Secure Video Streaming in Heterogeneous Small Cell Networks with Untrusted Cache Helpers

Lin Xiang*, Derrick Wing Kwan Ng†, Robert Schober*, and Vincent W.S. Wong‡
*University of Erlangen-Nuremberg, Germany
†University of New South Wales, Australia ‡University of British Columbia, Canada

***Abstract*—This paper studies secure video streaming in cache-enabled small cell networks, where some of the cache-enabled small cell base stations (BSs) helping in video delivery are *untrusted*. Unfavorably, caching improves the eavesdropping capability of these untrusted helpers as they may intercept both the cached and the delivered video files. To address this issue, we propose joint caching and scalable video coding (SVC) of video files to enable secure cooperative multiple-input multiple-output (MIMO) transmission and exploit the cache memory of all BSs for improving system performance. The caching and delivery design is formulated as a non-convex mixed-integer optimization problem to minimize the total BS transmit power required for secure video streaming. We develop an algorithm based on the modified generalized Benders decomposition (GBD) to solve the problem optimally. Inspired by the optimal algorithm, a low-complexity suboptimal algorithm is also proposed. Simulation results show that the proposed schemes achieve significant gains in power efficiency and secrecy performance compared to three baseline schemes.**

I. Introduction

Small cells are one of the most promising techniques proposed for spectral- and energy-efficient video streaming in fifth-generation (5G) radio access networks [1]. However, small cell networks typically require high-capacity secure backhaul links to transport the video files from the Internet to the small cell base stations (BSs). Although wireless backhauling is usually preferred for small cells due to its low cost and high flexibility in deployment, the capacity provided by wireless backhauling can be insufficient, which limits the maximum number of concurrent streaming users. Moreover, since wireless transmission is susceptible to eavesdropping, the security of wireless backhauling is a fundamental concern.

Recently, wireless caching has been proposed to enhance the capacity of small cell backhauling [2]–[7]. In wireless caching, the most popular contents are pre-stored at the access points and BSs in close proximity of the user equipments (UEs) [2]. Exploring caching as an alternative to small cell backhauling was investigated first in [3], where caching was shown to reduce the average downloading delay substantially. Caching for improving the energy efficiency of wireless backhauling systems was studied in [4]. In [5], [6], joint caching and buffering for small cell networks was proposed to overcome the backhaul capacity bottleneck and the half-duplex transmission constraint simultaneously. In [7], caching was optimized to facilitate power-efficient cooperative multiple-input multiple-output (MIMO) transmission in small cell networks.

Meanwhile, caching for enhancing communication secrecy was investigated in [8] and [9]. In [8], caching-enabled cooperative MIMO transmission was shown to be an effective physical layer security mechanism for increasing the secrecy rate of the system. However, a secure backhaul for cache placement was required, which cannot always be realized with wireless backhauling in practice. Considering an insecure backhaul, the authors of [9] developed a secure cache placement strategy to prevent the eavesdroppers from obtaining a sufficient number of coded packets for successful recovery of the video file.

In this paper, we investigate caching in small cell networks to facilitate secure video streaming. Due to the distributed small cell network architecture, the caching scheme needs to tackle the secrecy threat originating from *untrusted* cache helpers, i.e., cache-enabled small cell BSs which are potential eavesdroppers. For example, home-owned or open-access small cell BSs may attempt to obtain unpaid premium video streaming service by pretending to cooperate with a macro BS. Different from the case of cache-disabled eavesdroppers considered in [8], [9], these untrusted helpers do not cooperate altruistically, but may maliciously eavesdrop the cached and the delivered video data. In fact, if video data is cached at the helpers, it can be utilized as side information to improve eavesdropping. Therefore, two fundamental questions need to be addressed when cache helpers are untrusted: (1) Can cooperation with untrusted helpers still yield any secrecy benefits? That is, can the cache deployed at the untrusted helpers be utilized to improve the system performance? If so, (2) how to cache and cooperate intelligently to reap the possible performance gains?

We address the above issues as follows. To facilitate secure cooperative transmission with untrusted cache helpers, we propose a caching scheme that combines scalable video coding (SVC) [10] and cooperative MIMO transmission. Specifically, each video file is encoded by SVC into base-layer subfiles, containing basic-quality and independently decodable video information, and enhancement-layer subfiles, containing high-quality video information but decodable only after the base layer has been successfully decoded. By caching the enhancement-layer subfiles across all BSs and the base-layer subfiles only across trusted BSs, secure cooperation of all BSs for power-efficient secure video streaming is enabled. The contributions of this paper are as follows:

- We study a new secrecy threat in small cell networks originating from untrusted cache helpers, i.e., cache-enabled eavesdropping small cell BSs. To facilitate secure cooperative MIMO transmission of trusted and untrusted BSs in small cell networks, we propose a secure caching scheme based on SVC.
- We optimize caching and delivery for minimization of the transmit power under quality-of-service (QoS) and secrecy constraints. An optimal iterative algorithm is proposed based on the modified generalized Benders decomposition (GBD). To reduce the computational complexity, a polynomial-time suboptimal scheme is also proposed.
- Simulation results show that the proposed schemes can exploit the cache capacities of both trusted and untrusted BSs to significantly reduce the transmit power and the secrecy outage probability compared to three baseline schemes.

Notation: Throughout this paper, $\mathbb{C}$ denotes the set of complex numbers; $\mathbf{I}_L$ is an $L \times L$ identity matrix; $\mathbf{1}_{M\times N}$ and $\mathbf{0}_{M\times N}$ are $M \times N$ all-one and all-zero matrices, respectively; $(\cdot)^T$ and $(\cdot)^H$ are the transpose and complex conjugate transpose operators, respectively; $\|\cdot\|_\ell$ denotes the ℓ-norm of a vector;

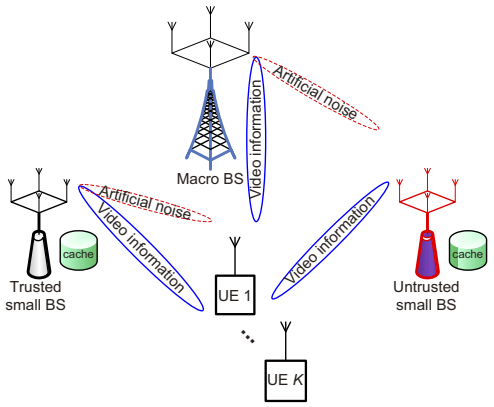

Fig. 1. System model for secure video delivery in a heterogeneous small cell cellular network with a trusted and an untrusted BS.

$\mathsf{tr}(\cdot)$, $\mathsf{rank}(\cdot)$, $\det(\cdot)$, and $\lambda_{\max}(\cdot)$ denote the trace, the rank, the determinant, and the maximum eigenvalue of a matrix, respectively; $\mathsf{diag}(\mathbf{x})$ is a diagonal matrix with the diagonal elements given by vector $\mathbf{x}$; $\mathsf{E}(\cdot)$ is the expectation operator; 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"; $|\mathcal{X}|$ represents the cardinality of set $\mathcal{X}$; $\mathbf{A} \succeq \mathbf{0}$ and $\mathbf{A} \succ \mathbf{0}$ indicate that matrix $\mathbf{A}$ is positive semidefinite and positive definite, respectively; finally, $[x]^+$ stands for $\max\{0, x\}$.

II. System Model

As shown in Fig. 1, we consider wireless video streaming in a heterogeneous small cell network, where M small cell BSs, each equipped with a cache memory of size $C_m^{\max}$ bits, are connected to the macro BS via wireless backhaul links. Let $m \in \mathcal{M} \triangleq \{0, 1, \ldots, M\}$ be the BS index, where $m = 0$ refers to the macro BS. The macro BS is connected to the Internet gateway via a secure optical fiber backhaul link. For simplicity of notation, the backhaul to the macro BS is modeled as a cache with an equivalent capacity of $C_0^{\max}$ bits. Assume that BS m, $m \in \mathcal{M}$, is equipped with N_m antennas. The total number of transmit antennas is denoted by $N \triangleq \sum_{m\in\mathcal{M}} N_m$.

The library owns F video files, indexed by $\mathcal{F} \triangleq \{1, \ldots, F\}$, to be streamed from the Internet to K single-antenna UEs, indexed by $\mathcal{K} \triangleq \{1, \ldots, K\}$. By adopting SVC coding, as employed e.g. in the H.264/Moving Picture Experts Group (MPEG)-4 standard [10], each video file $f \in \mathcal{F}$ is encoded into one base-layer subfile, $(f, 0)$, and $L-1$ enhancement-layer subfiles, (f, l), $l \in \{1, \ldots, L-1\}$. Let $\mathcal{L} \triangleq \{0, \ldots, L-1\}$ be the index set of all layers. The base layer can be decoded independent of the enhancement layers, while enhancement layer $l \in \mathcal{L}\backslash\{0\}$ can be decoded only after layer(s) $0, \ldots, l-1$ have already been decoded [10]. Therefore, the layers have to be decoded in a sequential manner. Due to the specific encoding and decoding structure, the base layer needs the most protection for ensuring secrecy.

The small cell BSs serve as helpers of the macro BS in delivering the video files. However, a subset of small cell BSs are untrusted as they may leak the cached video data and intercept the transmitted video data utilizing the cached data as side information. Let $\mathcal{M}_\mathcal{T} \triangleq \{0, 1, \ldots, J\}$ and $\mathcal{M}_\mathcal{U} \triangleq \{J+1, \ldots, M\}$ denote the sets of trusted and untrusted BSs having a total number of $N_\mathcal{T} \triangleq \sum_{m\in\mathcal{M}_\mathcal{T}} N_m$ and $N_\mathcal{U} \triangleq \sum_{m\in\mathcal{M}_\mathcal{U}} N_m$ antennas, respectively, where $J \leq M$.

