf{D}_{\text{II},2}$), can be efficiently solved for each choice of cooperation sets, cf.

Theorem 1, the overall computational complexity grows exponentially with the number of BSs due to the combinatorial nature of the problem. To be specific, we define $\overline{T}_m\triangleq\min\left\{\left\lfloor B_m^{\max}/\min_{f\in\mathcal{F}(\mathcal{S})}Q_{f,m}\right\rfloor,F(\mathcal{S})\right\}$ and $\underline{T}_m\triangleq\left\lfloor B_m^{\max}/\max_{f\in\mathcal{F}(\mathcal{S})}Q_{f,m}\right\rfloor$. According to Lemma 1, the optimal cooperation formation solutions are contained in the vertices of the polyhedral simplex defined by the intersection of hyperplanes $\sum_{f\in\mathcal{F}(\mathcal{S})}q_{f,l,m}\leq T_m$ and hypercubes $q_{f,l,m}\in[0,1]$, where $\underline{T}_m\leq T_m\leq\overline{T}_m,m\in\mathcal{M}$. As a result, the enumeration of approximately $\prod_{m=1}^M\binom{T_m}{F(\mathcal{S})}$ vertices (cooperation sets) is required in the worst case for solving problem R0 by exhaustive search (branch-and-bound).

Non-polynomial time enumeration methods are only applicable for systems of small size. For practical systems with medium and large sizes, however, effective polynomial-time algorithms are desired. Herein, a low-complexity iterative algorithm based on greedy heuristics is proposed to solve problem R0 and is summarized in Algorithm 1.

Let k be the iteration index. Define $\mathcal{Q}_k\triangleq\{(f,m)\mid q_{f,l,m}=1,m\in\mathcal{M},f\in F(\mathcal{S})\}$ as the BS cooperation solution set at iteration k. Algorithm 1 starts with the initialization $\mathcal{Q}_0=\prod_{f\in\mathcal{F}(\mathcal{S})}\mathcal{M}_{f,l}^{\text{F-Coop}}=\prod_{f\in\mathcal{F}(\mathcal{S})}\mathcal{M}$, that is, $q_{f,l,m}=1,\forall m\in\mathcal{M},\forall f\in F(\mathcal{S})$. At iteration $k=1,2,\ldots$, the $q_{f,l,m}$'s are fixed according to $\mathcal{Q}_{k-1}$ and, consequently, the cooperative beamforming solutions are obtained via problem R0($\mathbf{D}_{\text{II},2}$), whose optimal value is denoted by $f_{\text{II}}^*(\mathcal{Q}_{k-1})$. If $\mathcal{Q}_{k-1}$ fulfills the backhaul capacity constraint C3, the algorithm stops and returns the solutions for cooperative BS transmission. Otherwise, the greedy algorithm sets $q_{f',l,m'}=0$ for that $(f',m')\in\mathcal{Q}_{k-1}$ which incurs the smallest penalty on the total transmit power, i.e.,

$$(f',m')\in\underset{(f,m)\in\mathcal{F}(\mathcal{S})\times\mathcal{M}_k^{\text{vio}}}{\arg\min}\left[f_{\text{II}}^*(\mathcal{Q}_{k-1}\setminus\{(f,m)\})-f_{\text{II}}^*(\mathcal{Q}_{k-1})\right],$$
(23)

and updates $\mathcal{Q}_k=\mathcal{Q}_{k-1}\setminus\{(f',m')\}$. In (23), $\mathcal{M}_{k-1}^{\text{vio}}$ denotes the index set of BSs violating constraint C3 if $\mathcal{Q}_{k-1}$ is adopted for cooperative transmission, i.e.,

$$\mathcal{M}_{k-1}^{\text{vio}}\triangleq\left\{m\in\mathcal{M}\mid\sum_{(f,m)\in\mathcal{Q}_{k-1}}Q_{f,m}>B_m^{\max}\right\}.\quad(24)$$

The iteration process is repeated until C3 is fulfilled.

Note that during each iteration of Algorithm 1, (23) is solved by enumerating $F(\mathcal{S})\times\left|\mathcal{M}_{k-1}^{\text{vio}}\right|$ choices of (f,m). The total number of choices is bounded from above by $F(\mathcal{S})\times\left|\mathcal{M}_0^{\text{vio}}\right|\times T$ in the worst case, where $\sum_{m\in\mathcal{M}}\underline{T}_m\leq T\leq\sum_{m\in\mathcal{M}}\overline{T}_m$. Consequently, the overall computational complexity of Algorithm 1 is given by

$$\Theta^{\text{greedy}}=F(\mathcal{S})\times\left|\mathcal{M}_0^{\text{vio}}\right|\times T\Theta^{\text{sdp}},\quad(25)$$

which grows only polynomially with the number of BSs M and the number of LRs K, cf. (22). In general, the proposed

Algorithm 1 Greedy Iterative Algorithm for Solving R0

1: **Initialization**: $\mathcal{Q}_0 \leftarrow \prod_{f \in \mathcal{F}(\mathcal{S})} \mathcal{M}_{f,l}^{\text{F}-\text{Coop}}$, $k \leftarrow 1$;
2: Solve problem R0($\mathbf{D}_{\text{II},2}$) for $\mathcal{Q}_0$;
3: **while** $\mathcal{M}_{k-1}^{\text{vio}} \neq \emptyset$ (cf. (24)) **do**
4:    **for each** $(f, m) \in \mathcal{F}(\mathcal{S}) \times \mathcal{M}_k^{\text{vio}}$ **do**
5:       Solve problem R0($\mathbf{D}_{\text{II},2}$) for $\mathcal{Q}_{k-1} \setminus \{(f, m)\}$;
6:    **end for**
7:    $\mathcal{Q}_k \leftarrow \mathcal{Q}_{k-1} \setminus \{(f', m')\}$, where (f', m') solves (23);
8:    $k \leftarrow k + 1$,
9: **end while**

greedy algorithm is suboptimal. However, for large cache capacity, Algorithm 1 terminates without having to solve (23) and the obtained solution is globally optimal.

Remark 2. We note that alternative polynomial-time methods that could be used to solve problem R0 have certain pitfalls. For example, by reformulating the binary variables as ℓ_0-norms, problem R0 can be solved by the approximation methods proposed in [40], [41]. However, the solutions obtained by these approximation methods are generally (primal) infeasible for most of the cases under investigation. This is because, on the one hand, the big-M constraint $\overline{\text{C4}}$ would force $q_{f,l,m}$ to small non-zero fractions instead of binary solutions that fulfill C2. On the other hand, generating primal feasible solutions from approximate results is a non-trivial task for NP-hard problems. For example, a brute-force deterministic or random rounding of these continuous solutions usually leads to the violation of the backhaul capacity constraint C3. For similar reasons, primal feasibility is also not ensured when other polynomial-time methods such as convex relaxation (e.g., linear programming or SDP based relaxation of the binary constraints) and difference of convex programming (also referred to as successive convex approximation) [42] are adopted. Yet, with Algorithm 1, the likelihood of obtaining an infeasible primal solution is low due to the greedy iterative search, cf. (24). As will be shown in Section V, the solution obtained with Algorithm 1 is close to the optimal value in the medium and large cache capacity regimes.

C. Solution of Problem Q0

Problem Q0 has a considerably (Ω-times) larger problem size than R0. Hence, solving problem Q0 via enumeration methods seems impossible due to the overwhelming computational complexity. Besides, the greedy suboptimal method in Algorithm 1 cannot be directly applied for solving problem Q0 since constraint C1 becomes bilinear over the joint optimization space $\{\mathbf{C}_{\text{I}}\}$ of Q0. We address both issues by applying the following binary relaxation method.

In particular, the bilinear constraint C1 is transformed to

$$\widetilde{\text{C1}} : c_{f,m} + b_{f,l,m,\omega} \geq q_{f,l,m,\omega}. \qquad (26)$$

If the average backhaul capacity is insufficient, $\widetilde{\text{C1}}$ and C8 together lead to $b_{f,l,m,\omega} = (1 - c_{f,m}) \times q_{f,l,m,\omega}$ since

Q0 enjoys a similar monotonicity as R0, cf. Lemma 1; otherwise, $\widetilde{\text{C1}}$ is inactive. Thus, $\widetilde{\text{C1}}$ and C1 are equivalent.

Moreover, let $\widetilde{\text{C2}}$ be a relaxation of C2 where the binary constraints are replaced by $q_{f,l,m,\omega} \in [0, 1]$. By adopting $\widetilde{\text{C1}}$ and $\widetilde{\text{C2}}$, we arrive at a relaxed version of Q0:

Q1: minimize $\dfrac{1}{\Omega} \sum_{\omega=1}^{\Omega} f_{\text{I},\omega}$ (27)

    subject to $\overline{\text{C3}}$, C8, $\overline{\text{C9}}$: $\mathbf{D}_{\text{I},\omega} \in \widehat{\mathcal{D}}_{\text{I},\omega}$, $\omega \in \{1, \ldots, \Omega\}$,

    variables $\mathbf{C}_{\text{I}} = [c_{f,m}, q_{f,l,m,\omega}, b_{f,l,m,\omega}, \mathbf{w}_{\boldsymbol{\rho},\omega}, \mathbf{V}_{\omega}]$,

where $\widehat{\mathcal{D}}_{\text{I},\omega} \triangleq \left\{ \mathbf{D}_{\text{I},\omega} \mid \widetilde{\text{C1}}, \widetilde{\text{C2}}, \text{C4–C7} \right\}$. Although problem Q1 remains non-convex due to constraints C4, C6, and C7, an equivalent convex SDP can be obtained for Q1 in a similar manner as for problem R0, cf. Section IV-A. Thus, the relaxed problem Q1 can be solved efficiently (in polynomial time) using the interior point method [36].

