Contract-Based Small-Cell Caching for Data Disseminations in Ultra-Dense Cellular Networks

Jun Li, *Senior Member, IEEE*, Shunfeng Chu, Feng Shu, *Member, IEEE*, Jun Wu, *Senior Member, IEEE*, and Dushantha Nalin K. Jayakody, *Senior Member, IEEE*

Abstract—Evidence indicates that demands from mobile users (MU) on popular cloud content, e.g., video clips, account for a dramatic increase in data traffic over cellular networks. The repetitive downloading of hot content from cloud servers will inevitably bring a vast quantity of redundant data transmissions to networks. A strategy of distributively pre-storing popular cloud content in the memories of small-cell base stations (SBS), namely, small-cell caching, is an efficient technology for reducing the communication latency whilst mitigating the redundant data streaming substantially. In this paper, we establish a commercialized small-cell caching system consisting of a network service provider (NSP), several video providers (VP), and randomly distributed MUs. We conceive this system in the context of *5G* cellular networks, where the SBSs are ultra-densely deployed with the intensity much higher than that of the MUs. In such a system, the NSP, in charge of the SBSs, wishes to lease these SBSs to the VPs for the purpose of making profits, whilst the VPs, after pushing popular videos into the rented SBSs, can provide faster local video transmissions to the MUs, thereby gaining more profits. Specifically, we first model the MUs and SBSs as two independent Poisson point processes, and develop, via stochastic geometry theory, the probability of the specific event that an MU obtains the video of its choice directly from the memory of an SBS. Then, with the help of the probability derived, we formulate the profits of both the NSP and the VPs. Next, we solve the profit maximization problem based on the framework of contract theory, where the NSP acts as a monopolist setting up the optimal contract according to the statistical information of the VPs. Incentive mechanisms are also designed to motivate each VP to choose a proper resource-price item offered by the NSP. Numerical results validate the effectiveness of our proposed contract framework for the commercial caching system.

Index Terms—Wireless caching, ultra-dense small-cell networks, stochastic geometry, poisson point process, contract theory

1 Introduction

Wireless data traffic is expected to increase exponentially in the next few years, driven by a staggering growth of mobile users (MU) and their bandwidth-hungry mobile applications. Researchers and operators are now beginning to conceive of the next generation cellular networks beyond 2020, known as *5G*, to cope with the data avalanche [1]. A key disruptive aspect of *5G* is the exploitation of the enhanced intelligence of smart mobile devices to implement the *Mobile Cloud Computing* (MCC) infrastructure [2]. In essence, MCC shifts part of the data storage and processing from the mobile devices to the cloud, thereby alleviating the heavy burden of the storage and computational power on these devices.

In the conventional MCC structure, the cloud is centralised and composed of powerful server clusters within the Internet. However, these server clusters are far away from the MUs, leading to high latency of transmissions. At the same time, there are large numbers of duplicate demands on popular content, e.g., high quality videos, stored in the cloud [3]. As such, conventional MCC will suffer from redundant data traffic arising from repetitive data requests. Our vision of efficient cloud data dissemination in future *5G* networks is low transmission latency and near zero redundant data traffic in the backbone network. A key step towards our vision is transparently storing data from the cloud to the memories of local network nodes, forming a distributed caching system [3], [4], [5], such that the MUs can obtain data from adjacent nodes. The advantages of a distributed caching system include bringing content closer to the MUs and alleviating redundant data transmissions via redirecting the downloading requests to the local nodes.

Generally, the caching process consists of two stages: a data placement stage and a data delivery stage [6]. In the first stage, part of popular cloud data are cached into local

- *J. Li is with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China, the National Mobile Communications Research Laboratory, Southeast University, Nanjing 211189, China, and the School of Computer Science and Robotics, National Research Tomsk Polytechnic University, Tomsk 634050, Russia. E-mail: jun.li@njust.edu.cn.*
- *S. Chu is with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China. E-mail: shunfeng.chu@njust.edu.cn.*
- *F. Shu is with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China, and the College of Computer and Information Sciences, Fujian Agriculture and Forestry University, Fuzhou 350002, China. E-mail: shufeng@njust.edu.cn.*
- *J. Wu is with the Department of Computer Science and Technology, Tongji University, Shanghai 200092, China. E-mail: wujun@tongji.edu.cn.*
- *D.N.K. Jayakody is with the School of Science and Robotics, National Research Tomsk Polytechnic University, Tomsk 634050, Russia. E-mail: nalin@tpu.ru.*

Manuscript received 28 Oct. 2016, revised 24 Apr. 2018, accepted 2 July 2018. Date of publication 6 July 2018; date of current version 1 Apr. 2019.
(Corresponding author: Jun Li, Feng Shu, and Jun Wu.)
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TMC.2018.2853746

storages during off-peak time, while in the second stage, requested videos are delivered from the caching system to the MUs. Recent works advance the caching technology in device-to-device (D2D) networks and wireless sensor networks [7], [8], [9]. Specifically in [7], a caching scheme is proposed for a D2D based cellular network on the MUs' caching of popular video content. In this scheme, the D2D cluster size is optimized for reducing the downloading delay. In [8], [9], the authors propose novel caching schemes for wireless sensor networks, where the protocol model of [10] is adopted.

As small-cell embedded architectures will be prevailing in future cellular networks, known as heterogeneous networks (HetNet) [11], [12], [13], [14], [15], [16], caching relying on small-cell base stations (SBS), namely, small-cell caching, is a promising trend for the HetNets. The advantages brought by the small-cell caching are threefold. First, popular videos are pushed closer to the MUs when cached in SBSs, reducing the transmission latency. Second, redundant data transmissions over backbone networks, along with network congestions, are mitigated. Third, the majority of video traffic is offloaded from macro-cell base stations to SBSs. In [17], a small-cell caching scheme, called 'Femtocaching', is proposed for a cellular network embedded with SBSs, where the data placement at the SBSs is optimized in a centralized manner for reducing the transmission delay imposed. In [18], the small-cell caching is investigated in the context of stochastic networks. The average performance is developed via stochastic geometry [19], [20], where the distribution of network nodes are modeled by Poisson point process (PPP).

From above discussions, current research on wireless caching mainly considers the data placement issue. However, the whole caching system is coupled with many issues other than data placement. From commercial perspective, it will be more interesting to consider the topics such as pricing on video streaming, renting local storages, and so on. A commercialized caching system may consist of network service providers (NSP), video providers (VP), and MUs. The VR s, e.g., Youtube, purchase copyrights from video producers and publish these videos on their web-sites. The NSPs are typically operators of cellular networks, who are in charge of network facilities such as SBSs.

In such a commercial caching system, the VPs make profits by providing video streaming services to the MUs. As the central servers of the VPs, which store the popular videos, are usually located at backbone networks and far away from the MUs, an efficient solution is to locally cache these videos for reducing the transmission latency, thereby attracting more customers. These local caching demands raised by the VPs offer the NSPs profitable opportunities from leasing their resources, i.e., the SBSs. In this sense, both the VPs and NSPs are the beneficiaries from the local caching system. However, each entity is selfish and wishes to maximize its own benefit, rasing a competition problem. In [21], the authors propose a Stackelberg game to optimize the prices and resource allocations for maximizing the profits of both the NSP and the VPs. However, in the Stackelberg game, the NSP will need to know about detailed information, e.g., the optimal resource allocation solutions at the VPs, for working out an optimal pricing strategy. In the case that the NSP has incomplete information about the VPs, the Stackelberg game will not be functional.

While the NSP is the monopolist who controls all the trading resources, i.e., the SBSs, in the network, it is more reasonable to assume that the NSP unilaterally designs the incentive items for attracting the VPs renting the SBSs. As a well-known market driven mechanism, contract theory is effective to design incentive schemes under asymmetric information scenario [22], which has been widely adopted by the researchers in area of wireless communications. For example, in order to make dynamic pricing for idle spectrum resource, the authors in [23] model the spectrum trading process as monopoly market and come up with optimal contract for the primary user. In [24], the authors tackle the cooperative spectrum sharing between one primary user and multiple secondary users based on contract theory.

In this paper, we research on a commercialized caching prototype within a contract theory framework. The system consists of an NSP and multiple VPs, where the NSP, as the monopolist in the market in charge of the trading resource, wishes to lease its SBSs to the VPs for the purpose of making profits. Meanwhile the VPs, after pushing popular videos into the rented SBSs, can also gain profit from providing faster video downloading services to the MUs. In the context of contract theory, the NSP dominates the market and offers a series of contract items for the VRs to accept. These contract items are designed by the NSP for maximizing its own profit, while a VR has no choices but passively picks out the one that maximizes this VR's profit. We conceive this system for future *5G* cellular networks, with SBSs being much denser than the MUs.

To the best of the authors' knowledge, our work is the first of its kind that investigates wireless caching in the context of contract theory. The contributions of this paper are summarized as follows. First, we model the MUs and SBSs as two independent Poisson point processes, and develop, via stochastic geometry theory, the probability of the specific event that an MU obtains the video of its choice directly from the memory of an SBS. Then, with the help of the probability derived, we formulate the average profits, in terms of per unit area and unit period, of both the NSP and the VRs. We further classify the VRs with different types according to their popularity to the MUs. Next, we propose an incentive mechanism for motivating the VRs to participate into the market under the strongly incomplete information scenario, where the NSP only knows the statistical information regarding the type distribution of the VRs. Afterwards, we develop the optimal contract that can maximize the NSP's profit as well as guarantee the feasibility of for the VRs. The optimal contract is achieved via a serial of feasibility studies, followed by a novel algorithm for solving the profit maximization problem. Numerical results validate our contract framework by showing its effectiveness in improving the NSP's profit.

The rest of this paper is organized as follows. We describe the caching system model in Section 2 and analyze the caching performance of the system in Section 3. We then formulate contract based problem in Section 4. In Section 5, we design the optimal contract to maximize the NSP's profit along with effect incentive schemes for attracting the VPs. Our numerical results are detailed in Section 6, while our conclusions are provided in Section 7.