The system is assumed to be time slotted and the duration of a time slot is smaller than the channel coherence time. The video files in the cache are updated every T_0 time slots (e.g., once per day), where $T_0 \gg 1$. For notational simplicity, we consider the system only during one typical period $\mathcal{T}_0 \triangleq \{1, \ldots, T_0\}$ and the corresponding time slots are indexed by $t \in \mathcal{T}_0$.

A. Secure Video Caching and Delivery

As the cache helpers in set $\mathcal{M}_\mathcal{U}$ are untrusted, only the enhancement layers are cached at BS $m \in \mathcal{M}_\mathcal{U}$. Hence, the cached subfiles cannot be decoded by the untrusted BSs as long as they do not have access to the base-layer subfiles. Meanwhile, BSs, which have the same base-layer or enhancement-layer subfile cached, can employ cooperative transmission for power-efficient and secure delivery of the subfile to the UEs. On the other hand, video files that are uncached at the small cell BSs can be delivered only by the macro BS. Let $q_{f,l,m} = 1$ indicate that subfile (f, l) is cached at BS m, and $q_{f,l,m} = 0$ otherwise. We thus have the condition

$$\begin{aligned}\text{C1: } & q_{f,l,m} \in \{0, 1\},\ \forall (f, l) \in \mathcal{F} \times \mathcal{L},\ \forall m \in \mathcal{M} \text{ and} \\ & q_{f,0,m} = 0,\ \forall f \in \mathcal{F},\ \forall m \in \mathcal{M}_\mathcal{U}.\end{aligned} \tag{1}$$

Besides, assume that the size of subfile (f, l) is $V_{f,l}$ bits. Then, the cache placement has to satisfy the capacity constraint,

$$\text{C2: } \sum\nolimits_{(f,l)\in\mathcal{F}\times\mathcal{L}} q_{f,l,m} V_{f,l} \leq C_m^{\max},\ m \in \mathcal{M}. \tag{2}$$

During data delivery, the set of BSs cooperating for delivery of subfile (f, l) is denoted by $\mathcal{M}_{f,l}^{\text{Coop}} \triangleq \{m \in \mathcal{M} \mid q_{f,l,m} = 1\}$. Assume that a UE requests one file but possibly multiple layers of the file at a time. We denote a request from user κ for file f by $\boldsymbol{\rho} \triangleq (\kappa, f)$ and the set of requests by $\mathcal{S} \subseteq \mathcal{K} \times \mathcal{F}$. For convenience, the requesting user and the requested file corresponding to $\boldsymbol{\rho}$ are denoted by $\kappa(\boldsymbol{\rho})$ and $f(\boldsymbol{\rho})$, respectively. Moreover, user $\kappa(\boldsymbol{\rho})$ may request $L_{\boldsymbol{\rho}}$ layers, indexed by $\mathcal{L}_{\boldsymbol{\rho}} \triangleq \{0, 1, \ldots, L_{\boldsymbol{\rho}} - 1\}$.

The source symbols of file f for serving request $\boldsymbol{\rho}$ in time slot $t \in \mathcal{T}_0$, denoted by $s_{\boldsymbol{\rho},l,t} \in \mathbb{C}$, $l \in \mathcal{L}_{\boldsymbol{\rho}}$, are complex Gaussian random variables with $s_{\boldsymbol{\rho},l,t} \sim \mathcal{CN}(0, 1)$. Let $\mathbf{w}_{\boldsymbol{\rho},l,t} \triangleq [\mathbf{w}_{\boldsymbol{\rho},l,0,t}^H, \ldots, \mathbf{w}_{\boldsymbol{\rho},l,M,t}^H]^H \in \mathbb{C}^{N\times 1}$ denote the joint beamforming vector for sending symbol $s_{\boldsymbol{\rho},l,t}$, where $\mathbf{w}_{\boldsymbol{\rho},l,m,t} \in \mathbb{C}^{N_m\times 1}$ is the individual beamforming vector used by BS $m \in \mathcal{M}$ in time slot t. Then, the joint transmit signal of BS set $\mathcal{M}$ in time slot $t \in \mathcal{T}_0$ is given by

$$\mathbf{x}_t = \sum\nolimits_{\boldsymbol{\rho}\in\mathcal{S}} \sum\nolimits_{l\in\mathcal{L}_{\boldsymbol{\rho}}} \mathbf{w}_{\boldsymbol{\rho},l,t} s_{\boldsymbol{\rho},l,t} + \mathbf{v}_t, \tag{3}$$

where $\mathbf{v}_t \in \mathbb{C}^{N\times 1}$ is an artificial noise (AN) vector sent to proactively interfere the reception of the untrusted BSs [11]. We assume $\mathbf{v}_t \sim \mathcal{CN}(\mathbf{0}, \mathbf{V}_t)$, where $\mathbf{V}_t$ is the covariance matrix of the artificial noise, i.e., $\mathbf{V}_t \triangleq \mathsf{E}[\mathbf{v}_t \mathbf{v}_t^H] \succeq \mathbf{0}$. As $\mathbf{v}_t$ is injected cooperatively only by the trusted BS set $\mathcal{M}_\mathcal{T}$, we have $\boldsymbol{\Lambda}_\mathcal{U} \mathbf{v}_t = \mathbf{0}$ and $\boldsymbol{\Lambda}_\mathcal{U} \mathbf{V}_t = \mathbf{0}$, where $\boldsymbol{\Lambda}_\mathcal{U}$ is an $N \times N$ diagonal matrix given by $\boldsymbol{\Lambda}_\mathcal{U} = \mathsf{diag}(\mathbf{0}_{N_\mathcal{T}\times 1}^T, \mathbf{1}_{N_\mathcal{U}\times 1}^T)$. Moreover, cooperative transmission of $s_{\boldsymbol{\rho},l,t}$ is possible only if the requested subfile is cached at the BSs, i.e., we require

$$\begin{aligned}\text{C3: } & \mathsf{tr}\left(\boldsymbol{\Lambda}_m \mathbf{w}_{\boldsymbol{\rho},l,t} \mathbf{w}_{\boldsymbol{\rho},l,t}^H\right) \leq q_{f(\rho),l,m} P_m^{\max}, \\ & m \in \mathcal{M},\ \boldsymbol{\rho} \in \mathcal{S},\ l \in \mathcal{L}_{\boldsymbol{\rho}},\ t \in \mathcal{T}_0,\end{aligned} \tag{4}$$

where $P_m^{\max}$ is the maximum transmit power at BS m, and $\boldsymbol{\Lambda}_m$ is an $N \times N$ diagonal matrix given by $\boldsymbol{\Lambda}_m = \mathsf{diag}\big(\mathbf{0}_{(\sum_{j=0}^{m-1} N_j)\times 1}^T, \mathbf{1}_{N_m\times 1}^T, \mathbf{0}_{(\sum_{j=m+1}^{M} N_j)\times 1}^T\big)$ such that $\boldsymbol{\Lambda}_m \mathbf{w}_{\boldsymbol{\rho},l,t} \mathbf{w}_{\boldsymbol{\rho},l,t}^H \equiv \mathbf{w}_{\boldsymbol{\rho},l,m,t} \mathbf{w}_{\boldsymbol{\rho},l,m,t}^H$. C3 is a big-M constraint [12] and enforces $\boldsymbol{\Lambda}_m \mathbf{w}_{\boldsymbol{\rho},l,t} \mathbf{w}_{\boldsymbol{\rho},l,t}^H = \mathbf{0}$, i.e., $\mathbf{w}_{\boldsymbol{\rho},l,m,t} = \mathbf{0}$, whenever $q_{f,l,m} = 0$ or $m \notin \mathcal{M}_{f,l}^{\text{Coop}}$. Based on C1 and C3,

we have $\mathbf{w}_{\boldsymbol{\rho},0,m,t} \equiv \mathbf{0}$, $\forall m \in \mathcal{M}_{\mathcal{U}}$, i.e., the base-layer subfiles cannot be transmitted by untrusted BSs.

We assume a frequency flat fading channel for video data transmission. As a worst cast, we assume that the untrusted BSs are full-duplex, i.e., they can simultaneously eavesdrop the video information intended for the users and participate in the cooperative delivery of the cached files. At time $t \in \mathcal{T}_0$, the self-interference at BS j caused by simultaneous reception and transmission at the same frequency is denoted by $\mathbf{c}_{j,t}$. The received signals at user $\kappa(\boldsymbol{\rho})$ and the untrusted BSs, denoted by $y_{\boldsymbol{\rho},t} \in \mathbb{C}$ and $\mathbf{y}_{\mathcal{U},j,t} \in \mathbb{C}^{N_j \times 1}$, $j \in \mathcal{M}_{\mathcal{U}}$, respectively, are given by

$$y_{\boldsymbol{\rho},t} = \mathbf{h}_{\boldsymbol{\rho},t}^H \mathbf{x}_t + z_{\boldsymbol{\rho},t} \text{ and } \mathbf{y}_{\mathcal{U},j,t} = \mathbf{G}_{j,t}^H \mathbf{x}_t + \mathbf{c}_{j,t} + \mathbf{z}_{j,t}, \quad (5)$$

where $\mathbf{h}_{\boldsymbol{\rho},t} = [\mathbf{h}_{\boldsymbol{\rho},0,t}^H, \ldots, \mathbf{h}_{\boldsymbol{\rho},M,t}^H]^H \in \mathbb{C}^{N \times 1}$ and $\mathbf{G}_{j,t} = [\mathbf{G}_{j,0,t}^H, \ldots, \mathbf{G}_{j,J,t}^H, \mathbf{0}_{(M-J)\times N_j}^H]^H \in \mathbb{C}^{N \times N_j}$ are the channel vectors/matrices from BS set $\mathcal{M}$ to user $\kappa(\boldsymbol{\rho})$ and BS j, respectively. $\mathbf{h}_{\boldsymbol{\rho},m,t} \in \mathbb{C}^{N_m \times 1}$ and $\mathbf{G}_{j,m,t} \in \mathbb{C}^{N_m \times N_j}$ model the channels between BS $m \in \mathcal{M}$ and the respective receivers. Finally, $z_{\boldsymbol{\rho},t} \sim \mathcal{CN}(0, \sigma^2)$ and $\mathbf{z}_{j,t} \sim \mathcal{CN}(\mathbf{0}, \sigma_j^2 \mathbf{I}_{N_j})$ are the complex Gaussian noises at the users and the BSs, respectively.