In general, the proposed relaxation solution of Q1 provides a performance lower bound for Q0. However, the following theorem establishes that the relaxation solution is asymptotically optimal in the limiting case of large Ω.

Theorem 2 (Asymptotic Optimality of the Relaxation Solution). Problems Q1 and Q0 become equivalent as $\Omega \to \infty$ in the sense that their optimum values and the optimal caching decisions become identical if problem Q0 is feasible.

   Proof: Please refer to Appendix C.   ■

V. SIMULATION RESULTS

In this section, we evaluate the system performance for the proposed caching and secure delivery schemes. Consider a cluster of $M = 7$ hexagonal cells, where a BS is deployed at the center of each cell with an inter-BS distance of 500 m. Each BS is equipped with $N_{\text{t}} = 4$ antennas and the ER has $N_{\text{e}} = 2$ antennas. We assume that a library of $F = 10$ video files, each of duration 45 minutes and size 500 MB (Bytes), is delivered to $K = 5$ single-antenna LRs. Consequently, an estimated secrecy data rate of $R_{\boldsymbol{\rho}}^{\text{sec}} \approx Q_f = 500 \times 8.0 \times 10^6 / (45 \times 60) \approx 1.5$ Mbps is required at each LR for uninterrupted video streaming. The LRs and the ER are uniformly and randomly distributed in the system while the minimum distance between receiver and BS is 50 m. Each LR requests one file independent of the other LRs. Let θ_f be the probability of file $f \in \mathcal{F}$ being requested and let $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_F]$ be the probability distribution of the requests for the different files. We set $\theta_f = \frac{1}{f^{\kappa}} / \sum_{f \in \mathcal{F}} \frac{1}{f^{\kappa}}$ with $\kappa = 1.1$ according to the Zipf distribution [10]. Moreover, the 3GPP path loss model ("Urban Macro NLOS" scenario) in [43] is adopted. For channel estimation, we define the normalized estimation error of the ER as $\alpha^2 \triangleq \frac{\varepsilon_{\text{e}}^2}{\|\mathbf{G}\|_F^2}$. Unless otherwise specified, we assume $\alpha^2 = 0.05$. The capacities of the backhaul links are independently and identically distributed (i.i.d.) as $\Pr(B_m^{\max} = 0 \text{ Mbps}) = 0.3$, $\Pr(B_m^{\max} = 3 \text{ Mbps}) = 0.4$,

Parameters	Settings
System bandwidth	10 MHz
Duration of time slot	$\tau = 10$ ms
File splitting	$L = 45$ mins$/\tau = 2.7 \times 10^4$
BS transmit power	$P_m^{\max} = 48$ dBm
Noise power density	-172.6 dBm/Hz
Delivery QoS requirement	$R_{\boldsymbol{\rho}}^{\text{req}} = 1.1 \times R_{\boldsymbol{\rho}}^{\text{sec}} = 1.65$ Mbps
Delivery secrecy threshold	$R_{\text{e},\boldsymbol{\rho}}^{\text{tol}} = 0.1 \times R_{\boldsymbol{\rho}}^{\text{sec}} = 0.15$ Mbps

and $\Pr(B_m^{\max} = 6 \text{ Mbps}) = 0.3$, $\forall m$, which can be interpreted as the probabilities of high, medium, and low non-VoD traffic scenarios in the cellular network, respectively. The other relevant system parameters are given in Table II. Before video delivery starts, $\Omega = 50$ scenarios are randomly generated based on the adopted models for the user preference, CSI, and backhaul capacity to determine the initial cache status, cf. problem Q0.

A. Performance of the Proposed Caching Scheme

First, we study the performance of the proposed caching scheme. For comparison, we consider three heuristic caching schemes as baselines:

- Baseline 1 (Preference-based caching): The most popular files are cached. Assuming $\boldsymbol{\theta}$ is known, the cache control decision is made based on

$$\text{maximize} \quad \sum_{f \in \mathcal{F}, m \in \mathcal{M}} \theta_f c_{f,m} V_f \qquad (28)$$
$$\text{subject to} \quad c_{f,m} \in [0,1], f \in \mathcal{F}, m \in \mathcal{M}, \quad \text{C8},$$
$$\text{variables} \quad c_{f,m}.$$

- Baseline 2 (Uniform caching): The same amount of data is cached for each file, i.e.,

$$c_{f,m} V_f = \frac{1}{F} \min\{C_m^{\max}, \sum_{f'=1}^{F} V_{f'}\}, \ \forall f, \forall m, \quad (29)$$

and the users' preferences are not taken into account.
- Baseline 3 (Power-efficient caching): This scheme is identical to the proposed caching scheme, except that the secrecy constraint C7 in Q0 is excluded from $\mathcal{D}_{\text{I},\omega}$ during cache placement.

For Baselines 1, 2, and 3, the proposed delivery scheme, i.e., Algorithm 1, is adopted.

Figs. 2(a) and 2(b) show the total transmit power and the secrecy outage probability, defined as $p_{\text{out}} \triangleq \Pr(R_{\boldsymbol{\rho}}^{\text{sec}} < [R_{\boldsymbol{\rho}}^{\text{req}} - R_{\text{e},\boldsymbol{\rho}}^{\text{tol}}]^+)$, of the considered caching schemes as functions of the cache capacity, respectively. Herein, p_{out} corresponds to the probability that problem R0 is infeasible because either the scheme fails to satisfy the QoS constraint C6 or the secrecy constraint C7. As can be observed from Figs. 2(a) and 2(b), a larger cache capacity leads to both a lower total BS transmit power and a smaller secrecy outage

probability[10] as larger (virtual) transmit antenna arrays can be formed during video delivery. For example, for the proposed scheme, the average number of cooperating BSs is 4.0 for $C_m^{\max} = 4000$ MB (80% of library size) compared to 2.8 for $C_m^{\max} = 1000$ MB (20% of library size), which leads to a transmit power reduction of up to 4.8 dB. The performance gap between the considered caching schemes is negligible for small (large) cache capacities because of insufficient (saturated) BS cooperation. For medium cache capacities, however, the proposed caching scheme achieves considerable transmit power savings due to its ability to exploit the historical information regarding user requests, backhaul capacity, and CSI for resource allocation. For a similar reason, both the transmit power and the secrecy outage probability of Baseline 3 are lower than those of Baselines 1 and 2. However, the performance of Baseline 3 is generally worse than that of the proposed scheme. This indicates that including the secrecy constraint also for cache placement is necessary for maximizing the performance during video file delivery. Note also that when the increase in cache capacity is not sufficient to support additional BSs for cooperative transmission, the performance does not improve. As a result, the total transmit power in Fig. 2(a) decreases in a piece-wise constant manner as the cache capacity increases.

B. Performance of the Proposed Delivery Scheme

Next, we study the performance of the proposed delivery scheme. As a performance benchmark, the optimal solution of problem R0 is evaluated by an exhaustive search. Besides, the following naive delivery schemes are considered as baselines:

- Baseline 4 (Coordinated beamforming): The user is associated with the nearest BS which has sufficient backhaul capacity available. Each video (sub)file is only delivered from the associated BS, i.e., $\sum_{m \in \mathcal{M}} q_{f,l,m} = 1, \forall (f,l) \in \mathcal{F} \times \mathcal{L}$.
- Baseline 5 (Full BS cooperation): The backhaul capacity constraints are dropped and all BSs cooperate to serve all users, i.e., $q_{f,l,m} = 1, \forall f, l, m$. For Baselines 4 and 5, the optimal beamforming solutions are obtained based on R0($\mathbf{D}_{\text{II},2}$) where $q_{f,l,m}$s are fixed accordingly.
- Baseline 6 (Non-robust transmission): Different from the proposed delivery scheme, the CP treats the possibly erroneous channel estimate $\widehat{\mathbf{G}}$ as accurate. Consequently, the delivery decisions are made according to R0 by setting $\varepsilon_{\text{e}}^2 = 0$, irrespective of the channel estimation errors.

Figs. 3(a) and 3(b) illustrate the performance of the considered delivery schemes as functions of the cache

---

[10]In case of outage, in a practical system, a retransmission request would be triggered.