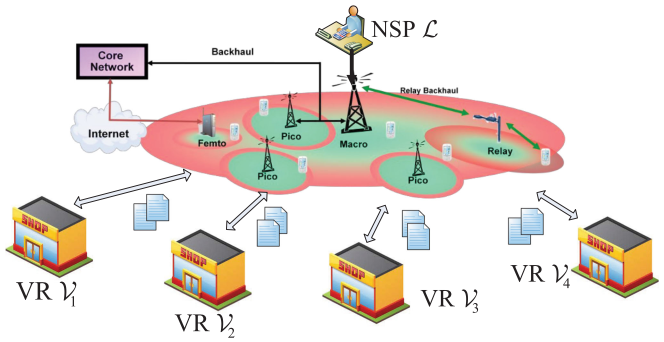


Fig. 1. An example of the small-cell caching system with four VPs.

2 System Model

We consider a commercial small-cell caching system consisting of an NSP, V VPs, and a number of MUs. Let us denote by $\mathcal{L}$ the NSP, by $\boldsymbol{\mathcal{V}} = \{\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_V\}$ the set of the VPs, and by $\mathcal{M}$ one of the MUs. Fig. 1 shows the prototype of a caching system with four VPs. In such a system, the VPs negotiate with the NSP for renting part of its SBSs and caching popular videos. Both the NSP and each VP aim for maximizing their profits.

2.1 Network Model

In future *5G* cellular networks with ultra-dense small cells, the number of SBSs is expected to be much greater than the number of MUs [25]. Let us consider an ultra-dense small-cell caching network composed of the MUs and densely deployed SBSs owned by $\mathcal{L}$, where each SBS is equipped with a fixed volume of storage for caching Q video files. We assume that the SBSs transmit over the channels that are orthogonal to those of the macro-cell base stations, and thus there is no interference incurred by the macro-cell base stations. Also, we assume that these SBSs are spatially distributed according to a homogeneous PPP (HPPP) Φ of intensity λ. Here, the intensity λ represents the number of the SBSs per unit area. Furthermore, we model the distribution of the MUs as an independent HPPP Ψ of intensity ζ.

The wireless down-link channels spanning from the SBSs to the MUs are independent and identically distributed (*i.i.d.*), and modeled as the combination of path-loss and Rayleigh fading. Without a loss of generality, we carry out our analysis for a typical MU located at the origin. The path-loss between an SBS located at x and the typical MU is denoted by $\|x\|^{-\alpha}$, where α is the path-loss exponent. The channel power of the Rayleigh fading between them is denoted by h_x, where $h_x \sim \exp(1)$. The noise at an MU is Gaussian distributed with a variance σ^2.

To alleviate interference among the densely deployed SBSs, we adopt the dynamic on-off architecture [26], where an SBS will switch to the idle mode, i.e., turn off its radio transmissions, if there is no MU associated with it for video downloading. This dynamic on-off architecture is proposed and under investigation in *3GPP* as an important candidate of *5G* technologies aiming to mitigate the severe inter-cell interference in the future ultra-dense small-cell networks.

2.2 Preference and Affiliation

It is natural that the MUs have different appetites or preferences for various types of videos: Some videos may be downloaded by the majority of the MUs, while others may be rarely requested. We now model the preference distribution, i.e., the distribution of request probabilities, among the popular videos to be cached. Let us denote by $\mathcal{F} = \{\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_F\}$ the file set consisting of F video files, where each video file contains an individual movie or video clip that is frequently requested by MUs. The popularity distribution of $\mathcal{F}$ is represented by a vector $\mathbf{p} = [p_1, p_2, \ldots, p_N]$. That is, the MUs make independent requests of the fth video $\mathcal{F}_f$, $f = 1, \ldots, F$, with the probability of p_f. Generally, $\mathbf{p}$ can be modeled by the Zipf distribution [27] as

$$p_f = \frac{1/f^{\beta}}{\sum_{j=1}^{F} 1/j^{\beta}}, \quad \forall f, \tag{1}$$

where the exponent β is a positive value, characterizing the video popularity. A higher β corresponds to a higher content reuse, where the most popular files account for the majority of download requests. Note that the use of Zipf is a lower bound of what a stochastic process is for modeling the arrival of files. From Eq. (1), a video file with a smaller f corresponds to a higher popularity.

At the same time, the MUs have imbalanced affiliations with regard to the V VPs, i.e., some VPs have more mobile customers than others. For example, the majority of the MUs may tend to access Youtube for video streaming. The affiliation distribution among the VPs is denoted by $\mathbf{q} = [q_1, q_2, \ldots, q_V]$, where q_v, $v = 1, \ldots, V$, represents the probability that an MU is affiliated with $\mathcal{V}_v$. The affiliation distribution $\mathbf{q}$ can also be modeled by the Zipf distribution. Hence, we have

$$q_v = \frac{1/v^{\gamma}}{\sum_{j=1}^{V} 1/j^{\gamma}}, \quad \forall v, \tag{2}$$

where γ is a positive value, characterizing the preference of the VPs. A higher γ corresponds to a higher probability of accessing the most popular VPs.

Note that Zipf distribution are widely used in many rankings, such as city populations in various countries, corporation sizes, numbers of people subscribing different TV channels [28], [29]. Therefore, in this paper we utilize it to represent the popularity ranking of both the network content and the VPs. As we will discuss in Section 4, the distribution of the popularity among the VPs will define the types of these VPs, which is an important parameter in the contract theory.

3 Analysis on Caching Performance

Consider three stages in the caching system. In the first stage, the VPs purchase the copyrights of popular videos from video producers and publish them on their web-sites. In the second stage, the VPs negotiate with the NSP on renting part of its SBSs. In the third stage, the MUs connect to the SBSs for downloading the desired videos. We will particulary focus our attention on the second and third stages within a contract theory-based framework.

In the second stage, upon obtaining the popular videos, the VPs negotiate with the NSP for renting its SBSs. Since the SBSs in the network are modeled stochastically, the VPs are not interested in a particular SBS, but a fraction of the SBSs. As $\mathcal{L}$ leases its SBSs to multiple VPs, we denote by $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_V]$ the intensity vector of the rented SBSs, where λ_v represents the intensity of the SBSs that are rented by $\mathcal{V}_v$, $\forall v$. We assume that the SBSs rented by each VP are uniformly distributed. Hence, the SBSs that are allocated to $\mathcal{V}_v$ can be modeled as a "thinned" HPPP Φ_v with intensity λ_v. We have $\sum_{v=1}^{V} \lambda_v \leq \lambda$. The data placements of the second stage commence during network off-peak time after the VPs obtain access to their rented SBSs. For simplicity, we assume that each of the rented SBSs will be required to cache the Q most popular video clips $\mathcal{F}_1, \ldots, \mathcal{F}_Q$.

We note that [30] considers the caching optimization problem with a limited storage size within one base station maintained by an NSP and shared by multiple VPs. The NSP focuses on the issue of storage resource allocation, i.e., partitioning the cache into multiple pieces and then allocating them to different content providers. However in our paper, the case is that there are a huge number of SBSs with limited storage, while the VPs have the same file library. If the storage in an SBS is shared by multiple VPs, it is unavoidable that the VPs will save the same files into the same SBS, thereby leading to a waste of the caching space and reducing the efficiency of resource allocation. Therefore, our assumption that a VP is interested in renting the whole SBS is more efficient.

We view the SBSs rented by $\mathcal{V}_v$ combined with the MUs affiliated with $\mathcal{V}_v$ as the vth tier, namely, Tier-v, where the intensities of these SBSs and MUs are λ_v and $q_v\zeta$, respectively. In the dynamic on-off architecture, some of the SBSs may be activated for transmitting data, while other SBSs may be in an idle mode. We denoted by $\mathcal{E}_v^{\text{active}}$ the event that an SBS in Tier-v is active, and its probability can be approximately expressed by [31]

$$\Pr(\mathcal{E}_v^{\text{active}}) \approx 1 - \left(1 + \frac{q_v\zeta}{3.5\lambda_v}\right)^{-3.5}. \tag{3}$$

We assume that the SBSs from $\mathcal{V}_v$ are allocated with a power P_v. Hence, the received signal-to-interference-plus-noise ratio (SINR) at the typical MU from an SBS in Φ_v located at x can be expressed as

$$\rho(x) = \frac{P_v h_x \|x\|^{-\alpha}}{I_1 + I_2 + \sigma^2}, \tag{4}$$

where

$$I_1 = \sum_{v' \neq v} \sum_{x' \in \Phi_{v'}} \Pr(\mathcal{E}_{v'}^{\text{active}}) P_{v'} h_{x'} \|x'\|^{-\alpha}, \tag{5}$$

is the interference imposed by the SBSs from other $V-1$ tiers, while

$$I_2 = \sum_{x' \in \Phi_v \setminus x} \Pr(\mathcal{E}_{v'}^{\text{active}}) P_v h_{x'} \|x'\|^{-\alpha}, \tag{6}$$

is the interference from other SBSs in the same tier.

The typical MU is considered to be "covered" by an SBS located at x as long as $\rho(x)$ is no lower than a pre-set SINR threshold δ, i.e.,

$$\rho(x) \geq \delta. \tag{7}$$

Generally, an MU can be covered by multiple SBSs. Note that the SINR threshold δ defines the highest delay of downloading a video file. Since the quality and code rate of a video clip have been specified within the video file, the download delay will be the major factor predetermining the QoS perceived by the mobile customers.

In the third stage, the MUs start to download videos from nearby SBSs. We note that in [3], [32] an MU can associate with multiple SBSs for searching its desired file. This is a good strategy for increasing the locally downloading probability. However, in practical cellular systems, an MU generally associates with only one SBS for data transmissions to simplify the protocol design and interference management. When an MU $\mathcal{M}$ affiliated with $\mathcal{V}_v$ requires a video clip from $\mathcal{V}_v$, it searches the SBSs in Φ_v and tries to connect to the nearest SBS that covers $\mathcal{M}$. Provided that such an SBS exists, the MU $\mathcal{M}$ will obtain this video directly from this SBS, and we thereby define this event by $\mathcal{E}_v^{\text{cover}}$. On the other hand, if such an SBS does not exist, $\mathcal{M}$ will be redirected to the central servers of $\mathcal{V}_v$ for downloading the requested file. Since the servers of $\mathcal{V}_v$ are located at the backbone network, this redirection of the demand will trigger a transmission via the back-haul channels of the NSP $\mathcal{L}$, hence leading to an extra cost.