B. Achievable Secrecy Rate

Each user employs successive interference cancellation (SIC) at the receiver [10]. The base-layer subfile is decoded first, as it is required for the decoding of the other layers. In decoding the subfile of layer $l \in \mathcal{L}_{\boldsymbol{\rho}} \backslash \{0\}$, the decoded lower layers $0, \ldots, l-1$ are first removed from the received signal for interference cancellation. The process continues until layer $L_{\boldsymbol{\rho}} - 1$ is decoded [10]. Define the interference cancellation coefficient $a_{\boldsymbol{\rho},l}^{\boldsymbol{\rho}',l'} \in \{0,1\}$, where $a_{\boldsymbol{\rho},l}^{\boldsymbol{\rho}',l'} = 0$ if $\boldsymbol{\rho} = \boldsymbol{\rho}'$, $l \geq l'$, and $a_{\boldsymbol{\rho},l}^{\boldsymbol{\rho}',l'} = 1$ otherwise. The instantaneous achievable rate (bits/s/Hz) for layer $l \in \mathcal{L}_{\boldsymbol{\rho}}$ at user $\kappa(\boldsymbol{\rho})$ at time slot t is given by

$$R_{\boldsymbol{\rho},l,t} = \log_2 \Big(1 + \frac{\frac{1}{\sigma^2} \left|\mathbf{h}_{\boldsymbol{\rho},t}^H \mathbf{w}_{\boldsymbol{\rho},l,t}\right|^2}{1 + \frac{1}{\sigma^2} I_{\boldsymbol{\rho},l,t} + \frac{1}{\sigma^2} \mathbf{h}_{\boldsymbol{\rho},t}^H \mathbf{V}_t \mathbf{h}_{\boldsymbol{\rho},t}}\Big), \quad (6)$$

where $I_{\boldsymbol{\rho},l,t} = \sum_{(\boldsymbol{\rho}',l') \neq (\boldsymbol{\rho},l)} a_{\boldsymbol{\rho},l}^{\boldsymbol{\rho}',l'} \left|\mathbf{h}_{\boldsymbol{\rho},t}^H \mathbf{w}_{\boldsymbol{\rho}',l',t}\right|^2$ is the residual interference term for decoding layer l of user $\kappa(\boldsymbol{\rho})$, and $(\boldsymbol{\rho}, l) \neq (\boldsymbol{\rho}', l')$ indicates $\boldsymbol{\rho} \neq \boldsymbol{\rho}'$ and/or $l \neq l'$.

On the other hand, the untrusted BSs may eavesdrop the video information intended for the users. We consider the worst case in terms of secrecy and assume that BS $j \in \mathcal{M}_{\mathcal{U}}$ can fully cancel the self-interference power $\mathbf{c}_{j,t}$ during eavesdropping, and hence, can achieve the capacity upper bound for full-duplex communication given by

$$\begin{aligned} R_{j,\boldsymbol{\rho},l,t} &= \log_2 \det \left(\mathbf{I}_{N_j} + \tfrac{1}{\sigma_j^2} \mathbf{Z}_{j,\boldsymbol{\rho},l,t}^{-1} \mathbf{G}_{j,t}^H \mathbf{w}_{\boldsymbol{\rho},l,t} \mathbf{w}_{\boldsymbol{\rho},l,t}^H \mathbf{G}_{j,t}\right), \\ \mathbf{Z}_{j,\boldsymbol{\rho},l,t} &= \mathbf{I}_{N_j} + \tfrac{1}{\sigma_j^2} \mathbf{G}_{j,t}^H \mathbf{V}_t \mathbf{G}_{j,t} + \tfrac{1}{\sigma_j^2} \boldsymbol{\Psi}_{j,\boldsymbol{\rho},l,t} \succ \mathbf{0}, \quad (7) \\ \boldsymbol{\Psi}_{j,\boldsymbol{\rho},l,t} &= \sum_{(\boldsymbol{\rho}',l') \neq (\boldsymbol{\rho},l)} a_{\boldsymbol{\rho},l}^{\boldsymbol{\rho}',l'} (1 - q_{f(\boldsymbol{\rho}'),l',j}) \mathbf{G}_{j,t}^H \mathbf{w}_{\boldsymbol{\rho}',l',t} \mathbf{w}_{\boldsymbol{\rho}',l',t}^H \mathbf{G}_{j,t}. \end{aligned}$$

Note that if subfile (f, l') is cached at BS $j \in \mathcal{M}_{\mathcal{U}}$, we have $1 - q_{f,l',j} = 0$ in (7). That is, BS $j \in \mathcal{M}_{\mathcal{U}}$ can utilize the cached video data as side information to suppress the interference caused by subfile (f, l'). The secrecy rate achievable at user $\kappa(\boldsymbol{\rho})$ for decoding layer $l \in \mathcal{L}_{\boldsymbol{\rho}}$ at time $t \in \mathcal{T}_0$ is

$$R_{\boldsymbol{\rho},l,t}^{\text{sec}} = \Big[R_{\boldsymbol{\rho},l,t} - \max_{j \in \mathcal{M}_{\mathcal{U}}} R_{j,\boldsymbol{\rho},l,t}\Big]^+. \quad (8)$$

Remark 1. A passive eavesdropper as in [8], [9] can be cast as an untrusted BS without cache memory in the above model.

III. Problem Formulation and Solution

In this paper, we assume that the global channel state information (CSI) is perfectly known at the macro BS during video delivery. The caching and the delivery decisions are optimized at two different time scales such that the total transmit power required for secure video streaming is minimized. In the following, we first present the caching optimization problem and its solution. The delivery optimization problem is addressed at the end of this section.

A. Caching Optimization

Let $\mathbf{q} \triangleq [q_{f,l,m}]$ and $\mathbf{w}_t \triangleq [\mathbf{w}_{\boldsymbol{\rho},l,m,t}]$ be the caching and the beamforming optimization vectors, respectively. Similar to the previous works [5], [6], [8], the caching policy is optimized offline and updated (at the end of) every T_0 time slots based on the profile of historical user requests and CSI collected during the time period. For the considered typical period $\mathcal{T}_0$, the caching optimization problem is formulated as:

$$\begin{aligned} \text{P0:} \min_{\mathbf{q},\mathbf{w}_t,\mathbf{V}_t} \ & \sum\nolimits_{t \in \mathcal{T}_0} U_{\text{TP}}(\mathbf{q}, \mathbf{w}_t, \mathbf{V}_t) \quad (9) \\ \text{s.t.} \ & \text{C1, C2, C3,} \\ & \text{C4: } \mathbf{V}_t \succeq \mathbf{0}, \ \boldsymbol{\Lambda}_{\mathcal{U}} \mathbf{V}_t = \mathbf{0}, \ t \in \mathcal{T}_0, \\ & \text{C5: } \mathsf{tr}\Big(\boldsymbol{\Lambda}_m \big(\sum\nolimits_{\boldsymbol{\rho} \in \mathcal{S}} \sum\nolimits_{l \in \mathcal{L}_{\boldsymbol{\rho}}} \mathbf{w}_{\boldsymbol{\rho},l,t} \mathbf{w}_{\boldsymbol{\rho},l,t}^H + \mathbf{V}_t\big)\Big) \\ & \quad \leq P_m^{\max}, \quad m \in \mathcal{M}, \ t \in \mathcal{T}_0, \\ & \text{C6: } R_{\boldsymbol{\rho},l,t} \geq R_{\boldsymbol{\rho},l}^{\text{req}}, \quad \boldsymbol{\rho} \in \mathcal{S}, \ l \in \mathcal{L}_{\boldsymbol{\rho}}, \ t \in \mathcal{T}_0, \\ & \text{C7: } \max_{j \in \mathcal{M}_{\mathcal{U}}} R_{j,\boldsymbol{\rho},0,t} \leq R_{\boldsymbol{\rho},0}^{\text{tol}}, \quad \boldsymbol{\rho} \in \mathcal{S}, \ t \in \mathcal{T}_0, \end{aligned}$$

where $U_{\text{TP}}(\mathbf{q}, \mathbf{w}_t, \mathbf{V}_t) \triangleq \mathsf{tr}(\sum_{\boldsymbol{\rho} \in \mathcal{S}} \sum_{l \in \mathcal{L}_{\boldsymbol{\rho}}} \mathbf{w}_{\boldsymbol{\rho},l,t} \mathbf{w}_{\boldsymbol{\rho},l,t}^H + \mathbf{V}_t)$ denotes the total BS transmit power at time $t \in \mathcal{T}_0$. Constraint C5 limits the maximal transmit power of BS $m \in \mathcal{M}$ to $P_m^{\max}$. C6 guarantees a constant minimum video delivery rate, $R_{\boldsymbol{\rho},l}^{\text{req}}$, to provide QoS in delivering layer $l \in \mathcal{L}_{\boldsymbol{\rho}}$ for serving request $\boldsymbol{\rho} \in \mathcal{S}$. C7 constrains the maximum data rate leaked to the untrusted BSs in set $\mathcal{M}_{\mathcal{U}}$ to $R_{\boldsymbol{\rho},0}^{\text{tol}}$ for ensuring communication secrecy. Since the untrusted BSs are unable to decode the enhancement layers without base layer information, the system secrecy can be ensured by imposing C7 only on the delivery of the base-layer subfiles. C6 and C7 together guarantee a minimum achievable secrecy rate of $R_{\boldsymbol{\rho},0,t}^{\text{sec}} = \big[R_{\boldsymbol{\rho},0}^{\text{req}} - R_{\boldsymbol{\rho},0}^{\text{tol}}\big]^+$ for delivering the base-layer subfiles for request $\boldsymbol{\rho}$.