Fig. 2. (a) Total BS transmit power and (b) secrecy outage probability versus cache capacity (in percentage of the total library size) for different caching schemes.
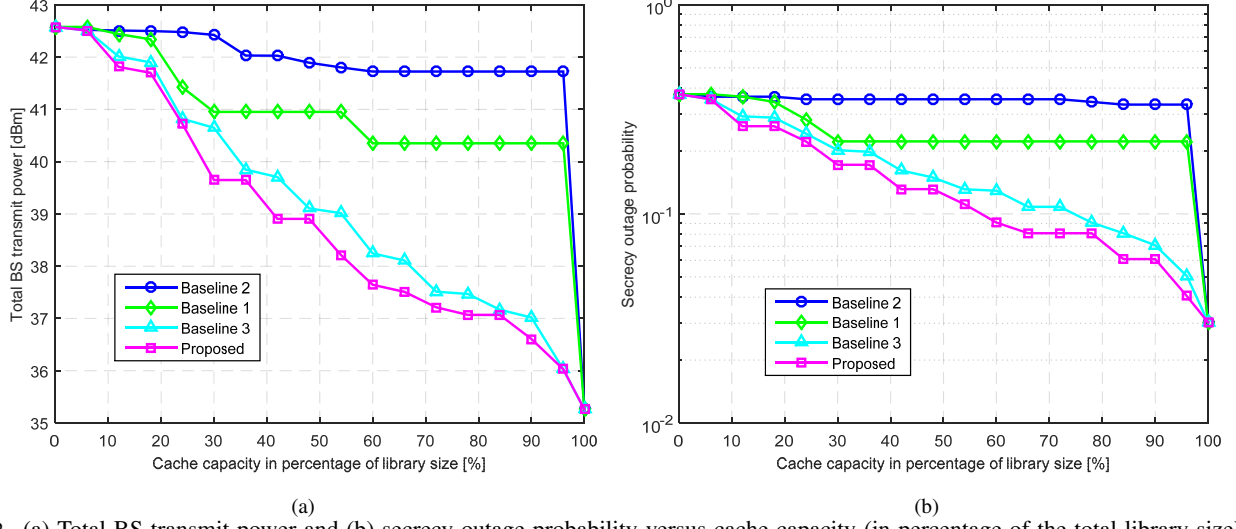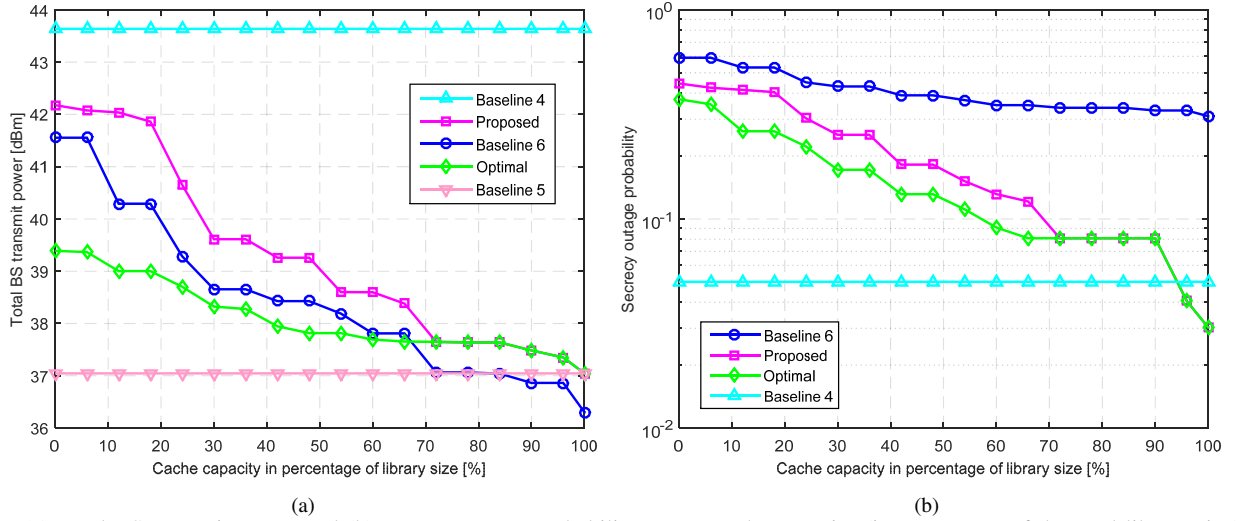

Fig. 3. (a) Total BS transmit power and (b) secrecy outage probability versus cache capacity (in percentage of the total library size) for different delivery schemes.

capacity[11]. The initial cache status is determined by solving problem Q0. From Fig. 3(a), we observe that, as expected, Baselines 4 and 5 constitute performance lower and upper bounds for the proposed delivery scheme, respectively. Comparing the proposed delivery scheme and the optimal delivery scheme, the performance gap between them reduces as the cache capacity increases. This is because, for large cache capacities, less backhaul traffic is generated and correspondingly the probability that C3 is active is reduced. It is interesting to observe that for a cache capacity of 3600 MB (70% of library size), the proposed scheme already

---

[11]The secrecy outage probability for Baseline 5 is not shown as Baseline 5 is infeasible for most cache capacity values.

achieves the same performance as the optimal delivery scheme. Meanwhile, Fig. 3(a) shows that the total transmit power of Baseline 6 is lower than that of the proposed scheme. In the large cache capacity regime, Baseline 6 even consumes less transmit power than full BS cooperation. This is because, under imperfect CSI, cooperative beamforming alone is not sufficient to prevent data leakage to the ER. Hence, the proposed scheme and Baseline 5 have to transmit a non-negligible amount of AN to degrade the reception of the ER for enhancing communication secrecy. On the other hand, as the AN can also leak into the LR channels, cf. (8), a higher transmit power is also needed for cooperative beamforming to fulfill the QoS constraint C6. Nevertheless, we observe from Fig. 3(b) that, by consuming extra transmit
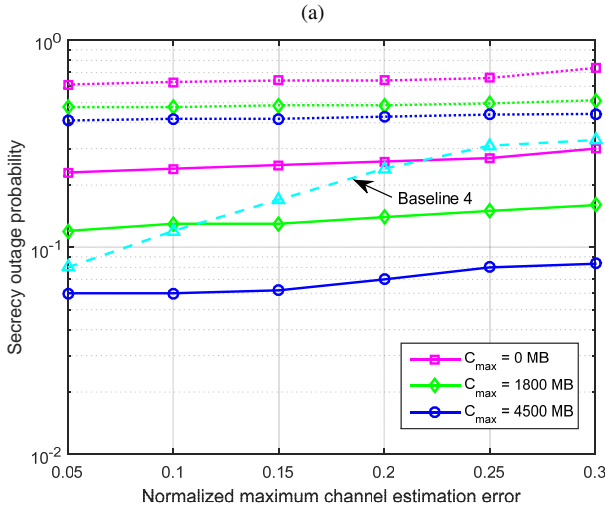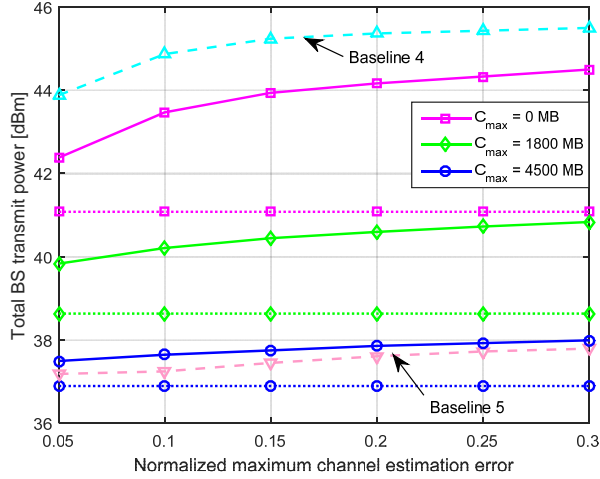
Fig. 4. (a) Total BS transmit power and (b) secrecy outage probability of different delivery schemes versus normalized channel estimation error: the performances of the proposed scheme and Baseline 6 are shown as solid and dotted lines, respectively; the performances of Baselines 4 and 5 are as dashed lines, respectively.

power under imperfect CSI, the secrecy outage probability of the proposed and the optimal schemes can be kept low and decreases significantly with the cache capacity. In contrast, for Baseline 6, the BSs transmit with insufficient power since the imperfect CSI is treated as perfect CSI. As a consequence, the secrecy outage probability for Baseline 6 is the highest among all the considered schemes and decreases much slower for increasing cache capacity than that of the proposed scheme. This implies that the imperfection of CSI has to be carefully taken into account for guaranteeing the secrecy of video delivery.

C. Impact of Channel Estimation Errors

In Figs. 4(a) and 4(b), we show the performance of the considered delivery scheme versus the normalized channel estimation error α^2 for different cache capacities. As can

be observed, for the proposed scheme, Baseline 4, and Baseline 5, both the total transmit power and the secrecy outage probability increase with the value of α^2. This is because, as the uncertainty region $\mathcal{U}_e$ is enlarged, the secrecy constraint C7 becomes more stringent, and to satisfy C7, a higher AN power is needed. Meanwhile, to satisfy the QoS constraint C6, a higher beamforming power is also needed to combat the leaked AN at the LRs. Nevertheless, by consuming extra transmit power, the proposed delivery scheme is able to maintain a reasonably low secrecy outage probability despite the deteriorating CSI quality. Moreover, for large cache capacities, the increase in transmit power is significantly reduced because a large cooperative transmit antenna array can be formed, and the performance gap between the proposed scheme and Baseline 5 becomes negligible. On the other hand, as Baseline 4 has only limited spatial degrees of freedom, its secrecy outage probability increases significantly with α^2 even though a large transmit power is consumed. As for Baseline 6, although the transmit power remains constant as α^2 increases, a high secrecy outage probability results, cf. Fig. 4(b).

D. Impact of the Number of Transmit Antennas

In Figs. 5(a) and 5(b), the performance of the proposed delivery scheme, Baseline 4, and Baseline 5 versus the cache capacity is evaluated for different numbers of transmit antennas, respectively. For a given cache capacity, more transmit antennas N_t equipped at the BSs, which increase the s.d.o.f., lead to transmit power savings for the considered schemes in Fig. 5(a). For similar reasons, the secrecy outage probability of the considered schemes is also decreased in Fig. 5(b) by using larger N_t. However, different from Baselines 4 and 5, a further reduction in both the secrecy outage probability and the total BS transmit power can be achieved for the proposed scheme by increasing the cache capacity at the BSs. Particularly, when the cache capacity is sufficiently large, the proposed scheme achieves the same power efficiency as Baseline 5, while simultaneously yielding a small secrecy outage probability.