When we consider coverage performance, we assume that the channels between the MUs and SBSs are stable or in slow fading, i.e., the change of SINR is not so frequent [20]. This is reasonable in the case that the users are static or move slowly. Furthermore, the Line-of-Sight (LoS) channels dominate in ultra-dense networks. Based on this assumption, we consider the stable channel during the transmission of a video file [21]. We are particularly interested in the analysis on the probability $\Pr(\mathcal{E}_v^{\text{cover}})$ based on stochastic geometry, and we have the following theorem.

Theorem 1. *The probability of the event $\mathcal{E}_v^{\text{cover}}$, i.e., the event that an MU in Tier v can successfully download a video file from its nearby SBS, can be expressed as*

$$\begin{aligned}\Pr(\mathcal{E}_v^{\text{cover}}) &= \sum_{f=1}^{Q} p_f \\ &\cdot \int_0^\infty \prod_{v'} \exp\left(-\pi \Pr(\mathcal{E}_{v'}^{\text{active}})\lambda_{v'} \left(\frac{P_{v'}}{P_v}\right)^{\frac{2}{\alpha}} C(\delta,\alpha) z^2\right) \\ &\quad \exp\left(-\pi \Pr(\mathcal{E}_v^{\text{active}})\lambda_v A(\delta,\alpha) z^2\right) \\ &\quad \exp\left(-\frac{z^\alpha \delta}{P_v}\sigma^2\right)\pi\lambda_v \exp\left(-\pi\lambda_v z^2\right) \mathrm{d}z^2,\end{aligned} \tag{8}$$

where we have

$$\begin{aligned} A(\delta,\alpha) &\triangleq \frac{2\delta}{\alpha-2}\, {}_2F_1\left(1, 1-\frac{2}{\alpha}; 2-\frac{2}{\alpha}; -\delta\right), \\ C(\delta,\alpha) &\triangleq \frac{2}{\alpha}\delta^{\frac{2}{\alpha}} B\left(\frac{2}{\alpha}, 1-\frac{2}{\alpha}\right). \end{aligned} \tag{9}$$

Furthermore, ${}_2F_1(\cdot)$ in the function $A(\delta,\alpha)$ is the hypergeometric function and the Beta function in $C(\delta,\alpha)$ is formulated as $B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\mathrm{d}t$.

Proof. The proof borrows the idea of [19], [20] for stochastic geometry analysis. Detailed can be refereed to Appendix A, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TMC.2018.2853746. □

Generally, the power of interference in a network is much greater than that of the noises. By assuming that $\frac{\sigma^2}{P_v}$ goes to zero, we have

$$\lim_{\frac{\sigma^2}{P_v}\to 0} \exp\left(-\frac{z^\alpha \delta \sigma^2}{P_v}\right) = 1. \tag{10}$$

Therefore, we can further simplify Eq. (8), in the case that $\frac{\sigma^2}{P_v}$ goes to zero, as

$$\Pr(\mathcal{E}_v^{\text{cover}}) = \frac{\sum_{f=1}^{Q} p_f \lambda_v}{(\Pr(\mathcal{E}_v^{\text{active}})A(\delta,\alpha)+1)\lambda_v + \sum_{v'} \Pr(\mathcal{E}_{v'}^{\text{active}})C(\delta,\alpha)\lambda_{v'} \frac{P_{v'}^{\frac{2}{\alpha}}}{P_v^{\frac{2}{\alpha}}}}. \tag{11}$$

The expression of $\Pr(\mathcal{E}_v^{\text{active}})$ in Eq. (3) can be transformed by its Taylor series. Since in our system, the SBSs are densely deployed, and we have $\lambda_v \gg q_v\zeta$, we ignore the high order items in the Taylor series. Then we further approximate $\Pr(\mathcal{E}_v^{\text{active}})$ as

$$\Pr(\mathcal{E}_v^{\text{active}}) \approx \frac{q_v \zeta}{\lambda_v}. \tag{12}$$

Generally, there is an uniform transmission power of P allocated to each SBS in a cellular network, i.e., $P_1 = P_2 = \cdots = P_V = P$. In this case, we substitute Eq. (12) into Eq. (10), and obtain a more concise form of $\Pr(\mathcal{E}_v^{\text{cover}})$ as

$$\begin{aligned}\Pr(\mathcal{E}_v^{\text{cover}}) &= \frac{\sum_{f=1}^{Q} p_f \lambda_v}{\lambda_v + q_v A(\delta,\alpha)\zeta + \sum_{v'} C(\delta,\alpha) q_{v'}\zeta} \\ &= \frac{\sum_{f=1}^{Q} p_f \lambda_v}{\lambda_v + q_v A(\delta,\alpha)\zeta + (1-q_v)C(\delta,\alpha)\zeta}.\end{aligned} \tag{13}$$

Remark 1. From Eq. (13) we can see that $\Pr(\mathcal{E}_v^{\text{cover}})$ is upper bounded by $\sum_{f=1}^{Q} p_f$, i.e., $\lim_{\lambda_v\to\infty} \Pr(\mathcal{E}_v^{\text{cover}}) = \sum_{f=1}^{Q} p_f$. That is, when the intensity of the SBSs goes to infinity, the MUs can download any file as long as it is cached by the SBSs.

In the following, we will design and optimize the caching system within the framework of the contract theory based on this successful downloading probability $\Pr(\mathcal{E}_v^{\text{cover}})$.

4 Contract-Based Problem Formulation

In our contract-based caching system, the NSP is in charge of the trading resources, i.e., the SBSs. Hence, trading on the SBSs is a monopoly market, where the NSP, as the monopolist, dominates the trading process, and the VPs act as the consumers.

4.1 Monopolist Model

The monopolist defines the trading commodities, as well as their qualities and prices. To be specific, the entire group of the SBSs in Φ_v, $\forall v$, is treated as a piece of commodity for lease. We denote the set of the V SBS groups by the vector $\mathbf{\Phi} \triangleq \{\Phi_1, \ldots, \Phi_V\}$. Here, for simplicity, we abuse the PPP Φ_v as the notation of the commodity rented by $\mathcal{V}_v$. Afterwards, the NSP sets the qualities and prices for these commodities. It is natural that we utilize the intensity λ_v of Φ_v to represent its quality. We assume that if a VP $\mathcal{V}_v$ rents the SBSs in Φ_v with the quality of λ_v, a payment $\pi(\lambda_v)$ will be charged by the NSP for a unit period. Consequently, we denote by $\mathbf{\Lambda} \triangleq \{\lambda_1, \ldots, \lambda_V\}$ the quality set, and by $\mathbf{\Pi} \triangleq \{\pi(\lambda_1), \ldots, \pi(\lambda_V)\}$ the price set for commodity set $\mathbf{\Phi}$.

We now investigate the profit of the NSP, S_v^{NSP}, gained by leasing its SBSs in Φ_v, and we have

$$S_v^{\text{NSP}}(\lambda_v) = \pi(\lambda_v) - s^{\text{cost}}\lambda_v, \tag{14}$$

where s^{cost} is the cost for the maintaining the cache (e.g., hard disks) of an SBS by the NSP during a unit period. The overall profit of the NSP can thus be expressed by

$$S^{\text{NSP}} = \sum_{v=1}^{V} S_v^{\text{NSP}}(\lambda_v). \tag{15}$$

The objective of the NSP is to maximize its profit via optimizing the qualities of its resources and the corresponding prices charged. Obviously, a rational NSP will not accept a negative S^{NSP}, and thus it always makes $\pi(\lambda_v) \geq s^{\text{cost}}\lambda_v$ when optimizing $\mathbf{\Lambda}$ and $\mathbf{\Pi}$. In some occasions, the NSP sets $\lambda_v = 0$ to denote a aborted trading process. This happens when either the NSP does not wish to lease the commodity Φ_v or the VP $\mathcal{V}_v$ gives up renting Φ_v, due to the negative profits. Correspondingly, the price will be set as $\pi(0) = 0$.

4.2 Consumer Model

We assume that each VP prefers a higher allocation of SBSs for achieving a better coverage on its affiliated MUs. According to [21], the revenue of $\mathcal{V}_v$, $R(\lambda_v)$, stems from providing fast and local video downloading services to these MUs. We have

$$\begin{aligned}R(\lambda_v) &= c q_v \zeta \Pr(\mathcal{E}_v^{\text{cover}}) s^{\text{cache}} \\ &= \frac{(c\zeta s^{\text{cache}} \sum_{f=1}^{Q} p_f) q_v \lambda_v}{\lambda_v + q_v A(\delta,\alpha)\zeta + (1-q_v)C(\delta,\alpha)\zeta},\end{aligned} \tag{16}$$

where c is the average number of video downloading requests of each MU during a unit period, and thus the item $c q_v \zeta \Pr(\mathcal{E}_v^{\text{cover}})$ represents the overall quantity of video downloading from $\mathcal{V}_v$'s local caching system per unit period. Furthermore, s^{cache} in Eq. (16) is the profit gained by $\mathcal{V}_v$ on an MU for providing local downloading of a video.

To facilitate the analysis, we further classify the VPs into different types according to the Zipf distribution. That is, the type of $\mathcal{V}_v$ is represented by the corresponding probability q_v, and the vector $\mathbf{q}$ can be viewed as the set of all types. Thus, there are overall V types, with each type containing one VP. From the second equation of Eq. (16), $R(\lambda_v)$ can also be viewed as a function of q_v. Note that in this paper, we focus on ultra-dense small-cell networks with a large enough area. Therefore, the density of SBSs can be viewed as continuous in the following contract-theoretic derivations. We have

$$\begin{aligned}\frac{\partial R(\lambda_v, q_v)}{\partial \lambda_v} &= \frac{c\zeta s^{\text{cache}} \sum_{f=1}^{Q} p_f q_v (q_v A(\delta,\alpha)\zeta + (1-q_v)C(\delta,\alpha)\zeta)}{(\lambda_v + q_v A(\delta,\alpha)\zeta + (1-q_v)C(\delta,\alpha)\zeta)^2}, \\ \frac{\partial R(\lambda_v, q_v)}{\partial q_v} &= \frac{c\zeta s^{\text{cache}} \sum_{f=1}^{Q} p_f \lambda_v (\lambda_v + C(\delta,\alpha)\zeta)}{(\lambda_v + q_v A(\delta,\alpha)\zeta + (1-q_v)C(\delta,\alpha)\zeta)^2}.\end{aligned} \tag{17}$$

Observing the above equations, it is easy to verify $\frac{\partial R(\lambda_v,q_v)}{\partial \lambda_v} > 0$ and $\frac{\partial R(\lambda_v,q_v)}{\partial q_v} > 0$. From the first equation of Eq. (17), the VP $\mathcal{V}_v$ prefers a higher λ_v, to achieve a greater revenue $R(\lambda_v)$, while from the second equation, considering two types of VPs V_{v_1} and V_{v_2} with $q_{v_1} > q_{v_2}$, if they rent the SBSs with the same quality, say $\lambda_{v_1} = \lambda_{v_2}$, there is $R(\lambda_{v_1}, q_{v_1}) > R(\lambda_{v_2}, q_{v_2})$.