Problem P0 is a non-convex mixed-integer nonlinear program (MINLP) due to the binary caching vector $\{\mathbf{q}\}$ and the non-convex constraints C6 and C7. We will show below that P0 can be transformed into an equivalent convex MINLP by semi-definite programming (SDP) relaxation and, subsequently, be solved optimally by an iterative algorithm.

Let $\mathbf{W}_{\boldsymbol{\rho},l,t} = \mathbf{w}_{\boldsymbol{\rho},l,t} \mathbf{w}_{\boldsymbol{\rho},l,t}^H \succeq \mathbf{0}$ with $\mathsf{rank}(\mathbf{W}_{\boldsymbol{\rho},l,t}) \leq 1$. Constraint C6 is equivalent to an affine inequality constraint,

$$\overline{\text{C6}}\text{: } \mathsf{tr}\left[\Big(\tfrac{\mathbf{W}_{\boldsymbol{\rho},l,t}}{\eta_{\boldsymbol{\rho},l}^{\text{req}}} - \sum\nolimits_{(\boldsymbol{\rho}',l') \neq (\boldsymbol{\rho},l)} a_{\boldsymbol{\rho},l}^{\boldsymbol{\rho}',l'} \mathbf{W}_{\boldsymbol{\rho}',l',t} - \mathbf{V}_t\Big) \mathbf{H}_{\boldsymbol{\rho},t}\right] \geq \sigma^2, \quad (10)$$

where $\eta_{\boldsymbol{\rho},l}^{\text{req}} \triangleq 2^{R_{\boldsymbol{\rho},l}^{\text{req}}} - 1$ and $\mathbf{H}_{\boldsymbol{\rho},t} \triangleq \mathbf{h}_{\boldsymbol{\rho},t} \mathbf{h}_{\boldsymbol{\rho},t}^H$. Meanwhile, let $\overline{\mathbf{W}}_{\boldsymbol{\rho},l,j,t}$ be an auxiliary optimization matrix subject to $\overline{\mathbf{W}}_{\boldsymbol{\rho},l,j,t} = (1 - q_{f(\boldsymbol{\rho}),l,j}) \mathbf{W}_{\boldsymbol{\rho},l,t} \succeq \mathbf{0}$. We have

$$\boldsymbol{\Psi}_{j,\boldsymbol{\rho},l,t} = \mathbf{G}_{j,t}^H \Big(\sum\nolimits_{(\boldsymbol{\rho}',l') \neq (\boldsymbol{\rho},l)} a_{\boldsymbol{\rho},l}^{\boldsymbol{\rho}',l'} \overline{\mathbf{W}}_{\boldsymbol{\rho}',l',j,t}\Big) \mathbf{G}_{j,t}, \quad (11)$$

if $\mathsf{rank}(\overline{\mathbf{W}}_{\boldsymbol{\rho},l,j,t}) \leq 1$ and the following constraints hold,

$$\text{C8: } \mathsf{tr}(\mathbf{W}_{\boldsymbol{\rho},l,t} - \overline{\mathbf{W}}_{\boldsymbol{\rho},l,j,t}) \preceq q_{f(\boldsymbol{\rho}),l,j} P_{\max}, \ j \in \mathcal{M}_{\mathcal{U}},$$

$$\text{C9: } \mathsf{tr}(\overline{\mathbf{W}}_{\boldsymbol{\rho},l,j,t}) \preceq (1-q_{f(\boldsymbol{\rho}),l,j})P_{\max},\ j\in\mathcal{M}_{\mathcal{U}}, \tag{12}$$
$$\text{C10: } \mathbf{W}_{\boldsymbol{\rho},l,t}\succeq\overline{\mathbf{W}}_{\boldsymbol{\rho},l,j,t},\ \overline{\mathbf{W}}_{\boldsymbol{\rho},l,j,t}\succeq\mathbf{0},\ \boldsymbol{\rho}\in\mathcal{S},\ j\in\mathcal{M}_{\mathcal{U}},$$

where $P_{\max}\triangleq\sum_{m\in\mathcal{M}}P_m^{\max}$. Here, C8 and C9 guarantee that $\overline{\mathbf{W}}_{\boldsymbol{\rho},l,j,t}=\mathbf{0}$ if $q_{f(\boldsymbol{\rho}),l,j}=1$, and $\overline{\mathbf{W}}_{\boldsymbol{\rho},l,j,t}=\mathbf{W}_{\boldsymbol{\rho},l,t}$ otherwise. By substituting $\overline{\mathbf{W}}_{\boldsymbol{\rho},l,j,t}$ and $\boldsymbol{\Psi}_{j,\boldsymbol{\rho},l,t}$, C7 can be reformulated into a linear matrix inequality (LMI) as follows

$$\begin{aligned}
\text{C7} &\iff \tfrac{1}{\sigma_j^2}\mathbf{w}_{\boldsymbol{\rho},0,t}^H\mathbf{G}_{j,t}\mathbf{Z}_{j,\boldsymbol{\rho},0,t}^{-1}\mathbf{G}_{j,t}^H\mathbf{w}_{\boldsymbol{\rho},0,t}\le\eta_{\boldsymbol{\rho},0}^{\text{tot}}\triangleq 2^{R_{\boldsymbol{\rho},0}^{\text{tot}}}-1,\\
&\iff \mathsf{tr}(\mathbf{Z}_{j,\boldsymbol{\rho},0,t}^{-1}\mathbf{G}_{j,t}^H\mathbf{W}_{\boldsymbol{\rho},0,t}\mathbf{G}_{j,t})\le\sigma_j^2\eta_{\boldsymbol{\rho},0}^{\text{tot}},\\
&\overset{\text{(a)}}{\iff} \lambda_{\max}(\mathbf{Z}_{j,\boldsymbol{\rho},0,t}^{-1/2}\mathbf{G}_{j,t}^H\mathbf{W}_{\boldsymbol{\rho},0,t}\mathbf{G}_{j,t}\mathbf{Z}_{j,\boldsymbol{\rho},0,t}^{-1/2})\le\sigma_j^2\eta_{\boldsymbol{\rho},0}^{\text{tot}},\\
&\iff \overline{\text{C7}}\text{: } \mathbf{G}_{j,t}^H\mathbf{W}_{\boldsymbol{\rho},0,t}\mathbf{G}_{j,t}\preceq\sigma_j^2\eta_{\boldsymbol{\rho},0}^{\text{tot}}\mathbf{Z}_{j,\boldsymbol{\rho},0,t},\ j\in\mathcal{M}_{\mathcal{U}},
\end{aligned}\tag{13}$$

where (a) holds due to the rank constraint on $\overline{\mathbf{W}}_{\boldsymbol{\rho},l,t}$.

Note that both $\overline{\text{C6}}$ and $\overline{\text{C7}}$ are jointly convex in $\mathbf{X}_t\triangleq\{\mathbf{W}_{\boldsymbol{\rho},l,t},\overline{\mathbf{W}}_{\boldsymbol{\rho},l,j,t},\mathbf{V}_t\}$. By applying the above transformations, problem P0 is equivalently reformulated as

$$\begin{aligned}
\min_{\mathbf{q},\mathbf{X}_t}\quad & \sum\nolimits_{t\in\mathcal{T}_0}U_{\text{TP}}(\mathbf{q},\mathbf{X}_t) \\
\text{s.t.}\quad & \text{C1, C2, C3, C4, C5, } \overline{\text{C6}}\text{, } \overline{\text{C7}}\text{, C8, C9, C10}\\
& \text{C11: } \mathsf{rank}(\mathbf{W}_{\boldsymbol{\rho},l,t})\le 1,\quad \boldsymbol{\rho}\in\mathcal{S},\ l\in\mathcal{L}_{\boldsymbol{\rho}},\ t\in\mathcal{T}_0,
\end{aligned}\tag{14}$$

where the rank constraint $\mathsf{rank}(\overline{\mathbf{W}}_{\boldsymbol{\rho},l,t})\le 1$ is dropped due to C10 and C11. Let P1 be the SDP relaxation of (14), obtained by dropping C11 in (14). Then, P1 is a convex MINLP, which is solved optimally by the modified GBD algorithm [12] below.

B. Optimal Iterative Solution

The optimal solution of P1 employs a two-layer decomposition: the binary caching optimization problem for $\mathbf{q}$ is solved in the outer layer, and the continuous delivery optimization problem for $\mathbf{X}_t$ in the inner layer. Note that $\mathbf{q}$ and $\mathbf{X}_t$ are coupled via constraints C3, C8, and C9. To simplify the decomposition, we perturb the right-hand side of C3, C8, and C9 by slack variables $s_{\boldsymbol{\rho},l,m,t}^{\text{C3}}\ge 0$, $s_{\boldsymbol{\rho},l,j,t}^{\text{C8}}\ge 0$, and $s_{\boldsymbol{\rho},l,j,t}^{\text{C9}}\ge 0$, respectively. Let $\mathbf{s}_t\triangleq[s_{\boldsymbol{\rho},l,m,t}^{\text{C3}},\ s_{\boldsymbol{\rho},l,j,t}^{\text{C8}},\ s_{\boldsymbol{\rho},l,j,t}^{\text{C9}}]$ be the perturbation vector and $\mathbf{s}_t\succeq\mathbf{0}$. Moreover, in the objective function, we add an ℓ_1-norm (exact) penalty cost function for $\mathbf{s}_t$

$$f_{\text{Pen}}(\mathbf{s}_t)\triangleq\mu\Big(\sum\nolimits_{\boldsymbol{\rho},l,m}s_{\boldsymbol{\rho},l,m,t}^{\text{C3}}+\sum\nolimits_{\boldsymbol{\rho},l,j}(s_{\boldsymbol{\rho},l,j,t}^{\text{C8}}+s_{\boldsymbol{\rho},l,j,t}^{\text{C9}})\Big),\tag{15}$$

with penalty factor $\mu\gg 1$. Consequently, P1 decomposes into T_0 SDP subproblems in the inner layer,

