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

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

***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, at the same time, exploit the cache memory of both the trusted and untrusted BSs for improving the system performance. Considering imperfect channel state information (CSI) at the transmitters, we formulate a two-timescale non-convex mixed-integer robust optimization problem to minimize the total transmit power required for guaranteeing the quality of service (QoS) and secrecy during video streaming. We develop an iterative algorithm based on a modified generalized Benders decomposition (GBD) to solve the problem optimally, where the caching and the cooperative transmission policies are determined via offline (long-timescale) and online (short-timescale) optimization, respectively. Furthermore, inspired by the optimal algorithm, a low-complexity suboptimal algorithm based on a greedy heuristic is proposed. Simulation results show that the proposed schemes achieve significant gains in power efficiency and secrecy performance compared to several baseline schemes.**

***Index Terms*—Physical layer security, untrusted nodes, wireless caching, MIMO, non-convex optimization, resource allocation.**

I. Introduction

SMALL cells are among the most promising solutions for meeting the enormous capacity requirements introduced by video streaming applications in the fifth-generation (5G) wireless networks [2]. By densely deploying low-power base stations (BSs), both the spectral and energy efficiencies of wireless communication systems can be improved significantly. However, to achieve these performance gains, high-capacity secure backhaul links are required to transport the video files from the Internet to the small cell BSs. While wireless backhauling is usually preferred for small cells due to its low cost and high flexibility in deployment [3], the capacity provided by wireless backhauling is often insufficient, which deteriorates the overall system performance [4] and limits the maximum number of concurrent streaming users/connections. Moreover, since wireless transmission is susceptible to eavesdropping, the security of wireless backhauling is a fundamental concern for 5G wireless networks.

Recently, wireless caching has been proposed as a viable solution to enhance the capacity of small cell backhauling [5]–[15]. Built upon the content-centric networking paradigm, in wireless caching, the most popular contents are pre-stored at the access points or BSs in close proximity of the user equipments (UEs). Consequently, the backhaul traffic is offloaded by reusing the cached content [5], [6]. Caching as an alternative to small cell backhauling was first investigated in [7], where caching was shown to also substantially reduce the average downloading delay. Besides, caching improves the energy efficiency of wireless backhauling systems as was shown in [8]. In [9], caching was optimized to facilitate power-efficient cooperative multiple-input multiple-output (MIMO) transmission in small cell networks. In [10], joint caching and buffering for small cell networks was proposed to overcome the backhaul capacity bottleneck and the half-duplex transmission constraint simultaneously to enable fast downloading of video files. In [11], [12], coded caching was introduced, which reduces the backhaul load by exploiting coded multicast transmission for simultaneous delivery of different files. Coded caching was extended to various network scenarios, see [13]–[15] and references therein.

On the other hand, although communication secrecy is of high importance in wireless networks, providing security for networks employing wireless caching has been a major challenge. This is because current video streaming applications, e.g. YouTube and Netflix, mainly rely on end-to-end encryption schemes such as the hypertext transfer protocol secure (HTTPS) [23] to ensure communication security. However, with such schemes, the benefits of content-centric networking vanish as encrypted contents are uniquely defined for each user request and cannot be reused to serve other user requests [5]. For this reason, caching was mainly considered for content without security restrictions in the literature [5]–[10]. To overcome this limitation, physical layer security (PLS) schemes for wireless caching were proposed in [16]–[18]. As PLS techniques rely on wiretap channel coding instead of source encryption, content can still be reused at the wireless edge for secure wireless transmission, and hence, caching and PLS are compatible. In [16], cache-enabled cooperative MIMO

The work of D. W. K. Ng was supported under Australian Research Council's Discovery Early Career Researcher Award funding scheme (DE170100137). The work of R. Schober was supported by the Alexander von Humboldt Professorship Program. The work of V.W.S. Wong was supported by the Natural Sciences and Engineering Research Council of Canada. Part of this work has been accepted for presentation at the *IEEE Global Commun. Conf. (Globecom)*, Singapore, Dec. 2017 [1].

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).

transmission was shown to be an effective physical layer mechanism for increasing the secrecy rate for video delivery in homogeneous cellular networks. However, a secure backhaul for cache placement was required in [16], which cannot always be guaranteed with wireless backhauling in practice. Considering an insecure backhaul, a secure cache placement strategy for heterogeneous cellular networks (HetNets) was developed in [17], whereby eavesdroppers tapping the insecure backhaul can be prevented from obtaining a sufficient number of coded packets for successful recovery of the video file. Assuming caching at the end users, the authors of [18] proposed a secure coded multicast scheme for relay networks to prevent end users and external eavesdroppers from intercepting the non-requested and the delivered files, respectively.

However, [16]–[18] optimistically assumed that the (cache) helpers can be *trusted* for cooperation and that the cache cannot be exploited for eavesdropping purposes. These assumptions may be unrealistic for HetNets. In particular, due to the distributed network architecture, cache-enabled small cell BSs can be *untrusted* helpers, i.e., they may be potential eavesdroppers[1], and hence, may not cooperate altruistically [20]–[22]. Examples of untrusted helpers include home-owned and open-access small cell BSs which can be easily manipulated by third parties to eavesdrop premium video streaming services, for which they have not paid, and/or users' private video files. In contrast to trusted small cell BSs deployed and owned by the service provider, at these untrusted small cell BSs, user data is left unprotected and prone to eavesdropping because the small cell BS itself is responsible for encrypting and decrypting the user data before forwarding it to the macro BS and the intended users, respectively [20]–[22]. Moreover, different from the case of cache-disabled eavesdroppers[2] considered in [16], [17], the cache memory equipped at the untrusted helpers unfavorably enhances their eavesdropping capability as they can intercept both the cached and the delivered video data, and utilize the cached video data as side information to improve reception.

Two fundamental questions need to be addressed when cache helpers are untrusted: (a) *Can cooperation with untrusted helpers still yield secrecy benefits? That is, can the cache deployed at the untrusted helpers be utilized to improve the system performance?* If so, (b) *how to cache and cooperate intelligently to reap the possible performance gains?* To our knowledge, state-of-the-art small cell networks perform only passive authentication of BSs and completely exclude untrusted BSs from participating in cooperative transmission [21], [22]. However, in this case, untrusted BS cannot be used to improve the system performance[3].

In [25], untrusted helpers without caching have been investigated for relay networks. It was shown that cooperation with compress-and-forward relays yields a positive secrecy rate even if the relays are untrusted. However, the problem studied in this paper is more challenging as the untrusted helpers can cache content to enhance their eavesdropping capability. Thus, the solutions proposed in [25] are not applicable and a new study is needed. In this paper, to facilitate secure cooperative transmission with untrusted cache helpers, we propose an advanced caching scheme that combines scalable video coding (SVC) 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 which is 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 is enabled by exploiting the encoding and decoding structure of SVC. Thereby, the large virtual transmit antenna array formed by all BSs that have cached the same subfile introduces additional degrees of freedom (DoFs) which may be utilized for secure and power-efficient video streaming.

To reap the cache-enabled secrecy benefits, a centralized framework for caching and delivery optimization is adopted. Hence, the proposed architecture follows the cloud radio access network (CRAN) philosophy [4], [26] which has been advocated for next-generation HetNets for cooperative MIMO transmission [27], [28] and advanced resource allocation [29]. In the conference version of this paper [1], we investigated cache-enabled secure transmission by assuming perfect knowledge of all channels. However, in practice, the channel state information (CSI) gathered at the central controller, e.g. the macro BS, is imperfect due to quantization noise and feedback delay, which deteriorates the system performance and has to be taken into account for system design. To mitigate the information leakage from the trusted BSs to the untrusted BSs, artificial noise (AN) based jamming is applied in this paper. In the literature [29]–[31], AN has been employed to effectively reduce the receive signal-to-interference-plus-noise ratio (SINR) at the eavesdropper without interfering the desired users. In this paper, we consider cooperative AN transmission by the trusted BSs for power-efficient jamming to combat the eavesdropping of the untrusted BSs. By considering untrusted BSs and imperfect CSI, we jointly optimize caching and cooperative data and AN transmission for a secure and power-efficient system design. In particular, a two-timescale robust optimization problem is formulated for minimization of the transmit power required for secure video streaming under imperfect CSI. The main contributions of this paper can be summarized 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 small cell BSs, we propose a secure caching scheme based on SVC.
- We optimize the caching and the cooperative delivery policies for minimization of the transmit power while

[1]In this paper, we only consider passive eavesdroppers which remain silent during eavesdropping. Studying the case of active eavesdroppers such as jamming and spoofing attackers [19] is an interesting topic for future work.

[2]The case considered in this paper is more general than that in [16], [17]. In fact, the eavesdroppers in [16], [17] can be viewed as untrusted helpers with zero cache capacity.

[3]We note that, as untrusted BSs present the man-in-the-middle threat to wireless networks, HTTPS also cannot facilitate secure cooperative transmission [24].

guaranteeing quality-of-service (QoS) and communication secrecy under imperfect CSI. We show that the optimal delivery decisions can be obtained by semidefinite programming (SDP) relaxation with probability one under mild conditions. For the optimal caching decisions, an optimal iterative algorithm is developed based on a modified generalized Benders decomposition (GBD). To reduce the computational complexity, a polynomial-time suboptimal greedy scheme is also proposed.

- Our simulation results show that the proposed robust schemes can efficiently exploit the cache capacities of both trusted and untrusted small cell BSs to enable power-efficient and secure video streaming in heterogeneous small cell networks.

The remainder of this paper is organized as follows. In Section II, we present the system model for cooperative secure video streaming in the presence of untrusted cache helpers. The formulation and solution of the proposed optimization problem are provided in Sections III and IV, respectively. In Section V, we evaluate the performance of the proposed schemes and compare it to that of several baseline schemes. Finally, Section VI concludes the paper.

Notations: $\mathbb{R}$ and $\mathbb{C}$ denote the sets of real and complex numbers, respectively; $\Re\{z\}$ denotes the real part of $z \in \mathbb{C}$; $\mathbf{I}_L$ 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; $\|\cdot\|_F$, $\mathrm{tr}(\cdot)$, $\mathrm{rank}(\cdot)$, $\det(\cdot)$, and $\lambda_{\max}(\cdot)$ denote the Frobenius norm, the trace, the rank, the determinant, and the maximal eigenvalue of a square matrix, respectively; $\mathbb{E}(\cdot)$ 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"; $\mathrm{diag}(\mathbf{x})$ is a diagonal matrix with the diagonal elements given by vector $\mathbf{x}$; $|\mathcal{X}|$ 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

A. Network Topology

We consider downlink 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 distributed in the coverage area of a macro BS, see Fig. 1(a). For convenience, a list with key notations used in this paper is provided in Table I. 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 video server on the Internet via a dedicated secure high-capacity backhaul link such as optical fiber. For simplicity of notation, the backhaul to the macro BS is modeled as a cache with an equivalent capacity of $C_0^{\max}$ bits. In contrast, the small cell BSs are connected to the macro BS via wireless backhaul links for convenience of deployment. Assume that BS_m, $m \in \mathcal{M}$, is equipped with N_m antennas. The total number of transmit antennas is denoted by $N \triangleq \sum_{m\in\mathcal{M}} N_m$.

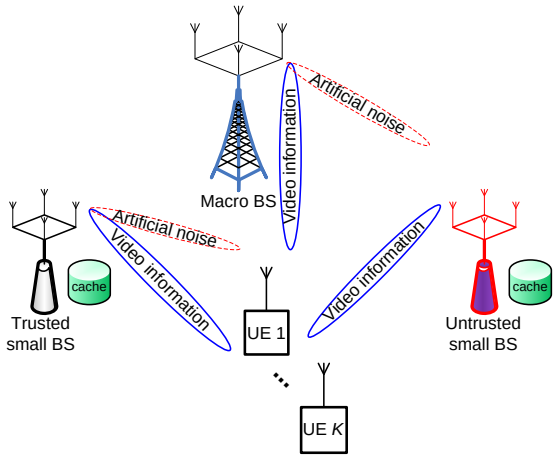
The video server owns a library of F video files, indexed by $\mathcal{F} \triangleq \{1, \ldots, F\}$, to be streamed to K single-antenna UEs, indexed by $\mathcal{K} \triangleq \{1, \ldots, K\}$. The size of file f is V_f bits. Employing SVC coding, as utilized e.g. for wireless video delivery in the H.264/Moving Picture Expert Group (MPEG)-4 standard [32]–[34], each video file $f \in \mathcal{F}$ is encoded into one base-layer subfile, $(f, 0)$, and $L-1$ enhancement-layer subfiles, (f, l), $l \in \{1, \ldots, L-1\}$, where the information embedded in enhancement layer l is used to refine the information contained in the previous layers $0, \ldots, l-1$. Let $\mathcal{L} \triangleq \{0, \ldots, L-1\}$ be the index set of all layers. The size of subfile (f, l) is $V_{f,l}$ bits. The base layer can be decoded independent of the enhancement layers. In contrast, enhancement layer $l \in \mathcal{L}\setminus\{0\}$ can be decoded only after layers $0, \ldots, l-1$ have already been decoded. Therefore, the layers have to be decoded in a sequential manner [32]. Due to this specific encoding and decoding structure, only the base layer has to be protected in order to ensure communication secrecy. An eavesdropper, who cannot decode the base layer, will also not be able to decode any of the enhancement layers.

The small cell BSs serve as helpers of the macro BS in delivering the video files. However, a subset of the small cell BSs are untrusted. These BSs may leak the cached video data and eavesdrop the transmitted video data while 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 $J \leq M$ and $N_\mathcal{T} + N_\mathcal{U} = N$. In this paper, we assume that the set of untrusted BSs, $\mathcal{M}_\mathcal{U}$, is known. In practice, untrusted BSs may largely be home-owned and open-access small cell BSs which have insufficient security protection and can easily be compromised by third parties. Due to the eavesdropping and intensive processing, untrusted BSs may consume a large power and/or experience a long end-to-end latency even if the uplink and downlink throughputs are small. Hence, the power/delay versus throughput pattern of untrusted BSs is statistically different from that of trusted operator-owned BSs such that they constitute outliers. Therefore, by exploiting the power, delay, and throughput records of all BSs collected by the service providers, the set of untrusted BSs can be estimated by applying state-of-the-art outlier detection methods, e.g., supervised and unsupervised learning techniques [35]–[37].

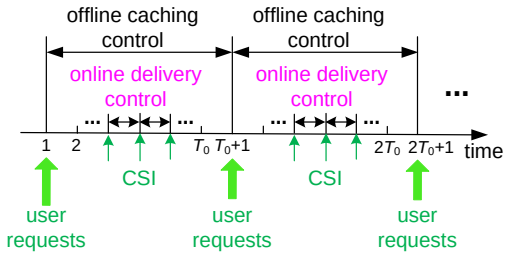
The considered system is time slotted and the duration of a time slot is smaller than the channel coherence time. We consider a two-timescale control for caching and delivery. As shown in Fig. 1(b), the video files in the cache are updated every T_0 time slots, referred as one period, based on the historical profiles of user preferences and CSI. In contrast, the video delivery decisions are determined in each time slot based on the actual requests of the users and instantaneous CSI. We have $T_0 \gg 1$, as the users' preferences vary on a much slower scale (e.g., from day to day) than the user requests. 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 $t \in \mathcal{T}_0$.

TABLE I
LIST OF KEY NOTATIONS.

