When Virtual Network Operator Meets E-Commerce Platform: Advertising via Data Reward

Qi Cheng, Hangguan Shan, *Senior Member, IEEE*, Weihua Zhuang, *Fellow, IEEE*, Tony Q. S. Quek, *Fellow, IEEE*, Zhaoyang Zhang, *Senior Member, IEEE*, Fen Hou, *Member, IEEE*

Abstract—In China, some e-commerce platform (EP) companies such as Alibaba and JD are now allowed to partner with network operators (NOs) to act as virtual network operators (VNOs) to provide mobile data services for mobile users (MUs). However, it is a question worth researching on how to generate more profits for all network players, with EP companies being VNOs, through appropriate integration of the VNO business and the companies' own e-commerce business. To address this issue, in this work we propose a novel incentive mechanism for advertising via mobile data reward, and model it as a three-stage static Stackelberg game. We obtain the closed-form optimal solution of the Nash equilibrium by backward induction. Besides, for the scenario lack of knowledge on the interaction between the NO and VNO in a dynamic game, we propose a deep Q-network (DQN) based algorithm to derive the optimal strategies of the NO and VNO. Simulation results show impact of system parameters on the utilities of game players and social welfare. We also study the impact of system parameters on different algorithms and discover that the proposed DQN-based algorithm can learn a good strategy as compared with the Stackelberg equilibrium solution.

Index Terms—Virtual network operator, e-commerce platform, data reward, Stackelberg game, deep Q-network, network economics.

1 Introduction

With the deployment of fifth-generation (5G) communication networks in the world, how people use information services have been changing not only from offline to online but also more and more from personal computers to mobile terminals [2]. In particular, during the COVID-19 pandemic in 2020, many people in China stayed home and relied on e-commerce platforms (EPs) such as Alibaba, JD, and Pinduoduo to purchase fruits, vegetables, and daily necessities. In fact, short video shopping, (interactive) live streaming, virtual reality/augmented reality (VR/AR) marketing, and other new applications have increasingly entered the daily lives of people. These applications are occupying more and more network traffic, and it is expected that their traffic in operators' networks will increase unprecedentedly in the future [3]. Correspondingly, e-commerce platforms have become major contributors to network traffic, and ordinary people are becoming more and more attached to these EPs.

On the other hand, similar to other countries, in 2018 China has allowed e-commerce platforms or technical companies such as Alibaba and JD to act as virtual network operators (VNOs) to participate in network operations. This gives the e-commerce platforms and traditional network operators (NOs) an opportunity to renew their business model for not only increasing their own revenues and the service satisfaction of mobile users (MUs) but also improving network resource utilization at the same time. Closely related, in recent years the network design that a VNO participates in has been one of the hot issues which the academic community concerns about, for example, spectrum trading among NOs and VNOs [4], cooperative games among VNOs and MUs [5], and competition strategies among multiple VNOs with complete and incomplete information [6]. However, in the research scenarios of the aforementioned works, VNOs do not have the second role of an e-commerce platform at the same time, so the proposed solutions cannot be directly applied to the scenario of interest here.

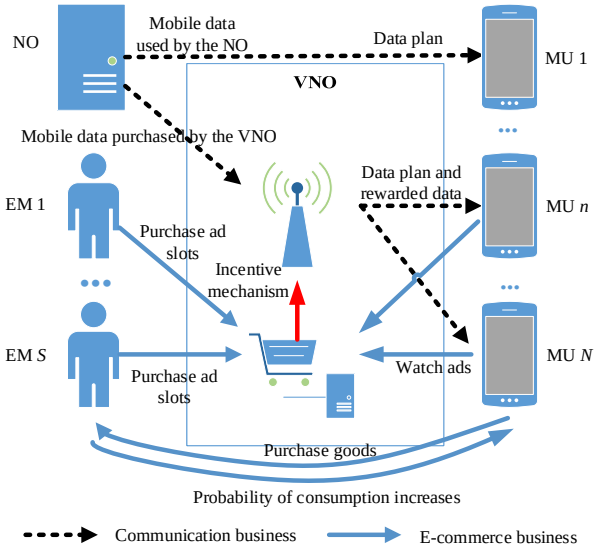


Fig. 1: Research scenario of an EP as a VNO.

In fact, when an e-commerce platform participates in network design and operation, the number of ways in which the benefits of all parties in the network are affected will be more than that when only the traditional NOs or NOs and

This paper was presented in part at IEEE/ACM IWQoS, 2021 [1].

non-EP VNOs participate together in network operations. As shown in Fig. 1, the EP serves both the e-commerce merchants (EMs) and MUs on its platform. When the EP serves as a VNO as well, it can attract MUs to spend on its platform through data rewards [7], thereby increasing the revenue of both the EMs and MUs. However, it needs to pay the NO for the data rewards, so the strategies of the NO, VNO, EMs, and MUs are interrelated in this network.

In the new scenario of an EP being a VNO participating in network operations, many challenging issues arise. In this work, we focus on the following problems:

- How should the NO price its mobile data or other network resources to maximize its revenue?
- How should the VNO customize its data plan or a data incentive mechanism to attract MUs, and provide services with EMs (e.g., advertising) to attract them to purchase its services for improving their revenue?
- How should the EMs purchase services from the EP to increase their revenue from the e-commerce business?
- How should the MUs select the data plan of the NO and VNO, and act according to the data incentive mechanism of the VNO to maximize their utility?

In response to the aforementioned questions raised in the new scenario, as shown in Fig. 1, in this work we consider a basic scene consisting of a single NO, a single EP which serves as a VNO, multiple EMs served by the EP, and multiple MUs who can choose mobile services from either the NO or VNO. To improve its revenue from the data market and advertisement (ad) market, the VNO proactively purchases mobile data from the NO and gives MUs certain data rewards to attract MUs to watch ads and consume in the EP. In this way, not only the NO and VNO can increase their profits, but also the MUs can obtain more mobile data and the EMs may benefit more from the advertisements, leading to a quadri-win effect. We illustrate the incentive mechanism for advertising via mobile data reward of this scenario in Fig. 1, where the black dotted arrow and the blue solid arrow represent the communication business and the e-commerce business, respectively.

In a static game, different players have different ownerships of information and resources, thus their priorities in the game are not equal. In this case, to address the aforementioned problems, we consider modeling the interactions among the NO, VNO, MUs, and EMs using a three-stage Stackelberg game. In Stage I, the NO decides the price of mobile data (i.e., the payment for purchasing unit mobile data) for the VNO. In Stage II, the VNO decides the data plan fee for the MUs and the ad price (i.e., the payment for purchasing one ad slot) for the EMs. In Stage III, the MUs with different valuations for the mobile service make their data plan subscription and ad watching decisions, in which we consider an α-fair data consumption utility function and uniform distribution of MUs' valuation. Meanwhile, the EMs decide the number of ad slots to purchase, considering the advertising's wear-out effect (i.e., the effectiveness of an ad decreases as the number of files that a user watches the same ad increases).

In practice, the NO is unlikely to own the complete information of the VNO. Therefore, the game between the NO and VNO becomes a non-cooperative dynamic game with incomplete information. In this case, to also address the aforementioned problems, we consider modeling it as a multi-leaders multi-followers two-stage Stackelberg game, with the NO and VNO as leaders, and the MUs and EMs as followers, and propose a reinforcement learning-based approach to address the strategy design. The main contributions of this paper are summarized as follows:

- For the scenario of a single NO, a single EP as a VNO, multiple EMs, and multiple MUs, we propose a novel incentive mechanism for advertising via mobile data reward, and formulate the interactions of the four types of decision-makers as a three-stage Stackelberg game;
- We obtain a closed-form optimal solution of the Nash equilibrium of the problem by backward induction. Using the optimal solution obtained, we can provide the optimal strategies for the NO, VNO, EMs, and MUs, respectively, in the network design or usage, such as the pricing of the NO's mobile data, the pricing of the VNO's data plan and advertising, the number of advertisements that an EM purchases from the VNO, and the data plan selection and advertisement watching strategies of an MU;
- For the scenario without sufficient knowledge of the interactions between the NO and VNO, we model it as a two-stage Stackelberg game in each cycle where the NO and VNO act as leaders and EMs and MUs act as followers. We propose a DQN-based algorithm to derive the optimal strategies of the NO and VNO via trial and error in the dynamic game;
- We conduct extensive simulations to study the impact of some important system parameters, such as the data amount of the VNO's data plan and data reward per ad, on the Nash equilibrium. We also compare the proposed DQN-based algorithm with other benchmarks under different system parameters, and demonstrate that the proposed DQN-based algorithm can learn a good strategy compared with the Stackelberg equilibrium (SE) solution.

The remainder of this paper is organized as follows. In Section II, we review related works. In Section III, we propose the incentive mechanism for advertising via mobile data reward, and introduce the static three-stage Stackelberg game. In Section IV, we obtain the closed-form optimal solution of the Nash equilibrium of the problem by backward induction, and analyze the impact of the system parameters on social welfare. In Section V, the DQN-based optimal strategy decision is elaborated for the scenario of the dynamic game. In Section VI, we discuss simulation results to reveal the impact of system parameters on the Nash equilibrium and compare it with the DQN-based algorithm. Finally, we draw the conclusions in Section VI.