$$\begin{aligned}
\nu_t(\mathbf{q})\triangleq\min_{\mathbf{X}_t,\mathbf{s}_t\succeq\mathbf{0}}\ & U_{\text{TP}}(\mathbf{q},\mathbf{X}_t)+f_{\text{Pen}}(\mathbf{s}_t)\\
\text{s.t.}\ & \mathbf{X}_t\in\mathcal{X}_t\triangleq\{\mathbf{X}_t\mid\text{C4, C5, }\overline{\text{C6}}\text{, }\overline{\text{C7}}\text{, C10}\},\\
& \overline{\text{C3}}\text{: } \mathsf{tr}(\boldsymbol{\Lambda}_m\mathbf{W}_{\boldsymbol{\rho},l,t})-q_{f(\boldsymbol{\rho}),l,m}P_m^{\max}\le s_{\boldsymbol{\rho},l,m,t}^{\text{C3}},\\
& \overline{\text{C8}}\text{: } \mathsf{tr}(\mathbf{W}_{\boldsymbol{\rho},l,t}-\overline{\mathbf{W}}_{\boldsymbol{\rho},l,j,t})-q_{f(\boldsymbol{\rho}),l,j}P_{\max}\le s_{\boldsymbol{\rho},l,j,t}^{\text{C8}},\\
& \overline{\text{C9}}\text{: } \mathsf{tr}(\overline{\mathbf{W}}_{\boldsymbol{\rho},l,j,t})-(1-q_{f(\boldsymbol{\rho}),l,j})P_{\max}\le s_{\boldsymbol{\rho},l,j,t}^{\text{C9}},
\end{aligned}\tag{16}$$

and a mixed-integer linear program (MILP) in the outer layer

$$\begin{aligned}
\min_{\mathbf{q},\alpha}\quad & \alpha\\
\text{s.t.}\quad & \alpha\ge\nu(\mathbf{q})\triangleq\sum\nolimits_{t\in\mathcal{T}_0}\nu_t(\mathbf{q}),\quad \mathbf{q}\in\mathcal{Q}\triangleq\{\mathbf{q}\mid\text{C1, C2}\}.
\end{aligned}\tag{17}$$

Problem (17) is referred as the master problem. By perturbation, the feasible set of $\nu(\mathbf{q})$ is extended to $\mathcal{Q}\subseteq\{0,1\}^{F\times L\times M}$. Meanwhile, the master problem is equivalent to P1 when $\mu\gg 1$, as stated in Proposition 1.

Proposition 1. For $\mu\gg 1$, we have i) if the SDP subproblem in (16) is infeasible, i.e., $\nu(\mathbf{q})=+\infty$, then the same holds true for problem P1; ii) if P1 is feasible, then (16) is always feasible for any $\mathbf{q}\in\mathcal{Q}$; iii) the optimal solution of $\mathbf{q}$ for P1 also solves the master problem (17); and iv) if the optimal solution of (16) satisfies $\mathbf{s}_t\neq\mathbf{0}$, then P1 is infeasible.

The proof of Proposition 1 closely follows the proof of [13, Theorem 9.3.1] and is omitted here because of space constraints.

After decomposition, problem (16) is an SDP, which can be efficiently solved by interior point methods using, e.g., CVX [14]. To solve the master problem, we further resort to the Lagrangian dual of problem (16). Let $\lambda_{\boldsymbol{\rho},l,m,t}^{\overline{\text{C3}}}\ge 0$, $\lambda_{\boldsymbol{\rho},l,j,t}^{\overline{\text{C8}}}\ge 0$, and $\lambda_{\boldsymbol{\rho},l,j,t}^{\overline{\text{C9}}}\ge 0$ be the Lagrange multipliers of $\overline{\text{C3}}$, $\overline{\text{C8}}$, and $\overline{\text{C9}}$, respectively. Define $\boldsymbol{\lambda}_t\triangleq[\lambda_{\boldsymbol{\rho},l,m,t}^{\overline{\text{C3}}},\lambda_{\boldsymbol{\rho},l,j,t}^{\overline{\text{C8}}},\lambda_{\boldsymbol{\rho},l,j,t}^{\overline{\text{C9}}}]$. The Lagrangian of (16) is separable with respect to $\{\mathbf{q}\}$ and $\{\mathbf{X}_t,\mathbf{s}_t\}$, i.e., it can be denoted as

$$\mathcal{L}_{\mathbf{q}}(\mathbf{X}_t,\mathbf{s}_t;\boldsymbol{\lambda}_t)=f_1(\mathbf{q};\boldsymbol{\lambda}_t)+f_2(\mathbf{X}_t,\mathbf{s}_t;\boldsymbol{\lambda}_t),\tag{18}$$

where $f_1(\mathbf{q};\boldsymbol{\lambda}_t)$ and $f_2(\mathbf{X}_t,\mathbf{s}_t;\boldsymbol{\lambda}_t)$ are the collections of the terms involving $\mathbf{q}$ and $\{\mathbf{X}_t,\mathbf{s}_t\}$, respectively. Since, for any given $\mathbf{q}$, problem (16) is a convex problem and fulfills Slater's condition, the following result holds due to strong duality [12]:

$$\nu_t(\mathbf{q})=\max_{\boldsymbol{\lambda}\succeq\mathbf{0}}\min_{\mathbf{X}_t\in\mathcal{X}_t,\,\mathbf{s}_t\succeq\mathbf{0}}\mathcal{L}_{\mathbf{q}}(\mathbf{X}_t,\mathbf{s}_t;\boldsymbol{\lambda}_t),\quad\forall\mathbf{q}\in\mathcal{Q}.\tag{19}$$

Consequently, the master problem is reformulated as

$$\begin{aligned}
\min_{\mathbf{q}\in\mathcal{Q},\alpha}\quad & \alpha\\
\text{s.t.}\quad & \alpha\ge\sum\nolimits_{t\in\mathcal{T}_0}\xi_t(\mathbf{q};\boldsymbol{\lambda}_t),\quad\forall\boldsymbol{\lambda}_t\succeq\mathbf{0},
\end{aligned}\tag{20}$$

where $\xi_t(\mathbf{q};\boldsymbol{\lambda}_t)\triangleq\min_{\mathbf{X}_t\in\mathcal{X}_t,\,\mathbf{s}_t\succeq\mathbf{0}}\mathcal{L}_{\mathbf{q}}(\mathbf{X}_t,\mathbf{s}_t;\boldsymbol{\lambda}_t)$.

We solve problem (20) by the iterative relaxation algorithm given in Algorithm 1. Let k be the iteration index. We start from one constraint at $k=1$. Then, the number of constraints are increased sequentially as the iteration proceeds. Specifically, for given dual variables $\boldsymbol{\lambda}_t^j$, $j=1,\dots k-1$, the following master problem is solved in iteration k,

$$\begin{aligned}
\min_{\mathbf{q}\in\mathcal{Q},\alpha}\quad & \alpha\\
\text{s.t.}\quad & \alpha\ge\sum\nolimits_{t\in\mathcal{T}_0}\xi_t\big(\mathbf{q};\boldsymbol{\lambda}_t^j\big),\quad j=1,\dots,k-1.
\end{aligned}\tag{21}$$

Problem (21) is a relaxation of problem (20). Due to the enlarged feasible set, the optimal value of problem (21) gives a lower bound on that of problem (20). The relaxation solution, denoted by $(\mathbf{q}^k,\alpha^k)$, becomes optimal for (20) and P1 if it is feasible for problem (20). Otherwise, a cutting plane (also referred as an optimality cut [12]) is added to the feasible set of (21) to tighten the relaxation solution. As this process continues, a non-decreasing sequence of lower bounds is generated until the relaxation solution becomes feasible, i.e., solves problem (20) optimally, or until the problem is known to be infeasible.

As shown in Algorithm 1 (lines 5−10), the feasibility or optimality of $\mathbf{q}^k$ is verified by solving the SDP subproblem in (16). This is because, if $\mathbf{q}^k$ is optimal, then solving (16) for $\mathbf{q}=\mathbf{q}^k$ in line 5 would return the optimal value of α^k, i.e., $\nu(\mathbf{q}^k)=\alpha^k$, owing to the strong duality of (16). Otherwise, $\nu(\mathbf{q}^k)$ gives an upper bound on the optimal value, and thus, $\nu(\mathbf{q}^k)\ge\alpha^k$. By keeping the current lowest upper bound, i.e., $UB\leftarrow\min\{UB,\nu(\mathbf{q}^k)\}$ (cf. line 9), the optimality condition is satisfied when the gap between UB and the lower bound vanishes. Note that the values of $\boldsymbol{\lambda}_t^k$ at iteration k can be chosen

Algorithm 1 Optimal iterative algorithm for solving P0/P1

1: **Initialization**: Given $\mathbf{q}^0 \leftarrow \mathbf{0}$. Solve the primal problem (16) for given $\mathbf{q}^0$ and determine $\mathbf{X}_t^1, \mathbf{s}_t^1, \boldsymbol{\lambda}_t^1$; set tolerance $\varepsilon \geq 0$, $UB \leftarrow \nu(\mathbf{q}^0)$, $LB \longleftarrow \infty$, $k \leftarrow 1$;
2: **while** $(UB > LB + \varepsilon)$ **do**
3: Solve the relaxed master problem (21) for given $\mathbf{X}_t^k$, $\mathbf{s}_t^k$, $\boldsymbol{\lambda}_t^k$ and determine the solutions $(\mathbf{q}^k, \alpha^k)$;
4: Update lower bound and solution: $LB \leftarrow \alpha^k$, $\mathbf{q}^* \leftarrow \mathbf{q}^k$;
5: Solve the primal problem (16) for given $\mathbf{q}^k$ and determine the primal and the dual solutions $\mathbf{X}_t^{k+1}$, $\mathbf{s}_t^{k+1}$, $\boldsymbol{\lambda}_t^{k+1}$;
6: **if** $(\nu(\mathbf{q}^k) = +\infty$, i.e., (16) is infeasible, **OR** $\nu(\mathbf{q}^k) \leq \alpha^k + \varepsilon)$ **then**
7: Set $\mathbf{X}_t^* \leftarrow \mathbf{X}_t^{k+1}$, $\mathbf{s}_t^* \leftarrow \mathbf{s}_t^{k+1}$ and exit the while loop;
8: **else if** $(\nu(\mathbf{q}^k) < UB)$ **then**
9: Update upper bound and solution: $UB \leftarrow \nu(\mathbf{q}^k)$, $\mathbf{X}_t^* \leftarrow \mathbf{X}_t^{k+1}$, $\mathbf{s}_t^* \leftarrow \mathbf{s}_t^{k+1}$;
10: **end if**
11: Update iteration index: $k \leftarrow k + 1$;
12: **end while**
13: **if** $(\mathbf{s}_t^* = \mathbf{0})$ **then**
14: Return the optimal solutions $\mathbf{q}^*$ and $\mathbf{X}_t^*$;
15: **else**
16: Return the infeasible problem P0/P1.
17: **end if**

as the optimal dual solutions of (19) to determine $\xi_t\big(\mathbf{q}; \boldsymbol{\lambda}_t^k\big)$ for $\mathbf{q} = \mathbf{q}^k$, cf. line 5, due to the following proposition.