VI. CONCLUSION

In this paper, caching was exploited to improve PLS for cellular video streaming. Caching and cooperative transmission were optimized based on a mixed-integer two-stage robust optimization problem for minimization of the total transmit power needed to secure video streaming under imperfect CSI knowledge. Caching was shown to reduce the backhaul capacity required for cooperative transmission among a large group of BSs and to increase the available s.d.o.f. As the problem was shown to be NP-hard, suboptimal polynomial-time algorithms were developed to solve the problem efficiently. The optimality of the proposed
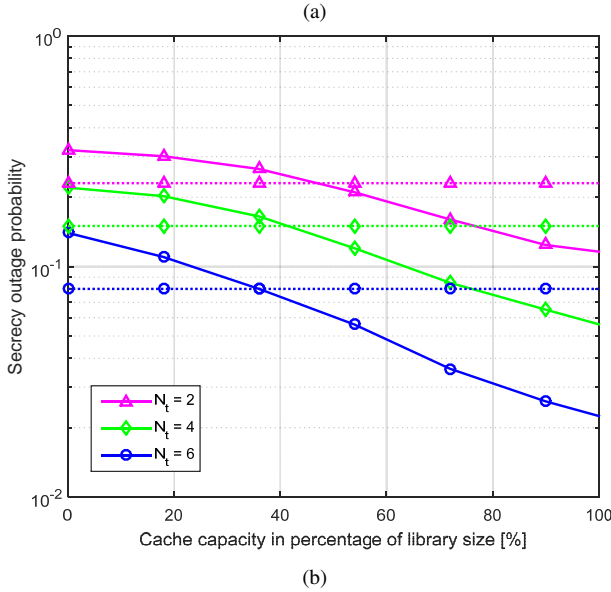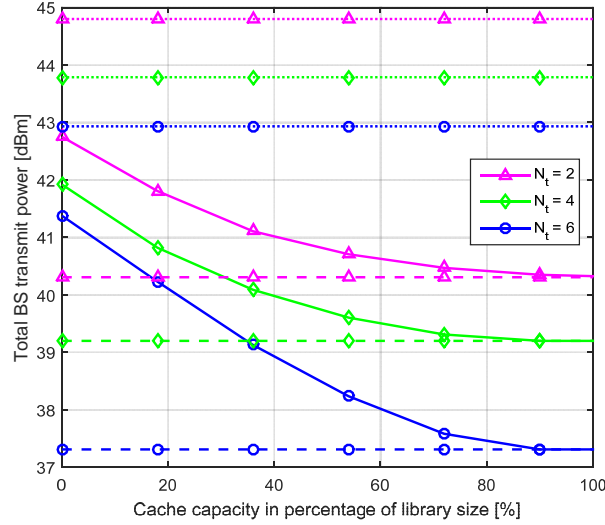
Fig. 5. (a) Total BS transmit power and (b) secrecy outage probability of the proposed delivery scheme (solid line), Baseline 4 (dotted line) and Baseline 5 (dashed line) versus cache capacity (in percentage of the total library size) for different numbers of transmit antennas at BSs.

algorithms was verified in the regimes of a large cache capacity and a large number of training scenarios, respectively. Simulation results showed that the proposed caching and delivery schemes can significantly enhance both the PLS and power efficiency of cellular video streaming. Moreover, by deploying caches with large capacities, both the performance advantages and the robustness of full BS cooperation can be reaped even for limited backhaul capacities and imperfect CSI knowledge.

APPENDIX A

PROOF OF PROPOSITION 1

Constraint C7 is first transformed to an LMI as follows,

$$
\text{C7} \overset{(a)}{\iff} 1 + \frac{1}{\sigma_{\text{e}}^2} \mathbf{w}_{\boldsymbol{\rho}}^H \mathbf{G} \mathbf{Z}_{\text{e},\boldsymbol{\rho}}^{-1} \mathbf{G}^H \mathbf{w}_{\boldsymbol{\rho}} \leq 2^{R_{\text{e},\boldsymbol{\rho}}^{\text{tol}}}, \; \forall \Delta \mathbf{G} \in \mathcal{U}_{\text{e}},
$$
$$
\overset{(b)}{\iff} \text{tr}\left(\mathbf{Z}_{\text{e},\boldsymbol{\rho}}^{-\frac{1}{2}} \mathbf{G}^H \mathbf{w}_{\boldsymbol{\rho}} \mathbf{w}_{\boldsymbol{\rho}}^H \mathbf{G} \mathbf{Z}_{\text{e},\boldsymbol{\rho}}^{-\frac{1}{2}} \right) \leq \kappa_{\boldsymbol{\rho}}^{\text{tol}}
$$
$$
\overset{(c)}{\implies} \lambda_{\max}\left(\mathbf{Z}_{\text{e},\boldsymbol{\rho}}^{-\frac{1}{2}} \mathbf{G}^H \mathbf{W}_{\boldsymbol{\rho}} \mathbf{G} \mathbf{Z}_{\text{e},\boldsymbol{\rho}}^{-\frac{1}{2}} \right) \leq \kappa_{\boldsymbol{\rho}}^{\text{tol}} \quad (30)
$$
$$
\iff \mathbf{Z}_{\text{e},\boldsymbol{\rho}}^{-\frac{1}{2}} \mathbf{G}^H \mathbf{W}_{\boldsymbol{\rho}} \mathbf{G} \mathbf{Z}_{\text{e},\boldsymbol{\rho}}^{-\frac{1}{2}} \preceq \kappa_{\boldsymbol{\rho}}^{\text{tol}} \mathbf{I}
$$
$$
\iff \overline{\text{C7}}: \mathbf{G}^H \mathbf{W}_{\boldsymbol{\rho}} \mathbf{G} \preceq \kappa_{\boldsymbol{\rho}}^{\text{tol}} \mathbf{Z}_{\text{e},\boldsymbol{\rho}}, \; \forall \Delta \mathbf{G} \in \mathcal{U}_{\text{e}},
$$

where (a) follows from Sylvester's determinant identity $\det(\mathbf{I} + \mathbf{AB}) = \det(\mathbf{I} + \mathbf{BA})$, (b) follows from the trace identity $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$, and (c) is due to the inequality $\lambda_{\max}(\mathbf{A}) \leq \text{tr}(\mathbf{A})$ for $\mathbf{A} \succeq \mathbf{0}$ [44] and the variable substitution $\mathbf{W}_{\boldsymbol{\rho}} = \mathbf{w}_{\boldsymbol{\rho}} \mathbf{w}_{\boldsymbol{\rho}}^H$. For (c), $\lambda_{\max}(\mathbf{A}) = \text{tr}(\mathbf{A})$ holds if $\text{rank}(\mathbf{A}) \leq 1$. Considering the variable transformation $\mathbf{W}_{\boldsymbol{\rho}} = \mathbf{w}_{\boldsymbol{\rho}} \mathbf{w}_{\boldsymbol{\rho}}^H$, we conclude that C7 and $\overline{\text{C7}}$ are equivalent if $\text{rank}(\mathbf{W}_{\boldsymbol{\rho}}) \leq 1$.

For perfect CSI with $\varepsilon_{\text{e}} = 0$, $\overline{\text{C7}}$ is a convex constraint that can be easily applied in SDP. On the other hand, when $\varepsilon_{\text{e}} > 0$, $\overline{\text{C7}}$ becomes semi-infinite, i.e., it represents infinitely many LMIs of $\mathbf{W}_{\boldsymbol{\rho}}$ and $\mathbf{V}$, due to the continuous uncertainty set $\mathcal{U}_{\text{e}}$. Hence, although the constraint is convex, the optimization problem is still computationally infeasible. To resolve this problem, transforming the semi-infinite constraint into a finite number of convex constraints is necessary and will be accomplished by the following lemma.

Lemma 2 (Robust Quadratic Matrix Inequality [45])**.** Let $f(\mathbf{X}) = \mathbf{X}^H \mathbf{A} \mathbf{X} + \mathbf{X}^H \mathbf{B} + \mathbf{B}^H \mathbf{X} + \mathbf{C}$, and $\mathbf{D} \succeq \mathbf{0}$. Then, the following two statements are equivalent:

(i) $f(\mathbf{X}) \succeq \mathbf{0}$ holds for any $\mathbf{X} \in \left\{ \mathbf{X} \mid \text{tr}\left(\mathbf{X}^H \mathbf{D} \mathbf{X} \right) \leq 1 \right\}$;

(ii) There exist some $\delta \geq 0$ satisfying the following LMI,

$$
\begin{bmatrix} \mathbf{C} - \delta \mathbf{I} & \mathbf{B}^H \\ \mathbf{B} & \mathbf{A} + \delta \mathbf{D} \end{bmatrix} \succeq \mathbf{0}. \quad (31)
$$

Substituting $\mathbf{G} = \widehat{\mathbf{G}} + \Delta \mathbf{G}$, constraint $\overline{\text{C7}}$ is equivalently reformulated as

$$
\Delta \mathbf{G}^H \mathbf{T}_{\text{e},\boldsymbol{\rho}} \Delta \mathbf{G} + \widehat{\mathbf{G}}^H \mathbf{T}_{\text{e},\boldsymbol{\rho}} \Delta \mathbf{G} + \Delta \mathbf{G}^H \mathbf{T}_{\text{e},\boldsymbol{\rho}} \widehat{\mathbf{G}} \quad (32)
$$
$$
+ \widehat{\mathbf{G}}^H \mathbf{T}_{\text{e},\boldsymbol{\rho}} \widehat{\mathbf{G}} - \kappa_{\boldsymbol{\rho}}^{\text{tol}} \mathbf{I}_{N_{\text{e}}} \preceq \mathbf{0}, \quad \forall \Delta \mathbf{G} \in \mathcal{U}_{\text{e}},
$$

where $\mathbf{T}_{\text{e},\boldsymbol{\rho}} \triangleq \mathbf{W}_{\boldsymbol{\rho}} - \frac{\kappa_{\boldsymbol{\rho}}^{\text{tol}}}{\sigma_{\text{e}}^2} \mathbf{V}$. Based on Lemma 2, constraint (32) is equivalent to

$$
\begin{bmatrix} \widehat{\mathbf{G}}^H \mathbf{T}_{\text{e},\boldsymbol{\rho}} \widehat{\mathbf{G}} + (\delta_{\text{e}} - \kappa_{\boldsymbol{\rho}}^{\text{tol}}) \mathbf{I}_{N_{\text{e}}} & \widehat{\mathbf{G}}^H \mathbf{T}_{\text{e},\boldsymbol{\rho}} \\ \mathbf{T}_{\text{e},\boldsymbol{\rho}} \widehat{\mathbf{G}} & \mathbf{T}_{\text{e},\boldsymbol{\rho}} - \frac{\delta_{\text{e}}}{\varepsilon_{\text{e}}^2} \mathbf{I}_{MN_{\text{t}}} \end{bmatrix} \preceq \mathbf{0},
$$
$$
\exists \delta_{\text{e}} \geq 0,
$$
$$
\iff \widetilde{\text{C7}}: \mathbf{U}_{\text{e}}^H \mathbf{T}_{\text{e},\boldsymbol{\rho}} \mathbf{U}_{\text{e}} \preceq \begin{bmatrix} (\kappa_{\boldsymbol{\rho}}^{\text{tol}} - \delta_{\text{e}}) \mathbf{I}_{N_{\text{e}}} & \mathbf{0} \\ \mathbf{0} & \frac{\delta_{\text{e}}}{\varepsilon_{\text{e}}^2} \mathbf{I}_{MN_{\text{t}}} \end{bmatrix},
$$
$$
\delta_{\text{e}} \geq 0. \quad (33)
$$