We now investigate the relationship between $R(\lambda_v, q_v)$, v, and the Zipf parameter γ. First, we have

$$\frac{\partial q_v}{\partial \gamma} = \frac{v^{-\gamma}\left(\sum_{j=1}^{V} j^{-\gamma}\ln j - \ln v \sum_{j=1}^{V} j^{-\gamma}\right)}{\left(\sum_{j=1}^{V} j^{-\gamma}\right)^2}. \tag{18}$$

Then combined with $\frac{\partial R(\lambda_v,q_v)}{\partial q_v}$ in Eq. (17), we have the following remark.

Remark 2. In the case of $v < \exp\left\{\frac{\sum_{j=1}^{V} j^{-\gamma}\ln j}{\sum_{j=1}^{V} j^{-\gamma}}\right\}$, we have $\frac{\partial R(\lambda_v,q_v)}{\partial \gamma} < 0$, i.e., $R(\lambda_v, q_v)$ is an increasing function of γ; Otherwise, $R(\lambda_v, q_v)$ is a decreasing function of γ.

We can further obtain

$$\begin{aligned}&\frac{\partial^2 R(\lambda_v, q_v)}{\partial \lambda_v^2}\\ &= \frac{-2c\zeta s^{\text{cache}}\sum_{f=1}^{Q} p_f q_v (q_v A(\delta,\alpha)\zeta + (1-q_v)C(\delta,\alpha)\zeta)}{(\lambda_v + q_v A(\delta,\alpha)\zeta + (1-q_v)C(\delta,\alpha)\zeta)^3}.\end{aligned} \tag{19}$$

Obviously, we have $\frac{\partial^2 R(\lambda_v,q_v)}{\partial \lambda_v^2} < 0$ which means $R(\lambda_v, q_v)$ is a concave function of λ_v.

On the other hand, the VP $\mathcal{V}_v$ needs to pay for renting the SBSs, which is $\pi(\lambda_v)$ per unit period. Therefore, the net profit of the VP $\mathcal{V}_v$ can be expressed as

$$S^{\text{VP}}(\lambda_v, q_v) = R(\lambda_v, q_v) - \pi(\lambda_v). \tag{20}$$

Each VP is selfish and rational, whose objective is to maximize its profit during the resource trading.

4.3 Social Welfare

The social welfare of a commodity Φ_v is defined as the sum profit of the NSP and $\mathcal{V}_v$ from trading on Φ_v, i.e.,

$$\begin{aligned}W(\lambda_v, q_v) &= S_v^{\text{NSP}}(\lambda_v) + S^{\text{VP}}(\lambda_v, q_v)\\ &= R(\lambda_v, q_v) - s^{\text{cost}}\lambda_v.\end{aligned} \tag{21}$$

Since there is $\frac{\partial^2 R(\lambda_v,q_v)}{\partial \lambda_v^2} < 0$ as shown in (19), we can obtain $\frac{\partial^2 W(\lambda_v,q_v)}{\partial \lambda_v^2} < 0$ as well, i.e., the social welfare $W(\lambda_v, q_v)$ is also a concave function of λ_v. Therefore, maximization of $W(\lambda_v, q_v)$ can be achieved by solving

$$\frac{\partial W(\lambda_v, q_v)}{\partial \lambda_v} = 0. \tag{22}$$

In an open market, the optimal scheme in terms of social welfare may not be adopted by the NSP and the VPs, since they are selfish and only focus on maximizing their own profits. Nevertheless, the social welfare provides a upper bound of the combined profit of the NSP and the VPs.

4.4 Contracts Formulation

When designing contracts with the VPs, the NSP optimizes the quality set $\mathbf{\Lambda}$ and the corresponding price set $\mathbf{\Pi}$ for maximizing its profit. Note that it is a one-to-one mapping between the quality set $\mathbf{\Lambda}$ and the type set $\mathbf{q}$. That is, for the VP $\mathcal{V}_v$, it will be assigned with the quality $\lambda_v(q_v)$ with a price $\pi(\lambda_v(q_v))$. We henceforth utilize π_v to denote the price $\pi(\lambda_v(q_v))$ for simplicity. We refer to the combination of the quality and price sets as quality-price contract, denoted by $\Xi = \{\lambda_v, \pi_v | \forall v\}$.

A feasible contract is a set of quality-price combinations, in which the VP $\mathcal{V}_v$ with the type q_v, $\forall v$, prefers the product with quality λ_v at price π_v to any other product with a different quality. To be specific, each VP finds it in its own interest to buy the product assigned to its type, which is called incentive compatible (IC), i.e.,

$$R(\lambda_v, q_v) - \pi_v \geq R(\lambda_{v'}, q_v) - \pi_{v'}, \quad \forall v' \neq v. \tag{23}$$

Additionally, each VP is assumed to be rational to not buy a product without a positive profit, i.e.,

$$R(\lambda_v, q_v) - \pi_v \geq 0, \quad \forall v. \tag{24}$$

This property is referred to as individual rationality (IR).

A feasible contract must satisfy the IC and IR constraints, and any contract satisfying the IC and IR must be feasible. The overall profit of the NSP can be rewritten as

$$S^{\text{NSP}} = \sum_{v=1}^{V} \pi_v - s^{\text{cost}}\lambda_v. \tag{25}$$

The optimal contract, denoted by $\{(\lambda_v^\star, \pi_v^\star), \forall v\}$, is thus defined as a feasible contract that maximizes the profit of the NSP. We have

$$\{(\lambda_v^\star, \pi_v^\star), \forall v\} = \arg\max_{\lambda_v, \pi_v} \sum_{v=1}^{V} \pi_v - s^{\text{cost}}\lambda_v, \tag{26}$$

subject to the IC and IR constraints in Eqs. (23) and (24), respectively.

5 Optimal Contract Design

We have previously provided the contract model and formulate the optimal contract problem. In this section, we will focus on how to develop the optimal contract that maximizes the profit of the NSP.

In our contract, each consumer type containing only one VP specifies one quality-price contract item, and these consumer types satisfy $q_1 > q_2 > \cdots > q_V$ according to the Zipf distribution. We consider the incomplete information scenario, where the NSP is only aware of the Zipf distribution parameter γ, while it does not know the type value of a given VP. The quality-price contract items can be denoted as $\{(\lambda_v, \pi_v),\ v = \{1, 2, \ldots, V\}\}$.

5.1 Feasibility of Contract

The profit maximization problem in Eq. (26) is nontrivial, which can, however, be solved by first simplifying the IR and IC constraints before the optimization.

Lemma 1. *Regarding the optimal contract in the incomplete information scenario, IR constraint can be replaced by*

$$\frac{(c\zeta s^{\text{cache}}\sum_{f=1}^{Q} p_f) q_V \lambda_V}{\lambda_V + q_V A(\delta,\alpha)\zeta + (1-q_V)C(\delta,\alpha)\zeta} - \pi_V = 0. \tag{27}$$

given that the IC constraint holds.

Proof. Please refer to Appendix B, available in the online supplemental material. □

From the above lemma, if Eq. (27) holds, all the user types will satisfy the IR constraint. Additionally, there are overall $V(V-1)$ constraints in the IC. We further reduce these constraints and obtain the following lemma.

Lemma 2. *If all the $V(V-1)$ constraints in the IC are satisfied, then the monotonicity constraint will hold, i.e., $\lambda_{v_1} \geq \lambda_{v_2}$ if and only if the user type $q_{v_1} \geq q_{v_2}$.*

Proof. Please refer to Appendix C, available in the online supplemental material. □

Lemma 3. *For any type $q_{j_1} \geq q_{j_2}$ and $\lambda_{v_1} \geq \lambda_{v_2}$, the revenue function satisfies the following condition:*

$$\begin{aligned} &R(\lambda_{v_1}, q_{j_1}) - R(\lambda_{v_1}, q_{j_2}) \\ &\quad \geq R(\lambda_{v_2}, q_{j_1}) - R(\lambda_{v_2}, q_{j_2}). \end{aligned} \tag{28}$$

Proof. Please refer to Appendix D, available in the online supplemental material. □

Lemma 4 (LDICs: Local Downward Incentive Constraints). *If the LDICs are satisfied for the type $q_v, \forall v$, i.e.,*

$$R(\lambda_v, q_v) - \pi_v \geq R(\lambda_{v+1}, q_v) - \pi_{v+1}, \tag{29}$$

with $\lambda_v \geq \lambda_{v+1}$, then the IC constraint will hold for any $v_1 \leq v_2$, i.e.,

$$R(\lambda_{v_1}, q_{v_1}) - \pi_{v_1} \geq R(\lambda_{v_2}, q_{v_2}) - \pi_{v_2}. \tag{30}$$

Proof. Please refer to Appendix E, available in the online supplemental material. □

Lemma 5 (LUICs: Local Upward Incentive Constraints). *If the LUICs are satisfied for the type $q_v, \forall v$, i.e.,*

$$R(\lambda_v, q_v) - \pi_v \geq R(\lambda_{v-1}, q_v) - \pi_{v-1}, \tag{31}$$

with $\lambda_v \leq \lambda_{v-1}$, then the IC constraint will hold for any $v_1 \geq v_2$, i.e.,

$$R(\lambda_{v_1}, q_{v_1}) - \pi_{v_1} \geq R(\lambda_{v_2}, q_{v_1}) - \pi_{v_2}. \tag{32}$$

Proof. The proof is similar to that in *Lemma 4*. Hence we omit the duplications here. □

Lemma 6. *If the profit of the NSP is maximized, i.e., the contract is at the optimum, then the LDICs must satisfy the following condition:*

$$R(\lambda_v, q_v) - \pi_v = R(\lambda_{v+1}, q_v) - \pi_{v+1}. \tag{33}$$

Proof. Please refer to Appendix F, available in the online supplemental material. □