$\mathcal{M}$, $\mathcal{M}_{\mathcal{T}}$, $\mathcal{M}_{\mathcal{U}}$	Sets of M BSs, J trusted BSs, and $M-J$ untrusted BSs
N, $N_{\mathcal{T}}$, $N_{\mathcal{U}}$	Total number of antennas at all BSs, trusted BSs, and untrusted BSs
$\mathcal{M}_{f,l}^{\text{Coop}}$	Subset of cooperating BSs for delivery of subfile (f,l)
$\mathcal{K}$, $\mathcal{F}$, $\mathcal{L}$, $\mathcal{T}_0$	Sets of K users, F video files, L layer subfiles per file, and T_0 time slots
$\boldsymbol{\rho} \triangleq (k,f)$, $\mathcal{S}$	Request of user k for file f and set of user requests
$\kappa(\boldsymbol{\rho})$, $f(\boldsymbol{\rho})$	Requesting UE and requested file corresponding to $\boldsymbol{\rho}$
$q_{f,l,m} \in \{0,1\}$	Binary cooperative delivery decisions for subfile (f,l) at BS m
$s_{\boldsymbol{\rho},l,t}$	Source symbol of subfile (f,l) for serving request $\boldsymbol{\rho}$ at time t
$\mathbf{w}_{\boldsymbol{\rho},l,m,t}$, $\mathbf{w}_{\boldsymbol{\rho},l,t}$	Beamforming vectors of BS m and BS set $\mathcal{M}$ for sending symbol $s_{\boldsymbol{\rho},l,t}$
$\mathbf{v}_t$, $\mathbf{V}_t$	AN and its covariance matrix at time t
$C_m^{\max}$	Cache size at BS m
$R_{\boldsymbol{\rho},l,t}$, $R_{\boldsymbol{\rho},l,t}^{\text{sec}}$	Achievable rate and achievable secrecy rate at user $\kappa(\boldsymbol{\rho})$ for decoding $s_{\boldsymbol{\rho},l,t}$
$R_{j,\boldsymbol{\rho},l,t}$	Capacity of untrusted BS j for eavesdropping symbol $s_{\boldsymbol{\rho},l,t}$



(a)



(b)

Fig. 1. (a) System model for secure video delivery in a heterogeneous network, where a trusted and an untrusted small cell BS are distributed in the coverage area of a macro BS; (b) caching and delivery control are performed in two timescales.

B. Secure Video Caching and Delivery

As the cache helpers in set $\mathcal{M}_{\mathcal{U}}$ are untrusted, only the enhancement layers are cached at BSs $m \in \mathcal{M}_{\mathcal{U}}$. Hence, the cached subfiles cannot be used by the untrusted BSs to reconstruct the original video files as long as they do not have access to the base-layer subfiles. Meanwhile, BSs that 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. The cache placement has to satisfy the condition

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

and the capacity constraint

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

During data delivery, the set of BSs cooperating for the 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 UE 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 $\mathcal{L}_{\boldsymbol{\rho}} \triangleq \{0,1,\ldots,L_{\boldsymbol{\rho}}-1\}$.

We assume a frequency flat fading channel for video data transmission. As the worst case, we assume that the untrusted BSs are full-duplex, i.e., they can simultaneously eavesdrop the video information intended for the UEs and participate in the cooperative delivery of the cached files. In time slot $t \in \mathcal{T}_0$, the self-interference [38], [39] at BS $j \in \mathcal{M}_{\mathcal{U}}$ caused by simultaneous reception and transmission at the same frequency is denoted by $\mathbf{c}_{j,t}$. Let $\mathbf{x}_t \in \mathbb{C}^{N\times 1}$ denote the joint transmit signal of BS set $\mathcal{M}$. 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

$$y_{\boldsymbol{\rho},t} = \mathbf{h}_{\boldsymbol{\rho},t}^H \mathbf{x}_t + z_{\boldsymbol{\rho},t} \text{ and } \mathbf{y}_{\mathcal{U},j,t} = \mathbf{G}_{j,t}^H \mathbf{x}_t + \mathbf{c}_{j,t} + \mathbf{z}_{j,t}, \tag{3}$$

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 $\mathbf{G}_{j,t} = [\mathbf{G}_{j,0,t}^H, \ldots, \mathbf{G}_{j,j-1,t}^H, \mathbf{0}_{N_j\times N_j}^H, \mathbf{G}_{j,j+1,t}^H, \ldots, \mathbf{G}_{j,M,t}^H]^H \in \mathbb{C}^{N\times N_j}$ 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. The term $\mathbf{0}_{N_j\times N_j}^H$ in the definition of $\mathbf{G}_{j,t}$ accounts for the fact that the self-interference at BS j is included in $\mathbf{c}_{j,t}$. Furthermore, $z_{\boldsymbol{\rho},t} \sim \mathcal{CN}(0,\sigma^2)$ and $\mathbf{z}_{j,t} \sim \mathcal{CN}(\mathbf{0}, \sigma_j^2 \mathbf{I}_{N_j})$ are the zero-mean complex Gaussian noises at the users and the BSs with variance σ^2 and covariance matrix $\sigma_j^2\mathbf{I}_{N_j}$, respectively.

The source symbols of subfile (f,l) for serving request $\boldsymbol{\rho}$ in time slot t, 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 $\mathbf{w}_{\boldsymbol{\rho},l,t} \triangleq [\mathbf{w}^H_{\boldsymbol{\rho},l,0,t},\ldots,\mathbf{w}^H_{\boldsymbol{\rho},l,M,t}]^H \in \mathbb{C}^{N\times 1}$ denote the joint beamforming vector for transmit 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

$$\mathbf{x}_t = \sum_{\boldsymbol{\rho}\in\mathcal{S}} \sum_{l\in\mathcal{L}_{\boldsymbol{\rho}}} \mathbf{w}_{\boldsymbol{\rho},l,t} s_{\boldsymbol{\rho},l,t} + \mathbf{v}_t, \tag{4}$$

where superposition coding is used to superimpose the $L_{\boldsymbol{\rho}}$ layers intended for user $\kappa(\boldsymbol{\rho})$ [40]. Herein, complex Gaussian distributed AN, $\mathbf{v}_t \in \mathbb{C}^{N\times 1}$, is sent cooperatively by the trusted BSs in set $\mathcal{M}_{\mathcal{T}}$ to proactively interfere the reception of the untrusted BSs in set $\mathcal{M}_{\mathcal{U}}$ [30]. We assume $\mathbf{v}_t \sim \mathcal{CN}(\mathbf{0}, \mathbf{V}_t)$, where $\mathbf{V}_t \in \mathbb{C}^{N\times N}$ is the covariance matrix of the AN, i.e., $\mathbf{V}_t \triangleq \mathbb{E}[\mathbf{v}_t\mathbf{v}_t^H] \succeq \mathbf{0}$. As $\mathbf{v}_t$ is cooperatively injected only by the trusted BS set $\mathcal{M}_{\mathcal{T}}$, we require $\mathbf{\Lambda}_{\mathcal{U}}\mathbf{V}_t = \mathbf{0}$, where $\mathbf{\Lambda}_{\mathcal{U}}$ is an $N \times N$ diagonal matrix given by $\mathbf{\Lambda}_{\mathcal{U}} = \mathrm{diag}\big(\mathbf{0}^T_{N_{\mathcal{T}}\times 1}, \mathbf{1}^T_{N_{\mathcal{U}}\times 1}\big)$, to ensure that the components of $\mathbf{V}_t$ which correspond to untrusted BSs are equal to zero. Moreover, for BS $m \in \mathcal{M}$, participating in cooperative transmission of $s_{\boldsymbol{\rho},l,t}$ is possible only if the requested subfile is cached at the BS, i.e., we require

$$\text{C3: } \mathrm{tr}\left(\mathbf{\Lambda}_m \mathbf{w}_{\boldsymbol{\rho},l,t}\mathbf{w}^H_{\boldsymbol{\rho},l,t}\right) \le q_{f(\boldsymbol{\rho}),l,m} P^{\max}_m, \quad m \in \mathcal{M},\, \boldsymbol{\rho} \in \mathcal{S},\, l \in \mathcal{L}_{\boldsymbol{\rho}},\, t \in \mathcal{T}_0, \tag{5}$$

where $P^{\max}_m$ is the maximum transmit power at BS m, and $\mathbf{\Lambda}_m$ is an $N \times N$ diagonal matrix given by $\mathbf{\Lambda}_m = \mathrm{diag}\Big(\mathbf{0}^T_{(\sum_{j=0}^{m-1} N_j)\times 1}, \mathbf{1}^T_{N_m\times 1}, \mathbf{0}^T_{(\sum_{j=m+1}^{M} N_j)\times 1}\Big)$ such that $\mathrm{tr}\big(\mathbf{\Lambda}_m \mathbf{w}_{\boldsymbol{\rho},l,t}\mathbf{w}^H_{\boldsymbol{\rho},l,t}\big) \equiv \mathrm{tr}\left(\mathbf{w}_{\boldsymbol{\rho},l,m,t}\mathbf{w}^H_{\boldsymbol{\rho},l,m,t}\right)$ holds. C3 enforces $\mathbf{\Lambda}_m \mathbf{w}_{\boldsymbol{\rho},l,t}\mathbf{w}^H_{\boldsymbol{\rho},l,t} = \mathbf{0}$, i.e., $\mathbf{w}_{\boldsymbol{\rho},l,m,t} = \mathbf{0}$, when $q_{f,l,m} = 0$, i.e., $m \notin \mathcal{M}^{\mathrm{Coop}}_{f,l}$. Otherwise, when $q_{f,l,m} = 1$, i.e., $m \in \mathcal{M}^{\mathrm{Coop}}_{f,l}$, C3 ensures that the maximum transmit power, $P^{\max}_m$, of BS $m \in \mathcal{M}$ is not exceeded. A constraint of the form of C3 is also referred to as a big-M constraint [41]. 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.

C. Achievable Secrecy Rate

Each user employs successive interference cancellation (SIC) at the receiver [40]. 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 previously decoded lower layers $0,\ldots,l-1$ are first removed from the received signal for interference cancellation. This process continues until layer $L_{\boldsymbol{\rho}} - 1$ is decoded [32]. Define the interference cancellation coefficient $a^{\boldsymbol{\rho}',l'}_{\boldsymbol{\rho},l} \in \{0,1\}$, where $a^{\boldsymbol{\rho}',l'}_{\boldsymbol{\rho},l} = 1$ indicates that the transmission of subfile $(f(\boldsymbol{\rho}'),l')$ interferes that of subfile $(f(\boldsymbol{\rho}),l)$, and $a^{\boldsymbol{\rho}',l'}_{\boldsymbol{\rho},l} = 0$ otherwise. By adopting SIC decoding for the SVC video files, we have

$$a^{\boldsymbol{\rho}',l'}_{\boldsymbol{\rho},l} = \begin{cases} 0, & \text{if } \boldsymbol{\rho} = \boldsymbol{\rho}',\, l \ge l'; \\ 1, & \text{otherwise.} \end{cases} \tag{6}$$

The instantaneous achievable rate (bits/s/Hz) for layer $l \in \mathcal{L}_{\boldsymbol{\rho}}$ at user $\kappa(\boldsymbol{\rho})$ is given by

$$R_{\boldsymbol{\rho},l,t} = \log_2\left(1 + \frac{\frac{1}{\sigma^2}\left|\mathbf{h}^H_{\boldsymbol{\rho},t}\mathbf{w}_{\boldsymbol{\rho},l,t}\right|^2}{1 + \frac{1}{\sigma^2}I_{\boldsymbol{\rho},l,t} + \frac{1}{\sigma^2}\mathbf{h}^H_{\boldsymbol{\rho},t}\mathbf{V}_t\mathbf{h}_{\boldsymbol{\rho},t}}\right), \tag{7}$$

$$I_{\boldsymbol{\rho},l,t} = \sum_{(\boldsymbol{\rho}',l')\neq(\boldsymbol{\rho},l)} a^{\boldsymbol{\rho}',l'}_{\boldsymbol{\rho},l}\left|\mathbf{h}^H_{\boldsymbol{\rho},t}\mathbf{w}_{\boldsymbol{\rho}',l',t}\right|^2, \tag{8}$$

where $I_{\boldsymbol{\rho},l,t}$ 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. For guaranteeing communication secrecy, the proposed secure delivery scheme is designed to avoid information leakage even under worst-case conditions. Specifically, we assume that BS $j \in \mathcal{M}_{\mathcal{U}}$ can fully cancel the self-interference power $\mathbf{c}_{j,t}$ during eavesdropping[4], and hence, achieves the full-duplex capacity upper bound for layer l of the signal intended for user $\kappa(\boldsymbol{\rho})$ given by

$$R_{j,\boldsymbol{\rho},l,t} = \log_2\det\left(\mathbf{I}_{N_j} + \tfrac{1}{\sigma^2_j}\mathbf{Z}^{-1}_{j,\boldsymbol{\rho},l,t}\mathbf{G}^H_{j,t}\mathbf{w}_{\boldsymbol{\rho},l,t}\mathbf{w}^H_{\boldsymbol{\rho},l,t}\mathbf{G}_{j,t}\right), \tag{9}$$

$$\mathbf{Z}_{j,\boldsymbol{\rho},l,t} = \mathbf{I}_{N_j} + \tfrac{1}{\sigma^2_j}\mathbf{G}^H_{j,t}\mathbf{V}_t\mathbf{G}_{j,t} + \tfrac{1}{\sigma^2_j}\mathbf{\Psi}_{j,\boldsymbol{\rho},l,t} \succ \mathbf{0}, \tag{10}$$

$$\mathbf{\Psi}_{j,\boldsymbol{\rho},l,t} = \sum_{(\boldsymbol{\rho}',l')\neq(\boldsymbol{\rho},l)} a^{\boldsymbol{\rho}',l'}_{\boldsymbol{\rho},l}\left(1 - q_{f(\boldsymbol{\rho}'),l',j}\right)\mathbf{G}^H_{j,t}\mathbf{w}_{\boldsymbol{\rho}',l',t}\mathbf{w}^H_{\boldsymbol{\rho}',l',t}\mathbf{G}_{j,t}. \tag{11}$$

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 (11). That is, in addition to SIC, BS $j \in \mathcal{M}_{\mathcal{U}}$ can also 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 $l \in \mathcal{L}_{\boldsymbol{\rho}}$ in time slot $t \in \mathcal{T}_0$ is given by

$$R^{\mathrm{sec}}_{\boldsymbol{\rho},l,t} = \Big[R_{\boldsymbol{\rho},l,t} - \max_{j\in\mathcal{M}_{\mathcal{U}}} R_{j,\boldsymbol{\rho},l,t}\Big]^+. \tag{12}$$

Remark 1. Note that a passive eavesdropper, as considered for non-caching networks in [29]–[31] and caching networks in [16], [17], can be cast as an untrusted BS having no cache memory or no data cached. Considering C1–C3, such an eavesdropper will not participate in the cooperative transmission of the video files. Thus, the untrusted cache helpers considered in this paper correspond to a more general eavesdropping model than that investigated in the literature [16], [17], [29]–[31].

III. Problem Formulation

In this section, we first present the adopted imperfect CSI model for video delivery. Then, a two-timescale robust optimization problem is formulated for minimization of the total BS transmit power required for video streaming under

[4]In practice, if self-interference is not perfectly canceled, the residual self-interference impairs the eavesdropping at the untrusted BSs, and hence, improves communication secrecy. However, estimating the residual self-interference at the central controller (e.g., the macro BS), which is responsible for resource allocation, may not be possible. Hence, we make the worst-case assumption of zero self-interference in this paper and the obtained results provide a lower bound on the performance for the case of imperfect self-interference cancellation.

QoS and secrecy constraints. Note that low transmit power is desirable to minimize the interference caused in other cells and to reduce the network operation cost. For a given cache status, the cooperative transmission decisions for each time slot are optimized online based on instantaneous CSI estimates. However, due to time causality and computational complexity constraints, the cache placement for each period is optimized offline based on historical user requests and CSI [10], [16].

A. Channel State Information

At the beginning of each time slot, the CSI $\mathbf{h}_{\boldsymbol{\rho},t}$ and $\mathbf{G}_{j,t}$ has to be acquired[5] at the centralized controller, i.e., the macro BS, for computing the resource allocation. The estimates of $\mathbf{h}_{\boldsymbol{\rho},t}$ and $\mathbf{G}_{j,t}$ gathered at the macro BS, denoted by $\widehat{\mathbf{h}}_{\boldsymbol{\rho},t} \in \mathbb{C}^{N\times 1}$ and $\widehat{\mathbf{G}}_{j,t} \in \mathbb{C}^{N\times N_j}$, respectively, will in general be imperfect. That is, the actual channels are given by $\mathbf{h}_{\boldsymbol{\rho},t} = \widehat{\mathbf{h}}_{\boldsymbol{\rho},t} + \Delta\mathbf{h}_{\boldsymbol{\rho},t}$ and $\mathbf{G}_{j,t} = \widehat{\mathbf{G}}_{j,t} + \Delta\mathbf{G}_{j,t}$, where $\Delta\mathbf{h}_{\boldsymbol{\rho},t}$ and $\Delta\mathbf{G}_{j,t}$ represent the respective channel estimation errors caused by quantization errors, imperfect feedback channels, as well as outdated and noisy estimates. In fact, the estimation errors $\Delta\mathbf{h}_{\boldsymbol{\rho},t}$ and $\Delta\mathbf{G}_{j,t}$ may be enhanced by the actions of the untrusted BSs which may not fully cooperate with the macro BS during channel estimation and feedback.