2 Related Works

In the literature, there exist many works studying the access and/or resource allocation problem for wireless MUs in the

traditional VNO scenarios, where the resources are usually defined in the frequency domain based on the orthogonal frequency division multiple access technology [4], [5], [6]. For example, in [4] the authors study a problem in which multiple VNOs under cellular networks can lease spectrum from a NO to provide data offloading services for MUs, and model the interaction among VNOs as a Cournot game and a Stackelberg game, respectively. In [5], a two-stage spectrum leasing framework is formulated as a tri-level nested optimization problem, where a VNO acquires spectrum resources through both advance reservation and on-demand request. The authors in [6] study the competition strategies among VNOs with complete or incomplete information about spectrum inventories, and use the Bayesian coalition formation game to formulate the pricing decision problem. The aforementioned works study the three-party game relationship among NOs, VNOs, and MUs from the perspective of communication services, without the potential second role of a VNO (i.e., an EP), thus without the interaction between the EMs and EP.

The changes that occur in the media mode of the e-commerce business have led to an increasing demand for mobile data by mobile users, making data plan fee an increasing concern for mobile users. In this case, the approach of sponsored data is considered as a possible solution to achieve multiple-win results [8]. Sponsored data is the provision of a certain amount of data or navigation functionality to a mobile application/website for free to a specific user. There are two main types of sponsored data: zero-rating [9] and data reward [8]. Depending on the role of content providers (CPs) in the provisioning of free services in a data plan, zero-rating implies that the CPs pay the NOs for offering free services individually to MUs [7]. An example of zero-rating applications is the "Free Basics" provided by Facebook which gives free access to a simplified version of some websites [10], [11]. Alternatively, data rewarding is also available for example by Aquto and Unlockd [8]. They develop mobile APPs that display ads and track the amount of rewarded data. The impact of value-added services is studied in [12] from the perspective of a VNO, but the model there is relatively simple and only the changes to a single VNO's utility is considered. The work in [9] provides a modeling of zero-rating services and their utility impact on NOs and subscribers, without taking VNO into account. The authors in [8] use a two-stage Stackelberg game to model a novel data rewarding ecosystem, where a NO offers users data rewards to create new revenue streams. Of all the sponsored data solutions, data reward is the most user-oriented, where users are the ones to decide when they want to engage with reward programs and how they want to use the data they won. Therefore, in this work we adopt the data reward scheme in our incentive mechanism design.

In the literature, there exist some works studying advertising through certain incentive mechanisms [13], [14], [15]. The work in [13] considers venue owners offering advertising to specific users with an access model of both paid access and viewing ads through Wi-Fi deployment in public places. The authors in [14] achieve incentives through cooperation between online APPs and offline venues to promote revenue for both parties. Moreover, many studies have investigated how to model the interactions among the Internet service provider (ISP), CP, and MUs. The work in [15] subsidizes users' data by CPs paying money to ISPs for the purpose of increasing their own profits. The work in [16] studies cooperation among ISPs and MUs and proposes a revenue sharing contract between ISPs by modeling them as a supply chain to deliver traffic in a two-sided market. The authors in [17] design incentive mechanisms with the goal of achieving high quality of crowdsourced data and low cost of incentivizing users participation in the scenario with both complete and incomplete information. However, as these works do not consider the fact that the advertising platforms, such as the venue owners, online APPs, and CPs, can also have the potential second role of a VNO, their research scenarios are different from the one which we are interested in.

Reinforcement learning is a branch of machine learning techniques that has been widely applied in the research of network economy, combined with game theory [18]. The authors in [19] formulate the competitive interactions between a sensing platform (SP) and MUs as a multi-stage Stackelberg game. The SP calculates the quantity of sensing time to purchase from each MU by solving a convex optimization problem, while each follower observes the trading records and iteratively adjusts its pricing strategy based on a multi-agent deep deterministic policy gradient (MADDPG) framework [20]. In [21], the interactions between a mobile crowdsensing (MCS) server and several smartphone users are formulated as a Stackelberg game, and the MCS system can apply the DQN algorithm to derive the optimal MCS policy against faked sensing attacks. The work in [22] provides a secure edge caching scheme for the content provider and mobile users in a mobile social network (MSN), models the interactions between the CP and edge caching devices as a Stackelberg game, and employs Q-leaning to derive the optimal strategies of agents. Since reinforcement learning learns strategies through the interaction between the agent and environment via trial and error, it is particularly suitable for the dynamic game scenario with incomplete information, for example, between the NO and VNO in our work.

3 System Model

In this section, we model the strategies of the four types of decision-makers in the incentive mechanism, NO, VNO, MUs, and EMs, according to their roles played in the scenario of an EP being a VNO. Specifically, as shown in Fig. 1, the NO is in charge of selling the mobile data and data plan to the VNO and MUs respectively. The VNO not only sells its data plan to MUs and charges EMs for advertising but also rewards MUs for watching ads of EMs, which may potentially stimulate users to spend on the EM. MUs have the right to choose the NO or VNO for data service and decide whether and how to earn data rewards by watching ads of EMs if subscribing to the VNOs data plan. EMs may purchase ad slots in the EP to improve business profits. Considering a static game scenario, in this section we formulate the aforementioned interactions of all four types of game participators as a three-stage Stackelberg game. The key notations are listed in Table 1.

TABLE 1: Key Notations

Notation	Definition
A	Parameter of EMs' utility function
B	Total network capacity of the NO
B_0	Amount of mobile data used by the NO itself
B_1	Amount of mobile data purchased by the VNO
c	Data price of the NO
F_0	Data plan fee of the NO
F_1	Data plan fee of the VNO
G	Total number of ads to be watched by all MUs
$g(\theta)$	Probability density function of θ
h	Users' average disutility of watching one ad
m_s	Number of ad slots purchased by EM s
N	Number of MUs
p	Price of each ad in the EP
Q_0	Data amount of the NO's data plan
Q_1	Data amount of the VNO's data plan
r	An MU's data plan subscription decision
S	Number of EMs
w	Amount of data reward for watching per ad
x	Number of ads that a user chooses to watch
z	Amount of data that a user obtains from its subscription and ad watching
θ_{m}	User type parameter
Γ^{MU}	Utility of an MU
Γ^{EM}	Utility of an EM
Γ^{VNO}	Utility of the VNO
Γ^{NO}	Utility of the NO

3.1 Network Operator

To gain insights into the system design, we focus on a single data plan scenario, which has been widely considered in the literature (e.g., [12], [15]). We consider a monopolistic NO offering a predetermined (monthly) flat-rate data plan, (F_0, Q_0), to MUs, where $F_0 > 0$ denoting the subscription fee, and $Q_0 > 0$ the data amount associated with a subscription. The NO decides the price of mobile data c (i.e., the payment for purchasing unit mobile data) for the VNO. Suppose that in the system the network has a total network capacity of B [5], [15], e.g., LTE base stations can handle only a finite amount of traffic at any given time. In addition, NOs generally bill their subscribers on a monthly basis, so we will mainly consider the non-cooperative game of participants within a billing cycle. Let the amount of mobile data used by the NO be B_0 and the amount of mobile data purchased by the VNO be B_1. Thus, we have

$$B_0 + B_1 \leq B. \tag{1}$$

3.2 Virtual Network Operator

For the communication business, the VNO offers a monthly flat-rate data plan, (F_1, Q_1), to MUs, where $F_1 > 0$ denoting the subscription fee, and $Q_1 > 0$ the data amount associated with a subscription. In practice, the data amount of the VNO's data plan is usually used for some exclusive applications with large traffic of its own enterprise or for attracting MUs. Hence, this type of data plan usually has the characteristics of more data amount but a lower price for unit mobile data as compared with the NO's, which however results in a higher data plan fee. Without loss of generality, we assume $Q_0 < Q_1$ and $F_0 < F_1$.

To incentivize MUs to choose its data plan while watching ads placed on its EP for EMs, we consider that the VNO will give w mobile data per ad to those subscribers who use its data plan while also watching ads on its EPs, similar to some data rewarding schemes in reality.

The VNO needs to decide two variables: (i) data plan fee F_1, while the amount of mobile data Q_1 is supposed to be predetermined; (ii) an ad price p, which is the price that the VNO charges the EMs for buying one ad slot. Here, we consider a price-based mechanism, where the VNO sells the ad slots in advance at a fixed price.

3.3 Mobile User

We consider a continuum of users, and denote the mass of users by N. Let θ denote a user's type, which parameterizes its valuation for mobile service [8], [13]. We assume that θ is a continuous random variable drawn from $[0, \theta_{\mathrm{m}}]$, and its probability density function $g(\theta)$ satisfies $g(\theta) > 0$ for all $\theta \in [0, \theta_{\mathrm{m}}]$. For convenience of analysis, similar to [23] we assume that θ is uniformly distributed in $[0, \theta_{\mathrm{m}}]$.

Let $r \in \{0, 1\}$ denote a user's data plan subscription decision, and $x \in [0, \infty)$ denote the number of ads that a user chooses to watch (during one month). We allow x and the advertisers' purchasing decisions to be fractional [13], [24]. Then, the amount of data that a user obtains from its subscription and ad watching is

$$z = (Q_1 + wx)r + Q_0(1 - r) \tag{2}$$

where the first term represents the data amount of the MU obtains if they subscribe to the data plan of the VNO and watch x ads, and the second item represents the data amount of the MU obtains if they subscribe to the data plan of the NO.