5.2 Optimality of Contract

According to the profit expression of the NSP, the profit maximization problem, subject to the IC and IR constraints for all types, can be written as

$$\begin{aligned} &\max_{(\lambda_v, \pi_v)} \sum_{v=1}^{V} (\pi_v - s^{\text{cost}} \lambda_v) \\ &\text{s.t. } R(\lambda_v, q_v) - \pi_v \geq 0, (\text{IR}) \\ &\quad R(\lambda_v, q_v) - \pi_v \geq R(\lambda_{v'}, q_v) - \pi_{v'}, (\text{IC}) \\ &\quad \forall v, \quad \forall v' \neq v. \end{aligned} \tag{34}$$

Based on the series lemmas in the previous section, the above problem can be further represented by

$$\begin{aligned} &\max_{(\lambda_v, \pi_v)} \sum_{v=1}^{V} (\pi_v - s^{\text{cost}} \lambda_v) \\ &\text{s.t. } R(\lambda_V, q_V) - \pi_V = 0, \\ &\quad \lambda_v \geq \lambda_{v'} \text{ iff } q_v \geq q_{v'}, \\ &\quad R(\lambda_v, q_v) - \pi_v = R(\lambda_{v+1}, q_v) - \pi_{v+1}, \\ &\quad \forall v, \quad \forall v' \neq v. \end{aligned} \tag{35}$$

From the first and third conditions of the above problem, we can obtain a series of equations as follows:

$$\begin{aligned} \pi_V &= R(\lambda_V, q_V), \\ \pi_{V-1} &= R(\lambda_V, q_V) + w_{V-1}, \\ \pi_{V-2} &= R(\lambda_V, q_V) + w_{V-1} + w_{V-2}, \\ &\vdots \\ \pi_1 &= R(\lambda_V, q_V) + w_{V-1} + \cdots + w_1, \quad \text{where} \\ &\begin{cases} w_v = 0 & v = V \\ w_v = R(\lambda_v, q_v) - R(\lambda_{v+1}, q_v) & v = 1, 2, \ldots, V-1. \end{cases} \end{aligned} \tag{36}$$

Based on the above equations, we can conclude

$$\begin{aligned} S^{\text{NSP}} &= \sum_{v=1}^{V} (\pi_v - s^{\text{cost}} \lambda_V) \\ &= VR(\lambda_V, q_V) + Vw_v + (V-1)w_{V-1} \\ &\quad + \cdots + w_1 - \sum_{v=1}^{V} s^{\text{cost}} \lambda_v \\ &= \sum_{v=1}^{V} \left(R(\lambda_V, q_V) + \sum_{k=v}^{V} w_k - s^{\text{cost}} \lambda_v \right) \\ &= VR(\lambda_V, q_V) + \sum_{k=1}^{V} w_k + \cdots + \sum_{k=V-1}^{V} w_k - \sum_{k=1}^{V} s^{\text{cost}} \lambda_v \\ &= VR(\lambda_V, q_V) + \sum_{k=1}^{V-1} (R(\lambda_k, q_k) - R(\lambda_{k+1}, q_k)) + \cdots \\ &\quad + \sum_{k=V-1}^{V-1} (R(\lambda_k, q_k) - R(\lambda_{k+1}, q_k)) - \sum_{v=1}^{V} s^{\text{cost}} \lambda_v \\ &= VR(\lambda_V, q_V) - \sum_{k=v}^{V} s^{\text{cost}} \lambda_v \\ &\quad + (V-1)R(\lambda_{V-1}, q_{V-1}) + \cdots + R(\lambda_1, q_1) \\ &\quad - (V-1)R(\lambda_V, q_{V-1}) - \cdots - R(\lambda_2, q_1). \end{aligned} \tag{37}$$

Furthermore, by defining $q_0 = 0$ for notation simplification, the profit S^{NSP} can be further written as

$$S^{\text{NSP}} = \sum_{v=1}^{V} \big(vR(\lambda_v, q_v) - (v-1)R(\lambda_v, q_{v-1}) - s^{\text{cost}} \lambda_v \big). \tag{38}$$

By introducing the function

$$S_v \triangleq vR(\lambda_v, q_v) - (v-1)R(\lambda_v, q_{v-1}) - s^{\text{cost}}\lambda_v, \quad (39)$$

we can find that S_v is only related to λ_v and independent of other qualities $\lambda_{v'}$, $\forall v' \neq v$. Thus for any $v \in \{1, 2, \ldots, V\}$, the optimal quality λ_v^* can be obtained by maximizing each of S_v separately, i.e.,

$$\begin{aligned}\lambda_v^* &= \arg\max_{\lambda_v} S_v \\ &= \arg\max_{\lambda_v}(vR(\lambda_v, q_v) - (v-1)R(\lambda_v, q_{v-1}) - s^{\text{cost}}\lambda_v).\end{aligned} \quad (40)$$

In networks with ultra-dense SBSs, the intensity of the SBSs λ can be assumed as infinity. Therefore, there is no constraint on the upper bound of λ_v during the optimization.

To solve the optimal problem in Eq. (40), we now investigate the function S_v. First, we calculate the first-order and second-order derivatives of S_v relative to λ_v, and we have

$$\begin{aligned}\frac{\partial S_v}{\partial \lambda_v} &= v\frac{\partial R(\lambda_v, q_v)}{\partial \lambda_v} - (v-1)\frac{\partial R(\lambda_v, q_{v-1})}{\partial \lambda_v} - s^{\text{cost}}, \\ \frac{\partial^2 S_v}{\partial \lambda_v^2} &= v\frac{\partial^2 R(\lambda_v, q_v)}{\partial \lambda_v^2} - (v-1)\frac{\partial^2 R(\lambda_v, q_{v-1})}{\partial \lambda_v^2}.\end{aligned} \quad (41)$$

We further define the following two functions

$$\chi_1(v) \triangleq v\frac{\partial R(\lambda_v, q_v)}{\partial \lambda_v}, \chi_2(v) \triangleq v\frac{\partial^2 R(\lambda_v, q_v)}{\partial \lambda_v^2}, \quad (42)$$

and after some manipulations, we can obtain

$$\begin{aligned}\frac{\partial \chi_1(v)}{\partial v} &\propto (1-\gamma)v^{-\gamma}(\lambda_v K + T_v) + v^{1-\gamma}\frac{\partial T_v}{\partial v}(\lambda_v K - T_v), \\ \frac{\partial \chi_2(v)}{\partial v} &\propto -\big((1-\gamma)v^{-\gamma}(\lambda_v K + T_v) + v^{1-\gamma}\frac{\partial T_v}{\partial v}(\lambda_v K - 2T_v)\big),\end{aligned} \quad (43)$$

where $K \triangleq \sum_{j=1}^{V} j^{-\gamma}$ and $T_v \triangleq v^{-\gamma}A(\delta, \alpha)\zeta + (K - v^{-\gamma})C(\delta, \alpha)\zeta$. It can be verified that $\frac{\partial T_v}{\partial v} > 0$.

Whether S_v is convex or concave depends on many factors, and we now have the following related remarks.

Remark 3. In the case that S_v is a convex function of λ_v, this means $v\frac{\partial^2 R(\lambda_v, q_v)}{\partial \lambda_v^2}$ is an increasing function of v, i.e., $\frac{\partial \chi_2(v)}{\partial v} > 0$. Comparing $\frac{\partial \chi_1(v)}{\partial v}$ with $\frac{\partial \chi_2(v)}{\partial v}$, we conclude that $\frac{\partial \chi_1(v)}{\partial v}$ is most probably less than zero, and thus we have $\frac{\partial S_v}{\partial \lambda_v} < 0$, i.e., S_v is a decreasing function of λ_v.

Remark 4. In the case that S_v is a concave function of λ_v, this means $v\frac{\partial^2 R(\lambda_v, q_v)}{\partial \lambda_v^2}$ is a decreasing function of v, i.e., $\frac{\partial \chi_2(v)}{\partial v} < 0$. Comparing $\frac{\partial \chi_1(v)}{\partial v}$ with $\frac{\partial \chi_2(v)}{\partial v}$, we conclude that $\frac{\partial \chi_1(v)}{\partial v} > 0$, and thus $v\frac{\partial R(\lambda_v, q_v)}{\partial \lambda_v}$ is an increasing function of v, i.e., $v\frac{\partial R(\lambda_v, q_v)}{\partial \lambda_v} - (v-1)\frac{\partial R(\lambda_v, q_{v-1})}{\partial \lambda_v} > 0$. Therefore, whether there is $\frac{\partial S_v}{\partial \lambda_v} < 0$ or not depends on the value of s^{cost}, i.e., S_v can be either a decreasing or an increasing function of λ_v.

It can be seen that there is no closed form of the optimal solution. In this case, we resort to MATLAB to solve the optimization problem in (40). Furthermore, note that λ_v^*, $\forall v$, are the solutions to the problem in (35) without considering the second constraint $\lambda_v \geq \lambda_{v'}$ iff $q_v \geq q_{v'}$. Therefore, we need to check whether these solutions satisfy the constraint. In the case that the constraint is not satisfied, the obtained solution λ_v^* should be adjusted according to "Bunching and Ironing" approach [22].

Substituting the solution λ_v^* into Eq. (36), we can obtain the corresponding optimal price π_v^* as follows:

$$\pi_v^* = R(\lambda_V^*, q_V) + \sum_{k=v}^{V} w_k^*, \quad \text{where,} \quad (44)$$

$$\begin{cases} w_v^* = 0 & v = V \\ w_v^* = R(\lambda_v^*, q_v) - R(\lambda_{v+1}^*, q_v) & v = 1, 2, \ldots, V-1. \end{cases} \quad (45)$$

The detailed process of obtaining the optimal solution is concluded in *Algorithm 1*, as shown in Table 1.

TABLE 1
Description of *Algorithm 1* for the Optimal Contract

Algorithm 1.