Proposition 2. Let $(\mathbf{X}_t^k, \mathbf{s}_t^k)$ and $\boldsymbol{\lambda}_t^k$ be the optimal primal and dual solutions of (19) for $\mathbf{q} = \mathbf{q}^k$ at iteration k, respectively. Then, $(\mathbf{X}_t^k, \mathbf{s}_t^k)$ also solves the minimization problem in the optimality cut generated in iteration $k+1$ (cf. (21)), that is, $(\mathbf{X}_t^k, \mathbf{s}_t^k) \in \arg\min_{\mathbf{X}_t \in \mathcal{X}_t,\, \mathbf{s}_t \succeq \mathbf{0}} \mathcal{L}_{\mathbf{q}}(\mathbf{X}_t, \mathbf{s}_t; \boldsymbol{\lambda}_t^k)$.

Proof: For problem (16), we have $(\mathbf{X}_t^k, \mathbf{s}_t^k) \in \arg\min_{\mathbf{X}_t, \mathbf{s}_t \succeq \mathbf{0}} \mathcal{L}_{\mathbf{q}^k}(\mathbf{X}_t, \mathbf{s}_t; \boldsymbol{\lambda}_t^k)$, i.e., $(\mathbf{X}_t^k, \mathbf{s}_t^k)$ minimizes the Lagrangian $\mathcal{L}_{\mathbf{q}^k}(\mathbf{X}_t, \mathbf{s}_t)$. Since $\mathcal{L}_{\mathbf{q}^k}(\mathbf{X}_t, \mathbf{s}_t) = f_1(\mathbf{q}^k; \boldsymbol{\lambda}_t^k) + f_2(\mathbf{X}_t, \mathbf{s}_t; \boldsymbol{\lambda}_t^k)$ and $f_1(\mathbf{q}^k; \boldsymbol{\lambda}_t^k)$ is constant, we also have $(\mathbf{X}_t^k, \mathbf{s}_t^k) \in \arg\min_{\mathbf{X}_t, \mathbf{s}_t \succeq \mathbf{0}} f_2(\mathbf{X}_t, \mathbf{s}_t; \boldsymbol{\lambda}_t^k)$. Finally, the proposition holds as $\mathcal{L}_{\mathbf{q}}(\mathbf{X}_t, \mathbf{s}_t; \boldsymbol{\lambda}_t)$ is separable with respect to $\{\mathbf{q}\}$ and $\{\mathbf{X}_t, \mathbf{s}_t\}$ in (21), cf. (18). This completes the proof. ∎

Based on Proposition 2, the relaxed master problem (21) is a mixed-integer linear program and can be solved efficiently by global optimization solvers such as MOSEK [15]. Algorithm 1 converges in a finite number of iterations [12]. The obtained solution is globally optimal for P1. In general, the solution of P1 gives a lower bound for P0. However, by inspecting the rank of $\mathbf{W}_{\boldsymbol{\rho},l,t}$ for the SDP solution, we can further show that the SDP relaxation is tight as detailed in the following proposition.

Proposition 3. Assume that the channel vectors $\mathbf{h}_{\boldsymbol{\rho},t}$, $\boldsymbol{\rho} \in \mathcal{S}$, $t \in \mathcal{T}_0$, can be modeled as statistically independent random vectors. Then P1 and P0 are equivalent in the sense that whenever P0 is feasible, the solution of P1 is also (globally) optimal for P0 with probability one; moreover, the optimal beamformer is given by the principal eigenvector of $\mathbf{W}_{\boldsymbol{\rho},l,t}$.

Proof: Please refer to the Appendix. ∎

C. Suboptimal Caching Scheme

Algorithm 1 has an exponential-time computational complexity in the worst case. For systems with limited computing resources, designing polynomial-time suboptimal schemes is necessary to leverage a better trade-off between system performance and computational complexity. Based on Proposition 3, P0 can be solved via its equivalent convex MINLP, P1. As is also evident from (16), if $\mathbf{q}$ is fixed, P1 reduces to an SDP and can be solved optimally in polynomial time. Therefore, by additionally adjusting $\mathbf{q}$ via greedy heuristics, we obtain the low-complexity suboptimal scheme in Algorithm 2.

Algorithm 2 Suboptimal iterative algorithm for solving P0/P1

1: **Initialization**: Given $\mathbf{q}_m^1 \leftarrow \mathbf{0}$, $\forall m \in \mathcal{M}$; $k \leftarrow 1$;
2: **while** $\mathcal{I}^k \neq \emptyset$ **do**
3: **for each** $i \in \mathcal{I}^k$ **do**
4: Solve primal problem (16) for each $\mathbf{q}_i$ in $\mathcal{Q}_i^k \cap \mathcal{Q}$;
5: Determine $\mathbf{q}_i^{k+1}$ via (23);
6: **end for**
7: $k \leftarrow k + 1$.
8: **end while**

Let $\mathcal{F}_{\mathcal{S}}$ and $\mathbf{q}_m$ be the set of files requested by $\mathcal{S}$ and the caching vector at BS m, respectively. We define

$$\mathcal{Q}_m^k \triangleq \Big\{\mathbf{q}_m \in \{0,1\}^{|\mathcal{F}_{\mathcal{S}}| \times L} \mid \big\|\mathbf{q}_m - \mathbf{q}_m^k\big\|_2^2 \leq 1\Big\} \tag{22}$$

as the set of binary vectors within a distance of one from $\mathbf{q}_m^k$. Besides, $\mathcal{I}^k \triangleq \big\{m \in \mathcal{M} \mid \big|\mathcal{Q}_m^k \cap \mathcal{Q}\big| > 1\big\}$ defines the set of BS indices where $\mathcal{Q}_m^k$ and $\mathcal{Q}$ have non-unique intersection points. During iteration k, the vector in set $\mathcal{Q}_i^k \cap \mathcal{Q}$ that minimizes the objective value of primal problem (16) is chosen as the new caching vector at BS $i \in \mathcal{I}^k$, i.e.,

$$\mathbf{q}_i^{k+1} \in \underset{\mathbf{q}_i \in \mathcal{Q}_i^k \cap \mathcal{Q}}{\arg\min}\, \nu\left(\mathbf{q}_1, \ldots, \mathbf{q}_M\right), \tag{23}$$

is selected to successively reduce the objective value. The iteration continues until $\mathcal{Q}_i^k \cap \mathcal{Q}$ becomes unique, i.e., no further reduction in the objective value is possible, which yields the solution. Hence, the number of problem instances of (16) to be solved is bounded by $ML^2\left|\mathcal{F}_{\mathcal{S}}\right|^2$. Therefore, Algorithm 2 has a polynomial-time computational complexity.

D. Delivery Optimization

Assume that the cache status $\mathbf{q}$ for $\mathcal{T}_0$ is determined, e.g., by solving problem P0 at the end of the time period prior to $\mathcal{T}_0$. The cooperative transmission policy $\{\mathbf{w}_t, \mathbf{V}_t\}$ for time $t \in \mathcal{T}_0$ is then optimized online by solving problem

$$\begin{aligned} \min_{\mathbf{w}_t, \mathbf{V}_t} \quad & U_{\text{TP}}\left(\mathbf{q}, \mathbf{w}_t, \mathbf{V}_t\right) \\ \text{s.t.} \quad & \text{C3, C4, C5, C6, C7,} \end{aligned} \tag{24}$$

where constraints C3–C7 are defined for time slot t and the availability of instantaneous CSI is assumed. Problem (24) is non-convex due to constraints C6 and C7. However, Propositions 1 and 3 imply that problem (24) can be optimally solved in the same manner as the SDP subproblem in (16).

IV. Simulation Results

In this section, we evaluate the performance of the proposed optimal and suboptimal schemes. Consider a cell of radius $R_1 = 1$ km, where the macro BS is located at the center of the cell. Besides, three small cell BSs are uniformly distributed on a circle of radius $R_2 = 0.5R_1$ centered at the macro BS. The macro BS is equipped with $N_0 = 4$ antennas while each small cell BS has $N_m = 2$ antennas. We assume that $F = 10$ video files, each of duration 45 minutes and size 500 MB (Bytes), are delivered to $K = 6$ single-antenna UEs. We adopt an SVC codec with $L = 2$. That is, each video file is encoded into a base-layer subfile and one enhancement-layer subfile, each of size $V_{f,l} = 250$ MB. The minimum streaming rate and the secrecy rate threshold for the base-layer subfiles are $R_{\boldsymbol{\rho},0}^{\text{req}} = 825$ kbps and $R_{\boldsymbol{\rho},0}^{\text{tol}} = 0.1R_{\boldsymbol{\rho},0}^{\text{req}} = 82.5$ kbps, respectively. Therefore, if P0/P1 is feasible, a secrecy streaming rate of $R_{\boldsymbol{\rho},0,t}^{\text{sec}} = 742.5$ kbps can be guaranteed for secure and uninterrupted video streaming for each user as $R_{\boldsymbol{\rho},0,t}^{\text{sec}} \geq$

TABLE I
SIMULATION PARAMETERS.