The specific values of $\Delta\mathbf{h}_{\boldsymbol{\rho},t}$ and $\Delta\mathbf{G}_{j,t}$ are not known at the macro BS. To model the imperfect CSI, we assume that the possible values of $\Delta\mathbf{h}_{\boldsymbol{\rho},t}$ and $\Delta\mathbf{G}_{j,t}$ lie in ellipsoidal uncertainty regions [42] given by

$$\boldsymbol{\Omega}_{\boldsymbol{\rho},t} \triangleq \left\{\Delta\mathbf{h}_{\boldsymbol{\rho},t} \in \mathbb{C}^{N\times 1} \mid \Delta\mathbf{h}_{\boldsymbol{\rho},t}^H \boldsymbol{\Xi}_{\boldsymbol{\rho}} \Delta\mathbf{h}_{\boldsymbol{\rho},t} \le \varepsilon_{\boldsymbol{\rho}}^2\right\}, \quad \boldsymbol{\rho}\in\mathcal{S},\, t\in\mathcal{T}_0, \tag{13}$$

$$\boldsymbol{\Omega}_{j,t} \triangleq \left\{\Delta\mathbf{G}_{j,t} \in \mathbb{C}^{N\times N_j} \mid \mathrm{tr}\left(\Delta\mathbf{G}_{j,t}^H \boldsymbol{\Xi}_j \Delta\mathbf{G}_{j,t}\right) \le \varepsilon_j^2\right\}, \quad j\in\mathcal{M}_{\mathcal{U}},\, t\in\mathcal{T}_0. \tag{14}$$

Here, $\varepsilon_{\boldsymbol{\rho}} > 0$ and $\varepsilon_j > 0$ represent the radii of uncertainty regions $\boldsymbol{\Omega}_{\boldsymbol{\rho},t}$ and $\boldsymbol{\Omega}_{j,t}$, respectively; $\boldsymbol{\Xi}_{\boldsymbol{\rho}} \in \mathbb{C}^{N\times N}$ and $\boldsymbol{\Xi}_j \in \mathbb{C}^{N\times N}$ denote the orientations of the uncertainty regions, respectively, where $\boldsymbol{\Xi}_{\boldsymbol{\rho}} \succ \mathbf{0}$ and $\boldsymbol{\Xi}_j \succ \mathbf{0}$. In practice, the values of $\varepsilon_{\boldsymbol{\rho}}$, ε_j, $\boldsymbol{\Xi}_{\boldsymbol{\rho}}$, and $\boldsymbol{\Xi}_j$ depend on the channel coherence time and the adopted channel estimation methods.

B. Caching Optimization

Let $\mathbf{q} \triangleq [q_{1,0,0},\ldots,q_{f,l,m},\ldots,q_{F,L-1,M}]$ and $\mathbf{w}_{\boldsymbol{\rho},t} \triangleq [\mathbf{w}_{\boldsymbol{\rho},0,t}^H,\ldots,\mathbf{w}_{\boldsymbol{\rho},L_{\boldsymbol{\rho}}-1,t}^H]^H$ be the caching and the transmitter beamforming optimization vectors, respectively. Considering the two-timescale control in Fig. 1(b), the caching decision $\mathbf{q}$ is made (at the end of) every T_0 time slots based on the historical profiles of user requests and CSI that have been collected during the time period[6]. For the considered typical period $\mathcal{T}_0$, the caching optimization problem is formulated as:

$$\text{P0:}\ \underset{\mathbf{q},\mathbf{w}_{\boldsymbol{\rho},t},\mathbf{V}_t}{\text{minimize}} \sum_{t\in\mathcal{T}_0} U_{\text{TP}}\left(\mathbf{q},\mathbf{w}_{\boldsymbol{\rho},t},\mathbf{V}_t\right) \tag{15}$$

$$\text{subject to}\ \ \text{C1, C2, C3, C4: } \mathbf{V}_t \succeq \mathbf{0},\ \boldsymbol{\Lambda}_{\mathcal{U}}\mathbf{V}_t = \mathbf{0},$$

$$\text{C5: } \mathrm{tr}\left(\boldsymbol{\Lambda}_m\left(\sum_{\boldsymbol{\rho},l}\mathbf{w}_{\boldsymbol{\rho},l,t}\mathbf{w}_{\boldsymbol{\rho},l,t}^H + \mathbf{V}_t\right)\right) \le P_m^{\max},$$

$$\text{C6: } \min_{\Delta\mathbf{h}_{\boldsymbol{\rho},t}\in\boldsymbol{\Omega}_{\boldsymbol{\rho},t}} R_{\boldsymbol{\rho},l,t} \ge R_{\boldsymbol{\rho},l}^{\text{req}},\ \boldsymbol{\rho}\in\mathcal{S},\ l\in\mathcal{L}_{\boldsymbol{\rho}},$$

$$\text{C7: } \max_{j\in\mathcal{M}_{\mathcal{U}}} \max_{\Delta\mathbf{G}_{j,t}\in\boldsymbol{\Omega}_{j,t}} R_{j,\boldsymbol{\rho},0,t} \le R_{\boldsymbol{\rho},0}^{\text{tol}},\ \boldsymbol{\rho}\in\mathcal{S},$$

where $U_{\text{TP}}\left(\mathbf{q},\mathbf{w}_{\boldsymbol{\rho},t},\mathbf{V}_t\right) \triangleq \mathrm{tr}\left(\sum_{m\in\mathcal{M}}\boldsymbol{\Lambda}_m\left(\sum_{\boldsymbol{\rho}\in\mathcal{S}}\sum_{l\in\mathcal{L}_{\boldsymbol{\rho}}}\mathbf{w}_{\boldsymbol{\rho},l,t}\mathbf{w}_{\boldsymbol{\rho},l,t}^H + \mathbf{V}_t\right)\right)$ denotes the total BS transmit power in time slot $t\in\mathcal{T}_0$. Constraint C5 limits the maximal transmit power of BS $m\in\mathcal{M}$ to $P_m^{\max}$. C6 guarantees the minimum video delivery rate, $R_{\boldsymbol{\rho},l}^{\text{req}}$, in each time slot $t\in\mathcal{T}_0$ 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 $R_{\boldsymbol{\rho},0}^{\text{tol}}$ in each time slot t to ensure communication secrecy. Since the untrusted BSs are unable to decode the enhancement layers without base-layer information, secrecy can be ensured by imposing C7 only on the delivery of the base-layer subfiles. Due to the imperfect CSI, the minimum/maximum data rate in C6/C7 is guaranteed/constrained for all possible estimation error vectors/matrices in the respective uncertainty sets in order to facilitate robustness with respect to (w.r.t.) communication secrecy. This robust optimization approach has been commonly adopted for studying PLS in the literature, see [29]–[31] and references therein. Constraints C6 and C7 jointly guarantee a minimum achievable secrecy rate of $R_{\boldsymbol{\rho},0,t}^{\text{sec}} = \left[R_{\boldsymbol{\rho},0}^{\text{req}} - R_{\boldsymbol{\rho},0}^{\text{tol}}\right]^+$, $t\in\mathcal{T}_0$, for delivering the base-layer subfiles for request $\boldsymbol{\rho}$, provided that problem P0 is feasible.

Problem P0 is a non-convex mixed-integer nonlinear program (MINLP)[7] due to the binary caching decision vector $\mathbf{q}$ and the non-convex constraints C6 and C7. This type of problem is NP-hard [41]. Yet, since P0 is solved offline for a large timescale, we adopt a global optimization method to solve P0 optimally in Section IV-A. The obtained solution defines a performance benchmark for low-complexity suboptimal schemes, cf. Section IV-B.

C. Delivery Optimization

Assume that the instantaneous CSI estimates are given. Moreover, the cache status $\mathbf{q}$ for $\mathcal{T}_0$ has been determined at the end of the previous time period. Then, the cooperative transmission policy $\{\mathbf{w}_{\boldsymbol{\rho},t},\mathbf{V}_t\}$ for time $t\in\mathcal{T}_0$ is optimized online by solving the following problem

[5]For example, by exploiting channel reciprocity in time division duplex systems, $\mathbf{h}_{\boldsymbol{\rho},t}$ and $\mathbf{G}_{j,t}$ can be estimated in the uplink at the small cell and macro BSs based on pilots emitted by the UEs and the untrusted BSs, respectively. Then, the estimated CSI obtained at the small cell BSs is fed back to the macro BS via the X2 interface [53].

[6]Prediction of the users' future requests based on historical user profiles can further improve the cache placement at the cost of an increased computational complexity.

[7]For a non-convex MINLP, even if the binary constraints are relaxed into convex ones, the problem remains non-convex [41].

$$\text{Q0: } \underset{\mathbf{w}_{\boldsymbol{\rho},t},\mathbf{V}_t}{\text{minimize}} \quad U_{\text{TP}}\left(\mathbf{q},\mathbf{w}_{\boldsymbol{\rho},t},\mathbf{V}_t\right) \tag{16}$$
$$\text{subject to} \quad \text{C3, C4, C5, C6, C7.}$$

Problem Q0 is non-convex due to constraints C6 and C7. However, we will show that Q0 can be optimally solved by employing SDP relaxation, cf. Section IV-C.

IV. Problem Solution

In this section, the caching problem P0 is tackled first. We show that P0 can be transformed into a convex MINLP by SDP relaxation and further solved optimally by an iterative algorithm. Inspired by the optimal algorithm, a low-complexity suboptimal caching scheme is developed to balance between optimality and computational complexity. Moreover, we show that the delivery problem Q0 can be optimally and efficiently solved.

A. Optimal Caching Scheme

1) Problem Transformation: To reformulate problem P0 as a convex MINLP, C6 and C7 have to be transformed into convex constraints. Let $\mathbf{W}_{\boldsymbol{\rho},l,t} = \mathbf{w}_{\boldsymbol{\rho},l,t}\mathbf{w}_{\boldsymbol{\rho},l,t}^H \succeq \mathbf{0}$ be the beamforming matrix subject to $\text{rank}\left(\mathbf{W}_{\boldsymbol{\rho},l,t}\right) \leq 1$. By substituting $\mathbf{W}_{\boldsymbol{\rho},l,t}$ and employing elementary arithmetic operations, C6 is equivalently reformulated as an affine inequality constraint that is jointly convex w.r.t. $\{\mathbf{W}_{\boldsymbol{\rho},l,t},\mathbf{V}_t\}$,

$$\overline{\text{C6}}\text{: } \mathbf{h}_{\boldsymbol{\rho},t}^H\mathbf{T}_{\boldsymbol{\rho},l,t}\mathbf{h}_{\boldsymbol{\rho},t} \geq \sigma^2, \quad \forall\Delta\mathbf{h}_{\boldsymbol{\rho},t} \in \boldsymbol{\Omega}_{\boldsymbol{\rho},t}, \tag{17}$$

where $\mathbf{T}_{\boldsymbol{\rho},l,t} \triangleq \frac{1}{\eta_{\boldsymbol{\rho},l}^{\text{req}}}\mathbf{W}_{\boldsymbol{\rho},l,t} - \sum_{(\boldsymbol{\rho}',l')\neq(\boldsymbol{\rho},l)} a_{\boldsymbol{\rho},l}^{\boldsymbol{\rho}',l'}\mathbf{W}_{\boldsymbol{\rho}',l',t} - \mathbf{V}_t$ and $\eta_{\boldsymbol{\rho},l}^{\text{req}} \triangleq 2^{R_{\boldsymbol{\rho},l}^{\text{req}}} - 1$. However, as $\boldsymbol{\Omega}_{\boldsymbol{\rho},t}$ is a continuous set, $\overline{\text{C6}}$ is semi-infinite, i.e., it represents infinitely many inequalities for $\mathbf{T}_{\boldsymbol{\rho},l,t}$, and hence, is still intractable for optimization. To overcome this issue, $\overline{\text{C6}}$ is transformed into a finite number of convex constraints. To this end, we substitute $\mathbf{h}_{\boldsymbol{\rho},t} = \widehat{\mathbf{h}}_{\boldsymbol{\rho},t} + \Delta\mathbf{h}_{\boldsymbol{\rho},t}$ in $\overline{\text{C6}}$ and apply the S-procedure from [43, Appendix B], which leads to

$$\begin{aligned}
&\overline{\text{C6}}\text{: } \Delta\mathbf{h}_{\boldsymbol{\rho},t}^H\mathbf{T}_{\boldsymbol{\rho},l,t}\Delta\mathbf{h}_{\boldsymbol{\rho},t} + 2\Re\Big\{\widehat{\mathbf{h}}_{\boldsymbol{\rho},t}^H\mathbf{T}_{\boldsymbol{\rho},l,t}\Delta\mathbf{h}_{\boldsymbol{\rho},t}\Big\} \\
&\qquad + \widehat{\mathbf{h}}_{\boldsymbol{\rho},t}^H\mathbf{T}_{\boldsymbol{\rho},l,t}\widehat{\mathbf{h}}_{\boldsymbol{\rho},t} - \sigma^2 \geq 0, \quad \forall\Delta\mathbf{h}_{\boldsymbol{\rho},t} \in \boldsymbol{\Omega}_{\boldsymbol{\rho},t}, \\
\Longleftrightarrow\ &\widetilde{\text{C6}}\text{: } \mathbf{U}_{\boldsymbol{\rho},t}^H\mathbf{T}_{\boldsymbol{\rho},l,t}\mathbf{U}_{\boldsymbol{\rho},t} \succeq \\
&\qquad \begin{bmatrix} -\delta_{\boldsymbol{\rho},l,t}\boldsymbol{\Xi}_{\boldsymbol{\rho}} & \mathbf{0} \\ \mathbf{0}^H & \sigma^2 + \delta_{\boldsymbol{\rho},l,t}\varepsilon_{\boldsymbol{\rho}}^2 \end{bmatrix}, \ \delta_{\boldsymbol{\rho},l,t} \geq 0,
\end{aligned} \tag{18}$$

where $\mathbf{U}_{\boldsymbol{\rho},t} \triangleq [\mathbf{I}_N, \widehat{\mathbf{h}}_{\boldsymbol{\rho},t}] \in \mathbb{C}^{(N+1)\times(N+1)}$ and $\delta_{\boldsymbol{\rho},l,t}$ is an auxiliary optimization variable.