We use $\theta u(z)$ to capture a type-θ user's utility of using the mobile service. Here, $u(z), z \geq 0$, is the same for all users, and can be any strictly increasing, strictly concave, and twice differentiable function that satisfies $u(0) = 0$ and $\lim_{z \to \infty} u'(z) = 0$ [8]. The concavity of $u(z)$ captures the diminishing marginal return with respect to the data amount. Different utility functions have been considered in several recent works, such as α-fair function [15], logarithmic function [23], and exponential function [25]. To simplify analysis, we consider a specific α-fair utility function: $u(z) = 2\sqrt{z}$. Although extension to other similar utility functions is straightforward, in Section IV to obtain insights analytically we focus on the aforementioned utility function. Further, in Section V, we study the DQN-based optimal strategies of the NO and VNO, where the constraint of the simple form of the utility function can be relieved.

Hence, a type-θ user's utility can be expressed as

$$\Gamma^{\mathrm{MU}} = \theta u(z) - (F_1 + hx)r - F_0(1 - r) \tag{3}$$

where h denotes the user's average disutility (e.g., inconvenience) of watching one ad. We assume that the total disutility of watching ads linearly increases with the number of watched ads [26], [27]. In Section III-A we analyze the user's optimal decisions r^* and x^*. In the Nash equilibrium, we denote the total number of ads decided by all MUs G as

$$G = N \int_0^{\theta_{\mathrm{m}}} x^*(\theta) g(\theta) d\theta. \tag{4}$$

3.4 E-commerce Merchant

We consider S homogeneous EMs,[1] which implies that the utilities of EMs are the same [8]. The EMs have an incentive to purchase ad slots in the EP to promote their own business by displaying ads for MUs. Hence, the EMs need to decide the required number of ad slots, m_s, where $s \in \{1, 2, ..., S\}$. Considering that each EP adopts a random strategy for all ads to N MUs and each time an MU watches one ad, the probability that the MU sees the ad of EM s is $\frac{m_s}{G}$ [8], [13]. Notice that the EP can affect the total purchase volume of EMs by adjusting the price per unit of ad p to make sure $\sum_s m_s \leq G$. If $\sum_s m_s = G$, a user will always see an EM's ad at random; if $\sum_s m_s < G$, there is a certain probability that a user will not see any EMs' ad, in which case we consider replacing it with the EP's ad, or not showing any ad to the user.

In the related works of advertising business model, the utility of EMs increases and then decreases with the increment of $\frac{m_s}{G}$, which reflects the advertising's wear-out effect. This is because too many repetitions may make the user have a bad impression of the product. Some studies, such as [26] and [28], explicitly consider a quadratic relation between the ad repetition and the advertising's effectiveness. Hence, we adopt a quadratic function as well to model the utility of the EMs, given by

$$\Gamma^{\mathrm{EM}} = A\frac{m_s}{G} - \left(\frac{m_s}{G}\right)^2 - pm_s \tag{5}$$

where $A > 0$ is a system parameter. Note that a smalller A in (5) reflects a stronger degree of wear-out effect.

3.5 Three-Stage Stackelberg Game

For the static game scenario, we can model the interactions among the NO, VNO, MUs, and EMs by a three-stage Stackelberg game. In Stage I, the NO decides the price of mobile data (i.e., the payment for purchasing unit mobile data) for the VNO. In Stage II, the VNO decides the data plan fee for the MUs and the ad price (i.e., the payment for purchasing one ad slot) for the EMs. In Stage III, the MUs with different valuations for the mobile service make their data plan subscription and ad watching decisions. Meanwhile, the EMs decide the number of ad slots to be purchased.

4 Three-Stage Game Analysis

In this section, we analyze the three-stage Stackelberg game by backward induction [29]. It is noteworthy that, in the static Stackelberg game, information is transparent in one direction, i.e., the leaders know all information of followers.

4.1 Stage III: MUs' Access and EMs' Advertising

In this subsection, we analyze each MU's optimal data plan choice and the number of ads chosen to watch in the EP, and the EMs' optimal advertising strategy in Stage III. The MUs and EMs make their decisions by responding to the NO's price of mobile data c in Stage I, and to the VNO's decisions of F_1 and p in Stage II.

4.1.1 MUs' Optimal Access

To maximize the utility of mobile service, a type-θ user's decision strategy can be formulated as the following optimization problem (OP):

$$\text{OP 1}: \max_{r\in\{0,1\}, x\in[0,\infty)} \Gamma^{\mathrm{MU}} \tag{6}$$

where Γ^{MU} is given in (3). Although OP 1 is a mixed-integer OP, we can characterize the MU's decision in the following proposition by detailed case-by-case analysis.

Proposition 1. The optimal decisions of a type-θ user ($\theta \in [0, \theta_{\mathrm{m}}]$) are as follows:

Case 1: If $F_0 < F_1 < 2\frac{h\sqrt{Q_1}}{w}(\sqrt{Q_1} - \sqrt{Q_0}) + F_0$, $x^* = \begin{cases} 0 & \theta \in [0, \theta_2) \\ 0 & \theta \in [\theta_2, \theta_1) \\ \frac{\theta^2 w}{h^2} - \frac{Q_1}{w} & \theta \in [\theta_1, \theta_{\mathrm{m}}] \end{cases}$, $r^* = \begin{cases} 0 & \theta \in [0, \theta_2) \\ 1 & \theta \in [\theta_2, \theta_1) \\ 1 & \theta \in [\theta_1, \theta_{\mathrm{m}}] \end{cases}$;

Case 2: If $2\frac{h\sqrt{Q_1}}{w}(\sqrt{Q_1} - \sqrt{Q_0}) + F_0 \leq F_1 \leq \frac{{\theta_{\mathrm{m}}}^2 w}{h} - 2\theta_{\mathrm{m}}\sqrt{Q_0} + \frac{hQ_1}{w} + F_0$, $x^* = \begin{cases} 0 & \theta \in [0, \theta_3) \\ \frac{\theta^2 w}{h^2} - \frac{Q_1}{w} & \theta \in [\theta_3, \theta_{\mathrm{m}}] \end{cases}$, $r^* = \begin{cases} 0 & \theta \in [0, \theta_3) \\ 1 & \theta \in [\theta_3, \theta_{\mathrm{m}}] \end{cases}$;

Case 3: If $F_1 > \frac{{\theta_{\mathrm{m}}}^2 w}{h} - 2\theta_{\mathrm{m}}\sqrt{Q_0} + \frac{hQ_1}{w} + F_0$, $x^* = 0$, $r^* = 0, \theta \in [0, \theta_{\mathrm{m}}]$, where

$$\theta_0 = \frac{h\sqrt{Q_0}}{w} \tag{7a}$$

$$\theta_1 = \frac{h\sqrt{Q_1}}{w} \tag{7b}$$

$$\theta_2 = \frac{F_1 - F_0}{2(\sqrt{Q_1} - \sqrt{Q_0})} \tag{7c}$$

$$\theta_3 = \frac{h}{w}\left(\sqrt{Q_0} + \sqrt{Q_0 - Q_1 - \frac{w}{h}(F_0 - F_1)}\right). \tag{7d}$$

In Case 1, there are three types of MUs—the first type of MUs choosing the NO's data plan, the second type of MUs choosing the VNO's data plan but without watching ads, and the last type choosing the VNO's data plan and watching ads. In Case 2, we can discover two types of MUs—the first type of MUs choosing the NO's data plan, and the second choosing the VNO's data plan and watching ads. In Case 3, all MUs choose to subscribe to the data plan of the NO, which is contrary to the motivation of the incentive mechanism we design. Therefore, in the following part of the analysis, we focus on Case 1 and Case 2.

4.1.2 EMs' Optimal Advertising

To maximize the utility of advertising, each EM should solve the following optimization problem:

$$\text{OP 2}: \max_{m_s} \ \Gamma^{\mathrm{EM}} \tag{8a}$$

$$\text{s.t. } m_s \geq 0 \tag{8b}$$

where Γ^{EM} is given in (5). Because OP 2 is a convex OP, we can find its optimal solution easily as follows.

1. In the static game scenario, if heterogeneous EMs are considered, the closed-form solution to the optimization problems in Section IV cannot be obtained. However, in the dynamic game scenario, the assumption can be relieved if a learning-based approach is utilized as to be studied in Section V.

Proposition 2. In both Case 1 and Case 2, the optimal decision for each EM can be expressed as

$$m_s^* = G\frac{A - pG}{2}. \tag{9}$$

4.2 Stage II: VNO's Prices of Data Plan and Advertising

The VNO obtains revenue from both the mobile data market and the ad market. Next, we analyze the VNO's strategy by considering the two cases on MUs being categorized, as found in Proposition 1.