Input: $Q, \gamma, \beta, \alpha, \delta, \zeta, s^{\text{cache}}$, and s^{cost}
Output: λ_v^* and π_v^*, $\forall v$, S^{NSP}
Steps:
1: **for** $v = 1$ to V **do**
2: set $S_v = vR(\lambda_v, q_v) - (v-1)R(\lambda_v, q_{v-1}) - s^{\text{cost}}\lambda_v$
3: set $\lambda_v^* = \arg\max_{\lambda_v} S_v$
4: **end for**
5: **while** λ_v^* is not feasible **do**
6: find an infesible subsequence $\hat{\lambda}_m, \hat{\lambda}_{m+1}, \ldots, \hat{\lambda}_n$
7: set $\lambda_v^* = \arg\max_{\lambda_v} \sum_{i=m}^{n} S_i$, $\forall v = m, m+1, \ldots, n$
8: **end while**
9: **for** $v = 1$ to V **do**
10: set $\pi_v^* = R(\lambda_V^*, q_V) + \sum_{k=v}^{V} w_k^*$
11: **end for**
12: set $S^{\text{NSP}} = \sum_{v=1}^{V}(R(\lambda_V, q_V) + \sum_{k=v}^{V} w_k - s^{\text{cost}}\lambda_v)$

6 Numerical Results

In this section, we will conduct numerical results to validate the proposed contract design, and guide the NSP for offering proper incentives to the VPs while keeping its own profit maximized. Note that the derivation of $\Pr(\mathcal{E}_v^{\text{cover}})$ in Theorem 1 follows the similar procedure as in Ref. [33], and has been verified by Monte-Carlo simulations in this reference. Therefore, we omit the verification on $\Pr(\mathcal{E}_v^{\text{cover}})$ by simulations here, and focus our attention on the contract-based numerical results.

Unless otherwise specified, parameter settings are given as follows. Note that the following parameters are adopted from or similar to [20], [21]. a) In the caching system, we set the number of files in $\mathcal{F}$ to $F = 100$, the number of VPs to $V = 8$, and the storage size $Q = 10$. The profit/UAP is considered to be the profit gained per month within an area of one square kilometer, i.e., $/\text{month} \cdot \text{km}^2$. We further set $c = 30/\text{month}$, which is the average number of video requests from an MU per month, set $s^{\text{cache}} = 1$, which is the profit gained by $\mathcal{V}_v$ on an MU for providing local downloading of a video, and set $s^{\text{cost}} = 0.1/\text{month}$, which is the cost for maintaining the cache of an SBS by the NSP each month;

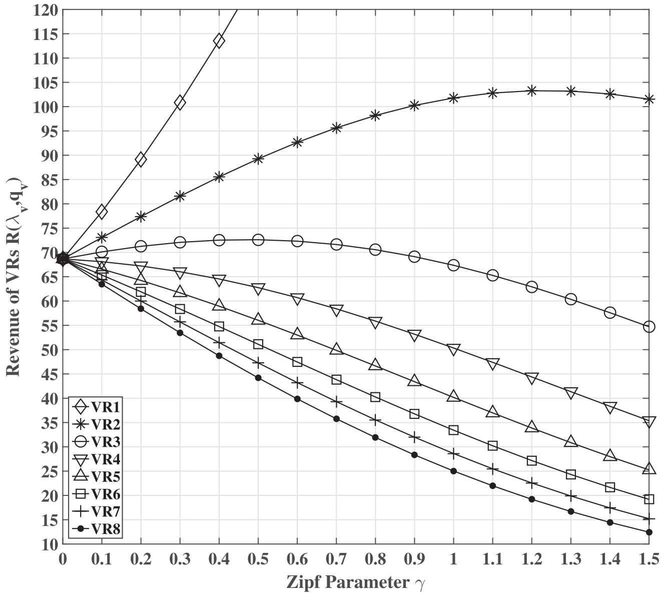


Fig. 2. The revenue $R(\lambda_v, q_v)$ versus γ for a fixed quality $\lambda_v = 200, \forall v$.

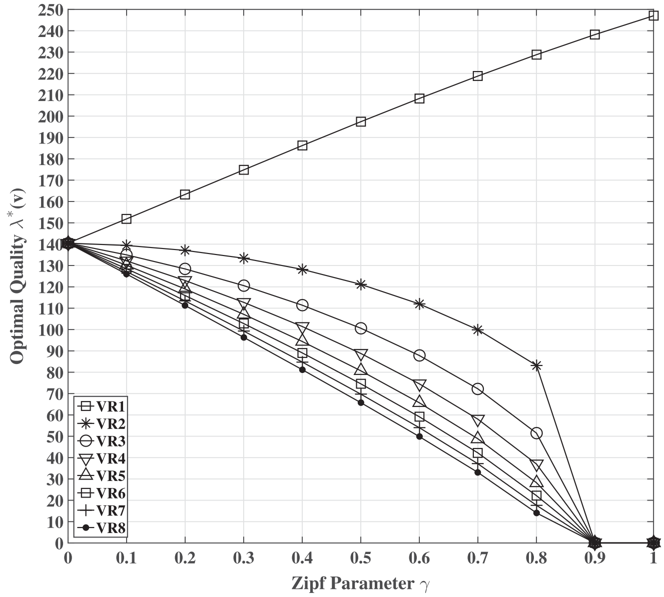


Fig. 4. The optimal quality $\lambda_v^*, \forall v$, versus γ.

b) For the underlying wireless network, we set the intensity of the MUs to $\zeta = 50/km^2$, and investigate the scenarios with ultra-densely deployed SBSs. Additionally, the path loss parameter is set to $\alpha = 4$, and the receiver SINR threshold is set to $\delta = 0.3$.

Note that the types of the VP follows the Zipf distribution, while γ reflects the imbalance property of the type distribution among the VPs. A larger γ represents a more uneven type distribution of the VPs. Thus, we will use γ as the x-axis in the following results.

First, to verify the relation between $R(\lambda_v, q_v)$, v, and γ, we fix $\lambda_v = 200, \forall v$, and study the trends of $R(\lambda_v, q_v)$ for all the VPs under different γ. Fig. 2 illustrates the the values of $R(\lambda_v, q_v)$ versus γ. It can be observed that $\mathcal{V}_1$'s revenue $R(\lambda_1, q_1)$ is always an increasing function of γ, while $R(\lambda_2, q_2)$ and $R(\lambda_3, q_3)$ first increase and then decrease with the growth of γ. Also, the other five VPs' revenues are always decreasing with γ. This complies with the assertion in Remark 2 that a small v may leading to an increasing function of γ.

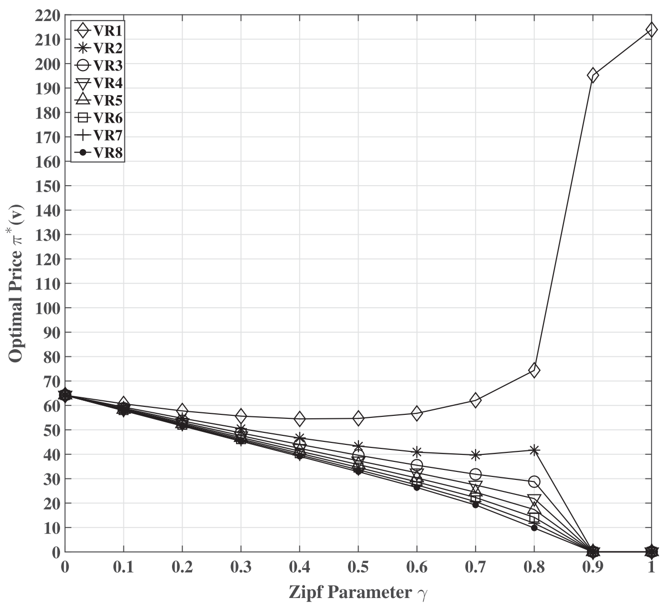


Fig. 3. The optimal price $q_v^*, \forall v$, versus γ.

Figs. 3 and 4 illustrate the optimal price and optimal quality with the varying of γ from 0 to 1, respectively. It is observed that both the optimal quality and optimal price are increasing with the types of the VPs. To be specific, the VP $\mathcal{V}_v$ with a larger type q_v will be charged with a higher price π_v^* for its commodity. Even so, this VP intends to rent more SBSs, i.e., a higher quality λ_v^*, as shown in Fig. 4, to achieve a higher profit. Note that with the increase of γ, the first VP $\mathcal{V}_1$ becomes more and more dominant in the value of its type, i.e., q_1 is much higher than other type values. This means that more and more MUs in the network will affiliate with $\mathcal{V}_1$, while the number of the MUs affiliated with other VPs keeps reducing. In this case, the VPs $\mathcal{V}_2, \ldots, \mathcal{V}_8$ have to reduce the quality of the commodity so as to make their profit non-negative. Particularly in the case of $\gamma \geq 0.9$, all the other VPs will quit the contract, except for $\mathcal{V}_1$, with zero price and zero quality. We can see from Fig. 3 that $\pi^*(q_1)$ has a dramatic growth when $\gamma = 0.9$, while the optimal prices for other VPs drop down to zero. Fig. 4 indicates that the quality of the commodity purchased by $\mathcal{V}_1$ goes up steadily, compared with those of other VPs, which decrease to zero with the increase of γ.

Fig. 5 shows the optimal profit $S^{\text{VP}}(\lambda_v^*, q_v^*), \forall v$, under the optimal quality-price pair (λ_v^*, q_v^*) versus γ. We can see from the figure that $S^{\text{VP}}(\lambda_v, q_v)$, for $v = 1, \ldots, 7$, are concave functions of γ in the case $0 \leq \gamma \leq 0.9$, while $S^{\text{VP}}(\lambda_8, q_8)$ always equals to zero according to Lemma 1. When $\gamma \geq 0.9$, all the VPs cannot gain profit from the system. Particularly for $\mathcal{V}_1$, although the optimal price and quality for the commodity it purchases are non-zero when $\gamma = 0.9$, its profit drops to zero. This is because that all the other VPs have quitted the market, and according to *Lemma 1*, the profit for the last one in the market has to be zero for meeting the IR constraint.