Next, let $\overline{\mathbf{W}}_{\boldsymbol{\rho},l,j,t} = (1 - q_{f(\boldsymbol{\rho}),l,j})\mathbf{W}_{\boldsymbol{\rho},l,t} \succeq \mathbf{0}$ be an auxiliary optimization matrix. We have

$$\boldsymbol{\Psi}_{j,\boldsymbol{\rho},l,t} = \mathbf{G}_{j,t}^H\Big(\sum_{(\boldsymbol{\rho}',l')\neq(\boldsymbol{\rho},l)} a_{\boldsymbol{\rho},l}^{\boldsymbol{\rho}',l'}\overline{\mathbf{W}}_{\boldsymbol{\rho}',l',j,t}\Big)\mathbf{G}_{j,t}, \tag{19}$$

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

$$\begin{aligned}
&\text{C8: } \text{tr}(\mathbf{W}_{\boldsymbol{\rho},l,t} - \overline{\mathbf{W}}_{\boldsymbol{\rho},l,j,t}) \preceq q_{f(\boldsymbol{\rho}),l,j}P_{\max}, \\
&\text{C9: } \text{tr}(\overline{\mathbf{W}}_{\boldsymbol{\rho},l,j,t}) \preceq (1 - q_{f(\boldsymbol{\rho}),l,j})P_{\max}, \\
&\text{C10: } \mathbf{W}_{\boldsymbol{\rho},l,t} \succeq \overline{\mathbf{W}}_{\boldsymbol{\rho},l,j,t}, \ \overline{\mathbf{W}}_{\boldsymbol{\rho},l,j,t} \succeq \mathbf{0},
\end{aligned} \tag{20}$$

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 an LMI as follows

$$\begin{aligned}
\text{C7} \ &\Longleftrightarrow \ \tfrac{1}{\sigma_j^2}\mathbf{w}_{\boldsymbol{\rho},0,t}^H\mathbf{G}_{j,t}\mathbf{Z}_{j,\boldsymbol{\rho},0,t}^{-1}\mathbf{G}_{j,t}^H\mathbf{w}_{\boldsymbol{\rho},0,t} \leq \eta_{\boldsymbol{\rho},0}^{\text{tot}} \\
&\Longleftrightarrow \ \text{tr}\left(\mathbf{Z}_{j,\boldsymbol{\rho},0,t}^{-1}\mathbf{G}_{j,t}^H\mathbf{W}_{\boldsymbol{\rho},0,t}\mathbf{G}_{j,t}\right) \leq \sigma_j^2\eta_{\boldsymbol{\rho},0}^{\text{tot}}, \\
&\overset{\text{(a)}}{\Longleftrightarrow} \ \lambda_{\max}\left(\mathbf{Z}_{j,\boldsymbol{\rho},0,t}^{-1/2}\mathbf{G}_{j,t}^H\mathbf{W}_{\boldsymbol{\rho},0,t}\mathbf{G}_{j,t}\mathbf{Z}_{j,\boldsymbol{\rho},0}^{-1/2}\right) \leq \sigma_j^2\eta_{\boldsymbol{\rho},0}^{\text{tot}}, \\
&\Longleftrightarrow \ \overline{\text{C7}}\text{: } \mathbf{G}_{j,t}^H\mathbf{T}_{\boldsymbol{\rho},0,j,t}\mathbf{G}_{j,t} \preceq \sigma_j^2\mathbf{I}_{N_j}, \ \forall j \in \mathcal{M}_{\mathcal{U}}, \\
&\qquad\qquad \forall\Delta\mathbf{G}_{j,t} \in \boldsymbol{\Omega}_{j,t},
\end{aligned} \tag{21}$$

where $\eta_{\boldsymbol{\rho},0}^{\text{tot}} \triangleq 2^{R_{\boldsymbol{\rho},0}^{\text{tot}}} - 1$, $\mathbf{T}_{\boldsymbol{\rho},0,j,t} \triangleq \frac{1}{\eta_{\boldsymbol{\rho},0}^{\text{tot}}}\mathbf{W}_{\boldsymbol{\rho},0,t} - \sum_{(\boldsymbol{\rho}',l')\neq(\boldsymbol{\rho},0)} a_{\boldsymbol{\rho},0}^{\boldsymbol{\rho}',l'}\overline{\mathbf{W}}_{\boldsymbol{\rho}',l',j,t} - \mathbf{V}_t$, and (a) holds due to $\text{rank}(\overline{\mathbf{W}}_{\boldsymbol{\rho},l,t}) \leq 1$. In fact, as $\boldsymbol{\Omega}_{j,t}$ is a continuous set, $\overline{\text{C7}}$ in (21) represents infinitely many LMIs that are jointly convex w.r.t. $\left\{\mathbf{W}_{\boldsymbol{\rho},0,t},\mathbf{V}_t,\overline{\mathbf{W}}_{j,t}\right\}$. For tractability, $\overline{\text{C7}}$ has to be transformed into a finite number of convex constraints. This can be accomplished by exploiting the robust quadratic matrix inequality in [45, Theorem 3.3]. Thereby, we obtain

$$\begin{aligned}
&\overline{\text{C7}} \ \Longleftrightarrow \ \widetilde{\text{C7}}\text{: } \mathbf{U}_{j,t}^H\mathbf{T}_{\boldsymbol{\rho},0,j,t}\mathbf{U}_{j,t} \preceq \\
&\qquad \begin{bmatrix} (\sigma_j^2 - \delta_{\boldsymbol{\rho},0,j,t})\mathbf{I}_{N_j} & \mathbf{0} \\ \mathbf{0} & \frac{\delta_{\boldsymbol{\rho},0,j,t}}{\varepsilon_j^2}\boldsymbol{\Xi}_j \end{bmatrix}, \delta_{\boldsymbol{\rho},0,j,t} \geq 0,
\end{aligned} \tag{22}$$

where $\mathbf{U}_{j,t} \triangleq [\widehat{\mathbf{G}}_{j,t}\mathbf{I}_N] \in \mathbb{C}^{N\times(N+N_j)}$ and $\delta_{\boldsymbol{\rho},0,j,t}$ is an auxiliary optimization variable.

Finally, by defining the delivery variable $\mathbf{D}_t \triangleq [\mathbf{W}_{\boldsymbol{\rho},l,t}, \overline{\mathbf{W}}_{\boldsymbol{\rho},l,j,t}, \mathbf{V}_t]$ and applying the above transformations, the original problem P0 is equivalently reformulated as

$$\begin{aligned}
&\underset{\mathbf{q},\mathbf{D}_t}{\text{minimize}} \quad \sum_{t\in\mathcal{T}_0} U_{\text{TP}}\left(\mathbf{q},\mathbf{D}_t\right) \\
&\text{subject to} \quad \text{C1, C2, C4, } \widetilde{\text{C6}}\text{, } \widetilde{\text{C7}}\text{, C8, C9, C10,} \\
&\qquad \text{C3: } \text{tr}(\boldsymbol{\Lambda}_m\mathbf{W}_{\boldsymbol{\rho},l,t}) \leq q_{f(\boldsymbol{\rho}),l,m}P_m^{\max}, \\
&\qquad \text{C5: } \text{tr}\Big(\boldsymbol{\Lambda}_m\Big(\sum_{\boldsymbol{\rho},l}\mathbf{W}_{\boldsymbol{\rho},l,t} + \mathbf{V}_t\Big)\Big) \leq P_m^{\max}, \\
&\qquad \text{C11: } \text{rank}(\mathbf{W}_{\boldsymbol{\rho},l,t}) \leq 1, \ \boldsymbol{\rho} \in \mathcal{S}, \ l \in \mathcal{L}_{\boldsymbol{\rho}}.
\end{aligned} \tag{23}$$

Here, constraint $\text{rank}(\overline{\mathbf{W}}_{\boldsymbol{\rho},l,t}) \leq 1$ is dropped due to C10 and C11. Let P1 denote the SDP relaxation of problem (23), obtained by dropping C11 in (23). Then, problem P1 is a convex MINLP, i.e., by relaxing the binary constraints of P1 into convex ones, we arrive at a convex problem.

The GBD algorithm is a simple iterative method to handle convex MINLPs [41, Section 6.3]. In each GBD iteration, upper and lower bounds on the optimal value are generated by solving a primal subproblem and a master problem, respectively. To ensure convergence, optimality and feasibility cuts are successively added to tighten the bounds and eliminate the infeasible solutions possibly obtained during the iterations, respectively. The GBD algorithm is attractive for solving P1 as it can be efficiently implemented exploiting the structure of P1. In particular, the resulting primal subproblem is a convex problem where strong duality holds while the master problem is a mixed-integer linear program (MILP), and both problems are easy to handle using off-the-shelf numerical solvers such

as CVX [46] and MOSEK [47]. However, the GBD algorithm typically suffers from slow convergence. This is because, when an infeasible solution is obtained in an iteration of the GBD, the resulting feasibility cut is usually ineffective in improving the solution. If the problem is infeasible, the GBD algorithm terminates only after having performed an exhaustive search over all possible candidate solutions. To remedy this issue, an improved GBD algorithm[8] is proposed below.

2) Problem Decomposition: The proposed modified GBD algorithm applies a two-layer decomposition of problem P1 and solves a binary caching optimization problem for $\mathbf{q}$ in the outer layer and a continuous delivery optimization problem for $\mathbf{D}_t$ in the inner layer. However, $\mathbf{q}$ and $\mathbf{D}_t$ are coupled via constraints C3, C8, and C9. To facilitate the decomposition, we perturb the right-hand sides of C3, C8, and C9 by introducing slack variables $s^{\text{C3}}_{\boldsymbol{\rho},l,m,t} \geq 0$, $s^{\text{C8}}_{\boldsymbol{\rho},l,j,t} \geq 0$, and $s^{\text{C9}}_{\boldsymbol{\rho},l,j,t} \geq 0$, respectively. Let $\mathbf{s}_t \triangleq [s^{\text{C3}}_{\boldsymbol{\rho},l,m,t},\, s^{\text{C8}}_{\boldsymbol{\rho},l,j,t},\, s^{\text{C9}}_{\boldsymbol{\rho},l,j,t}]$ 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 $\mathbf{s}_t$,

$$f_{\text{Pen}}(\mathbf{s}_t) \triangleq \mu \left(\sum_{\boldsymbol{\rho},l,m} s^{\text{C3}}_{\boldsymbol{\rho},l,m,t} + \sum_{\boldsymbol{\rho},l,j} \left(s^{\text{C8}}_{\boldsymbol{\rho},l,j,t} + s^{\text{C9}}_{\boldsymbol{\rho},l,j,t} \right) \right), \quad (24)$$

with penalty factor $\mu \gg 1$. Consequently, problem P1 decomposes into T_0 SDP subproblems in the inner layer, i.e., one subproblem for each time slot $t \in \mathcal{T}_0$, and an MILP in the outer layer, which are shown in (25) and (26) at the top of the next page, respectively. Problem (26) is referred as the master problem. Thereby, problems (25) and (26) are equivalent to P1 when $\mu \gg 1$, as stated in Proposition 1.

Proposition 1. For $\mu \gg 1$, problems P1 and (25), (26) are equivalent such that: i) if P1 is feasible, then SDP subproblem (25) is *always* feasible for $\mathbf{q} \in \mathcal{Q}$; moreover, the optimal solution of $\mathbf{q}$ for P1 solves the master problem (26); ii) if problem (25) is infeasible, i.e., $\nu(\mathbf{q}) = +\infty$, or its optimal solution satisfies $\mathbf{s}_{t'} \neq \mathbf{0}$ for some $t' \in \mathcal{T}_0$, then problem P1 is infeasible.

Proof: Please refer to Appendix A. ∎

Remark 2. By perturbation, the feasible set of $\nu(\mathbf{q})$ in (25), (26) is extended to $\mathcal{Q} \subseteq \{0,1\}^{F \times L \times M}$. Consequently, based on Proposition 1, infeasible solutions can be avoided if problem P1 is feasible, cf. i), or identified easily if problem P1 is infeasible, cf. ii). These properties facilitate an efficient implementation of the GBD algorithm in the sequel to optimally solve (25) and (26).

SDP subproblem (25) is convex and can be solved by interior point methods [43], [48], i.e., numerical solvers such as CVX [46] are applicable. Meanwhile, exploiting the convexity (and strong duality) of (25), we can further simplify the formulation of the master problem (26). Let $\lambda^{\overline{\text{C3}}}_{\boldsymbol{\rho},l,m,t} \geq 0$, $\lambda^{\overline{\text{C8}}}_{\boldsymbol{\rho},l,j,t} \geq 0$, and $\lambda^{\overline{\text{C9}}}_{\boldsymbol{\rho},l,j,t} \geq 0$ be the Lagrange multipliers for $\overline{\text{C3}}$, $\overline{\text{C8}}$, and $\overline{\text{C9}}$, respectively, and define $\boldsymbol{\lambda}_t \triangleq [\lambda^{\overline{\text{C3}}}_{\boldsymbol{\rho},l,m,t}, \lambda^{\overline{\text{C8}}}_{\boldsymbol{\rho},l,j,t}, \lambda^{\overline{\text{C9}}}_{\boldsymbol{\rho},l,j,t}] \succeq \mathbf{0}$. The Lagrangian of (25) can be written as

$$\mathcal{L}_{\mathbf{q}}(\mathbf{D}_t, \mathbf{s}_t; \boldsymbol{\lambda}_t) = f_1(\mathbf{q}; \boldsymbol{\lambda}_t) + f_2(\mathbf{D}_t, \mathbf{s}_t; \boldsymbol{\lambda}_t), \quad (27)$$

where $f_1(\mathbf{q}; \boldsymbol{\lambda}_t) = \sum_{\boldsymbol{\rho},l,j} \left(\lambda^{\overline{\text{C9}}}_{\boldsymbol{\rho},l,j,t} - \lambda^{\overline{\text{C8}}}_{\boldsymbol{\rho},l,j,t} \right) q_{f(\boldsymbol{\rho}),l,j} P_{\max} - \sum_{\boldsymbol{\rho},l,m,t} \lambda^{\overline{\text{C3}}}_{\boldsymbol{\rho},l,m,t} q_{f(\boldsymbol{\rho}),l,m} P^{\max}_m$, and

$$\begin{aligned} f_2(\mathbf{D}_t, \mathbf{s}_t; \boldsymbol{\lambda}_t) &= f_{\text{Pen}}(\mathbf{s}_t) + \\ &\sum_{\boldsymbol{\rho},l} \text{tr}\Big[\sum_m \lambda^{\overline{\text{C3}}}_{\boldsymbol{\rho},l,m,t} \boldsymbol{\Lambda}_m + \big(1 + \sum_j \lambda^{\overline{\text{C8}}}_{\boldsymbol{\rho},l,j,t}\big) \mathbf{I}_N \Big] \mathbf{W}_{\boldsymbol{\rho},l,t} \\ &+ \sum_{\boldsymbol{\rho},l,j} \big(\lambda^{\overline{\text{C9}}}_{\boldsymbol{\rho},l,j,t} - \lambda^{\overline{\text{C8}}}_{\boldsymbol{\rho},l,j,t} \big) \text{tr}\left(\overline{\mathbf{W}}_{\boldsymbol{\rho},l,j,t} \right) - \sum_{\boldsymbol{\rho},l,j} \lambda^{\overline{\text{C9}}}_{\boldsymbol{\rho},l,j,t} P_{\max} \\ &- \sum_{\boldsymbol{\rho},l,m} \lambda^{\overline{\text{C3}}}_{\boldsymbol{\rho},l,m,t} s^{\text{C3}}_{\boldsymbol{\rho},l,m,t} + \sum_{\boldsymbol{\rho},l,j} \big(\lambda^{\overline{\text{C8}}}_{\boldsymbol{\rho},l,j,t} s^{\text{C8}}_{\boldsymbol{\rho},l,j,t} - \lambda^{\overline{\text{C9}}}_{\boldsymbol{\rho},l,j,t} s^{\text{C9}}_{\boldsymbol{\rho},l,j,t} \big). \end{aligned}$$

$\mathcal{L}_{\mathbf{q}}(\mathbf{D}_t, \mathbf{s}_t; \boldsymbol{\lambda}_t)$ is separable w.r.t. $\{\mathbf{q}\}$ and $\{\mathbf{D}_t, \mathbf{s}_t\}$. Since, for given $\mathbf{q} \in \mathcal{Q}$, problem (25) is convex and fulfills Slater's condition, the following result holds due to strong duality:

$$\nu_t(\mathbf{q}) = \max_{\boldsymbol{\lambda}_t \succeq \mathbf{0}} \min_{\mathbf{D}_t \in \mathcal{D}_t,\, \mathbf{s}_t \succeq \mathbf{0}} \mathcal{L}_{\mathbf{q}}(\mathbf{D}_t, \mathbf{s}_t; \boldsymbol{\lambda}_t), \quad \forall \mathbf{q} \in \mathcal{Q}. \quad (28)$$

Consequently, the master problem is reformulated as

$$\begin{aligned} \underset{\mathbf{q} \in \mathcal{Q}, \alpha}{\text{minimize}} \quad & \alpha \\ \text{subject to} \quad & \alpha \geq \sum\nolimits_{t \in \mathcal{T}_0} \xi_t(\mathbf{q}; \boldsymbol{\lambda}_t), \quad \forall \boldsymbol{\lambda}_t \succeq \mathbf{0}, \end{aligned} \quad (29)$$

where $\xi_t(\mathbf{q}; \boldsymbol{\lambda}_t) \triangleq \min_{\mathbf{D}_t \in \mathcal{D}_t,\, \mathbf{s}_t \succeq \mathbf{0}} \mathcal{L}_{\mathbf{q}}(\mathbf{D}_t, \mathbf{s}_t; \boldsymbol{\lambda}_t)$. Although problem (29) still contains an infinite number of constraints (w.r.t. $\boldsymbol{\lambda}_t$) and undetermined functions $\xi_t(\cdot;\cdot)$, it is readily solvable by an iterative relaxation method as will be explained in the following.

3) Optimal Iterative Solution: The proposed iterative algorithm is given in Algorithm 1. Let k be the iteration index. We start from one constraint at $k = 1$, which defines a cutting plane (also referred as an optimality cut [41]). Then, the number of constraints/cuts are increased sequentially as the iteration proceeds. Specifically, for given dual variables $\boldsymbol{\lambda}^j_t$, $j = 1, \ldots k-1$, the following master problem is solved in iteration k,

$$\begin{aligned} \underset{\mathbf{q} \in \mathcal{Q}, \alpha}{\text{minimize}} \quad & \alpha \\ \text{subject to} \quad & \alpha \geq \sum\nolimits_{t \in \mathcal{T}_0} \xi_t\big(\mathbf{q}; \boldsymbol{\lambda}^j_t\big), \quad j = 1, \ldots, k-1. \end{aligned} \quad (30)$$

Problem (30) is a relaxation of problem (29). Due to the enlarged feasible set, the optimal value of problem (30) gives a lower bound on that of problem (29). The relaxation solution, denoted by $(\mathbf{q}^k, \alpha^k)$, is optimal for problem (29) if it is feasible for problem (29). Otherwise, we add another optimality cut to the feasible set of (30) in the next iteration to tighten the relaxation. As this process continues, we obtain a non-decreasing sequence of lower bounds until the relaxed solution becomes feasible, i.e., solves problem (29) optimally, or until the problem is known to be infeasible.