4.2.1 Case 1

In the mobile data market, each MU with $\theta \in [\theta_2, \theta_{\rm m}]$ who subscribes to the data plan of the VNO should pay F_1 to it, and the VNO should pay for mobile data to the NO. The VNO's corresponding revenue is

$$\Gamma^{\rm data} = N\left(1 - \frac{\theta_2}{\theta_{\rm m}}\right)F_1 - cB_1. \tag{10}$$

In the ad market, each EM pays p for each purchased ad slot. The VNO's corresponding revenue is

$$\Gamma^{\rm tax} = Spm_s^* = SpG\frac{A - pG}{2}. \tag{11}$$

Recall that B_1 denotes the total data demand of the VNO's subscribers, i.e., the total amount of mobile data that MUs request (by subscription and watching ads) given reward w. Based on Proposition 1, we can compute B_1 as

$$B_1 = N\frac{\theta_1 - \theta_2}{\theta_{\rm m}}Q_1 + N\int_{\theta_1}^{\theta_{\rm m}} \frac{\theta^2 w^2}{h^2} g(\theta)d\theta \tag{12}$$

where the first item means that the MUs with $\theta \in [\theta_2, \theta_1)$ will get Q_1 amount of data from the VNO and the second item means that the MUs with $\theta \in [\theta_1, \theta_{\rm m}]$ will get $\frac{\theta^2 w^2}{h^2}$ amount of data. Because only the MUs with $\theta \in [\theta_1, \theta_{\rm m}]$ will choose to watch ads, we can compute G as

$$G = N\int_{\theta_1}^{\theta_{\rm m}} (\frac{\theta^2 w}{h^2} - \frac{Q_1}{w})\frac{1}{\theta_{\rm m}}d\theta. \tag{13}$$

To maximize the total utility of both data and ad markets, the VNO's strategy design in Case 1 can be formulated as the following optimization problem:

$$\text{OP 3}: \quad \max_{F_1, p} \quad \Gamma^{\rm VNO} = \Gamma^{\rm data} + \Gamma^{\rm tax} \tag{14a}$$

$$\text{s.t.} \quad S\sum_s m_s^* \le G. \tag{14b}$$

Here, constraint (14b) implies that the total number of ads purchased by the EMs should not exceed the total number of ads decided by the MUs. As OP 3 is a convex optimization problem, with some manipulation, we have the following proposition.

Proposition 3. In Case 1, the optimal strategy for VNO can be expressed as

$$F_1^* = \theta_{\rm m}\left(\sqrt{Q_1} - \sqrt{Q_0}\right) + \frac{F_0}{2} + \frac{cQ_1}{2} \triangleq F_{\rm A} \tag{15a}$$

$$p^* = \frac{A - \frac{2}{S}}{G}. \tag{15b}$$

4.2.2 Case 2

We study the VNO's strategy design in Case 2 below. In the mobile data market, each MU with $\theta \in [\theta_3, \theta_{\rm m}]$ who subscribes to the data plan of the VNO should pay F_1 to it, and the VNO should pay for mobile data to the NO. The VNO's mobile data revenue is

$$\Gamma^{\rm data} = N\left(1 - \frac{\theta_3}{\theta_{\rm m}}\right)F_1 - cB_1. \tag{16}$$

In the ad market, the VNO's revenue for selling advertising is the same with (11). In Case 2, each MU with $\theta \in [\theta_3, \theta_{\rm m}]$ who subscribes to the data plan of the VNO will get $\frac{\theta^2 w^2}{h^2}$ amount of data. Then, we can compute B_1 as

$$B_1 = N\int_{\theta_3}^{\theta_{\rm m}} \frac{\theta^2 w^2}{h^2} g(\theta)d\theta. \tag{17}$$

Here, because only the MUs with $\theta \in [\theta_3, \theta_{\rm m}]$ will choose to watch ads, we can compute G as

$$G = N\int_{\theta_3}^{\theta_{\rm m}} (\frac{\theta^2 w}{h^2} - \frac{Q_1}{w})\frac{1}{\theta_{\rm m}}d\theta. \tag{18}$$

Similar to OP 3, to maximize the total utility of both data and ad markets, the VNO's OP in Case 2 can be formulated as

$$\text{OP 4}: \quad \max_{F_1, p} \quad \Gamma^{\rm VNO} = \Gamma^{\rm data} + \Gamma^{\rm tax} \tag{19a}$$

$$\text{s.t.} \quad S\sum_s m_s^* \le G. \tag{19b}$$

OP 4 is a non-convex optimization problem. However, recalling that $\theta_3 = \frac{h}{w}\left[\sqrt{Q_0} + \sqrt{Q_0 - Q_1 - \frac{w}{h}(F_0 - F_1)}\,\right]$, we can denote F_1 with a function of θ_3 as

$$F_1 = \frac{h}{w}\left[\left(\frac{\theta_3 w}{h} - \sqrt{Q_0}\right)^2 + Q_1 - Q_0\right] + F_0 \triangleq \phi(\theta_3). \tag{20}$$

Then, we can transform OP 4 into a non-convex optimization problem about variables θ_3 and p as

$$\text{OP 5}: \max_{\theta_3, p} \quad N\left(1 - \tfrac{\theta_3}{\theta_{\rm m}}\right)\phi(\theta_3) - cB_1 + Spm_s^* \tag{21a}$$

$$\text{s.t.} \quad S\sum_s m_s^* \le G \tag{21b}$$

$$\theta_3 \in [\theta_1, \theta_{\rm m}]. \tag{21c}$$

The objective function of OP 5 is a cubic function and constraints (21b)-(21c) are linear constraints. Then, its optimal solution must be obtained at the endpoints of interval $[\theta_1, \theta_{\rm m}]$ or at the stationary points. Thus, we have the following proposition.

Proposition 4. In Case 2, the optimal strategy for the VNO can be expressed as

$${\theta_3}^* = \arg\max\{\hat{\Gamma}^{\rm VNO}(\theta_1), \hat{\Gamma}^{\rm VNO}(\theta_4)\} \tag{22a}$$

$$p^* = \frac{A - \frac{2}{S}}{G} \tag{22b}$$

where $\hat{\Gamma}^{\rm VNO}(\theta_3) = \Gamma^{\rm VNO}(F_1 = \phi(\theta_3))$ and θ_4 is the root of equation

$$\frac{d\hat{\Gamma}^{\rm VNO}(\theta_3)}{d\theta_3} = 0. \tag{23}$$

For the sake of convenience to express, we define $\phi(\theta_1) \triangleq F_{\rm B}$ and $\phi(\theta_4) \triangleq F_{\rm C}$. Given an arbitrary price c of the NO, the VNO should compare its optimal utilities in Case 1 and Case 2 to decide the data plan fee F_1, so the optimal decision of the VNO is affected by price c of the NO and we have the following proposition.

Proposition 5. From the perspective of the NO, the optimal strategy of the VNO in Stage II can be summed up as

$$F_1^* = \begin{cases} F_{\rm A} & c \in \Omega_1 \ \text{ Case A} \\ F_{\rm B} & c \in \Omega_2 \ \text{ Case B} \\ F_{\rm C} & c \in \Omega_3 \ \text{ Case C} \end{cases} \tag{24a}$$

$$p^* = \frac{A - \frac{2}{S}}{G} \tag{24b}$$

where intervals $\Omega_i, i \in \{1,2,3\}$ are divided by comparing the VNO's utilities $\Gamma^{\rm VNO}(F_1^* = F_{\rm A})$ in Case 1, $\Gamma^{\rm VNO}(F_1^* = F_{\rm B})$ and $\Gamma^{\rm VNO}(F_1^* = F_{\rm C})$ in Case 2. For example, when $c \in \Omega_1$, we have $\Gamma^{\rm VNO}(F_1^* = F_{\rm A}) > \max\{\Gamma^{\rm VNO}(F_1^* = F_{\rm B}), \Gamma^{\rm VNO}(F_1^* = F_{\rm C})\}$. Intervals Ω_i's can be found numerically with Newton's method or by *fzeros()* function in MATLAB.

Substituting the VNO's optimal strategy in formula (24b) into the EMs' optimal strategy in formula (9) and utility function of formula (5), we can obtain the following proposition.

Proposition 6. Given the optimal strategy of the VNO, the optimal strategy and utility of the EMs can be derived as

$$m_s^* = \frac{G}{S} \tag{25a}$$

$$\Gamma^{\rm EM}(F_1^*, p^*) = \frac{1}{S^2}. \tag{25b}$$

Equation (25a) implies that, as long as both VNO and MUs make their optimal decisions in the proposed business model, the optimal strategies of all homogeneous EMs are the same and only affected by the number of competing EMs and the total number of ads decided by all MUs. Equation (25b) implies that the optimal utilities of all homogeneous EMs are the same and only affected by the number of EMs, regardless of any other system parameters. It is also noteworthy that, although the optimal utility of each EM is relatively small, the EMs can always achieve their goals of positive revenues in the Nash equilibrium.

4.3 Stage I: NO's Price of Mobile Data

The NO obtains revenue from both the MUs who subscribe to it and the VNO in the mobile data market. Next, we analyze the NO's strategy by considering the three cases of the VNO's optimal strategy being categorized, as found in Proposition 5.

4.3.1 Case A

In this case, to maximize the revenue of the NO, we study the following optimization problem:

$$\text{OP 6}: \quad \max_{c} \quad \Gamma^{\rm NO} = N\frac{\theta_2}{\theta_{\rm m}}F_0 + cB_1 \tag{26a}$$

$$\text{s.t.} \quad N\frac{\theta_2}{\theta_{\rm m}}Q_0 + B_1 \le B \tag{26b}$$

$$c \in \Omega_1 \tag{26c}$$

where θ_2 given in Proposition 1 is a function of $F_1^* = F_{\rm A} = \theta_{\rm m}\left(\sqrt{Q_1} - \sqrt{Q_0}\right) + \frac{F_0}{2} + \frac{cQ_1}{2}$ that is derived for the VNO in Case 1 of Stage II, and B_1 given in (12) is a function of θ_2. Constraint (26b) implies that the total demand of mobile data cannot exceed the network capacity B of the NO. OP 6 is convex, thus can be solved by CVX-toolbox in MATLAB. Let c_1^* denote its optimal solution.