Fig. 6 depicts the profit of the NSP versus γ under different storage sizes, i.e., $Q = 5, 10, 20, 30, 40, 50$, and 100. First, it is obvious that a large storage size Q corresponds to a higher profit S^{NSP}. Second, we can see that the profit S^{NSP} is a convex function of γ regardless of the storage size. A

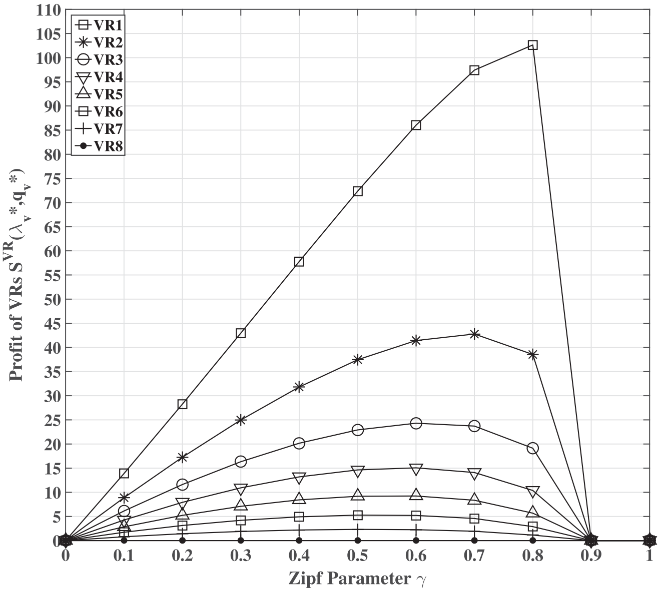


Fig. 5. $\mathcal{V}_v$'s profit $S^{\mathrm{VP}}(\lambda_v^*, q_v^*)$, $\forall v$, under the optimal quality-price pair.

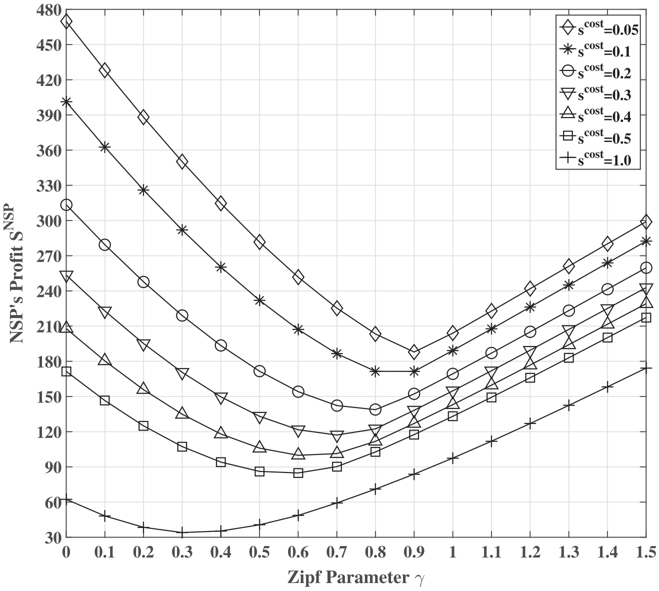


Fig. 7. NSP's profit under different costs s^{cost}.

typical minimum point of the NSP's profit appears when γ is around 0.9, which is insensitive to the value of Q. On the other hand, Fig. 7 illustrates the NSP's profit under various cost, i.e., $s^{\mathrm{cost}} = 0.05, 0.1, 0.2, 0.3, 0.4, 0.5$, and 1.0. It can be observed that less cost is helpful to increase S^{NSP}. Again, this figure proves that S^{NSP} is a convex function of γ regardless of the value of s^{cost}. The value of γ that minimizes S^{NSP} varies according to s^{cost}, i.e., a larger value of s^{cost} will leads to a smaller value of γ that minimizes S^{NSP}.

Finally, we focus on the social welfare $W(\lambda_v, q_v)$ defined in Eq. (21). Fig. 8 shows the overall social welfare $\sum_v W(\lambda_v, q_v)$ under three different storage size Q. Note that we consider two kinds of method for calculating the social welfare as follows. The solid line in Fig. 8 represents the case that there is no contract between the NSP and VPs, and the optimal λ_v that maximizes the social welfare is calculated according to Eq. (22). The dashed line in the figure represents the contract-based case with the best λ_v^* and π_v^* calculated in Eqs. (40) and (44), respectively. We can see that a large storage size contributes a lot to the social welfare performance. When γ is small, i.e., $\gamma \leq 0.4$, the contract-based and optimal-based welfare have the similar performance. When γ becomes large, the social welfare in the optimal scheme steadily increases, while the welfare of the contract-based method decreases first and then increases. Particularly, when γ is around 0.9 or 1.0, contract-based scheme reaches the minimum point of the performance.

Fig. 9 illustrates the overall social welfare under different values of cost s^{cost}. Similar to Fig. 8, this figure considers contract-based and optimal-based schemes for the social welfare. We can see that a high cost will jeopardize the performance of the social welfare. Also, when γ is small, two schemes have the similar performance. When γ becomes large, the social welfare in the optimal scheme increases

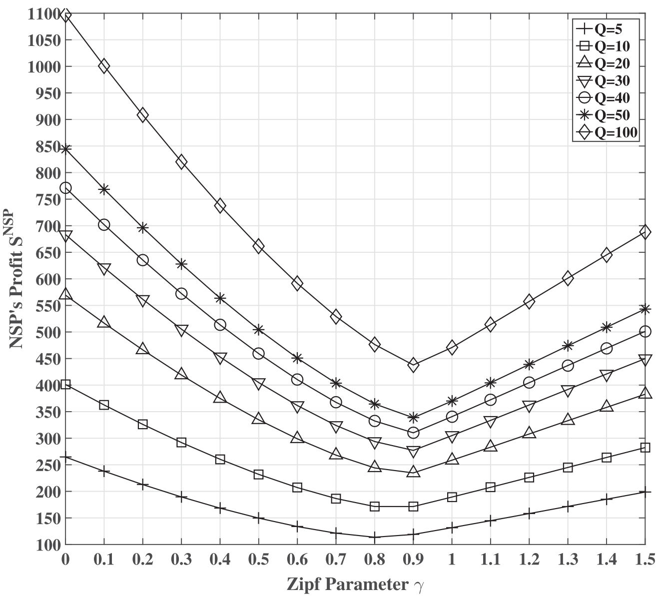


Fig. 6. NSP's profit under different storage sizes of Q.

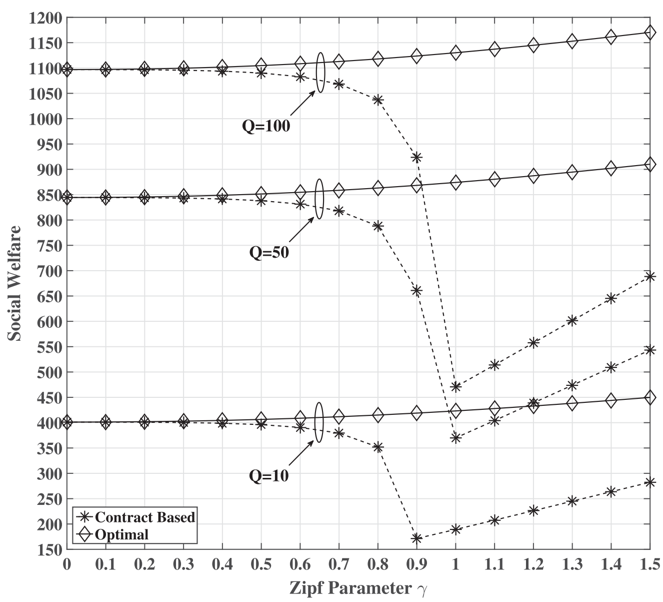


Fig. 8. Social welfare under different storage sizes of Q.

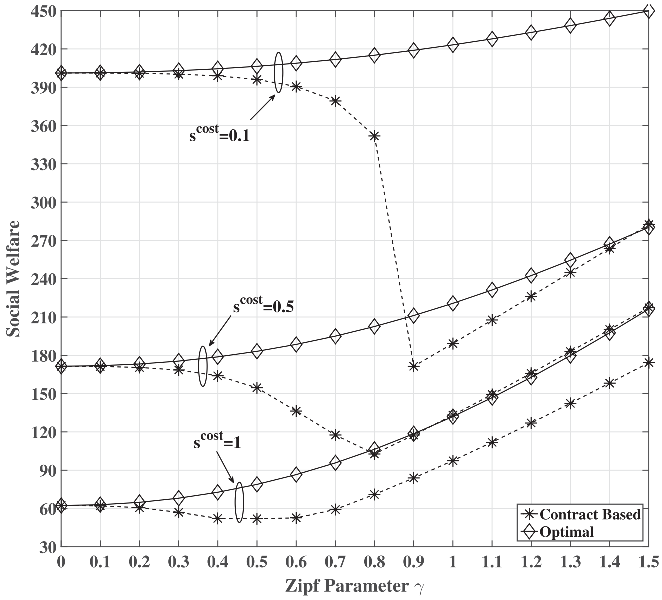


Fig. 9. Social welfare under different costs s^{cost}.

steadily, while the welfare of the contract-based method decreases first and then increases. This is similar to the situation in the previous figure. Also, it can be observed that in the cases of $s^{\text{cost}} = 0.1, 0.5$, and 1.0, the values of γ that minimizes the social welfare appears at 0.9, 0.8, and 0.5, respectively.

7 Conclusions

In this paper, we considered a commercial small-cell caching system consisting of an NSP and multiple VPs, where the NSP leases its SBSs to the VPs for gaining profits, while the VPs, after storing popular videos to the rented SBSs, can provide faster transmissions to the MUs, hence gaining more profits. We proposed a contract-based theoretic framework by viewing the SBSs as a type of resources. We first modeled the MUs and SBSs using two independent PPPs with the aid of stochastic geometry, and developed the probability expression of direct downloading. Then, based on the probability derived, we formulated a contract-based model for maximizing the average profit of the NSP as well as individual VPs. Next, we solved the associated non-convex optimization problem for maximizing the profit of the NSP, with the IC and IR constraints that provides the incentives for the VPs to participate. Finally, we provided several numerical results for showing that the proposed schemes are effective in both pricing and resource allocation.

Acknowledgments

This work is supported in part by the National Key R&D Program under Grants 2018YFB1004802 and 2016YFE0108000, in part by the National Natural Science Foundation of China under Grants 61727802, 61771244, 61501238, 61571329, in part by the Jiangsu Provincial Science Foundation under Project BK20150786, in part by the Specially Appointed Professor Program in Jiangsu Province 2015, in part by the Fundamental Research Funds for the Central Universities under Grant 30916011205, and in part by the Open Research Fund of National Mobile Communications Research Laboratory, Southeast University, under Grant 2017D04, and in part by the Competitiveness Enhancement Program of the National Research Tomsk Polytechnic University.