Two remarks regarding Algorithm 1 are in order. First, as can be observed in Algorithm 1 (lines 5–10), the feasibility or optimality of $\mathbf{q}^k$ is verified by solving SDP subproblem (25). This is because, if $\mathbf{q}^k$ is optimal, we know that solving problem (25) for $\mathbf{q} = \mathbf{q}^k$ in line 5 will return the optimal value of α^k, i.e., $\nu(\mathbf{q}^k) = \alpha^k$, owing to the strong duality of problem (25). Otherwise, $\nu(\mathbf{q}^k)$ gives an upper bound on the optimal value, and thus, $\nu(\mathbf{q}^k) \geq \alpha^k$. By keeping the lowest upper bound obtained so far, i.e., $UB \leftarrow \min\{UB, \nu(\mathbf{q}^k)\}$

[8] A similar approach as in the proposed improved GBD algorithm has been successfully applied to accelerate the outer approximation algorithm in [41, Section 6.6].

$$\nu_t(\mathbf{q}) \triangleq \underset{\mathbf{D}_t,\mathbf{s}_t\succeq\mathbf{0}}{\text{minimize}}\ U_{\text{TP}}(\mathbf{q},\mathbf{D}_t)+f_{\text{Pen}}(\mathbf{s}_t) \tag{25}$$
$$\text{subject to } \mathbf{D}_t\in\mathcal{D}_t\triangleq\left\{\mathbf{D}_t\,|\,\text{C4, C5, }\overline{\text{C6}}\text{, }\overline{\text{C7}}\text{, C10}\right\},\quad \overline{\text{C3}}\text{: tr}(\mathbf{\Lambda}_m\mathbf{W}_{\boldsymbol{\rho},l,t})-q_{f(\boldsymbol{\rho}),l,m}P_m^{\max}\le s^{\text{C3}}_{\boldsymbol{\rho},l,m},$$
$$\overline{\text{C8}}\text{: tr}(\mathbf{W}_{\boldsymbol{\rho},l,t}-\overline{\mathbf{W}}_{\boldsymbol{\rho},l,j,t})-q_{f(\boldsymbol{\rho}),l,j}P_{\max}\le s^{\text{C8}}_{\boldsymbol{\rho},l,j},\quad \overline{\text{C9}}\text{: tr}(\overline{\mathbf{W}}_{\boldsymbol{\rho},l,j,t})-(1-q_{f(\boldsymbol{\rho}),l,j})P_{\max}\le s^{\text{C9}}_{\boldsymbol{\rho},l,j},$$
$$\underset{\mathbf{q},\alpha}{\text{minimize}}\quad \alpha \tag{26}$$
$$\text{subject to}\quad \alpha\ge\nu(\mathbf{q})\triangleq\sum\nolimits_{t\in\mathcal{T}_0}\nu_t(\mathbf{q}),\qquad \mathbf{q}\in\mathcal{Q}\triangleq\{\mathbf{q}\,|\,\text{C1, C2}\}.$$

Algorithm 1 Optimal iterative algorithm for solving P1 and P0

1: **Initialization**: Given $\mathbf{q}^0\leftarrow\mathbf{0}$. Solve the SDP subproblem (25) for given $\mathbf{q}^0$ and determine $\mathbf{D}_t^1,\mathbf{s}_t^1,\boldsymbol{\lambda}_t^1$; set tolerance $\varepsilon\ge 0$, $UB\leftarrow\nu(\mathbf{q}^0)$, $LB\longleftarrow\infty$, $k\leftarrow 1$;
2: **while** $(UB>LB+\varepsilon)$ **do**
3: Solve the relaxed master problem (30) for given $\mathbf{D}_t^k$, $\mathbf{s}_t^k$, $\boldsymbol{\lambda}_t^k$ and determine the solution $(\mathbf{q}^k,\alpha^k)$;
4: Update lower bound and solution: $LB\leftarrow\alpha^k$, $\mathbf{q}^*\leftarrow\mathbf{q}^k$;
5: Solve SDP subproblem (25) for given $\mathbf{q}^k$ and determine the primal and the dual solutions $\mathbf{D}_t^{k+1}$, $\mathbf{s}_t^{k+1}$, $\boldsymbol{\lambda}_t^{k+1}$;
6: **if** $(\nu(\mathbf{q}^k)=+\infty$, i.e., (25) is infeasible, **OR** $\nu(\mathbf{q}^k)\le\alpha^k+\varepsilon)$ **then**
7: Set $\mathbf{D}_t^*\leftarrow\mathbf{D}_t^{k+1}$, $\mathbf{s}_t^*\leftarrow\mathbf{s}_t^{k+1}$ and exit the while loop;
8: **else if** $(\nu(\mathbf{q}^k)<UB)$ **then**
9: Update upper bound and solution: $UB\leftarrow\nu(\mathbf{q}^k)$, $\mathbf{D}_t^*\leftarrow\mathbf{D}_t^{k+1}$, $\mathbf{s}_t^*\leftarrow\mathbf{s}_t^{k+1}$;
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 $\mathbf{D}_t^*$;
15: **else**
16: Return the infeasible problem P0/P1.
17: **end if**

(cf. line 9), the optimality condition is satisfied when the gap between UB and the lower bound vanishes.

Second, for computational convenience, the values of $\boldsymbol{\lambda}_t^k$ in iteration k can be intelligently chosen as the optimal dual solutions of problem (28) or (25) for $\mathbf{q}=\mathbf{q}^k$ in line 5. In this case, the constraint function $\xi_t(\cdot;\cdot)$ can be easily computed as explained in the following proposition.

Proposition 2. Let $(\mathbf{D}_t^k,\mathbf{s}_t^k)$ and $\boldsymbol{\lambda}_t^k$ be the optimal primal and dual solutions of (25) for $\nu(\mathbf{q}^k)$ at iteration k, respectively. Then, we have i) $(\mathbf{D}_t^k,\mathbf{s}_t^k)$ also solves the minimization problem in the optimality cut of iteration $k+1$ (cf. (30)), i.e., $(\mathbf{D}_t^k,\mathbf{s}_t^k)\in\arg\min_{\mathbf{D}_t\in\mathcal{D}_t,\,\mathbf{s}_t\ge\mathbf{0}}\mathcal{L}_{\mathbf{q}}(\mathbf{D}_t,\mathbf{s}_t;\boldsymbol{\lambda}_t^k)$.

ii) By choosing $\boldsymbol{\lambda}_t=\boldsymbol{\lambda}_t^k$, function $\xi_t\left(\mathbf{q};\boldsymbol{\lambda}_t^k\right)$ reduces to an affine function given by

$$\xi_t\left(\mathbf{q};\boldsymbol{\lambda}_t^k\right)=\sum_{\boldsymbol{\rho},l,j}(\lambda^{\overline{\text{C9}},k}_{\boldsymbol{\rho},l,j,t}-\lambda^{\overline{\text{C8}},k}_{\boldsymbol{\rho},l,j,t})P_{\max}\Big(q_{f(\boldsymbol{\rho}),l,j}-q^k_{f(\boldsymbol{\rho}),l,j}\Big)$$
$$-\sum_{\boldsymbol{\rho},l,m}\lambda^{\overline{\text{C3}},k}_{\boldsymbol{\rho},l,m,t}P_m^{\max}\left(q_{f(\boldsymbol{\rho}),l,m}-q^k_{f(\boldsymbol{\rho}),l,m}\right)$$
$$+U_{\text{TP}}\left(\mathbf{q}^k,\mathbf{D}_t^k\right)+f_{\text{Pen}}\left(\mathbf{s}_t^k\right). \tag{31}$$

Proof: Please refer to Appendix B. ∎

Based on Proposition 2, the relaxed master problem (30) is an MILP and can be solved optimally, e.g., using the numerical solver MOSEK [47]. Similar to the conventional GBD method, Algorithm 1 converges in a finite number of iterations as shown in Proposition 3. The obtained solution is globally optimal for problem P1. In general, the solution of P1 gives a lower bound for problem P0. However, by inspecting the rank of the SDP solution of problem P1, we can further show that the SDP relaxation is tight.

Proposition 3. Algorithm 1 converges in a finite number iterations. Moreover, assuming that the channel vectors $\widehat{\mathbf{h}}_{\boldsymbol{\rho},t}$, $\boldsymbol{\rho}\in\mathcal{S}$, can be modeled as statistically independent random vectors, problems 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, and the optimal beamformer is given by the principal eigenvector of $\mathbf{W}_{\boldsymbol{\rho},l,t}$.

Proof: Please refer to Appendix C. ∎

Due to perturbation, only optimality cuts, cf. (30), need to be generated by Algorithm 1 in each iteration, cf. Remark 2. This is different from the classical GBD algorithm [41, Section 6.3] where feasibility cuts are also required to exclude infeasible solutions during intermediate iterations. Since the optimality cuts can successively improve the lower bounds, Algorithm 1 is expected to converge faster than the classic GBD algorithm if P1 is feasible. On the other hand, even if P1 is infeasible, the perturbed problem is generally feasible. Then, optimality cuts can be still generated to iteratively improve the solutions and reduce the required number of iterations with a high probability.

4) Computational Complexity: Assume that the interior-point method [43], [48] is applied to solve the SDP subproblems in each iteration of the GBD algorithm. The computational complexity of solving each SDP subproblem w.r.t. the number of UEs, K, the number of BSs, M, the number of BS antennas, N, and the number of SVC layers, L, can be approximated as [49, Theorem 3.12]

$$\Theta^{\text{sdp}}=\mathcal{O}\Bigg(\underbrace{\left((MKL)^4\left(N^3+N^2+2\right)+(MKL)^3\right)}_{\text{Complexity per iteration}}$$
$$\times\underbrace{\sqrt{MKLN}\log\left(\epsilon^{-1}\right)}_{\text{Number of iterations}}\Bigg)$$
$$=\mathcal{O}\left((MKL)^{4.5}N^{3.5}\log\left(\epsilon^{-1}\right)\right), \tag{32}$$

where $\epsilon>0$ is the solution accuracy specified by the numerical solver and $\mathcal{O}(\cdot)$ is the big-O notation. Although the SDP

Algorithm 2 Suboptimal iterative algorithm for solving P1 and P0

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 SDP subproblem (25) for each given $\{\mathbf{q}_m\}$ satisfying $\mathbf{q}_i \in \mathcal{Q}_i^k \cap \mathcal{Q}$;
5: Determine $\mathbf{q}_i^{k+1}$ in (34);
6: **end for**
7: $k \leftarrow k+1$.
8: **end while**

subproblems can be solved in polynomial time in line 5, cf. (32), the overall computational complexity of Algorithm 1 grows non-polynomially with the size of problem P0. This is because the MILP solver in line 3 may incur an exponential-time computational complexity, $\mathcal{O}(2^{FLM})$, in the worst case [41], even though the likelihood that the worst case occurs is low due to the employed perturbation, cf. Remark 2. Thus, only offline cache optimization may be feasible in practical implementations.

B. Suboptimal Caching Scheme

For systems with limited computing resources, Algorithm 1 may not be applicable due to its worst-case exponential-time computational complexity. Instead, polynomial-time suboptimal schemes facilitating a better trade-off between system performance and computational complexity may be preferable. Based on Proposition 3, P0 can be solved via its equivalent convex MINLP, P1. As is also evident from (25), for given $\mathbf{q}$, P1 reduces to an SDP and can be solved optimally in polynomial time, cf. (32). Therefore, by additionally adjusting $\mathbf{q}$ via a greedy iterative search, we obtain the low-complexity suboptimal scheme in Algorithm 2.

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

$$\mathcal{Q}_m^k \triangleq \left\{ \mathbf{q}_m \in \{0,1\}^{|\mathcal{F}_\mathcal{S}| \times L} \mid \left\| \mathbf{q}_m - \mathbf{q}_m^k \right\|_2^2 \leq 1 \right\} \tag{33}$$

as the set of binary vectors within a distance of one from $\mathbf{q}_m^k$. Besides, $\mathcal{I}^k \triangleq \left\{ m \in \mathcal{M} \mid \left| \mathcal{Q}_m^k \cap \mathcal{Q} \right| > 1 \right\}$ 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 (25) is chosen as the new caching vector at BS $i \in \mathcal{I}^k$, i.e.,

$$\mathbf{q}_i^{k+1} = \underset{\mathbf{q}_i \in \mathcal{Q}_i^k \cap \mathcal{Q}}{\arg\min} \quad \nu\left(\mathbf{q}_1, \ldots, \mathbf{q}_M\right). \tag{34}$$

That is, the cache vectors $\mathbf{q}_i^{k+1}$, within a distance of one from $\mathbf{q}_i^k$, are iteratively updated 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 function is possible, which yields the solution. Hence, the number of problem instances of (25) to be solved is bounded by $ML^2 \left|\mathcal{F}_\mathcal{S}\right|^2$. Consequently, the overall computational complexity of Algorithm 2 is approximated as $\mathcal{O}\left(ML^2 \left|\mathcal{F}_\mathcal{S}\right|^2 \Theta^{\mathrm{sdp}}\right)$, which grows only polynomially with the problem size.

Remark 3. By adopting the greedy heuristic, searching over the non-convex set $\mathcal{Q}$ of P1 can be done in polynomial time. The obtained solution is ensured to be feasible for P1. Moreover, it is often close-to-optimal due to the iterative minimization in (34) [50], as will be shown in Section V.

C. Optimal Delivery Solution

By applying the same transformation techniques as for problem P0 in Section IV-A and relaxing the rank constraint, cf. C11, problem Q0 can be reformulated as an SDP, which is equivalent to problem (25). Since, based on Proposition 3, the solution of problem (25) fulfills rank constraint C11 with probability one, delivery problem Q0 can be solved optimally via SDP subproblem (25), as stated in Corollary 1.

Corollary 1. SDP subproblem (25) (for the respective instantaneous CSI) and the delivery optimization problem Q0 are equivalent in the sense that the solution of (25) is also optimal for Q0 whenever Q0 is feasible.

Proof: The proof is similar to that of Propositions 1 and 3 and is omitted for brevity. ■

Therefore, delivery optimization problem Q0 can be solved in polynomial time with computational complexity Θ^{sdp}, cf. (32), which is desirable for online implementation [42]. Moreover, delivery optimization incurs a signaling overhead of $\mathcal{O}(MKLN)$ for collecting the CSI at the macro BS and distributing the optimization results to the small cell BSs.

Remark 4. Although the total number of BSs in dense small cell networks may be large [44], the coverage areas of most small cell BSs will not overlap. Therefore, the proposed caching and delivery algorithms may be applied to several small groups of small cell BSs with overlapping coverage areas rather than jointly to all small cell BSs. This considerably reduces the computational complexity and signaling overhead.

Remark 5. The proposed caching and delivery optimization framework can be extended to integrate (centralized and decentralized) coded caching at the end users [11]–[15]. For example, multicast codewords can be cooperatively transmitted by a subset of the BSs if these BSs have cached the subfiles required for coded multicast transmission. Such a design reaps the performance gains of both coded caching and cache-enabled cooperative MIMO transmission. However, the resulting cache optimization problem would involve a large number of binary caching variables, which have to be defined per subfile, and SDP relaxation of the delivery optimization problem may not yield the optimal solution anymore [51]. Hence, extending the proposed framework to coded caching is an interesting topic for future research.

V. 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 and three small cell BSs are uniformly distributed within the cell. The number of untrusted small cell BSs is set to $M_u \triangleq |\mathcal{M}_\mathcal{U}| = 1$, unless stated otherwise. To gain insight, in Figs. 2–5, we consider a small network with only three small cell BSs. A larger network is considered in Fig. 6. The macro BS is equipped with $N_0 = 6$ 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

TABLE II
SIMULATION PARAMETERS.