4.3.2 Case B

In this case, to find the NO's optimal strategy we solve the following optimization problem similar to OP 6:

$$\text{OP 7}: \quad \max_{c} \quad \Gamma^{\rm NO} = N\frac{\theta_1}{\theta_{\rm m}}F_0 + cB_1 \tag{27a}$$

$$\text{s.t.} \quad N\frac{\theta_1}{\theta_{\rm m}}Q_0 + B_1 \le B \tag{27b}$$

$$c \in \Omega_2. \tag{27c}$$

Notice that, in OP 7, θ_1 is given in Proposition 1, and B_1 is given in (17) which is a function of θ_3 thus a function of $F_1^* = F_{\rm B} = \phi(\theta_1)$ given in (24a) in Case 2 of Stage II. As OP 7 is a linear optimization problem, we can solve it by CVX-toolbox in MATLAB as well. Let c_2^* denote the optimal solution to OP 7.

4.3.3 Case C

In Case C, to derive the optimal strategy of the NO, we study the following optimization problem:

$$\text{OP 8}: \quad \max_{c} \quad \Gamma^{\rm NO} = N\frac{\theta_4}{\theta_{\rm m}}F_0 + cB_1 \tag{28a}$$

$$\text{s.t.} \quad N\frac{\theta_4}{\theta_{\rm m}}Q_0 + B_1 \le B \tag{28b}$$

$$c \in \Omega_3 \tag{28c}$$

where θ_4 is given in (23), and B_1 given in (17) is a function of θ_3 thus a function of $F_1^* = F_{\rm C} = \phi(\theta_4)$ given in (24a) in Case 2 of Stage II. Unfortunately, different from OPs 6 and 7, OP 8 is a non-convex optimization problem. However, similar to OP 4, we can transform it into a non-convex optimization problem with a cubic objective function of θ_4. Then, the optimal solution to OP 8 must be obtained at the endpoints of interval Ω_3 or at the stationary points. Let c_3^* denote its optimal solution.

Finally, the NO can compare its local optimal utilities in Cases A, B, and C to find its global optimal strategy, as summarized in the following proposition.

Proposition 7. The global optimal solution of the NO in Stage I can be derived as

$$c^* = \arg\max_{c} \left\{\Gamma^{\rm NO}(c_1^*), \Gamma^{\rm NO}(c_2^*), \Gamma^{\rm NO}(c_3^*)\right\}. \tag{29}$$

4.4 Social Welfare

Here, we study the social welfare (SW) of the whole system at the equilibrium, which consists of the NO's revenue, the VNO's revenue, the MUs' total utility, and the EMs' total payoff. The social welfare analysis is important for understanding how much the entire system benefits from the proposed incentive mechanism for advertising via mobile data reward, and how it is affected by different system parameters. Specifically, we can compute the SW as

$$\begin{aligned} SW =& \Gamma^{\mathrm{NO}}\left(c^*\right) \\ &+ \Gamma^{\mathrm{VNO}}\left(c^*, F_1^*, p^*\right) + S\Gamma^{\mathrm{EM}}\left(c^*, F_1^*, p^*\right) \\ &+ N\int_0^{\theta_{\mathrm{m}}} \Gamma^{\mathrm{MU}}\left(\theta, c^*, F_1^*, p^*\right) g\left(\theta\right) d\theta. \end{aligned} \tag{30}$$

Since it is difficult to analytically judge the monotonicity of the SW regarding system parameters, such as the VNO's data plan fee, data reward per ad, and data amount of the VNO's data plan, we study the impact of system parameters on the SW based on simulation results in Section VI.

5 DQN-based Optimal Strategy Decision for Dynamic Stackelberg Game

In practice, the interaction among the NO, VNO, EMs, and MUs is often a dynamic game process, i.e., the NO, VNO, EMs, and MUs can re-examine their optimal choices in each cycle. At the same time, the game players, especially the NO and VNO are lack of the other's knowledge on interaction. Hence, the game between the NO and VNO becomes a non-cooperative dynamic game with incomplete information. In this section, we consider applying the DQN method in multi-agent reinforcement learning to solve this problem.

Consider a dynamic game process and the duration of each decision cycle is the same as the traffic billing cycle, e.g., on a monthly basis. In this game scenario, considering that MUs and EMs are relatively rational, i.e., they can make their own optimal decisions according to Proposition 1 and Proposition 2, we can model the problem in each cycle as a two-stage Stackelberg game, i.e., the NO and the VNO are considered as leaders, while the MUs and the EMs are considered as followers.

Since the EMs and MUs are relatively rational, we model the EMs and MUs as the environment, and NO and VNO as two agents in reinforcement learning, both of whom are required to learn their optimal strategies in a long-term game.

5.1 DQN-based Dynamic Game Model

5.1.1 Environment

In the following we discuss the information ownership of the NO and VNO about the MUs and EMs. The information ownership of the NO and VNO about the MUs is considered as follows:

- $g(\theta)$: The distribution of the MU's type parameter θ is unknown to both the NO and VNO.
- G: The number of ads to be watched by all MUs is unknown to the NO. But, we consider that it is known to the VNO from, for example, its MU subscription management system.
- N: The total number of MUs, which is public and fixed, is known to both the NO and VNO. The NO knows the number of subscribers to its data plan N_0 and the VNO knows the number of subscribers to its data plan N_1. So they can infer the number of subscribers to each other's data plan.
- Γ^{MU}: The utility function of MUs is neither known to the NO nor VNO, but the final data plan choice of each MU is publicly available.

The information ownership of the NO and VNO about the EMs is considered as follows:

- m_s^*: The number of ads purchased by each EM is unknown to the NO but known to the VNO, thanks to the VNO's EM management system.
- Γ^{EM}: Utility functions for EMs are neither known to the NO nor the VNO.

5.1.2 State

The state of an agent is defined by the information it has. For the NO, state S_0^t in cycle t includes the number of subscribers N_0^{t-1} to the NO's data plan, the amount of mobile data, B_1^{t-1}, purchased by the VNO from the NO, and the data plan fee of the VNO, F_1^{t-1}, all at the beginning of the cycle, i.e.,

$$S_0^t = \left(N_0^{t-1}, B_1^{t-1}, F_1^{t-1}\right). \tag{31}$$

Similarly, state S_1^t of the VNO in cycle t includes the number of the NO's subscribers, N_1^{t-1}, the total number of ads, G^{t-1}, that all MUs decide to watch, the number of ads, m_s^{t-1}, that each EM decides to buy, and price c^{t-1} of mobile data set by the NO, all at the beginning moment of the cycle, i.e.,

$$S_1^t = \left(G^{t-1}, N_1^{t-1}, m_s^{t-1}, c^{t-1}\right). \tag{32}$$

5.1.3 Action

The actions of the agents in DQN method are discrete variables, so we consider discrete decision variables for the NO and VNO. The action of the NO includes only the price of mobile data sold to the VNO, $A_0^t = c^t$, where $c^t \in \mathbf{C} = \{c_1, c_2, ..., c_I\}$ and I denotes the dimension of $\mathbf{C}$.

The VNO's action includes its data plan fee and the price of the ad sold to the EMs, $A_1^t = (F_1^t, p^t)$ where $F_1^t \in \mathbf{F} = \{F_{1,1}, F_{1,2}, ..., F_{1,J}\}$ with J denoting the dimension of $\mathbf{F}$ and $p^t \in \mathbf{p} = \{p_1, p_2, ..., p_K\}$ with K denoting the dimension of $\mathbf{p}$.

5.1.4 Reward Function

In each cycle, the NO and the VNO jointly take actions and both agents update their states by observing the environment while receiving rewards. For the NO, the reward consists of two components, the revenue of data plan $N_0^t F_0^t$ paid by subscribers and the fee paid by the VNO $c^t B_1^t$ for the mobile data, i.e.,

$$\Gamma_0^t = N_0^t F_0 + c^t B_1^t. \tag{33}$$

For the VNO, the reward consists of three components, namely the fee paid by subscribers $N_1^t F_1^t$ to its data plan, the fee paid by all EMs who buy its ads $Sp^t m_s^t$, and the fee paid to the NO for mobile data $c^t B_1^t$, i.e.,

$$\Gamma_1^t = N_1^t F_1^t + Sp^t m_s^t - c^t B_1^t. \tag{34}$$

5.1.5 Updating Approach

In reinforcement learning, the Q function represents the expected long-term discounted utility of an agent. For the NO and VNO, according to the Bellman equation, the Q function can be expressed as

$$Q(S_i^t, A_i^t) = (1-\lambda_i)\, Q(S_i^t, A_i^t) + \lambda_i \left(\Gamma_i^t + \gamma_i V\left(S_i^{t+1}\right)\right) \tag{35}$$

where $i \in \{0, 1\}$ with $i = 0$ representing the NO and $i = 1$ representing the VNO, λ_i is the learning rate, γ_i is the discount factor of agent i, and V function represents the maximum value that can be achieved by the Q function, denoted as

$$V\left(S_i^{t+1}\right) = \max_{A_i} Q(S_i^{t+1}, A_i^{t+1}). \tag{36}$$

In the training process, when each agent takes an action in each cycle, the ε-greedy strategy is used, i.e., the action that maximizes the Q function is selected with probability $1-\varepsilon$, and an action in the action space is randomly selected with probability ε, which can be expressed as

$$\Pr\left\{A_i^t = A_i^*\right\} = \begin{cases} 1-\varepsilon, & A_i^* = \arg\max_{A_i} Q(S_i^t, A_i) \\ \varepsilon, & \text{random policy.} \end{cases} \tag{37}$$

With the DQN method, we can use well-trained convolution neural networks (CNNs) to estimate the Q-values [30], i.e.,

$$Q\left(S_i^t, A_i^t\right) \approx \widehat{Q}\left(S_i^t, A_i^t; \omega_i\right) \tag{38}$$

where ω_i indicates the CNN's parameters of agent i.