Parameters	Settings
System bandwidth	5 MHz
Duration of a time slot	10 ms
Duration of delivery period	45 min
Macro BS transmit power	$P_0^{\max}$= 48 dBm
Small BS transmit power	$P_m^{\max}$= 42 dBm
Noise power density	-172.6 dBm/Hz
Cache capacity at macro BS	$C_0^{\max}$= 500 MB

$250\times 8\times 10^6/(45\times 60)=741$ kbps. The streaming rate of the enhancement-layer subfiles is $R^{\rm req}_{\boldsymbol{\rho},1}=2R^{\rm sec}_{\boldsymbol{\rho},0,t}=1.5$ Mbps. The users are randomly distributed in the system. Each user requests the files independent of the other users. Let $\boldsymbol{\theta}=[\theta_1,\ldots,\theta_F]$ be the probability distribution of the requests for different files. We set $\boldsymbol{\theta}$ according to a Zipf distribution with parameter $\gamma=1.1$. In particular, assuming that file $f\in\mathcal{F}$ is the σ_fth most popular file for UEs, the probability of file $f\in\mathcal{F}$ being requested is given by $\theta_f=\frac{1}{\sigma_f^\gamma}/\sum_{f\in\mathcal{F}}\frac{1}{\sigma_f^\gamma}$ [16]. Moreover, the 3GPP path loss model for the "Urban Macro NLOS" scenario is adopted [17]. The other relevant system parameters are given in Table I.

For comparison, we consider two heuristic caching schemes and one non-cooperative delivery scheme as baselines:

- Baseline 1 (Random caching): The video (sub)files are randomly cached until the cache capacity is reached.
- Baseline 2 (Preference based caching): The most popular (sub)files are cached. In trusted BSs, the base-layer subfiles are cached with higher priority than the enhancement layer subfiles of the same video file. For Baselines 1 and 2, the optimal delivery decisions are obtained by solving (24).
- Baseline 3 (No cooperation with untrusted BSs): No video files are cached at untrusted BSs, which act as pure eavesdroppers. The optimal caching and delivery decisions are obtained from problems P0 and (24), respectively, with $C_m^{\max}=0$, $\forall m\in\mathcal{M}_{\mathcal{U}}$.

Fig. 2 show the total BS transmit power of the considered caching and delivery schemes as functions of the cache capacity for one untrusted BS, i.e., $M_u\triangleq|\mathcal{M}_{\mathcal{U}}|=1$. In particular, the system performance is evaluated during the online delivery of video files. As can be observed from Fig. 2, a larger cache capacity leads to a lower total BS transmit power as larger virtual transmit antenna arrays can be formed among the trusted and untrusted BSs for cooperative transmission (beamforming and AN jamming) of the base-layer and enhancement-layer subfiles, respectively. For example, the average numbers of cooperating BSs for optimally transmitting the base-layer and enhancement-layer subfiles are 0.5 and 0.7 for $C_m^{\max}=300$ MB, $m\in\mathcal{M}\backslash\{0\}$, respectively. These numbers increase to 1.6 and 2.5 for $C_m^{\max}=1200$ MB, $m\in\mathcal{M}\backslash\{0\}$, respectively, which yields a transmit power reduction of about 6 dB. Note also that there exists a non-negligible performance gap between the optimal scheme and Baseline 3, particularly in the high cache capacity regime. This is because, with the proposed scheme, the cache resources of the untrusted helpers can be exploited for performance improvement. The performance gap between the optimal scheme and Baselines 1 and 2 is negligible for small (large) cache capacities due to insufficient (saturated) BS cooperation. For medium cache capacities, however, the proposed scheme achieves considerable performance gains due to its ability to exploit information about user requests and CSI for resource allocation. Note also that the suboptimal scheme attains good performance in all regimes despite its low computational cost.

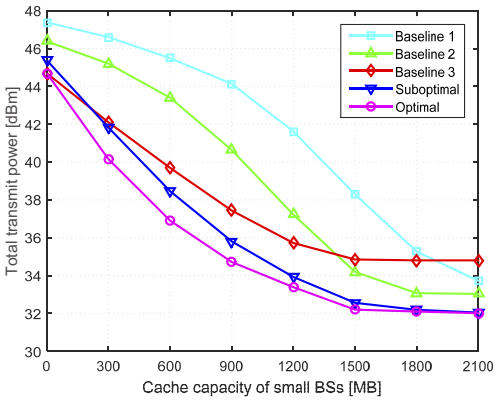

Fig. 2. Total BS transmit power versus cache capacity for different caching and delivery schemes.

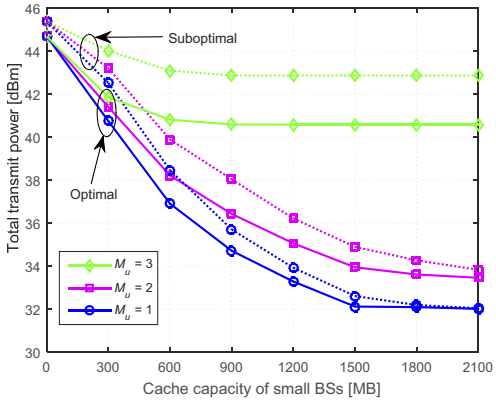

Fig. 3. Total BS transmit power versus cache capacity for different number of untrusted BSs.

Fig. 3 shows the total BS transmit power of the proposed schemes for different numbers of untrusted cache helpers, M_u. We observe that the optimal transmit power increases with M_u because, as M_u increases, fewer (trusted) BSs are available for cooperative transmission of the base-layer subfiles and, at the same time, the trusted BSs have to transmit a larger amount of AN to combat the increasing number of eavesdroppers. On the other hand, the base layers are not cached at the untrusted BSs; that is, for a larger M_u, more cache capacity can be utilized to transmit the enhancement-layer subfiles. Hence, the transmit power of the proposed schemes is only enlarged moderately when M_u increases from 1 to 2. However, when M_u increases from 2 to 3, the transmit power is increased significantly, particularly for cache capacities exceeding 600 MB. This is because, for $M_u=3$, the total number of antennas equipped at the untrusted BSs exceeds that at the trusted BSs, and hence, the available degrees of freedom for secure transmission of the base-layer subfiles are limited. To prevent potential leakage of the base-layer subfiles, the system has to allocate a large amount of power for transmitting AN to degrade the reception quality of the untrusted BSs.

In Fig. 4, the secrecy outage probability, defined as $p_{\rm out}\triangleq\Pr\left(\sum_{\boldsymbol{\rho}}R^{\rm sec}_{\boldsymbol{\rho},0,t}<\sum_{\boldsymbol{\rho}}\left[R^{\rm req}_{\boldsymbol{\rho},0}-R^{\rm tol}_{\boldsymbol{\rho},0}\right]^+\right)$, is evaluated for different numbers of untrusted helpers. Here, $p_{\rm out}$ characterizes the likelihood that problem P0/P1 fails to satisfy either the QoS constraint C6 or the secrecy constraint C7. As the cache capacity increases, the secrecy outage probability monotonically decreases for $M_u\leq 2$ due to the cooperative transmission of the base-layer and enhancement-layer subfiles. However, for $M_u=3$, the secrecy outage probability quickly saturates at a high level as the available degrees of freedom for secure trans-

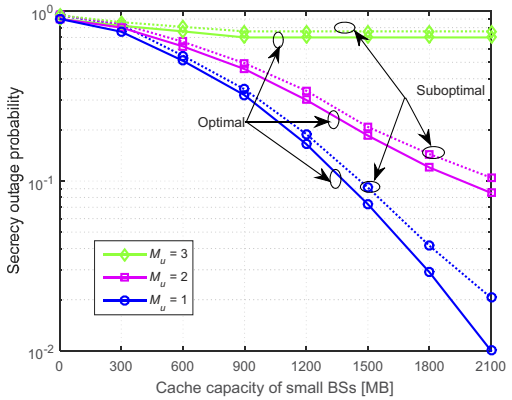

Fig. 4. Secrecy outage probability versus cache capacity for different number of untrusted BSs.

mission of the base-layer subfiles are limited, and consequently, the benefits introduced by the cache of the small cell BSs are limited to the transmission of the enhancement-layer subfiles only.

V. Conclusion

In this paper, secure video streaming was investigated in small cell networks with untrusted small cell BSs which can intercept both cached and delivered video data. In particular, SVC coding and caching were jointly exploited to facilitate secure cooperative MIMO transmission to mitigate the impact of untrusted BSs. Caching and delivery were optimized based on a non-convex mixed-integer problem, which was optimally solved by an iterative algorithm. To reduce the computational complexity, a suboptimal algorithm was also studied. Simulation results revealed that the proposed optimal and suboptimal schemes can significantly enhance both the secrecy and the power efficiency of the video streaming system as long as the total number of antennas at the trusted BSs exceeds that at the untrusted BSs.