Parameters	Settings
System bandwidth	5 MHz
Duration of time slot	10 ms
Duration of delivery period	45 mins
Macro BS transmit power	$P_0^{\max}$= 46 dBm
Small BS transmit power	$P_m^{\max}$= 39 dBm
Noise power density	-172.6 dBm/Hz
Cache capacity at macro BS	$C_0^{\max}$= 1000 MB

(Bytes), are delivered to $K = 6$ single-antenna UEs. 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 $\gamma = 1.1$. In particular, assuming that file $f \in \mathcal{F}$ is the σ_fth most popular file for the UEs, the probability of file $f \in \mathcal{F}$ being requested is given by $\theta_f = \frac{1}{\sigma_f^\gamma} / \sum_{f\in\mathcal{F}} \frac{1}{\sigma_f^\gamma}$ [52]. We adopt an SVC codec with $L=2$. That is, each video file is encoded into a base-layer subfile ($l = 0$) and an enhancement-layer subfile ($l = 1$), each of size $V_{f,l} = 250$ MB. The minimum streaming rate and the secrecy rate threshold for the base-layer subfiles are $R^{\text{req}}_{\boldsymbol{\rho},0} = 825$ kbps and $R^{\text{tol}}_{\boldsymbol{\rho},0} = 0.1 R^{\text{req}}_{\boldsymbol{\rho},0} = 82.5$ kbps, respectively. Therefore, if problem Q0 is feasible, a secrecy streaming rate of $R^{\text{sec}}_{\boldsymbol{\rho},0,t} = 742.5$ kbps can be guaranteed for secure and uninterrupted video streaming for each user as $R^{\text{sec}}_{\boldsymbol{\rho},0,t} \geq 250{\times}8{\times}10^6/(45{\times}60) = 741$ kbps. The streaming rate of the enhancement-layer subfiles is $R^{\text{req}}_{\boldsymbol{\rho},1} = 2R^{\text{sec}}_{\boldsymbol{\rho},0,t} = 1.5$ Mbps. The users are randomly distributed in the system. Based on the locations of the BSs and users, the path loss is calculated using the 3GPP model for the "urban macro non-line-of-sight" scenario [53]. The small-scale fading coefficients are independent and identically distributed (i.i.d.) Rayleigh random variables. We employ Euclidean spheres for modeling the uncertainty regions $\boldsymbol{\Omega}_{\boldsymbol{\rho},t}$ and $\boldsymbol{\Omega}_{j,t}$ by setting $\boldsymbol{\Xi}_{\boldsymbol{\rho}} = \boldsymbol{\Xi}_j = \mathbf{I}_N$. Meanwhile, we define the maximum normalized channel estimation error variances of $\mathbf{h}_{\boldsymbol{\rho},t}$ and $\mathbf{G}_{j,t}$ as $\sigma^2_{\boldsymbol{\rho}} = \frac{\varepsilon^2_{\boldsymbol{\rho}}}{\|\mathbf{h}_{\boldsymbol{\rho},t}\|_2^2}$ and $\sigma_j^2 = \frac{\varepsilon_j^2}{\|\mathbf{G}_{j,t}\|_F^2}$, respectively. Unless otherwise specified, we assume $\sigma^2_{\boldsymbol{\rho}} = 0.01$, $\boldsymbol{\rho} \in \mathcal{S}$ and $\sigma_j^2 = 0.05$, $j \in \mathcal{M}_{\mathcal{U}}$. All other relevant system parameters are given in Table II.

For comparison, we consider two heuristic caching schemes and a non-cooperative and a non-robust 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, since the base-layer subfiles are more important, they 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 problem Q0.
- Baseline 3 (No cooperation with untrusted BSs): No video files are cached at the untrusted BSs, which act as pure eavesdroppers. Hence, the untrusted BSs are not allowed to cooperate for delivery of the video files. This approach is adopted in state-of-the-art cellular networks. The optimal caching and delivery decisions are obtained from problems P0 and Q0, respectively, with $C_m^{\max} = 0$, $\forall m \in \mathcal{M}_{\mathcal{U}}$.
- Baseline 4 (Non-robust transmission): Different from the proposed schemes, the macro BS treats the channel estimates $\widehat{\mathbf{h}}_{\boldsymbol{\rho},t}$ and $\widehat{\mathbf{G}}_{j,t}$ as accurate. The optimal caching and delivery decisions are obtained by solving problems P0 and Q0, respectively, after setting $\mathbf{h}_{\boldsymbol{\rho},t} = \widehat{\mathbf{h}}_{\boldsymbol{\rho},t}$ and $\mathbf{G}_{j,t} = \widehat{\mathbf{G}}_{j,t}$.

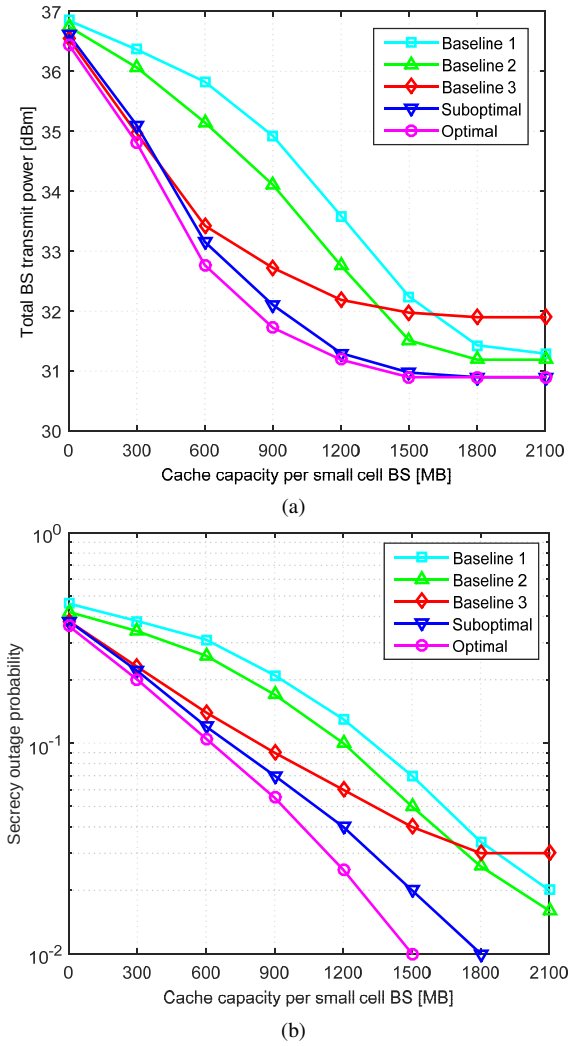


Fig. 2. (a) Total BS transmit power and (b) secrecy outage probability versus cache capacity for different caching and delivery schemes.

Figs. 2(a) and 2(b) illustrate the performance of the considered caching and delivery schemes as functions of the cache capacity. Herein, the system performance is evaluated during the online delivery of the video files, cf. problem Q0. The secrecy outage probability, defined as $p_{\text{out}} \triangleq \Pr\left(\sum_{\boldsymbol{\rho}} R^{\text{sec}}_{\boldsymbol{\rho},0,t} < \sum_{\boldsymbol{\rho}} \left[R^{\text{req}}_{\boldsymbol{\rho},0} - R^{\text{tol}}_{\boldsymbol{\rho},0}\right]^+\right)$, characterizes the likelihood that problem Q0 is infeasible because either the QoS constraint C6 or the secrecy constraint C7 cannot be satisfied. As can be observed from Figs. 2(a) and 2(b), for all considered schemes, a larger cache capacity leads to both a lower total BS transmit power and a smaller secrecy outage probability as larger virtual transmit antenna arrays can be formed among the trusted and

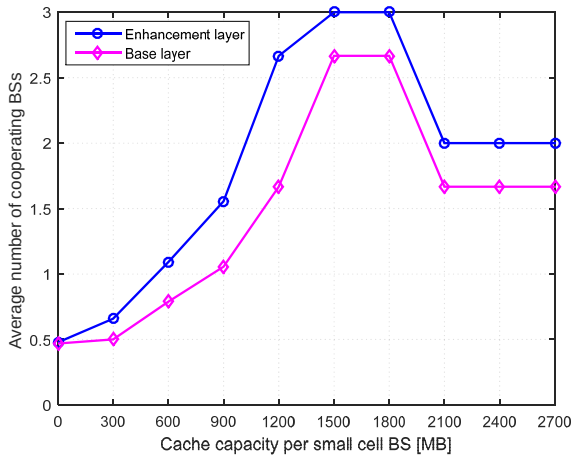


Fig. 3. Average number of cooperating BSs versus cache capacity for transmission of the base-layer and enhancement-layer subfiles when Algorithm 1 is employed for cache optimization.

untrusted BSs for cooperative beamforming transmission of the base-layer and enhancement-layer subfiles, respectively. There is a non-negligible performance gap between the optimal scheme and Baseline 3, particularly in the high cache capacity regime. This is because the proposed caching scheme can exploit the cache resources of the untrusted helpers for delivering enhancement-layer subfiles while Baseline 3 cannot. The performance gap between the proposed optimal scheme and Baselines 1 and 2 is small for small (large) cache capacities because of insufficient (saturated) BS cooperation. For medium cache capacities, however, the proposed optimal scheme achieves considerable performance gains due to its ability to exploit information regarding the user requests and CSI for cache placement. We note that the proposed suboptimal scheme attains good performance in all regimes despite its low computational cost.

To provide more insight into how the BSs cooperate, Fig. 3 shows the average numbers of cooperating (small cell and macro) BSs for transmission of the base-layer and enhancement-layer subfiles, denoted as $\overline{N}_{\mathrm{BL}}$ and $\overline{N}_{\mathrm{EL}}$, respectively, if Algorithm 1 is employed for cache optimization. Recall that the proposed caching scheme does not cache base-layer subfiles at untrusted BSs, cf. constraint C1. Consequently, for a given cache capacity, $\overline{N}_{\mathrm{BL}} \leq \overline{N}_{\mathrm{EL}}$ holds. Interestingly, the behavior of $\overline{N}_{\mathrm{BL}}$ and $\overline{N}_{\mathrm{EL}}$ is not monotonic as the cache capacity increases. In the small and medium cache capacity regime, $\overline{N}_{\mathrm{BL}}$ and $\overline{N}_{\mathrm{EL}}$ monotonically increase with the cache capacity since the number of subfiles that can be cached at the small cell BSs and the number of BSs that can participate in cooperative transmission increase. For example, $\overline{N}_{\mathrm{BL}}$ and $\overline{N}_{\mathrm{EL}}$ are 0.5 and 0.7 for 300 MB cache capacity per small cell BS, respectively, and increase to 1.7 and 2.7 for 1200 MB cache capacity per small cell BS, respectively. In the large cache capacity regime, the performance gains saturate as the available DoFs for the transmission of the base-layer and enhancement-layer subfiles saturate, cf. Fig. 2. However, $\overline{N}_{\mathrm{BL}}$ and $\overline{N}_{\mathrm{EL}}$ decrease by 1 before reaching the stationary BS cooperation topology. This is because, when the cache-enabled DoFs are sufficient, the optimal caching scheme

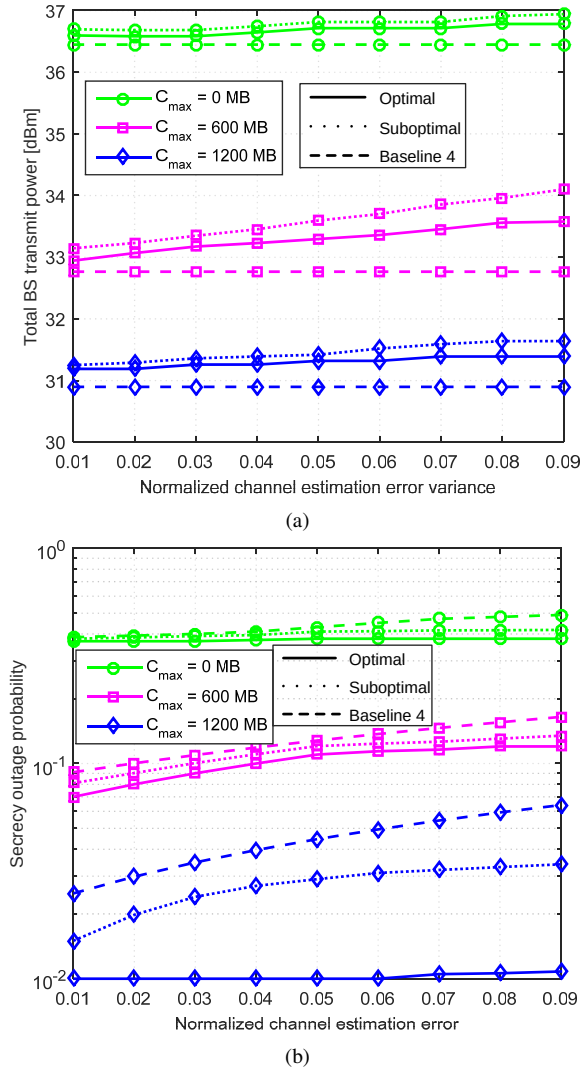


Fig. 4. (a) Total BS transmit power and (b) secrecy outage probability versus normalized channel estimation error variance for the proposed optimal scheme (solid line), the suboptimal scheme (dotted line), and Baseline 4 (dashed line).

selects preferred trusted cooperating (small cell and/or macro) BSs, e.g., based on their channel conditions and cache status, instead of exploiting all BSs available for cooperation.

Next, we evaluate the robustness of the proposed schemes w.r.t. channel estimation errors. Figs. 4(a) and 4(b) show the performance of the proposed schemes and Baseline 4 as functions of the normalized channel estimation error variances, σ_j^2, where $C_{\max}$ denotes the cache capacity per small cell BS. We observe that, compared to Baseline 4, the proposed schemes achieve a lower secrecy outage probability at the cost of a slightly higher transmit power consumption. Specifically, to achieve robustness in meeting QoS constraint C6 under imperfect CSI, the proposed schemes employ wide beams for transmitting the base-layer subfiles, which may lead to information leakage to the untrusted BSs. Hence, to ensure communication secrecy in constraint C7, the proposed schemes also have to transmit a non-negligible amount of AN to degrade the reception of the untrusted BSs. On the other hand, the wide beams and the interference caused by the AN to the

legitimate users, cf. (7), have to be compensated by increasing the transmit power of the beamforming vectors. Therefore, the total transmit power increases as the CSI uncertainty increases. In contrast, by treating the imperfect CSI as perfect in C6, Baseline 4 employs narrow transmit beams to save transmit power but this leads to the highest secrecy outage probability in Fig. 4(b).

The impact of the number of trusted and untrusted BSs on cache-enabled secrecy is studied in Figs. 5(a) and 5(b). Fig. 5(a) reveals that the required transmit power increases with the number of untrusted BSs M_u, if the total number of BSs is kept constant. This is because, as more helpers become untrusted, 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 potential eavesdroppers. On the other hand, as the base layers are not cached at the untrusted BSs, for a larger M_u, more cache capacity can be utilized to transmit the enhancement-layer subfiles. Hence, the transmit power of the optimal/suboptimal scheme is only enlarged moderately when M_u increases from 1 to 2; in the high cache capacity regime, the increase in transmit power is even negligible. Due to the cooperative transmission of the base-layer and enhancement-layer subfiles, in Fig. 5(b), the secrecy outage probability monotonically decreases with the cache capacity for $M_u \leq 2$. However, when M_u increases from 2 to 3, both the transmit power and the secrecy outage probability are increased significantly; particularly, their values saturate at high levels for cache capacities exceeding 600 MB. This is because, for $M_u = 3$, the total number of antennas equipped at the untrusted BSs equals the total number of antennas equipped at the trusted BSs; hence, the available DoFs for secure transmission of the base-layer subfiles are limited, irrespective of the cache capacities, as the cache of the untrusted small cell BSs can only facilitate the cooperative transmission of enhancement-layer subfiles. Hence, for secure delivery of the base-layer subfiles, the system has to allocate large amounts of power to AN transmission to degrade the reception quality of the untrusted BSs and, at the same time, to the users' signals to mitigate the impact of the interference caused by AN to the legitimate users.