5.1.6 Loss Function

During each iteration of training, we train the CNN of each agent by decreasing the loss function, which is represented as follows

$$\begin{aligned} L\left(\omega_i\right) = & \mathbb{E}\Bigg\{\Bigg[\Gamma_i^t + \gamma_i \max_{A_i^t} \widehat{Q}\left(S_i^{t+1}, A_i^t; {\omega_i}'\right) \\ & - \widehat{Q}\left(S_i^t, A_i^t; \omega_i\right)\Bigg]^2\Bigg\}. \end{aligned} \tag{39}$$

The training process uses back-propagation (BP) to train the network, and the loss function's gradient of the agent i can be expressed as

$$\begin{aligned} \nabla_{\omega_i} L\left(\omega_i\right) = & -\mathbb{E}\Bigg\{\Bigg[\Gamma_i^t + \gamma_i \max_{A_i^t} \widehat{Q}\left(S_i^{t+1}, A_i^t; {\omega_i}'\right) \\ & - \widehat{Q}\left(S_i^t, A_i^t; \omega_i\right)\Bigg] \nabla_{\omega_i} \widehat{Q}\left(S_i^t, A_i^t; \omega_i\right)\Bigg\}. \end{aligned} \tag{40}$$

5.2 Proposed DQN-based Algorithm

In our proposed DQN-based algorithm, in addition to s-trategies such as ε-greedy mentioned above, we use the following methods to make the DQN-based algorithm more efficient.

- First, in the proposed DQN-based algorithm, there are two neural networks for each agent, for example, for the NO, one neural network called evaluated Q-network changes parameter w_0 to decrease the loss function during the training process based on the gradient of the loss function, the other neural network called target Q-network changes parameter w_0' to calculate the estimated target Q-value during training, and its parameters are updated every few steps. These two networks can be used to reduce the oscillation or divergence of strategies during the training process [31];
- The DQN-based algorithm of each agent stores its own experience during training through an experience replay, for example, for the NO, whose experience can be represented as a tuple, $(S_0^t, A_0^t, \Gamma_0^t, S_0^{t+1})$. During the training process, the neural network is timely updated based on the gradient descent of the randomly sampled mini-batch to minimize the loss function;
- The training methods can be divided into synchronous training and asynchronous training. Synchronous training refers to the joint action of multiple agents. In contrast, asynchronous training means that agents are trained one by one, and the strategies of other agents are not updated when one agent is trained. However, in our scenario, notice that the NO's reward is affected directly by the VNO instead of the environment, so it is difficult to establish a direct relationship between its action and reward, thus increasing the difficulty of training. Hence, we can consider to train both the NO and VNO at the same time in the initial period of training process and

Algorithm 1 DQN-based optimal strategies of the NO and VNO

Initialize states S_0^t and S_1^t;
Initialize the evaluated Q-network's and target Q-network's parameters of the NO and VNO, i.e., ω_0, ${\omega_0}'$, ω_1, and ${\omega_1}'$;
Initialize replay memory D_0 of the NO and D_1 of the VNO to capacity N_D;
Set the maximum training time of the network as T;
for $t = 1,\ 2,\ ...,\ T$ **do**
    Obtain state $S_0^t = \left(N_0^{t-1}, B_1^{t-1}, F_1^{t-1}\right)$ of the NO and state $S_1^t = \left(G^{t-1}, N_1^{t-1}, m_s^{t-1}, c^{t-1}\right)$ of the VNO;
    The NO selects c^t via the ε-greedy algorithm, executes action, and sends c^t to the VNO;
    The VNO selects F_1^t and p^t via the ε-greedy algorithm, executes action, and sends F_1^t and p^t to the MUs and EMs, respectively;
    The NO observes and obtains N_0^t, B_1^t, and F_1^t, while the VNO observes and obtains the G^t, N_1^t, m_s^t, and c^t;
    The NO and VNO calculate their own rewards Γ_0^t and Γ_1^t;
    for $i = 0,\ 1$ **do**
        Store experience $(S_i^t,\ A_i^t,\ \Gamma_i^t,\ S_i^{t+1})$ in D_i;
        Sample random minibatch of experience from D_i;
        Update the CNN weights via gradients of loss function (40);
        Set ${\omega_i}' = \omega_i$ every few steps;
    end for
end for

then train both the NO and VNO asynchronously.

The details of the DQN-based algorithm are given in Algorithm 1. It is noteworthy that, in the dynamic game scenario, both the NO and VNO have some information that the other does not have. For example, the NO does not know how many ads the VNO's subscribers will watch, and the VNO does not know the total amount of mobile data that the NO possesses. Further, the NO and VNO can make decisions only based on their observations or experience obtained from the previous cycles. As such, in the proposed algorithm we consider simultaneous decision-making for the NO and VNO.

The computational complexity of the proposed algorithm depends on the structure of the implemented neural networks, and increases with the number of hidden layers and the number of corresponding neurons. Let N_S, N_A, H_l, and L represent the dimension of the input (observation space), the dimension of the output (action space), the number of neuron nodes in the l-th hidden layer, and the total number of the hidden layers in the neural network, respectively. Then, in each episode of the training process, the computational complexity of the proposed algorithm expressed in Big-O notation is given by $O(N_S H_1 + \sum_{l=1}^{L} H_l + \sum_{l=1}^{L-1} H_l H_{l+1} + H_L N_A)$ [32], [33].

TABLE 2: SIMULATION DEFAULT PARAMETERS

Variable	Default value
Amount of data reward for watching per ad w	0.275 GB
Data plan fee of the NO F_0	30 CNY
Data amount of the NO's data plan Q_0	30 GB
Data amount of the VNO's data plan Q_1	50 GB
Total network capacity of the NO B	10^{11} GB
Number of MUs N	10^8
Number of EMs S	10
Parameter A of EMs' utility function	10^{10}
User type parameter θ_{m}	50
Users' average disutility of watching one ad h	1

TABLE 3: TRAINING PARAMETERS

Variable	Default value
Learning rate	0.0001
Discount factor	0.95
Exploration rate	$1.0 \rightarrow 0.1$
Hysteretic rate	$0.2 \rightarrow 0.8$
Number of exploration episodes	70000
Number of training episodes	140000
Trace length	20
Batch size	32
Target network update frequency	4
Experience replay buffer size	1000
Number of hidden layers in the neural network	2
Number of neuron nodes in the first hidden layer	64
Number of neuron nodes in the second hidden layer	128

6 Numerical Results

In this section, we first provide numerical results to study the NO's revenue, the VNO's revenue, the MUs' utilities, the EMs' utilities, and the social welfare with different values of the VNO's data plan fee, data reward per ad, and data amount of the VNO's data plan for the static Stackelberg game. Then, for the scenario without public information in the dynamic game, we show the convergence process of the proposed DQN-based algorithm and study the impact of system parameters on the proposed algorithm and other benchmark algorithms.

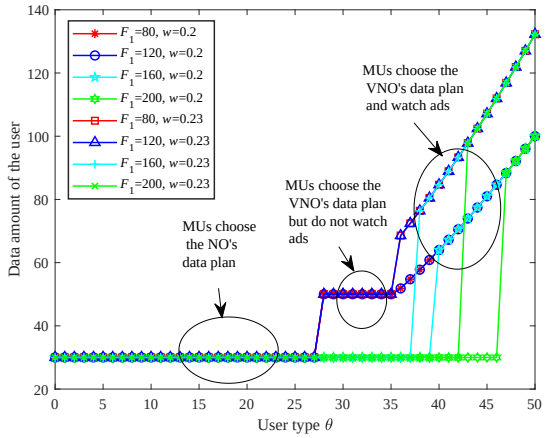


Fig. 2: Data amount of an MU with type θ.

6.1 Simulation Setup

The fees and data amounts of the NO's data plan and VNO's data plan are set based on the typical data plan of China Mobile and Tencent Data Card [34]. The total network capacity of the NO, number of MUs, number of EMs, user's type parameter θ_m, and users' average disutility of watching one ad are set according to [8]. The default values of important variables used in the simulation are summarized in Table 2.

Besides, the neural network setting and the tuned hyper-parameters adopted in the training procedure are listed in Table 3. During training, we gradually decrease the exploration rate to balance the exploration and exploitation, and increase the hysteretic rate to balance the updating between positive and negative samples since the accuracy of the evaluation is more critical after each agent has a fairly good strategy in the late stage of training. In the execution phase, each agent perceives local observation and selects an action with the maximum Q-value according to the individual trained model.