References

[1] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

[2] L. Lei, Z. Zhong, K. Zheng, J. Chen, and H. Meng, "Challenges on wireless heterogeneous networks for mobile cloud computing," *IEEE Wireless Commun.*, vol. 20, no. 3, pp. 34–44, Jun. 2013.

[3] N. Golrezaei, A. Molisch, A. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Commun. Mag.*, vol. 51, no. 4, pp. 142–149, Apr. 2013.

[4] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. Leung, "Cache in the air: Exploiting content caching and delivery techniques for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 131–139, Feb. 2014.

[5] X. Wang, Y. Zhang, V. C. M. Leung, N. Guizani, and T. Jiang, "D2D big data: Content deliveries over wireless device-to-device sharing in realistic large scale mobile networks," *IEEE Wireless Commun. Mag.*, vol. 25, no. 1, pp. 32–38, Feb. 2018.

[6] M. Maddah-Ali and U. Niesen, "Decentralized coded caching attains order-optimal memory-rate tradeoff," in *Proc. 51st Annu. Allerton Conf. Commun. Control Comput.*, Oct. 2013, pp. 421–427.

[7] N. Golrezaei, P. Mansourifard, A. Molisch, and A. Dimakis, "Base-station assisted device-to-device communications for high-throughput wireless video networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, pp. 3665–3676, Jul. 2014.

[8] M. Ji, G. Caire, and A. F. Molisch, "Wireless device-to-device caching networks: Basic principles and system performance," *IEEE J. Select. Areas Commun.*, vol. 34, no. 1, pp. 176–189, Jan. 2016.

[9] M. Ji, G. Caire, and A. Molisch, "Optimal throughput-outage trade-off in wireless one-hop caching networks," in *Proc. IEEE Int. Symp. Inf. Theory Proc.*, Jul. 2013, pp. 1461–1465.

[10] P. Gupta and P. Kumar, "The capacity of wireless networks," *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 388–404, Mar. 2000.

[11] F. Boccardi, R. Heath, A. Lozano, T. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

[12] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A survey on 3GPP heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 3, pp. 10–21, Jun. 2011.

[13] J. Akhtman and L. Hanzo, "Heterogeneous networking: An enabling paradigm for ubiquitous wireless communications," *Proc. IEEE*, vol. 98, no. 2, pp. 135–138, Feb. 2010.

[14] S. Bayat, R. Louie, Z. Han, B. Vucetic, and Y. Li, "Distributed user association and femtocell allocation in heterogeneous wireless networks," *IEEE Trans. Commun.*, vol. 62, no. 8, pp. 3027–3043, Aug. 2014.

[15] M. Mirahmadi, A. Al-Dweik, and A. Shami, "Interference modeling and performance evaluation of heterogeneous cellular networks," *IEEE Trans. Commun.*, vol. 62, no. 6, pp. 2132–2144, Jun. 2014.

[16] A. Gupta, H. Dhillon, S. Vishwanath, and J. Andrews, "Downlink multi-antenna heterogeneous cellular network with load balancing," *IEEE Trans. Commun.*, vol. 62, no. 11, pp. 4052–4067, Nov. 2014.

[17] 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.

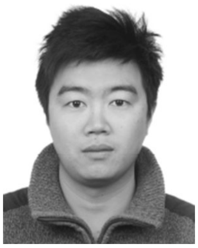
[18] E. Bastug, M. Bennis, and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs," in *Proc. 11th Int. Symp. Wireless Commun. Syst.*, Aug. 2014, pp. 649–653.

[19] D. Stoyan, W. Kendall, and M. Mecke, *Stochastic Geometry and Its Applications*, 2nd ed. Hoboken, NJ, USA: Wiley, 2003.

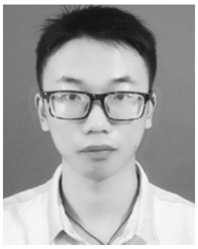
[20] M. Haenggi, J. Andrews, F. Baccelli, O. Dousse, and M. Franceschetti, "Stochastic geometry and random graphs for the analysis and design of wireless networks," *IEEE J. Select. Areas Commun.*, vol. 27, no. 7, pp. 1029–1046, Sep. 2009.

[21] J. Li, H. Chen, Y. Chen, Z. Lin, B. Vucetic, and L. Hanzo, "Pricing and resource allocation via game theory for a small-cell video caching system," *IEEE J. Select. Areas Commun.*, vol. 34, no. 8, pp. 2115–2129, Aug. 2016.

[22] P. Bolton and M. Dewatripont, *Contract Theory*. Cambridge, MA, USA: MIT Press, 2005.
[23] L. Gao, X. Wang, Y. Xu, and Q. Zhang, "Spectrum trading in cognitive radio networks: A contract-theoretic modeling approach," *IEEE J. Select. Areas Commun.*, vol. 29, no. 4, pp. 843–855, Apr. 2011.
[24] L. Duan, L. Gao, and J. Huang, "Cooperative spectrum sharing: A contract-based approach," *IEEE Trans. Mobile Comput.*, vol. 13, no. 1, pp. 174–187, Jan. 2014.
[25] D. Lopez-Perez, M. Ding, H. Claussen, and A. H. Jafari, "Towards 1 Gbps/UE in cellular systems: Understanding ultra-dense small cell deployments," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2078–2101, Oct.–Dec. 2015.
[26] I. Ashraf, L. Ho, and H. Claussen, "Improving energy efficiency of femtocell base stations via user activity detection," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Apr. 2010, pp. 1–5.
[27] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system," in *Proc. 7th ACM SIGCOMM Conf. Internet Meas.*, 2007, pp. 1–14.
[28] M. Eriksson, S. M. H. Rahman, F. Fraile, and M. Sjöström, "Efficient interactive multicast over DVB-T2: Utilizing dynamic SFNs and PARPS," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast.*, Jun. 2013, pp. 1–7.
[29] U. Jennehag and S. Pettersson, "On synchronization frames for channel switching in a GOP-based IPTV environment," in *Proc. 5th IEEE Consum. Commun. Netw. Conf.*, Jan. 2008, pp. 638–642.
[30] A. Araldo, G. Dán, and D. Rossi, "Stochastic dynamic cache partitioning for encrypted content delivery," in *Proc. 28th Int. Teletraffic Congr.*, Sep. 2016, pp. 139–147.
[31] S. Lee and K. Huang, "Coverage and economy of cellular networks with many base stations," *IEEE Commun. Lett.*, vol. 16, no. 7, pp. 1038–1040, Jul. 2012.
[32] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2015, pp. 3358–3363.
[33] J. Li, Y. Chen, Z. Lin, W. Chen, B. Vucetic, and L. Hanzo, "Distributed caching for data dissemination in the downlink of heterogeneous networks," *IEEE Trans. Commun.*, vol. 63, no. 10, pp. 3553–3568, Oct. 2015.



Jun Li (M'09-SM'16) received the PhD degree in electronic engineering from Shanghai Jiao Tong University, Shanghai, P. R. China, in 2009. From January 2009 to June 2009, he worked with the Department of Research and Innovation, Alcatel Lucent Shanghai Bell as a research scientist. From June 2009 to April 2012, he was a postdoctoral fellow with the School of Electrical Engineering and Telecommunications, University of New South Wales, Australia. From April 2012 to June 2015, he was a research fellow with the School of Electrical Engineering, University of Sydney, Australia. From June 2015 to now, he has been a professor with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, China. His research interests include network information theory, ultra-dense wireless networks, and mobile edge computing. He is a senior member of the IEEE.



Shunfeng Chu received the BS degree in electronic and information engineering from the Nanjing University of Science and Technology, Zijing College, Jiangsu, China, in 2011. He is currently working toward the master's degree in the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Jiangsu, China. His research interests include network economics, D2D communications, and small-cell caching.

Feng Shu (M'09) received the BS degree from Fuyang Teaching College, Fuyang, China, in 1994, the MS degree from XiDian University, Xi'an, China, in 1997, and the PhD degree from Southeast University, Nanjing, in 2002. From Sept. 2009 to Sept. 2010, he was a visiting postdoctor with the University of Texas at Dallas. In October 2005, he joined the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, China, where he is currently a professor and supervisor of PhD and graduate students. He is also with Fujian Agriculture and Forestry University and was awarded with the Mingjian Scholar chair professor in Fujian Province. His research interests include wireless networks, wireless location, and array signal processing. He is currently an associate editor for the journal *IEEE Access*. He has published about 200 papers, of which more than 100 are in archival journals including more than 25 papers in IEEE Journals and more than 50 SCI-indexed papers. He holds four Chinese patents. He is a member of the IEEE.

Jun Wu (M'01-SM'14) received the BS degree in information engineering and the MS degree in communication and electronic system from Xidian University, in 1993 and 1996, respectively, and the PhD degrees in signal and information procesing from the Beijing University of Posts and Telecommunications, in 1999. He is a professor with the Computer Science and Technology Department, Tongji University, Shanghai, China. He joined Tongji University as an professor in 2010. He was a principal scientist with Huawei and Broadcom before he joined Tongji. His research interests include wireless communication, information theory, machine learning, and signal processing. He is a senior member of the IEEE.

Dushantha Nalin K. Jayakody (M'14) received the BEng degree (with first-class honors) in Pakistan, in 2009 and was ranked as the merit position holder of the University (under SAARC Scholarship), the MSc degree in electronics and communications engineering from the Department of Electrical and Electronics Engineering, Eastern Mediterranean University, Cyprus, in 2010 (under the University full graduate scholarship) and ranked as the first merit position holder of the department, and the PhD degree in electronics and communications engineering from the University College Dublin, Ireland, in 2014. From Sep. 2014- Sep. 2016, he held a postdoc position with the Coding and Information Transmission Group, University of Tartu, Estonia, and the University of Bergen, Norway. Since 2016, he has been a professor with the Department of Control System Optimization, Institute of Cybernetics, National Research Tomsk Polytechnic University, Russia. He is a member of the IEEE and he has served as a session chair or technical program committee member for various international conferences, such as IEEE PIMRC 2013/2014, IEEE WCNC 2014/2016, IEEE VTC 2015, etc.