Finally, Figs. 6(a) and 6(b) show the performance of the proposed schemes and Baseline 3 for larger networks, where the number of users, K, the number of small cells, $M-1$, and the number of untrusted small cell BSs, M_u, satisfy $M = K+1$ and $M_u = K/10$. We consider a system bandwidth of 7 MHz to ensure that, despite the large number of users, the QoS requirements of each user can be fulfilled with high probability. For Baseline 3, the caching decisions are determined by employing Algorithms 1 and 2, and the resulting schemes are referred to as "Baseline 3 optimal" and "Baseline 3 suboptimal", respectively. For $K = 50$, only the performance of the proposed suboptimal and baseline suboptimal schemes is shown because of the high computational complexity of the optimal schemes. As K increases, more untrusted BSs, $M_u = K/10$, are present in the network and each untrusted BS can eavesdrop a larger number of users. Hence, the available DoFs per BS for the delivery of the base-layer subfiles is reduced and the likelihood of data leakage is increased. Therefore, to ensure secure video streaming, for larger K, a larger average transmit power per BS is needed for AN transmission. On the other hand, exploiting both trusted and untrusted BSs for cooperative transmission, the proposed schemes significantly outperform Baseline 3, particularly for networks with large numbers of users and large cache capacities.

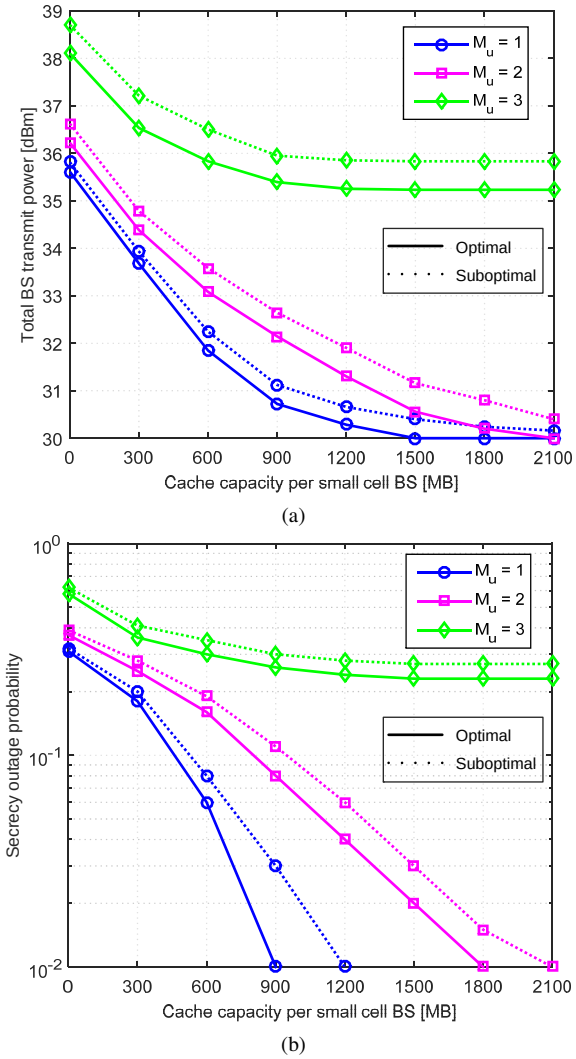
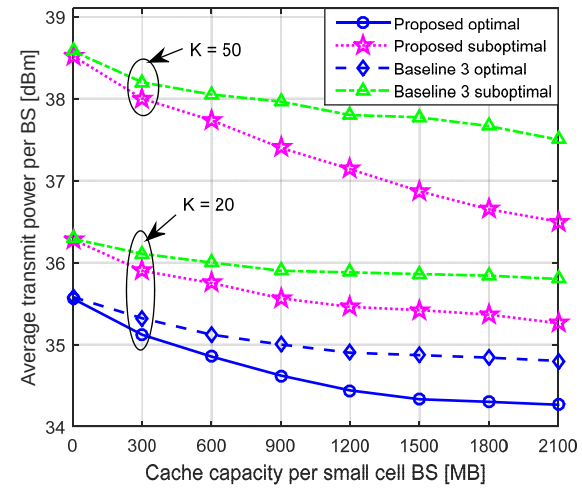


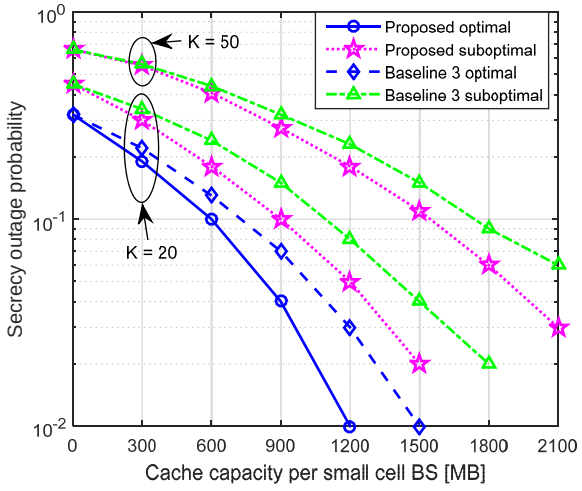
Fig. 5. (a) Total BS transmit power and (b) secrecy outage probability of the proposed optimal (solid line) and suboptimal (dotted line) schemes versus cache capacity for different numbers of untrusted BSs.

VI. Conclusion

In this paper, secure video streaming was investigated for small cell networks with untrusted small cell BSs which can intercept both cached and delivered video data. SVC coding and caching were jointly exploited to facilitate secure cooperative MIMO transmission and to not only mitigate the negative impact of the untrusted BSs but to exploit them for secrecy enhancement. A two-timescale non-convex robust optimization problem was formulated to optimize caching and delivery for minimization of the total BS transmit power required for secure video streaming with imperfect CSI knowledge. In the



(a)



(b)

Fig. 6. (a) Average transmit power per BS and (b) secrecy outage probability versus cache capacity for the proposed optimal and suboptimal schemes and Baseline 3, where $M = K + 1$ and $M_u = K/10$.

large timescale, the caching optimization problem was solved offline by a modified GBD algorithm. To reduce the computational complexity, a suboptimal caching algorithm was also studied. In the short timescale, the delivery optimization problem for a given cache status was solved online by SDP. Simulation results revealed that, compared to several baseline schemes, the proposed optimal and suboptimal schemes can significantly enhance both the secrecy and the power efficiency of video streaming in small cell networks as long as the total number of antennas at the trusted BSs exceeds that at the untrusted BSs.

APPENDIX

A. Proof of Proposition 1

We begin the proof by defining an auxiliary optimization problem:

$$\text{P2:} \quad \underset{\mathbf{D}_t,\,\mathbf{s}_t,\,\mathbf{q}}{\text{minimize}} \quad \sum\nolimits_{t\in\mathcal{T}_0} \left[U_{\text{TP}}\left(\mathbf{q},\mathbf{D}_t\right) + f_{\text{Pen}}\left(\mathbf{s}_t\right)\right] \tag{35}$$
$$\text{subject to} \quad \overline{\text{C3}},\ \overline{\text{C8}},\ \overline{\text{C9}},\ \text{C12},\ \mathbf{D}_t \in \mathcal{D}_t,\ \mathbf{s}_t \succeq \mathbf{0},\ \mathbf{q} \in \mathcal{Q},$$

where $\mathcal{D}_t$ is given in (25). Problems P2 and $\{(25), (26)\}$ are equivalent as $\min_{x\in\mathcal{X},\,y\in\mathcal{Y}} f(x,y) = \min_{y\in\mathcal{Y}}\left[\min_{x\in\mathcal{X}} f(x,y)\right]$ [43].

Because of the perturbation, the feasible set of problem P2 is a superset of the feasible set of problem P1. Thus, if P1 is feasible, so is P2 (and (25)). Moreover, the inequality constraint functions on the left hand sides of the big-M constraints $\overline{\text{C3}}$, $\overline{\text{C8}}$, and $\overline{\text{C9}}$ are bounded from above, e.g., $\text{tr}(\mathbf{\Lambda}_m\mathbf{W}_{\boldsymbol{\rho},l,t}) - q_{f(\boldsymbol{\rho}),l,m}P_m^{\max} \le P_m^{\max}$ for $\overline{\text{C3}}$. Considering $\mathbf{s}_t \succeq \mathbf{0}$ on the right hand sides of $\overline{\text{C3}}$, $\overline{\text{C8}}$, and $\overline{\text{C9}}$, the feasibility statement in part i) of Proposition 1 thus always holds for any $\mathbf{q} \in \mathcal{Q}$.

Next, we show the optimality statement in part i) by contradiction. Assume that $(\mathbf{q}^*, \mathbf{D}_t^*)$ solves P1. Then $(\mathbf{q}^*, \mathbf{D}_t^*, \mathbf{s}_t)$ with $\mathbf{s}_t \succeq \mathbf{0}$ is feasible for P2. Denote the objective function of P2 by $f(\mathbf{q}, \mathbf{D}_t, \mathbf{s}_t)$. We have $f(\mathbf{q}, \mathbf{D}_t, \mathbf{0}) \ge f(\mathbf{q}^*, \mathbf{D}_t^*, \mathbf{0}), \forall(\mathbf{q}, \mathbf{D}_t)$. Besides, let $(\mathbf{q}^+, \mathbf{D}_t^+, \mathbf{s}_t^+)$ be the optimal solution of P2. If $(\mathbf{q}^+, \mathbf{D}_t^+) \ne (\mathbf{q}^*, \mathbf{D}_t^*)$, then $\mathbf{s}_t^+ \ne \mathbf{0}$ necessarily holds. However, since $\mu \gg 1$, we have $f_{\text{Pen}}\left(\mathbf{s}_t^+\right) > f(\mathbf{q}^*, \mathbf{D}_t^*, \mathbf{0})$ and thus $f\left(\mathbf{q}^+, \mathbf{D}_t^+, \mathbf{s}_t^+\right) > f(\mathbf{q}^*, \mathbf{D}_t^*, \mathbf{0})$, which contradicts the optimality of $\left(\mathbf{q}^+, \mathbf{D}_t^+, \mathbf{s}_t^+\right)$. Therefore, part i) is proved.

Finally, we prove part ii). Obviously, i) implies that problem P1 is infeasible when $\nu(\mathbf{q}) = +\infty$. Assume that the optimal solution of problem P2, $\left(\mathbf{q}^+, \mathbf{D}_t^+, \mathbf{s}_t^+\right)$, satisfies $\mathbf{s}_t^+ \ne \mathbf{0}$. Then, a feasible solution of the form $(\mathbf{q}, \mathbf{D}_t, \mathbf{0})$ does not exist for P2, since otherwise, $f\left(\mathbf{q}^+, \mathbf{D}_t^+, \mathbf{s}_t^+\right) \le f(\mathbf{q}, \mathbf{D}_t, \mathbf{0})$ has to hold. Therefore, P1 is also infeasible, which completes the proof.

B. Proof of Proposition 2

As strong duality holds for (25), we know that $(\mathbf{D}_t^k, \mathbf{s}_t^k)$ minimizes the Lagrangian, i.e., $(\mathbf{D}_t^k, \mathbf{s}_t^k) \in \arg\min_{\mathbf{D}_t\in\mathcal{D}_t,\,\mathbf{s}_t\succeq\mathbf{0}} \mathcal{L}_{\mathbf{q}^k}(\mathbf{D}_t, \mathbf{s}_t; \boldsymbol{\lambda}_t^k)$. Since $\mathcal{L}_{\mathbf{q}^k}(\mathbf{D}_t, \mathbf{s}_t) = f_1(\mathbf{q}^k; \boldsymbol{\lambda}_t^k) + f_2(\mathbf{D}_t, \mathbf{s}_t; \boldsymbol{\lambda}_t^k)$, where $f_1(\mathbf{q}^k; \boldsymbol{\lambda}_t^k)$ is a constant, we also have $(\mathbf{D}_t^k, \mathbf{s}_t^k) \in \arg\min_{\mathbf{D}_t\in\mathcal{D}_t,\,\mathbf{s}_t\succeq\mathbf{0}} f_2(\mathbf{D}_t, \mathbf{s}_t; \boldsymbol{\lambda}_t^k)$. Finally, i) has to hold as $\mathcal{L}_{\mathbf{q}}(\mathbf{D}_t, \mathbf{s}_t; \boldsymbol{\lambda}_t)$ is separable w.r.t. $\mathbf{q}$ and $\{\mathbf{D}_t, \mathbf{s}_t\}$ in (30), cf. (27).

Meanwhile, according to the Karush-Kuhn-Tucker (KKT) conditions for (25), we have

$$\lambda_{\boldsymbol{\rho},l,m,t}^{\overline{\text{C3}},k}[\text{tr}(\mathbf{\Lambda}_m\mathbf{W}_{\boldsymbol{\rho},l,t}^k) - q_{f(\boldsymbol{\rho}),l,m}^k P_m^{\max} - s_{\boldsymbol{\rho},l,m,t}^{\text{C3},k}] = 0,$$
$$\lambda_{\boldsymbol{\rho},l,j,t}^{\overline{\text{C8}},k}[\text{tr}(\mathbf{W}_{\boldsymbol{\rho},l,t}^k - \overline{\mathbf{W}}_{\boldsymbol{\rho},l,j,t}^k) - q_{f(\boldsymbol{\rho}),l,j}^k P_{\max} - s_{\boldsymbol{\rho},l,j,t}^{\text{C8},k}] = 0,$$
$$\lambda_{\boldsymbol{\rho},l,j,t}^{\overline{\text{C9}},k}[\text{tr}(\overline{\mathbf{W}}_{\boldsymbol{\rho},l,j,t}^k) - (1 - q_{f(\boldsymbol{\rho}),l,j}^k)P_{\max} - s_{\boldsymbol{\rho},l,j,t}^{\text{C9},k}] = 0,$$

Therefore,

$$\begin{aligned}&\min_{\mathbf{D}_t\in\mathcal{D}_t,\,\mathbf{s}_t\succeq\mathbf{0}} f_2(\mathbf{D}_t, \mathbf{s}_t; \boldsymbol{\lambda}_t^k)\\ &= f_2(\mathbf{D}_t^k, \mathbf{s}_t^k; \boldsymbol{\lambda}_t^k)\\ &= U_{\text{TP}}\left(\mathbf{q}^k, \mathbf{D}_t^k\right) + f_{\text{Pen}}\left(\mathbf{s}_t^k\right) - f_1(\mathbf{q}^k; \boldsymbol{\lambda}_t^k).\end{aligned} \tag{36}$$

By substituting (36) into $\mathcal{L}_{\mathbf{q}}(\mathbf{D}_t, \mathbf{s}_t; \boldsymbol{\lambda}_t^k)$, (31) is established. This completes the proof.

C. Proof of Proposition 3

Let $\widehat{\mathbf{q}} \in \mathcal{Q}$ be a solution of (30). By Algorithm 1, the optimality cut $\alpha \ge \xi(\mathbf{q}; \widehat{\boldsymbol{\lambda}}_t)$ is then generated for $\mathbf{q} = \widehat{\mathbf{q}}$. We have $\xi(\widehat{\mathbf{q}}; \widehat{\boldsymbol{\lambda}}_t) = \nu(\widehat{\mathbf{q}})$ due to the strong duality of the

respective SDP subproblem (25) for $\mathbf{q} = \widehat{\mathbf{q}}$. If $(\overline{\mathbf{q}}, \overline{\alpha})$ were to solve (30) again in another iteration with $\overline{\mathbf{q}} = \widehat{\mathbf{q}}$, then $\overline{\alpha} \geq \xi(\overline{\mathbf{q}}; \widehat{\boldsymbol{\lambda}}_t)$ and $\overline{\alpha} \geq \nu(\widehat{\mathbf{q}}) = \nu(\overline{\mathbf{q}})$ would hold, which lead to the termination of Algorithm 1 since $LB \geq UB$, cf. line 2. This implies that $\widehat{\mathbf{q}}$ does not repeat itself in intermediate iterations. Since set $\mathcal{Q} \subseteq \{0,1\}^{F\times L\times M}$ is finite, Algorithm 1 has to converge in a finite number of iterations.