6.2 MUs' Optimal Decision

In Subsections 6.2 - 6.4, we study the revenue of the game players and the social welfare with different system parameters for the static three-stage Stackelberg game. Fig. 2 shows the data amount different MUs obtain under the VNO's different data plan fees F_1's and data rewards per ad w's. In general, there are three types of MUs in the figure–the first type of MUs choosing the NO's data plan, the second type of MUs choosing the VNO's data plan but without watching ads, and the last ones choosing the VNO's data plan and watching ads. If the VNO's data plan fee is relatively low (e.g., $F_1 = 80$ or 120 CNY), the MUs' decisions belong to Case 1 in Proposition 1, i.e., all three types of MUs appear in the network. If the VNO's data plan fee F_1 increases

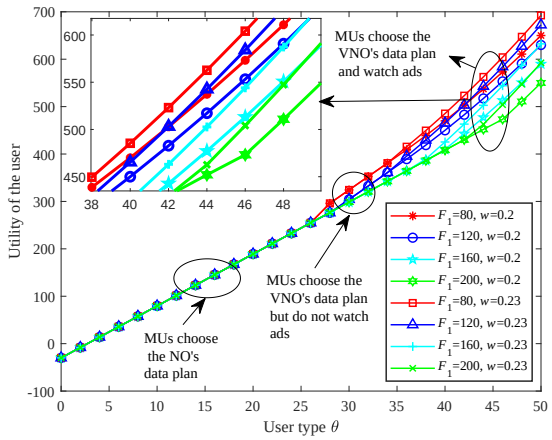


Fig. 3: Utility of an MU with type θ.

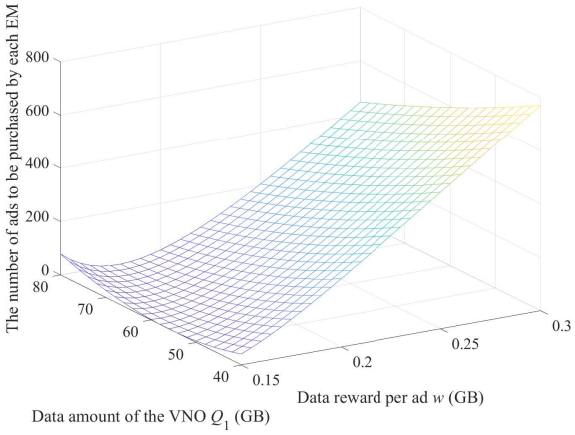


Fig. 4: The number of ad slots to be purchased by each EM.

for example to 160 or 200 CNY, the MUs' decisions fall into Case 2 in Proposition 1, i.e., only the first and third types of MUs appear in the network. Further, the larger the VNO's data plan fee, the more the MUs subscribe to the NO but the more the ads an MU who subscribes to the VNO will watch. This is rational because the increase of the VNO's data plan fee makes MUs who subscribe to it tend to earn data rewards by watching ads to compensate the high data plan fee. It is also noteworthy that, when MUs decide to watch ads, the number of ads they will watch mainly depends on the data reward w per ad and their user type (see Proposition 1), rather than network data plan fee.

Fig. 3 shows the relationship between MUs' utility and their user type θ under the VNO's different data plan fees F_1's and data rewards per ad w's. It can be seen from the figure that, given a fixed data reward, if user type $\theta \leq \theta_2$ ($\theta > \theta_2$), the utility of an MU remains unchanged (reduces) as the VNO's data plan fee increases, for example, $\theta_2 = 27$ when $F_1 = 80$ CNY and $w = 0.2$ GB in Case 1. This is rational because the MUs with $\theta \leq \theta_2$ will subscribe to the NO's data plan and the MUs with $\theta > \theta_2$ will choose the VNO's data plan according to Proposition 1. Furthermore, one can notice that, given a fixed VNO's data plan fee, before user type θ increases to a certain threshold (e.g., $\theta = 35$ for $F_1 = 80$ CNY), the utility of an MU remains unchanged, if the data reward increases from 0.2 GB to 0.23 GB. This is because the MUs with $\theta_2 < \theta \leq \theta_1$ will choose the VNO's data plan but do not watch ads, so the data amount of these MUs is not affected by the data reward per ad. However, after the user type beyond the threshold, the utility of an MU increases as the data reward per ad increases, since for these MUs, they will choose the VNO's data plan and watch ads, thus can obtain more mobile data according to Proposition 1.

6.3 Impact of System Parameters on the Nash Equilibrium

Because we are mainly concerned about the impact of an EP as a VNO in the proposed system, in this subsection we study the trends of the number of ad slots to be purchased by each EM, the optimal data price of the NO, the optimal utility of the VNO, the optimal total utility of the MUs, the optimal utility of the NO, and the social welfare under the VNO's different data amounts of the data plan Q_1's and data rewards per ad w's.

Fig. 4 shows the number of ad slots to be purchased by each EM in the Nash equilibrium. For example, when $Q_1 = 50$ GB and $w = 0.2$ GB, the number of ad slots purchased by each EM is about 3.45×10^8. From Fig. 2, it can be found that about 3×10^7 MUs choose to watch ads at this setting. Therefore, every month each MU who chooses to watch ads will watch 11.5 ads on average. It can be observed that the number of ad slots to be purchased by each EM has different changing trends with the VNO's data reward per ad if the VNO chooses different data amounts of its data plan. Specifically, for a smaller data amount of the VNO's data plan (e.g., 40 GB), the number of ad slots to be purchased by each EM increases when the VNO's data reward per ad increases. This is because the data amount of the VNO's data plan is small. So more MUs are willing to choose the VNO's data plan and these MUs will watch more ads to earn more mobile data. However, for a larger data amount of the VNO's data plan (e.g., 80 GB), the number of ad slots to be purchased by each EM decreases and then increases when the data reward per ad increases. This is because at the beginning, as the data reward per ad increases, the MUs are reluctant to watch too many ads because the rewarded data is enough to use. When the data reward per ad further increases, the MUs are attached to watch more ads for more revenue. It is also noteworthy that the number of ad slots to be purchased by each EM has different changing trends with the data amount of the VNO's data plan given the VNO's different data rewards per ad. Specifically, for a larger data reward per ad (e.g., 0.3 GB), the number of ad slots to be purchased by each EM decreases with the increase of the data amount of the VNO's data plan. It is rational because according to Proposition 1, fewer MUs are willing to watch ads and these MUs will watch fewer ads. On the other hand, for a smaller data reward per ad (e.g., 0.15 GB), the number of ad slots to be purchased by each EM decreases and then increases when the data amount of the VNO's data plan increases. This is because at first, fewer MUs are willing to watch ads and these MUs will watch fewer ads as the rewarded

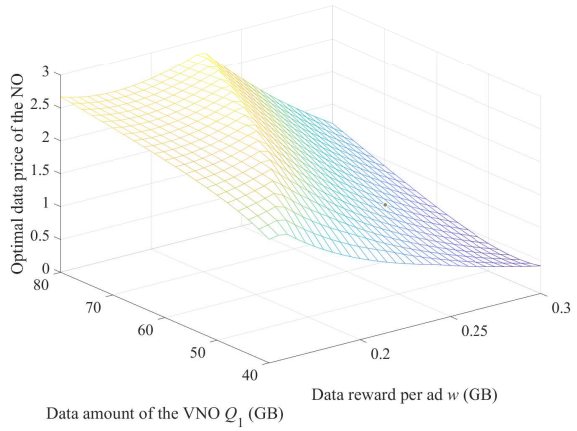


Fig. 5: The optimal data price of the NO under the VNO's different data amounts and data rewards per ad.

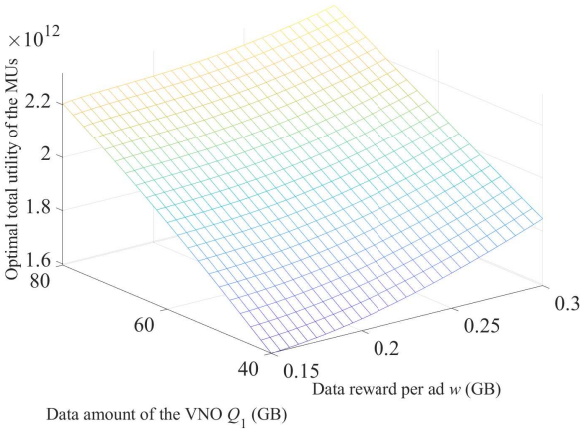


Fig. 7: Optimal total utility of the MUs.

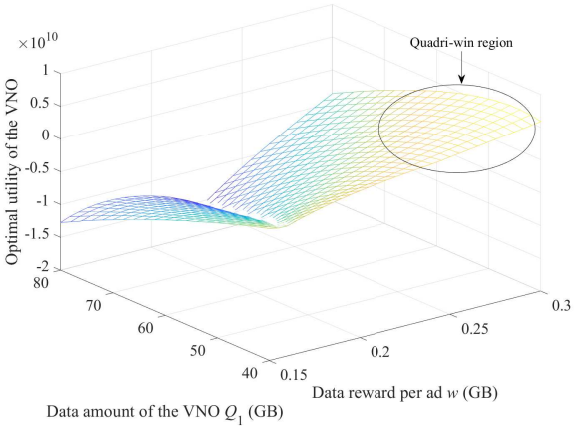


Fig. 6: The optimal utility of the VNO with its different data amounts and data rewards per ad.

data is unable to make up for the disutility of watching ads. However, when the data amount of the VNO's data plan further increases, the subscribers of the VNO have to watch more ads to compensate the high data plan fee.