Besides, constraints C7 and $\widetilde{\text{C7}}$ are equivalent if $\text{rank}(\mathbf{W}_{\boldsymbol{\rho}}) \leq 1$. This completes the proof.

APPENDIX B
PROOF OF THEOREM 1

Note that problems $\text{R0}(\mathbf{D}_{\text{II},2})$ and R1 are equivalent if and only if the rank constraint $\text{rank}(\mathbf{W}_{\boldsymbol{\rho}}^{*}) \leq 1$ is fulfilled. Below we only prove the result for problem R1 with $\varepsilon_{\text{e}} > 0$ and $\widetilde{\text{C7}}$. The result for $\varepsilon_{\text{e}} = 0$ and $\overline{\text{C7}}$ can be proved similarly; see [1].

Let $\boldsymbol{\alpha} = [\alpha_{m\boldsymbol{\rho}}]$, $\boldsymbol{\beta} = [\beta_m]$, $\boldsymbol{\lambda} = [\lambda_{\boldsymbol{\rho}}]$, $\boldsymbol{\Phi}_{\boldsymbol{\rho}}$, and $\boldsymbol{\Theta}_{\boldsymbol{\rho}} = [\boldsymbol{\Theta}_{1\boldsymbol{\rho}}, \boldsymbol{\Theta}_{2\boldsymbol{\rho}}]$ be the Lagrangian multipliers associated with constraints $\overline{\text{C4}}$, $\overline{\text{C5}}$, $\overline{\text{C6}}$, $\widetilde{\text{C7}}$, and $\overline{\text{C10}}$, respectively, where

$$\alpha_{m\boldsymbol{\rho}} \geq 0, \beta_m \geq 0, \lambda_{\boldsymbol{\rho}} \geq 0, \boldsymbol{\Phi}_{\boldsymbol{\rho}} \succeq \mathbf{0}, \boldsymbol{\Theta}_{1\boldsymbol{\rho}} \succeq \mathbf{0}, \text{ and } \boldsymbol{\Theta}_{2\boldsymbol{\rho}} \succeq \mathbf{0}.$$

The Lagrangian of problem R1 is

$$\mathcal{L}(\mathbf{W}_{\boldsymbol{\rho}}, \mathbf{V}; \boldsymbol{\Upsilon}) = \sum_{\boldsymbol{\rho}} \text{tr}\big[\big(\mathbf{B}_{1\boldsymbol{\rho}} - 2\lambda_{\boldsymbol{\rho}}\mathbf{H}_{\boldsymbol{\rho}} - \boldsymbol{\Theta}_{1\boldsymbol{\rho}}\big)\mathbf{W}_{\boldsymbol{\rho}} \quad (34)$$
$$+ \big(\mathbf{B}_{2\boldsymbol{\rho}} - \kappa_{\boldsymbol{\rho}}^{\text{tol}}\widehat{\mathbf{G}}\boldsymbol{\Phi}_{\boldsymbol{\rho}}\widehat{\mathbf{G}}^H - \boldsymbol{\Theta}_{2\boldsymbol{\rho}}\big)\mathbf{V}\big] + \Delta_0,$$

where $\boldsymbol{\Upsilon} \triangleq [\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\Phi}_{\boldsymbol{\rho}}, \boldsymbol{\Theta}_{\boldsymbol{\rho}}]$, Δ_0 is a collection of terms irrelevant for the proof, and

$$\mathbf{B}_{1\boldsymbol{\rho}} \triangleq \mathbf{I} + \boldsymbol{\Lambda}_{\boldsymbol{\rho}}^{\boldsymbol{\alpha},\boldsymbol{\beta}} + \widehat{\mathbf{G}}\boldsymbol{\Phi}_{\boldsymbol{\rho}}\widehat{\mathbf{G}}^H + \sum_{\boldsymbol{\rho}\in\mathcal{S}}(1 + \kappa_{\boldsymbol{\rho}}^{\text{req}})\lambda_{\boldsymbol{\rho}}\mathbf{H}_{\boldsymbol{\rho}} \succ \mathbf{0},$$

$$\mathbf{B}_{2\boldsymbol{\rho}} \triangleq \mathbf{I} + \boldsymbol{\Lambda}^{\boldsymbol{\beta}} + \lambda_{\boldsymbol{\rho}}\mathbf{H}_{\boldsymbol{\rho}} \succ \mathbf{0}, \quad (35)$$

with $\boldsymbol{\Lambda}_{\boldsymbol{\rho}}^{\boldsymbol{\alpha},\boldsymbol{\beta}} \triangleq \sum_{m\in\mathcal{M}}(\alpha_{m\boldsymbol{\rho}} + \beta_m)\boldsymbol{\Lambda}_m$ and $\boldsymbol{\Lambda}^{\boldsymbol{\beta}} \triangleq \sum_{m\in\mathcal{M}}\beta_m\boldsymbol{\Lambda}_m$. It can be verified that R1 is a convex optimization problem and fulfills Slater's constraint qualification. Thus, strong duality holds for problem R1 and the Karush–Kuhn–Tucker (KKT) conditions are both necessary and sufficient for a primal-dual point $(\mathbf{W}_{\boldsymbol{\rho}}, \mathbf{V}; \boldsymbol{\Upsilon})$ to be optimal. The KKT conditions for problem R1 are given by

$$\nabla_{\mathbf{W}_{\boldsymbol{\rho}}}\mathcal{L} = \mathbf{B}_{1\boldsymbol{\rho}} - 2\lambda_{\boldsymbol{\rho}}\mathbf{H}_{\boldsymbol{\rho}} - \boldsymbol{\Theta}_{1\boldsymbol{\rho}} = \mathbf{0}, \quad (36)$$

$$\mathbf{W}_{\boldsymbol{\rho}} \succeq \mathbf{0}, \quad \lambda_{\boldsymbol{\rho}} \geq 0, \quad \mathbf{W}_{\boldsymbol{\rho}}\boldsymbol{\Theta}_{1\boldsymbol{\rho}} = \mathbf{0}. \quad (37)$$

Based on (36) and (37), we have $\mathbf{W}_{\boldsymbol{\rho}}\mathbf{B}_{1\boldsymbol{\rho}} = 2\lambda_{\boldsymbol{\rho}}\mathbf{W}_{\boldsymbol{\rho}}\mathbf{H}_{\boldsymbol{\rho}}$. Besides, constraint $\overline{\text{C6}}$ is satisfied with equality for the optimal solution and thus $\lambda_{\boldsymbol{\rho}} > 0$. Moreover, since $\text{rank}(\mathbf{H}_{\boldsymbol{\rho}}) \leq 1$, the rank of the optimal $\mathbf{W}_{\boldsymbol{\rho}}$ can be determined as

$$\text{rank}(\mathbf{W}_{\boldsymbol{\rho}}) \overset{(a)}{=} \text{rank}(\mathbf{W}_{\boldsymbol{\rho}}\mathbf{B}_{1\boldsymbol{\rho}}) \overset{(b)}{=} \text{rank}(\lambda_{\boldsymbol{\rho}}\mathbf{W}_{\boldsymbol{\rho}}\mathbf{H}_{\boldsymbol{\rho}}) \quad (38)$$
$$\overset{(c)}{\leq} \min\{\text{rank}(\lambda_{\boldsymbol{\rho}}\mathbf{W}_{\boldsymbol{\rho}}), \text{rank}(\mathbf{H}_{\boldsymbol{\rho}})\} \leq 1,$$

where (a) is due to $\mathbf{B}_{1\boldsymbol{\rho}} \succ \mathbf{0}$, (b) is a result of (36) and (37), and (c) follows from the rank inequality $\text{rank}(\mathbf{AB}) \leq \min\{\text{rank}(\mathbf{A}), \text{rank}(\mathbf{B})\}$ [44]. Thus, $\text{rank}(\mathbf{W}_{\boldsymbol{\rho}}^{*}) \leq 1$ has to hold if problem R1 is feasible. This completes the proof.