To prove the equivalence between P1 and P0, we show here that the solution of the relaxed problem P1 satisfies $\mathrm{rank}(\mathbf{W}_{\boldsymbol{\rho},l,t}) = 1$ with probability one. Let $\boldsymbol{\Upsilon}_{\boldsymbol{\rho},l,t} \succeq \mathbf{0}$, $\boldsymbol{\Phi}_{\boldsymbol{\rho},l,t} \succeq \mathbf{0}$, and $\boldsymbol{\Theta}_{\boldsymbol{\rho},l,t} \succeq \mathbf{0}$ be the Lagrange multipliers associated with constraints $\widetilde{\text{C6}}$, C10, and C12: $\mathbf{W}_{\boldsymbol{\rho},l,t} \succeq \mathbf{0}$, respectively, where C12 is implied by C10. The Lagrangian of problem P1 is given by,

$$\mathcal{L}(\mathbf{W}_{\boldsymbol{\rho},l,t}; \boldsymbol{\Upsilon}_{\boldsymbol{\rho},l,t}, \boldsymbol{\Phi}_{\boldsymbol{\rho},l,t}, \boldsymbol{\Theta}_{\boldsymbol{\rho},l,t}) = \sum_{\boldsymbol{\rho},l} \mathrm{tr}\left[\left(\mathbf{B}_{\boldsymbol{\rho},l,t} - \boldsymbol{\Theta}_{\boldsymbol{\rho},l,t} - \tfrac{1}{\eta^{\text{req}}_{\boldsymbol{\rho},l}}\mathbf{U}_{\boldsymbol{\rho},t}\boldsymbol{\Upsilon}_{\boldsymbol{\rho},l,t}\mathbf{U}^H_{\boldsymbol{\rho},t}\right)\mathbf{W}_{\boldsymbol{\rho},l,t}\right] + \Delta_2, \tag{37}$$

where $\mathbf{B}_{\boldsymbol{\rho},l,t} \triangleq \mathbf{I} + \boldsymbol{\Delta}_1 - \boldsymbol{\Phi}_{\boldsymbol{\rho},l,t}$; and $\boldsymbol{\Delta}_1 \succeq \mathbf{0}$ and $\Delta_2 \in \mathbb{R}$ denote the collection of terms that are relevant and irrelevant to $\mathbf{W}_{\boldsymbol{\rho},l,t}$, respectively. Hence, the dual problem of P1 is given by

$$\max_{\boldsymbol{\Upsilon}_{\boldsymbol{\rho},l,t}\succeq\mathbf{0}, \boldsymbol{\Phi}_{\boldsymbol{\rho},l,t}\succeq\mathbf{0}, \boldsymbol{\Theta}_{\boldsymbol{\rho},l,t}\succeq\mathbf{0}} \ \min_{\mathbf{W}_{\boldsymbol{\rho},l,t}} \mathcal{L}(\mathbf{W}_{\boldsymbol{\rho},l,t}; \boldsymbol{\Upsilon}_{\boldsymbol{\rho},l,t}, \boldsymbol{\Phi}_{\boldsymbol{\rho},l,t}, \boldsymbol{\Theta}_{\boldsymbol{\rho},l,t}). \tag{38}$$

We now define

$$\boldsymbol{\Upsilon}_{\boldsymbol{\rho},l,t} \triangleq \begin{bmatrix} \overline{\boldsymbol{\Upsilon}}_{\boldsymbol{\rho},l,t} & \overline{\gamma}_{\boldsymbol{\rho},l,t} \\ \overline{\gamma}^H_{\boldsymbol{\rho},l,t} & \alpha_{\boldsymbol{\rho},l,t} \end{bmatrix} \in \mathbb{C}^{(N+1)\times(N+1)}, \tag{39}$$

where $\overline{\boldsymbol{\Upsilon}}_{\boldsymbol{\rho},l,t} \succeq \mathbf{0}$, $\alpha_{\boldsymbol{\rho},l,t} \geq 0$ and $\overline{\gamma}_{\boldsymbol{\rho},l,t}$ is chosen to ensure $\boldsymbol{\Upsilon}_{\boldsymbol{\rho},l,t} \succeq \mathbf{0}$. If P0 is feasible, so is P1, and then $\mathbf{W}_{\boldsymbol{\rho},l,t} \neq \mathbf{0}$. Moreover, as strong duality holds for SDP subproblem (25), the optimal beamformers and the optimal dual solutions satisfy the KKT optimality conditions. In particular, by substituting $\mathbf{U}_{\boldsymbol{\rho},t} = [\mathbf{I}_N, \widehat{\mathbf{h}}_{\boldsymbol{\rho},t}]$ and (39) into (37), we have

$$\overline{\mathbf{B}}_{\boldsymbol{\rho},l,t} - \tfrac{\alpha_{\boldsymbol{\rho},l,t}}{\eta^{\text{req}}_{\boldsymbol{\rho},l}}\widehat{\mathbf{h}}_{\boldsymbol{\rho},t}\widehat{\mathbf{h}}^H_{\boldsymbol{\rho},t} = \boldsymbol{\Theta}_{\boldsymbol{\rho},l,t}, \tag{40}$$

$$\boldsymbol{\Theta}_{\boldsymbol{\rho},l,t}\mathbf{W}_{\boldsymbol{\rho},l,t} = \mathbf{0}, \tag{41}$$

where $\overline{\mathbf{B}}_{\boldsymbol{\rho},l,t} = \mathbf{B}_{\boldsymbol{\rho},l,t} - \frac{1}{\eta^{\text{req}}_{\boldsymbol{\rho},l}}\left(\overline{\boldsymbol{\Upsilon}}_{\boldsymbol{\rho},l,t} + \overline{\gamma}_{\boldsymbol{\rho},l,t}\widehat{\mathbf{h}}^H_{\boldsymbol{\rho},t} + \overline{\gamma}^H_{\boldsymbol{\rho},l,t}\widehat{\mathbf{h}}_{\boldsymbol{\rho},t}\right)$.

Next, we show by contradiction that $\overline{\mathbf{B}}_{\boldsymbol{\rho},l,t} \succ \mathbf{0}$ holds with probability one. Assume that $\overline{\mathbf{B}}_{\boldsymbol{\rho},l,t}$ has at least one non-positive eigenvalue $\tau \leq 0$ and the corresponding eigenvector is $\widetilde{\mathbf{w}}_{\boldsymbol{\rho},l,t}$, i.e., $(\overline{\mathbf{B}}_{\boldsymbol{\rho},l,t} - \tau\mathbf{I})\widetilde{\mathbf{w}}_{\boldsymbol{\rho},l,t} = \mathbf{0}$. 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 (38), we further have

$$\mathcal{L}(\mathbf{W}_{\boldsymbol{\rho},l,t}; \boldsymbol{\Upsilon}_{\boldsymbol{\rho},l,t}, \boldsymbol{\Phi}_{\boldsymbol{\rho},l,t}, \boldsymbol{\Theta}_{\boldsymbol{\rho},l,t}) = \underbrace{\beta\,\tau\sum_{\boldsymbol{\rho},l}\widetilde{\mathbf{w}}^H_{\boldsymbol{\rho},l,t}\widetilde{\mathbf{w}}_{\boldsymbol{\rho},l,t}}_{\leq 0} - \beta\sum_{\boldsymbol{\rho},l}\tfrac{\alpha_{\boldsymbol{\rho},l,t}}{\eta^{\text{req}}_{\boldsymbol{\rho},l}}\widehat{\mathbf{h}}^H_{\boldsymbol{\rho},t}\widetilde{\mathbf{w}}_{\boldsymbol{\rho},l,t}\widetilde{\mathbf{w}}^H_{\boldsymbol{\rho},l,t}\widehat{\mathbf{h}}_{\boldsymbol{\rho},t} + \Delta_2. \tag{42}$$

Hence, if $\alpha_{\boldsymbol{\rho},l,t} = 0$, then $\widetilde{\mathbf{w}}_{\boldsymbol{\rho},l,t} = \mathbf{0}$ necessarily holds to ensure $\mathcal{L} > -\infty$; yet this contradicts the condition that $\mathbf{W}_{\boldsymbol{\rho},l,t} \neq \mathbf{0}$. On the other hand, if $\alpha_{\boldsymbol{\rho},l,t} > 0$, then the minimum value of (38) is obtained for $\beta \to \infty$, since $\mathbf{h}_{\boldsymbol{\rho},t}$ is statistically independent and $-\beta\frac{\alpha_{\boldsymbol{\rho},l,t}}{\eta^{\text{req}}_{\boldsymbol{\rho},l}}\widehat{\mathbf{h}}^H_{\boldsymbol{\rho},t}\widetilde{\mathbf{w}}_{\boldsymbol{\rho},l,t}\widetilde{\mathbf{w}}^H_{\boldsymbol{\rho},l,t}\widehat{\mathbf{h}}_{\boldsymbol{\rho},t} \to -\infty$ with probability one. That is, the dual problem (38) is unbounded from below, and consequently, the primal problem is infeasible, which is also a contradiction. Therefore, $\overline{\mathbf{B}}_{\boldsymbol{\rho},l,t} \succ \mathbf{0}$ is proved.

Finally, based on (40), (41), and $\overline{\mathbf{B}}_{\boldsymbol{\rho},l,t} \succ \mathbf{0}$, we have,

$$\begin{aligned}\mathrm{rank}(\mathbf{W}_{\boldsymbol{\rho},l,t}) &\overset{\text{(a)}}{=} \mathrm{rank}(\overline{\mathbf{B}}_{\boldsymbol{\rho},l,t}\mathbf{W}_{\boldsymbol{\rho},l,t}) \\ &\overset{\text{(b)}}{=} \mathrm{rank}\big(\tfrac{\alpha_{\boldsymbol{\rho},l,t}}{\eta^{\text{req}}_{\boldsymbol{\rho},l}}\mathbf{W}_{\boldsymbol{\rho},l,t}\widehat{\mathbf{h}}_{\boldsymbol{\rho},t}\widehat{\mathbf{h}}^H_{\boldsymbol{\rho},t}\big) \\ &\overset{\text{(c)}}{\leq} \min\left\{\mathrm{rank}\big(\tfrac{\alpha_{\boldsymbol{\rho},l,t}}{\eta^{\text{req}}_{\boldsymbol{\rho},l}}\mathbf{W}_{\boldsymbol{\rho},l,t}\big), \mathrm{rank}\big(\widehat{\mathbf{h}}_{\boldsymbol{\rho},t}\widehat{\mathbf{h}}^H_{\boldsymbol{\rho},t}\big)\right\} \\ &\leq 1,\end{aligned} \tag{43}$$

where (a) is due to $\overline{\mathbf{B}}_{\boldsymbol{\rho},l,t} \succ \mathbf{0}$, (b) is a result of (40) and (41), and (c) follows from the basic rank inequality $\mathrm{rank}(\mathbf{AB}) \leq \min\{\mathrm{rank}(\mathbf{A}), \mathrm{rank}(\mathbf{B})\}$. On the other hand, since $\mathbf{W}_{\boldsymbol{\rho},l,t} \neq \mathbf{0}$, the condition $\mathrm{rank}(\mathbf{W}_{\boldsymbol{\rho},l,t}) = 1$ holds with probability one. This completes the proof.

REFERENCES

[1] L. Xiang, D. W. K. Ng, R. Schober, and V. W. S. Wong, "Secure video streaming in heterogeneous small cell networks with untrusted cache helpers," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Singapore, Dec. 2017.

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

[3] X. Ge, H. Cheng, M. Guizani, and T. Han, "5G wireless backhaul networks: Challenges and research advances," *IEEE Netw.*, vol. 28, no. 6, pp. 6–11, Nov./Dec. 2014.

[4] P. Rost *et al.*, "Cloud technologies for flexible 5G radio access networks," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 68–76, May 2014.

[5] 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.

[6] M. Chen, M. Mozaffari, W. Saad, C. Yin, M. Debbah, and C. S. Hong, "Caching in the sky: Proactive deployment of cache-enabled unmanned aerial vehicles for optimized quality-of-experience," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 5, pp. 1046–1061, May 2017.

[7] 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, no. 12, pp. 8402–8413, Dec. 2013.

[8] D. Liu and C. Yang, "Energy efficiency of downlink networks with caching at base stations," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 907–922, Apr. 2016.

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

[10] 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. 66, no. 12, pp. 11366–11382, Dec. 2017..

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

[12] ——, "Decentralized coded caching attains order-optimal memory-rate tradeoff," *IEEE/ACM Trans. Netw.*, vol. 23, no. 4, pp. 1029–1040, Aug. 2015.

[13] J. Zhang, X. Lin, and X. Wang, "Coded caching under arbitrary popularity distributions," in *Proc. Inf. Theory and App. Workshop (ITA), 2015*, San Diego, CA, Feb. 2015.

[14] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr, "On the optimality of separation between caching and delivery in general cache networks," in *Proc. Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, Jun. 2017.

[15] A. A. Zewail and A. Yener, "Coded caching for combination networks with cache-aided relays," in *Proc. Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, Jun. 2017.

[16] 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," *to be published in IEEE Trans. Wireless Commun.*.

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

[18] A. A. Zewail and A. Yener, "Coded caching for resolvable networks with security requirements," in *Proc. Commun. and Netw. Security (CNS)*, Philadelphia, PA, Oct. 2016.

[19] L. Xiao, C. Xie, T. Chen, H. Dai, and H. V. Poor, "A Mobile Offloading Game Against Smart Attacks", *IEEE Access*, vol. 4, pp. 2281–2291, 2016.

[20] J. Wright and J. Cache, *Hacking Exposed Wireless: Wireless Security Secrets & Solutions*, 3rd ed. McGraw-Hill Education Group, 2015.

[21] G. Mantas, N. Komninos, J. Rodriuez, E. Logota, and H. Marques, "Security for 5G communications," in *Fundamentals of 5G Mobile Networks*, J. Rodriguez, Ed. John Wiley & Sons, Ltd., 2015.

[22] 3GPP TR 33.820 v8.3.0, "Technical specification group service and system aspects: Security of H(e)NB (Release 8)," Dec. 2009.

[23] E. Rescorla, "HTTP over TLS," *IETF RFC 2818*, May 2000.

[24] F. Callegati, W. Cerroni, and M. Ramilli, "Man-in-the-middle attack to the HTTPS protocol," *IEEE Security Privacy*, vol. 7, no. 1, pp. 78–81, Jan./Feb. 2009.

[25] X. He and A. Yener, "Cooperation with an untrusted relay: A secrecy perspective," *IEEE Trans. Inf. Theory*, vol. 56, no. 8, pp. 3807–3827, Aug. 2010.

[26] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for mobile networks: A technology overview," *IEEE Commun. Surveys & Tuts.*, vol. 17, no. 1, pp. 405–426, First Quarter 2015.

[27] 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.

[28] 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.

[29] 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, Sept 2015.

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

[31] 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.

[32] J.-R. Ohm, "Advances in scalable video coding," *Proc. IEEE*, vol. 93, no. 1, pp. 42–56, Jan. 2005.

[33] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H. 264/AVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.

[34] S. de la Fuente *et al.*, "iDASH: improved dynamic adaptive streaming over HTTP using scalable video coding," in *Proc. ACM MMSys*, California, CA, Feb. 2011.

[35] Y. Zhang, N. Meratnia, and P. Havinga, "Outlier detection techniques for wireless sensor networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 12, no. 2, pp. 159–170, 2010.

[36] P. Gogoi, D. Bhattacharyya, B. Borah, and J. K. Kalita, "A survey of outlier detection methods in network anomaly identification," *The Comput. J.*, vol. 54, no. 4, pp. 570–588, Apr. 2011.

[37] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artif. Intell. Review*, vol. 22, no. 2, pp. 85–126, Oct. 2004.

[38] D. Bharadia, E. McMilin, and S. Katti, "Full duplex radios," *ACM SIGCOMM Comput. Commun. Review*, vol. 43, no. 4, pp. 375–386, 2013.

[39] B. P. Day, A. R. Margetts, D. W. Bliss, and P. Schniter, "Full-duplex bidirectional MIMO: Achievable rates under limited dynamic range," *IEEE Trans. Signal Process.*, vol. 60, no. 7, pp. 3702–3713, Jul. 2012.

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

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

[42] D. P. Palomar and Y. C. Eldar, *Convex Optimization in Signal Processing and Communications*. Cambridge University Press, 2010.

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

[44] X. Ge, S. Tu, G. Mao, C.-X. Wang and T. Han, "5G Ultra-Dense Cellular Networks," *IEEE Wireless Commun.*, vol. 23, no. 1, pp. 72–79, Feb. 2016.

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

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

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

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

[50] B. Korte and J. Vygen, *Combinatorial Optimization*, 5th ed. Springer, 2012.

[51] N. D. Sidiropoulos, T. .N. Davidson, and Z. Luo, "Transmit beamforming for physical-layer multicasting," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2239–2251, Jun. 2006.

[52] 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.

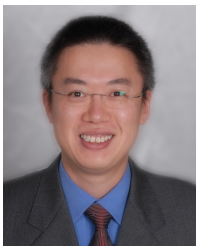
[53] 3GPP TR 36.814 v9.0.0, "Further advancements for E-UTRA physical layer aspects (Release 9)," Mar. 2010.

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 currently 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, SM'17) 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.