Appendix
Proof of Proposition 3

We show here that the solution of the relaxed problem P1 satisfies $\mathsf{rank}(\mathbf{W}_{\boldsymbol{\rho},l,t}) = 1$, with probability one. Let $\alpha_{\boldsymbol{\rho},l,t} \geq 0$, $\mathbf{\Phi}_{\boldsymbol{\rho},l,t} \succeq \mathbf{0}$, and $\mathbf{\Theta}_{\boldsymbol{\rho},l,t} \succeq \mathbf{0}$ be the Lagrange multipliers associated with constraints $\overline{\text{C6}}$, C10, and C12: $\mathbf{W}_{\boldsymbol{\rho},l,t} \succeq \mathbf{0}$, respectively, where C12 is implied by C10. Define $\mathbf{\Upsilon}_{\boldsymbol{\rho},l,t} = [\alpha_{\boldsymbol{\rho},l,t}, \mathbf{\Phi}_{\boldsymbol{\rho},l,t}, \mathbf{\Theta}_{\boldsymbol{\rho},l,t}]$. The Lagrangian of P1 is given by,

$$\mathcal{L}(\mathbf{W}_{\boldsymbol{\rho},l,t};\mathbf{\Upsilon}_{\boldsymbol{\rho},l,t}) = \sum_{\boldsymbol{\rho},l} \mathsf{tr}\left[\left(\mathbf{B}_{\boldsymbol{\rho},l,t} - \mathbf{\Theta}_{\boldsymbol{\rho},l,t} - \tfrac{\alpha_{\boldsymbol{\rho},l,t}}{\eta^{\text{req}}_{\boldsymbol{\rho},l}}\mathbf{H}_{\boldsymbol{\rho},t}\right)\mathbf{W}_{\boldsymbol{\rho},l,t}\right] + \mathbf{\Delta}_2,$$

where $\mathbf{B}_{\boldsymbol{\rho},l,t} \triangleq \mathbf{I} + \mathbf{\Delta}_1 - \mathbf{\Phi}_{\boldsymbol{\rho},l,t}$; and $\mathbf{\Delta}_1 \succeq \mathbf{0}$ and $\mathbf{\Delta}_2$ denote the collection of terms that are relevant and irrelevant to $\mathbf{W}_{\boldsymbol{\rho},l,t}$, respectively. Hence, the dual problem is given by

$$\max_{\mathbf{\Upsilon}_{\boldsymbol{\rho},l,t} \succeq \mathbf{0}} \min_{\mathbf{W}_{\boldsymbol{\rho},l,t}} \mathcal{L}(\mathbf{W}_{\boldsymbol{\rho},l,t};\mathbf{\Upsilon}_{\boldsymbol{\rho},l,t}). \tag{25}$$

If P0/P1 is feasible, then $\mathbf{W}_{\boldsymbol{\rho},l,t} \neq \mathbf{0}$. By exploiting the strong duality of the SDP subproblem (16), the optimal beamformers and the optimal dual solutions satisfy the following Karush-Kuhn-Tucker (KKT) optimality conditions,

$$\mathbf{B}_{\boldsymbol{\rho},l,t} - \tfrac{\alpha_{\boldsymbol{\rho},l,t}}{\eta^{\text{req}}_{\boldsymbol{\rho},l}}\mathbf{H}_{\boldsymbol{\rho},t} = \mathbf{\Theta}_{\boldsymbol{\rho},l,t}, \tag{26}$$

$$\mathbf{\Theta}_{\boldsymbol{\rho},l,t}\mathbf{W}_{\boldsymbol{\rho},l,t} = \mathbf{0}. \tag{27}$$

Next, we show by contradiction that $\mathbf{B}_{\boldsymbol{\rho},l,t} \succ \mathbf{0}$ holds with probability one. Assume that $\mathbf{B}_{\boldsymbol{\rho},l,t}$ has at least one non-positive eigenvalue $\tau \leq 0$ and the corresponding eigenvector is $\widetilde{\mathbf{w}}_{\boldsymbol{\rho},l,t}$, i.e., $(\mathbf{B}_{\boldsymbol{\rho},l,t} - \tau\mathbf{I})\widetilde{\mathbf{w}}_{\boldsymbol{\rho},l,t} = \mathbf{0}$. Let $\mathbf{W}_{\boldsymbol{\rho},l,t} = \beta\widetilde{\mathbf{w}}_{\boldsymbol{\rho},l,t}\widetilde{\mathbf{w}}^H_{\boldsymbol{\rho},l,t} \succeq \mathbf{0}$, where $\beta > 0$. By substituting $\mathbf{W}_{\boldsymbol{\rho},l,t}$ into (25), we further have

$$\mathcal{L}(\mathbf{W}_{\boldsymbol{\rho},l,t};\mathbf{\Upsilon}_{\boldsymbol{\rho},l,t}) = \underbrace{\beta\tau\sum\nolimits_{\boldsymbol{\rho},l}\widetilde{\mathbf{w}}^H_{\boldsymbol{\rho},l,t}\widetilde{\mathbf{w}}_{\boldsymbol{\rho},l,t}}_{\leq 0} - \beta\sum\nolimits_{\boldsymbol{\rho},l}\tfrac{\alpha_{\boldsymbol{\rho},l,t}}{\eta^{\text{req}}_{\boldsymbol{\rho},l}}\widetilde{\mathbf{w}}_{\boldsymbol{\rho},l,t}\mathbf{H}_{\boldsymbol{\rho},t}\widetilde{\mathbf{w}}^H_{\boldsymbol{\rho},l,t} + \mathbf{\Delta}_2. \tag{28}$$

Since P1 is feasible, constraint $\overline{\text{C6}}$ has to be satisfied with equality for the optimal solution. Consequently, $\alpha_{\boldsymbol{\rho},l,t} > 0$ and $\widetilde{\mathbf{w}}_{\boldsymbol{\rho},l,t}\mathbf{H}_{\boldsymbol{\rho}}\widetilde{\mathbf{w}}^H_{\boldsymbol{\rho},l,t} > 0$. We can further show that the minimum of (25) is obtained at $\beta \to \infty$, since $\mathbf{h}_{\boldsymbol{\rho},t}$ is statistically independent and $-\frac{\alpha_{\boldsymbol{\rho},l,t}}{\eta^{\text{req}}_{\boldsymbol{\rho},l}}\widetilde{\mathbf{w}}_{\boldsymbol{\rho},l,t}\mathbf{H}_{\boldsymbol{\rho},t}\widetilde{\mathbf{w}}^H_{\boldsymbol{\rho},l,t} \to -\infty$ with probability one. Thus, the dual problem (25) becomes unbounded from below, which implies that the primal problem is infeasible. This is a contradiction and $\mathbf{B}_{\boldsymbol{\rho},l,t} \succ \mathbf{0}$ is thus proved.

Finally, based on (26), (27), and $\mathbf{B}_{\boldsymbol{\rho},l,t} \succ \mathbf{0}$, we have,

$$\mathsf{rank}(\mathbf{W}_{\boldsymbol{\rho},l,t}) \overset{\text{(a)}}{=} \mathsf{rank}(\mathbf{B}_{\boldsymbol{\rho},l,t}\mathbf{W}_{\boldsymbol{\rho},l,t}) \overset{\text{(b)}}{=} \mathsf{rank}(\tfrac{\alpha_{\boldsymbol{\rho},l,t}}{\eta^{\text{req}}_{\boldsymbol{\rho},l}}\mathbf{H}_{\boldsymbol{\rho},t}\mathbf{W}_{\boldsymbol{\rho},l,t}) \overset{\text{(c)}}{\leq} \min\left\{\mathsf{rank}\left(\tfrac{\alpha_{\boldsymbol{\rho},l,t}}{\eta^{\text{req}}_{\boldsymbol{\rho},l}}\mathbf{W}_{\boldsymbol{\rho},l,t}\right), \mathsf{rank}(\mathbf{H}_{\boldsymbol{\rho},t})\right\} \leq 1,$$

where (a) is due to $\mathbf{B}_{\boldsymbol{\rho},l,t} \succ \mathbf{0}$, (b) is a result of (26) and (27), and (c) follows from the basic rank inequality $\mathsf{rank}(\mathbf{AB}) \leq \min\{\mathsf{rank}(\mathbf{A}), \mathsf{rank}(\mathbf{B})\}$. On the other hand, since $\mathbf{W}_{\boldsymbol{\rho},l,t} \neq \mathbf{0}$, the condition $\mathsf{rank}(\mathbf{W}_{\boldsymbol{\rho},l,t}) = 1$ holds with probability one. This completes the proof.

References

[1] V. W. S. Wong, R. Schober, D. W. K. Ng, and L.-C. Wang, *Key Technologies for 5G Wireless Systems*. Cambridge University Press, 2017.

[2] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. Leung, "Cache in the air: Exploiting content caching and delivery techniques for 5G systems," *IEEE Commun. Mag.*, vol. 52, pp. 131–139, Feb. 2014.

[3] K. Shanmugam, N. Golrezaei, A. Dimakis, A. Molisch, and G. Caire, "FemtoCaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, pp. 8402–8413, Dec. 2013.

[4] D. Liu and C. Yang, "Will caching at base station improve energy efficiency of downlink transmission?" in *Proc. IEEE GlobalSIP*, Atlanta, GA, Dec. 2014.

[5] L. Xiang, D. W. K. Ng, T. Islam, R. Schober, and V. W. S. Wong, "Cross-layer optimization of fast video delivery in cache-enabled relaying networks," in *Proc. IEEE GLOBECOM*, San Diego, CA, Dec. 2015.

[6] 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, Jun. 2017.

[7] A. Liu and V. Lau, "Cache-enabled opportunistic cooperative MIMO for video streaming in wireless systems," *IEEE Trans. Signal Process.*, vol. 62, pp. 390–402, Jan. 2014.

[8] 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 GLOBECOM*, Washington, DC, Dec. 2016.

[9] F. Gabry, V. Bioglio, and I. Land, "On edging caching with secrecy constraints," in *Proc. ICC*, Kuala Lumpur, Malaysia, May 2016.

[10] J.-R. Ohm, "Advances in scalable video coding," *Proc. IEEE*, vol. 93, pp. 42–56, Jan. 2005.

[11] S. Goel and R. Negi, "Guaranteeing secrecy using artificial noise," *IEEE Trans. Wireless Commun.*, vol. 7, pp. 2180–2189, Jun. 2008.

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

[13] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*, 3rd ed. John Wiley & Sons, 2013.

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

[15] Mosek ApS, "The MOSEK optimization software, version 8.0.0.45," [Online] Available: http://www.mosek.com, Nov. 2016.

[16] 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.

[17] 3GPP TR 36.814, "Further advancements for E-UTRA physical layer aspects (Release 9)," Mar. 2010.