APPENDIX C
PROOF OF THEOREM 2

We first assume that the caching decisions are given in problems Q0 and Q1, where the resulting problems are denoted by $\text{Q0}(\mathbf{D}_{\text{I},\omega})$ and $\text{Q1}(\mathbf{D}_{\text{I},\omega})$, respectively. Without loss of generality, $\text{Q0}(\mathbf{D}_{\text{I},\omega})$ can be written in general form as

$$\text{Q0}(\mathbf{D}_{\text{I},\omega}): \text{ minimize } \frac{1}{\Omega}\sum_{\omega=1}^{\Omega} f_{\text{I},\omega}(\mathbf{D}_{\text{I},\omega}) \quad (39)$$
$$\text{subject to } \sum_{\omega=1}^{\Omega} \mathbf{g}_{\text{I},\omega}(\mathbf{D}_{\text{I},\omega}) \preceq \mathbf{0},$$
$$\mathbf{D}_{\text{I},\omega} \in \mathcal{D}_{\text{I},\omega}, \omega \in \{1,\ldots,\Omega\},$$
$$\text{variables } \mathbf{D}_{\text{I},\omega} = [q_{f,l,m,\omega}, b_{f,l,m,\omega}, \mathbf{w}_{\boldsymbol{\rho},\omega}, \mathbf{V}_{\omega}],$$

where $\mathbf{g}_{\text{I},\omega} \triangleq [g_{\text{I},\omega,1},\ldots,g_{\text{I},\omega,M}]: \mathbb{R}^{\Omega\times 1} \to \mathbb{R}^{M\times 1}$ is an affine vector-valued function and represents the backhaul constraint $\overline{\text{C3}}$ with $g_{\text{I},\omega,m}(\cdot) \triangleq \sum_{f\in\mathcal{F}} b_{f,l,m,\omega}Q_f - B_{m,\omega}^{\max}$. Herein, $\mathcal{D}_{\text{I},\omega}$ is a non-convex set because of C2. We assume that $\text{Q0}(\mathbf{D}_{\text{I},\omega})$ is feasible. Let f^* and q^* denote the primal and the dual optimal values of $\text{Q0}(\mathbf{D}_{\text{I},\omega})$, respectively.

Meanwhile, $\text{Q1}(\mathbf{D}_{\text{I},\omega})$ is obtained from $\text{Q0}(\mathbf{D}_{\text{I},\omega})$ by relaxing the binary constraint C2. Due to the convexity of Q1, strong duality holds for problem $\text{Q1}(\mathbf{D}_{\text{I},\omega})$. Based on Lagrangian duality theory, it can be further shown that the dual problems of $\text{Q1}(\mathbf{D}_{\text{I},\omega})$ and $\text{Q0}(\mathbf{D}_{\text{I},\omega})$ are identical [46, Chapter 5.5.3]. Consequently, the optimal value of $\text{Q1}(\mathbf{D}_{\text{I},\omega})$ is also given by q^*. Then, Theorem 2 can be proved by resorting to the following proposition, which estimates the duality gap for $\text{Q0}(\mathbf{D}_{\text{I},\omega})$.

Proposition 2. For problem $\text{Q0}(\mathbf{D}_{\text{I},\omega})$, the duality gap, $f^* - q^*$, is bounded and satisfies

$$0 \leq f^* - q^* \leq \mathcal{O}\left(\frac{M+1}{\Omega}\right). \quad (40)$$

Specifically, according to (40), the difference between the optimal values of $\text{Q1}(\mathbf{D}_{\text{I},\omega})$ and $\text{Q0}(\mathbf{D}_{\text{I},\omega})$ becomes negligible for a sufficiently large value of Ω, i.e.,

$$\lim_{\Omega\to\infty}(f^* - q^*) = 0. \quad (41)$$

Since (41) holds for arbitrary caching decisions, the performance gap between Q1 and Q0 also vanishes as $\Omega \to \infty$. Therefore, the remainder of the proof will be focused on establishing (40) in Proposition 2.

Proof of Proposition 2: The left hand side inequality in (40) is simply due to the weak duality property for general nonlinear optimization problems [46, Chapter 5.1.2]. To prove the right hand side of (40), let us define the sets $\mathcal{X}_{\omega} \triangleq \{\mathbf{x}_{\omega} \triangleq [\mathbf{g}_{\text{I},\omega}(\mathbf{D}_{\text{I},\omega}), f_{\text{I},\omega}(\mathbf{D}_{\text{I},\omega})] \in \mathbb{R}^{M+1} \mid \mathbf{D}_{\text{I},\omega} \in \mathcal{D}_{\text{I},\omega}\}$, $\omega \in \{1,\ldots,\Omega\}$, and their vector (Minkowski) sum $\mathcal{X} \triangleq \{\mathbf{x} = \sum_{\omega=1}^{\Omega}\mathbf{x}_{\omega} \mid \mathbf{x}_{\omega} \in \mathcal{X}_{\omega}\} \subseteq \mathbb{R}^{M+1}$. Here, $\mathbf{x}_{\omega}$ defines an achievable constraint-objective value pair. For simplicity of notation, we also write $\mathcal{X} = \sum_{\omega=1}^{\Omega}\mathcal{X}_{\omega}$. Using $\mathcal{X}$ and its convex hull, $\text{conv}(\mathcal{X})$, the primal optimum and the dual optimum of $\text{Q0}(\mathbf{D}_{\text{I},\omega})$ are given by

$$f^* = \min\{z \mid (\mathbf{y}, z) \in \mathcal{X} \text{ with } \mathbf{y} \preceq \mathbf{0}\} \text{ and} \quad (42)$$
$$q^* = \min\{z \mid (\mathbf{y}, z) \in \text{conv}(\mathcal{X}) \text{ with } \mathbf{y} \preceq \mathbf{0}\},$$

respectively. Assume that $(\mathbf{y}^*, q^*) \in \text{conv}(\mathcal{X})$ obtains the dual optimum with $\mathbf{y}^* \preceq \mathbf{0}$.

The estimation of the duality gap is feasible due to the Shapley-Folkman theorem [47], [36, Proposition 5.26]. Specifically, constrained by the dimension of its subspace $\mathbb{R}^{(M+1) \times 1}$, each point of set $\text{conv}(\mathcal{X}) \subseteq \mathbb{R}^{(M+1) \times 1}$ can be represented as the vector sum of at least $(\Omega - M - 1)$ out of Ω points in $\mathcal{X}_\omega$. This means, for each $(\mathbf{y}^*, q^*) \in \text{conv}(\mathcal{X})$, there exist two index subsets $\mathcal{I}, \overline{\mathcal{I}} \subset \{1, \ldots, \Omega\}$ satisfying $\mathcal{I} \cap \overline{\mathcal{I}} = \emptyset$, $\mathcal{I} \cup \overline{\mathcal{I}} = \{1, \ldots, \Omega\}$, and $|\mathcal{I}| \leq M + 1$, such that

$$\sum_{\omega \in \mathcal{I}} \mathbf{y}_\omega^* + \sum_{\omega \in \overline{\mathcal{I}}} \mathbf{g}_{\mathrm{I},\omega}(\overline{\mathbf{D}}_{\mathrm{I},\omega}) \preceq \mathbf{0} \quad \text{and} \tag{43}$$
$$\sum_{\omega \in \mathcal{I}} q_\omega^* + \sum_{\omega \in \overline{\mathcal{I}}} f_{\mathrm{I},\omega}(\overline{\mathbf{D}}_{\mathrm{I},\omega}) = q^* \Omega$$

hold for $(\mathbf{y}_\omega^*, q_\omega^*) \in \text{conv}(\mathcal{X}_\omega)$, $\omega \in \mathcal{I}$ and $\overline{\mathbf{D}}_{\mathrm{I},\omega} \in \mathcal{D}_{\mathrm{I},\omega}$, $\omega \in \overline{\mathcal{I}}$. In other words, at most $M + 1$ vectors obtained from the dual (relaxed) problem can be infeasible for the primal problem Q0$(\mathbf{D}_{\mathrm{I},\omega})$, i.e., $(\mathbf{y}_\omega^*, q_\omega^*) \in \text{conv}(\mathcal{X}_\omega)$ but $(\mathbf{y}_\omega^*, q_\omega^*) \notin \mathcal{X}_\omega$ for $\omega \in \mathcal{I}$.

Therefore, we construct suboptimal solutions for Q0$(\mathbf{D}_{\mathrm{I},\omega})$ from the dual (relaxed) solutions in (43). Let $\alpha_\omega^i \in [0, 1]$ and $\sum_{i=1}^{M+2} \alpha_\omega^i = 1$. For each $\omega \in \mathcal{I}$, we can express $\mathbf{y}_\omega^*$ and q_ω^* as convex combinations of $\hat{\mathbf{D}}_{\mathrm{I},\omega}^i \in \mathcal{D}_{\mathrm{I},\omega}$, $i \in \{1, \ldots, M + 2\}$, i.e.,

$$\mathbf{y}_\omega^* = \sum_{i=1}^{M+2} \alpha_\omega^i \mathbf{g}_{\mathrm{I},\omega}(\hat{\mathbf{D}}_{\mathrm{I},\omega}^i) \tag{44}$$
$$\succeq \overline{\mathbf{g}}_{\mathrm{I},\omega} \triangleq \min\left\{ \sum_{i=1}^{M+2} \alpha_\omega^i \mathbf{g}_{\mathrm{I},\omega}(\hat{\mathbf{D}}_{\mathrm{I},\omega}^i) \mid \hat{\mathbf{D}}_{\mathrm{I},\omega}^i \in \mathcal{D}_{\mathrm{I},\omega} \right\},$$
$$q_\omega^* = \sum_{i=1}^{M+2} \alpha_\omega^i f_{\mathrm{I},\omega}(\hat{\mathbf{D}}_{\mathrm{I},\omega}^i)$$
$$\geq \overline{f}_{\mathrm{I},\omega} \triangleq \min\left\{ \sum_{i=1}^{M+2} \alpha_\omega^i f_{\mathrm{I},\omega}(\hat{\mathbf{D}}_{\mathrm{I},\omega}^i) \mid \hat{\mathbf{D}}_{\mathrm{I},\omega}^i \in \mathcal{D}_{\mathrm{I},\omega} \right\}.$$

We define the primal solution $\overline{\mathbf{D}}_{\mathrm{I},\omega} \in \mathcal{D}_{\mathrm{I},\omega}$, $\omega \in \mathcal{I} \cup \overline{\mathcal{I}}$, with

$$\overline{\mathbf{D}}_{\mathrm{I},\omega} \in \arg\min\left\{ f_{\mathrm{I},\omega}(\mathbf{D}_{\mathrm{I},\omega}) \mid \mathbf{g}_{\mathrm{I},\omega}(\mathbf{D}_{\mathrm{I},\omega}) \preceq \overline{\mathbf{g}}_{\mathrm{I},\omega}, \tag{45}$$
$$\mathbf{D}_{\mathrm{I},\omega} \in \mathcal{D}_{\mathrm{I},\omega} \right\},$$

for $\omega \in \mathcal{I}$. The feasible set of (45) is generally nonempty and thus $\overline{\mathbf{D}}_{\mathrm{I},\omega}$ usually exists. Meanwhile, $\overline{\mathbf{D}}_{\mathrm{I},\omega}$ is primal feasible since $\mathbf{g}_{\mathrm{I},\omega}(\overline{\mathbf{D}}_{\mathrm{I},\omega}) \preceq \mathbf{y}_\omega^* \preceq \mathbf{0}$, but generally suboptimal for (42) since $\sum_{\omega \in \mathcal{I} \cup \overline{\mathcal{I}}} f_{\mathrm{I},\omega}(\overline{\mathbf{D}}_{\mathrm{I},\omega}) \geq f^* \Omega$.

However, $\overline{\mathbf{D}}_{\mathrm{I},\omega}$ incurs only a bounded penalty on the objective value,

$$f_{\mathrm{I},\omega}(\overline{\mathbf{D}}_{\mathrm{I},\omega}) - q_\omega^* \overset{(a)}{\leq} f_{\mathrm{I},\omega}(\overline{\mathbf{D}}_{\mathrm{I},\omega}) - \overline{f}_{\mathrm{I},\omega} \leq \varrho_\omega \overset{(b)}{\leq} f_{\mathrm{I},\omega}^{UB} - f_{\mathrm{I},\omega}^{LB}, \tag{46}$$

where $f_{\mathrm{I},\omega}^{UB} \triangleq \max\left\{ f_{\mathrm{I},\omega}(\mathbf{D}_{\mathrm{I},\omega}) \mid \mathbf{D}_{\mathrm{I},\omega} \in \mathcal{D}_{\mathrm{I},\omega} \right\}$ and $f_{\mathrm{I},\omega}^{LB} \triangleq \min\left\{ f_{\mathrm{I},\omega}(\mathbf{D}_{\mathrm{I},\omega}) \mid \mathbf{D}_{\mathrm{I},\omega} \in \mathcal{D}_{\mathrm{I},\omega} \right\}$. In (46), (a) is due to (44), and (b) holds since $f_{\mathrm{I},\omega}^{UB} \geq f_{\mathrm{I},\omega}(\mathbf{D}_{\mathrm{I},\omega}) \geq f_{\mathrm{I},\omega}^{LB}$, $\forall \mathbf{D}_{\mathrm{I},\omega} \in \mathcal{D}_{\mathrm{I},\omega}$. We have $\varrho_\omega < +\infty$ since $f_{\mathrm{I},\omega}(\cdot)$ is a continuous function and $\mathcal{D}_{\mathrm{I},\omega} \neq \emptyset$ if problem Q0$(\mathbf{D}_{\mathrm{I},\omega})$ is feasible. Let $\varrho_{\max} \triangleq \max\left\{ \varrho_\omega \mid \omega = 1, \ldots, \Omega \right\}$. The duality gap finally

satisfies

$$f^* \leq \frac{1}{\Omega} \sum_{\omega \in \mathcal{I} \cup \overline{\mathcal{I}}} f_{\mathrm{I},\omega}(\mathbf{D}_{\mathrm{I},\omega}) \tag{47}$$
$$\overset{(a)}{\leq} q^* + \frac{1}{\Omega} \sum_{\omega \in \mathcal{I}} \left(\overline{f}_{\mathrm{I},\omega} - q_\omega^* + \varrho_\omega \right)$$
$$\leq q^* + \frac{M + 1}{\Omega} \varrho_{\max},$$

where (a) is due to (43) and (46). This completes the proof. ∎

Remark 3. Eq. (41) implies that the vector sum of Ω sets, $\mathcal{X}$, in subspace $\mathbb{R}^{(M+1) \times 1}$ tends to be convexified as $\Omega \to \infty$ or $M \ll \Omega$, in the sense that any vector in its convex hull, $\text{conv}(\mathcal{X})$, can be closely approximated by a vector in $\mathcal{X}$ itself due to the underlying geometry. This property has been exploited to solve MINLPs in several disciplines [48], [49].

References

[1] L. Xiang, D. W. K. Ng, R. Schober, and V. W. S. Wong, "Cache-enabled physical-layer security for video streaming in backhaul-limited cellular networks," in *Proc. IEEE Global Comm. Conf. (GLOBECOM) - Workshop on Trusted Communications with Physical Layer Security*, Washington, DC, Dec. 2016.

[2] R. Knutson, "Video boom forces Verizon to upgrade network," *The Wall Street Journal*, Dec. 2013.

[3] P. Rost, C. Bernardos, A. Domenico, M. Girolamo, M. Lalam, A. Maeder, D. Sabella *et al.*, "Cloud technologies for flexible 5G radio access networks," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 68–76, May 2014.

[4] G. Paschos, E. Baştuğ, I. Land, G. Caire, and M. Debbah, "Wireless caching: Technical misconceptions and business barriers," *IEEE Commun. Mag.*, pp. 16–22, Aug. 2016.

[5] A. Liu and V. Lau, "Mixed-timescale precoding and cache control in cached MIMO interference network," *IEEE Trans. Signal Process.*, vol. 61, no. 24, pp. 6320–6332, Dec. 2013.

[6] ——, "Cache-enabled opportunistic cooperative MIMO for video streaming in wireless systems," *IEEE Trans. Signal Process.*, vol. 62, no. 2, pp. 390–402, Jan. 2014.

[7] Z. Chen, J. Lee, T. Q. Quek, and M. Kountouris, "Cooperative caching and transmission design in cluster-centric small cell networks," in *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 3401-3415, May 2017.

[8] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6118–6131, Sep. 2016.

[9] L. Xiang, D. W. K. Ng, T. Islam, R. Schober, V. W. S. Wong, and J. Wang, "Cross-layer optimization of fast video delivery in cache- and buffer-enabled relaying networks," *IEEE Trans. Veh. Technol.*, vol. PP, no. 99, 2017.

[10] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *Proc. IEEE INFOCOM*, New York, NY, Mar. 1999.

[11] M. Peng, S. Yan, K. Zhang, and C. Wang, "Fog-computing-based radio access networks: Issues and challenges," *IEEE Netw.*, vol. 30, no. 4, pp. 46–53, Jul.-Aug. 2016.

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

[13] S. M. Azimi, O. Simeone, and R. Tandon, "Fundamental limits on latency in small-cell caching systems: An information-theoretic analysis," in *Proc. IEEE Global Comm. Conf. (GLOBECOM)*, Washington, DC, Dec. 2016.

[14] S. M. Azimi, O. Simeone, A. Sengupta, and R. Tandon, "Online edge caching in fog-aided wireless network," in *Proc. IEEE Int. Sym. Inf. Theory (ISIT)*, Aachen, Germany, Jun. 2017.

[15] A. Sengupta, R. Tandon, and T. C. Clancy, "Fundamental limits of caching with secure delivery," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 2, pp. 355–370, Feb. 2015.

[16] Z. H. Awan and A. Sezgin, "Fundamental limits of caching in D2D networks with secure delivery," in *Proc. IEEE Int. Conf. Comm. (ICC) - Workshop on Wireless Physical Layer Security*, London, UK, Jun. 2015.

[17] F. Gabry, V. Bioglio, and I. Land, "On edging caching with secrecy constraints," in *Proc. IEEE Int. Conf. Comm. (ICC)*, Kuala Lumpur, Malaysia, May 2016.

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

[19] A. Khisti and G. W. Wornell, "Secure transmission with multiple antennas I: The MISOME wiretap channel," *IEEE Trans. Inf. Theory*, vol. 56, no. 7, pp. 3088–3104, Jul. 2010.

[20] ——, "Secure transmission with multiple antennas II: The MIMOME wiretap channel," *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5515–5532, Nov. 2010.

[21] T.-Y. Liu, P. Mukherjee, S. Ulukus, S.-C. Lin, and Y.-W. P. Hong, "Secure degrees of freedom of MIMO Rayleigh block fading wiretap channels with no CSI anywhere," *IEEE Trans. Wireless Commun.*, vol. 14, no. 5, pp. 2655–2669, May 2015.

[22] A. G. Davoodi and S. A. Jafar, "Aligned image sets under channel uncertainty: Settling conjectures on the collapse of degrees of freedom under finite precision CSIT," *IEEE Trans. Inf. Theory*, vol. 62, no. 10, pp. 5603–5618, Oct. 2016.

[23] D. Gesbert, S. Hanly, H. Huang, S. S. Shitz, O. Simeone, and W. Yu, "Multi-cell MIMO cooperative networks: A new look at interference," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1380–1408, Dec. 2010.

[24] S. Goel and R. Negi, "Guaranteeing secrecy using artificial noise," *IEEE Trans. Wireless Commun.*, vol. 7, no. 6, pp. 2180–2189, Jun. 2008.

[25] Q. Li and W.-K. Ma, "Optimal and robust transmit designs for MISO channel secrecy by semidefinite programming," *IEEE Trans. Signal Process.*, vol. 59, no. 8, pp. 3799–3812, Aug. 2011.

[26] D. W. K. Ng and R. Schober, "Secure and green SWIPT in distributed antenna networks with limited backhaul capacity," *IEEE Trans. Wireless Commun.*, vol. 14, no. 9, pp. 5082–5097, Sep. 2015.

[27] K. J. Ma, R. Bartos, S. Bhatia, and R. Nair, "Mobile video delivery with HTTP," *IEEE Commun. Mag.*, vol. 49, no. 4, pp. 166–175, Apr. 2011.

[28] P. Blasco and D. Gündüz, "Learning-based optimization of cache content in a small cell base station," in *Proc. IEEE Int. Conf. Comm. (ICC)*, Sydney, Australia, Jun. 2014.

[29] J. R. Birge and F. Louveaux, *Introduction to Stochastic Programming*. Springer Science & Business Media, 2011.

[30] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2009.

[31] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[32] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, 2012.

[33] J. M. Cioffi, "A multicarrier primer," *ANSI T1E1. Committee Contribution*, 1991.

[34] A. Mukherjee and A. L. Swindlehurst, "Detecting passive eavesdroppers in the MIMO wiretap channel," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Tokyo, Japan, Mar. 2012.

[35] C. A. Floudas, *Nonlinear and Mixed Integer Optimization: Fundamentals and Applications*. Oxford University Press, 1995.

[36] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press, 1982.

[37] Y. Ye, *Interior Point Algorithms: Theory and Analysis*. John Wiley & Sons, 1997.

[38] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," [Online] Available: http://cvxr.com/cvx, Dec. 2016.

[39] I. Pólik and T. Terlaky, "Interior point methods for nonlinear optimization," in *Nonlinear Optimization*. Springer Berlin Heidelberg, 2010, pp. 215–276.

[40] J. Zhao, T. Q. Quek, and Z. Lei, "Coordinated multipoint transmission with limited backhaul data transfer," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2762–2775, Jun. 2013.

[41] F. Zhuang and V. K. Lau, "Backhaul limited asymmetric cooperation for MIMO cellular networks via semidefinite relaxation," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 684–693, Feb. 2014.

[42] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint optimization of radio and computational resources for multicell mobile-edge computing," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 1, no. 2, pp. 89–103, Jun. 2015.

[43] 3GPP TR 36.814, "Further advancements for E-UTRA physical layer aspects (Release 9)," Mar. 2010.

[44] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 2012.

[45] Z.-Q. Luo, J. F. Sturm, and S. Zhang, "Multivariate nonnegative quadratic mappings," *SIAM J. Optimization*, vol. 14, no. 4, pp. 1140–1162, Jul. 2004.

[46] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Athena Scientific, 1999.

[47] J. P. Aubin and I. Ekeland, "Estimates of the duality gap in nonconvex optimization," *Mathematics of Operations Research*, vol. 1, no. 3, pp. 225–245, Aug. 1976.

[48] D. Bertsekas, G. Lauer, N. Sandell, and T. A. Posbergh, "Optimal short-term scheduling of large-scale power systems," *IEEE Trans. Autom. Control*, vol. 28, no. 1, pp. 1–11, Jan. 1983.

[49] S.-J. Kim and G. Giannakis, "Scalable and robust demand response with mixed-integer constraints," *IEEE Trans. Smart Grid*, vol. 4, no. 4, pp. 2089–2099, Apr. 2013.

Lin Xiang (S'14) received his Bachelor and Master degrees in electronics and information engineering from Huazhong University of Science and Technology (HUST), China, in 2009 and 2012, respectively. From Aug. 2010 to Feb. 2011, he was an exchange student at University of Bologna, Italy, under support from Erasmus Mundus programme. He is now a PhD candidate at Institute for Digital Communication (IDC), Friedrich-Alexander-University of Erlangen-Nuremberg (FAU), Germany. His research work includes resource allocation for cache-aided wireless communication systems, performance analysis of wireless networks based on stochastic geometry theory, and renewable energy integration and electric vehicle charging in smart grid. He is a recipient of the Best Paper Award for IEEE Globecom 2010.

Derrick Wing Kwan Ng (S'06, M'12, SM17) received the bachelor degree with first class honors and the Master of Philosophy (M.Phil.) degree in electronic engineering from the Hong Kong University of Science and Technology (HKUST) in 2006 and 2008, respectively. He received his Ph.D. degree from the University of British Columbia (UBC) in 2012. In the summer of 2011 and spring of 2012, he was a visiting scholar at the Centre Tecnològic de Telecomunicacions de Catalunya - Hong Kong (CTTC-HK). He was a senior postdoctoral fellow at the Institute for Digital Communications, Friedrich-Alexander-University Erlangen-Nürnberg (FAU), Germany. He is now working as a Senior Lecturer and an ARC DECRA Research Fellow at the University of New South Wales, Sydney, Australia. His research interests include convex and non-convex optimization, physical layer security, wireless information and power transfer, and green (energy-efficient) wireless communications. Dr. Ng received the Best Paper Awards at the IEEE International Conference on Computing, Networking and Communications (ICNC) 2016, IEEE Wireless Communications and Networking Conference (WCNC) 2012, the IEEE Global Telecommunication Conference (Globecom) 2011, and the IEEE Third International Conference on Communications and Networking in China 2008. He has served as an editorial assistant to the Editor-in-Chief of the IEEE Transactions on Communications since Jan. 2012. He is currently an Editor of the IEEE Communications Letters, the IEEE Transactions on Wireless Communications, and the IEEE Transactions on Green Communications and Networking. He was honoured as an Exemplary Reviewer of the IEEE Transactions on Communications in 2015, the top reviewer of IEEE Transactions on Vehicular Technology in 2014 and 2016, and an Exemplary Reviewer of the IEEE Wireless Communications Letters for 2012, 2014, and 2015.

Robert Schober (S'98, M'01, SM'08, F'10) was born in Neuendettelsau, Germany, in 1971. He received the Diplom (Univ.) and the Ph.D. degrees in electrical engineering from the University of Erlangen-Nuermberg in 1997 and 2000, respectively. From May 2001 to April 2002 he was a Postdoctoral Fellow at the University of Toronto, Canada, sponsored by the German Academic Exchange Service (DAAD). From May 2002 to December 2011, he was a Professor and Canada Research Chair at the University of British Columbia (UBC), Vancouver, Canada. Since January 2012 he is an Alexander von Humboldt Professor and the Chair for Digital Communication at the Friedrich Alexander University (FAU), Erlangen, Germany. His research interests fall into the broad areas of Communication Theory, Wireless Communications, and Statistical Signal Processing.

Dr. Schober received several awards for his work including the 2002 Heinz Maier-Leibnitz Award of the German Science Foundation (DFG), the 2004 Innovations Award of the Vodafone Foundation for Research in Mobile Communications, the 2006 UBC Killam Research Prize, the 2007 Wilhelm Friedrich Bessel Research Award of the Alexander von Humboldt Foundation, the 2008 Charles McDowell Award for Excellence in Research from UBC, a 2011 Alexander von Humboldt Professorship, and a 2012 NSERC E.W.R. Steacie Fellowship. In addition, he received best paper awards from the German Information Technology Society (ITG), the European Association for Signal, Speech and Image Processing (EURASIP), IEEE WCNC 2012, IEEE Globecom 2011, IEEE ICUWB 2006, the International Zurich Seminar on Broadband Communications, and European Wireless 2000. Dr. Schober is a Fellow of the Canadian Academy of Engineering and a Fellow of the Engineering Institute of Canada. From 2012 to 2015, he served as the Editor-in-Chief of the IEEE Transactions on Communications. Currently, he serves as the Chair of the Steering Committee of the IEEE Transactions on Molecular, Biological and Multiscale Communication, as Member-at-Large on the Board of Governors of the IEEE Communication Society (ComSoc), and as ComSoc Distinguished Lecturer.
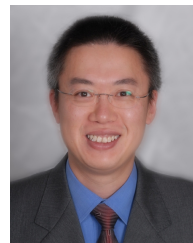
Vincent W.S. Wong (S'94, M'00, SM'07, F'16) received the B.Sc. degree from the University of Manitoba, Winnipeg, MB, Canada, in 1994, the M.A.Sc. degree from the University of Waterloo, Waterloo, ON, Canada, in 1996, and the Ph.D. degree from the University of British Columbia (UBC), Vancouver, BC, Canada, in 2000. From 2000 to 2001, he worked as a systems engineer at PMC-Sierra Inc. (now Microsemi). He joined the Department of Electrical and Computer Engineering at UBC in 2002 and is currently a Professor. His research areas include protocol design, optimization, and resource management of communication networks, with applications to wireless networks, smart grid, mobile cloud computing, and Internet of Things. Dr. Wong is an Editor of the *IEEE Transactions on Communications*. He has served as a Guest Editor of *IEEE Journal on Selected Areas in Communications* and *IEEE Wireless Communications*. He has also served on the editorial boards of *IEEE Transactions on Vehicular Technology* and *Journal of Communications and Networks*. He was a Technical Program Co-chair of *IEEE SmartGridComm*'14, as well as a Symposium Co-chair of *IEEE SmartGridComm* ('13, '17) and *IEEE Globecom*'13. He is the Chair of the IEEE Vancouver Joint Communications Chapter and has served as the Chair of the IEEE Communications Society Emerging Technical Sub-Committee on Smart Grid Communications. Dr. Wong received the 2014 UBC Killam Faculty Research Fellowship.