Fig. 5 shows the NO's optimal data price c^* in the Nash equilibrium. It can be observed that given a fixed data reward per ad the NO's optimal data price increases and then decreases with the increase of the data amount of the VNO's data plan. This is because at the beginning, as the data amount of the VNO's data plan increases, the utility that subscribers can get from the VNO's data plan increases, thus more MUs are willing to choose the VNO's data plan, making the VNO willing to buy more data from the NO at a relatively high price. However, when the data amount of the VNO's data plan increases to a certain threshold (e.g., $Q_1 = 62$ GB for $w = 0.2$ GB), a further increase in the NO's data price makes the VNO keep increasing its data plan fee, which, on the other hand, makes fewer MUs willing to subscribe to the VNO's data plan and in turn makes the total utility of the VNO decrease. So, the NO's data price decreases correspondingly. Similarly, given a fixed data amount of the VNO's data plan, the optimal data price of the NO also increases and then decreases with the increase of the VNO's data reward per ad.

Fig. 6 shows the VNO's optimal utility in the Nash equilibrium. It can be found that given a fixed data reward per ad the VNO's optimal utility decreases with the increase of its data amount of the data plan. This is because the VNO needs to pay more data costs to the NO as its data amount of the data plan increases. On the other hand, given a fixed value of the VNO's data amount of data plan, the utility of the VNO decreases and then increases with the increase of the VNO's data reward per ad. This is because, as seen in Fig. 5, the optimal data price of the NO increases and then decreases when the VNO's data reward per ad increases. So the data cost paid by the VNO to the NO also increases and then decreases. It is also noteworthy that the optimal utility of the VNO may be negative when the data reward per ad is small and the data amount of the data plan is large. Therefore, to avoid the situation that the VNO will not join this business model, the system parameters need to be reasonably tuned to achieve a quadri-win result as shown in Fig. 6. In contrast to the VNO's optimal utility, Fig. 7 shows that the optimal total utility of the MUs in the Nash equilibrium always increases with the data amount of the VNO's data plan or its data reward per ad, simply because such a generous behavior of the VNO makes the MUs enjoy more mobile services, as illustrated in (3).

In Fig. 8, we study the NO's optimal utility in the Nash equilibrium. It can be seen that the NO's optimal utility increases with the increment of the VNO's data amount of data plan given the latter's any fixed data reward per ad. This is because as the VNO's data amount of the data plan increases, the utility that an MU can get from the VNO's data plan increases, and therefore more MUs are willing to choose the VNO's data plan, making the VNO willing to buy more data from the NO at a high price. It is also noteworthy that the optimal utility of the NO has different changing trends with the VNO's data reward per ad if the VNO chooses different data amounts of its data plan. Specifically, for a smaller data amount of the VNO's data plan (e.g., 40 GB), the optimal utility of the NO increases and then decreases when the VNO's data reward per ad increases. This is because at the beginning, as the data reward per ad

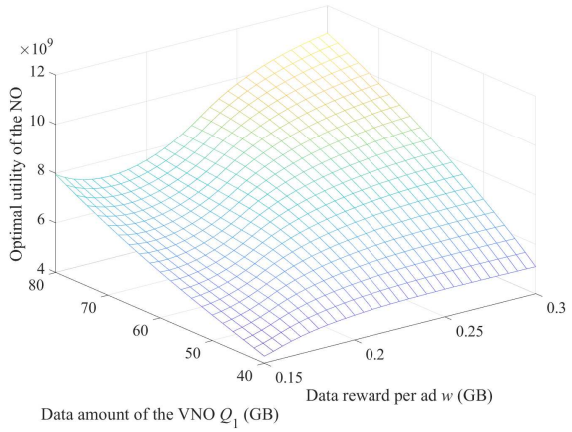


Fig. 8: The optimal utility of the NO under the VNO's different data amounts and data rewards per ad.

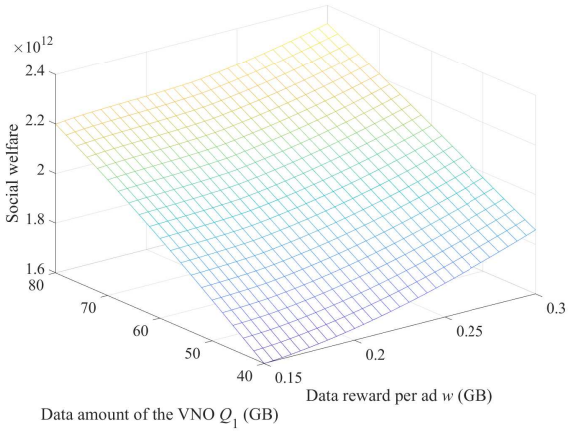


Fig. 9: The social welfare with the VNO's different data amounts and data rewards per ad.

increases, the utility of MUs increases and they are more willing to choose the VNO's data plan, thus increasing the NO's revenue in the data market. However, a larger data reward per ad makes the VNO's data cost increase and profit decrease, which in turn increases its data plan fee and makes the number of its subscribers decrease and eventually the NO's profit in the data market decrease. On the other hand, for a large data amount of the VNO's data plan (e.g., 80 GB), the NO's optimal utility decreases and then increases as the VNO's data reward per ad increases. This is because at first, as the data reward per ad increases, the data price of the NO also increases, making the VNO's utility decrease and thus reduce the number of its subscribers, which eventually reduces the NO's profit. However, as the VNO's data reward per ad further increases, the data price of the NO starts to decrease. Therefore, the VNO's utility starts to increase and gradually starts to adjust the data plan fee to increase the number of subscribers, which makes the utility of the NO's data market increase.

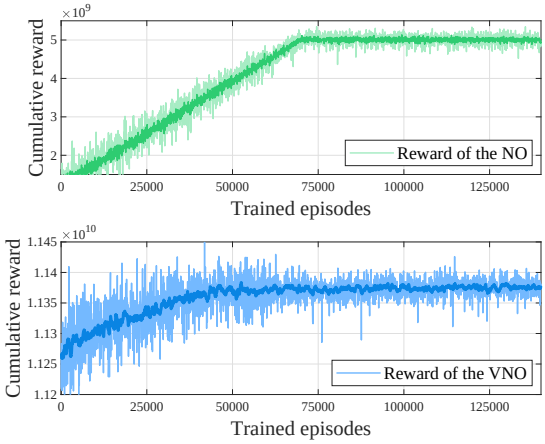


Fig. 10: Cumulative reward of the NO and VNO during training.

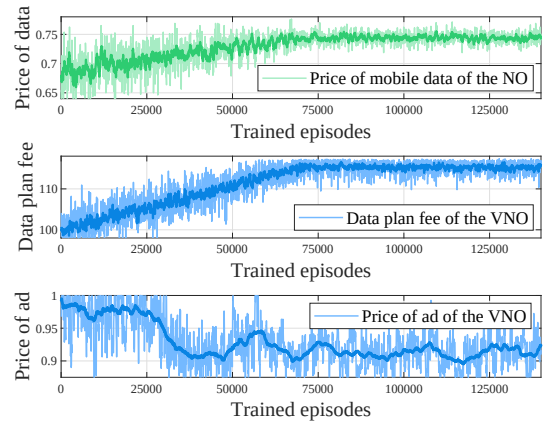


Fig. 11: Decision variables of the NO and VNO during training.

6.4 Impact of System Parameters on Social Welfare

Fig. 9 shows how the social welfare changes in the Nash equilibrium. It can be observed that the social welfare increases with the increase of the VNO's data amount of the data plan for the same data reward per ad, and it aslo increases with the increase of the VNO's data reward per ad for the same data amount of data plan. As can be observed from Figs. 6-8 and Proposition 6, this is because the social welfare is mainly composed of the total utility of MUs, instead of the utilities of other game players. When the VNO's data amount of data plan and data reward per ad are larger, the total utility of MUs becomes larger. However, as can be observed from Fig. 6, for the VNO, making its data amount of data plan and data reward per ad as large as possible is not the optimal strategy for itself. In our system parameter setting, the VNO should increase the data reward per ad and decrease the data amount of its data plan to increase its optimal utility in Nash equilibrium. At the same time, as shown in Fig. 6, the four types of decision-makers in the incentive mechanism can have a quadri-win result only in this region.

6.5 Convergence Process of the Proposed DQN-based Algorithm

In Subsections 6.5 - 6.6, we show the convergence process of the proposed DQN-based algorithm and compare the proposed algorithm with other benchmarks for the scenario lack of public information in the dynamic game. Fig. 10 shows the cumulative rewards of the NO and VNO during the training process. It can be seen that the NO's reward increases smoothly and finally converges around 5.1×10^9 after about 7×10^4 epochs, which is very close to the theoretical value 4.9×10^9 of the SE (see Fig. 8). This indicates that during the training process, the NO makes its average reward slowly increase by adjusting its own mobile data price c, and finally reaches a stable state. It can be observed that the reward of the VNO increases and finally converges to about 1.14×10^{10} after about 7×10^4 epochs, which is very close to the theoretical value of 1.15×10^{10} of the SE (see Fig. 6). Notice that the variance of the VNO's reward curve in the training process is larger than that of the NO. This indicates that during the training process, the VNO needs to adapt to the changes of the NO's mobile data price c on the one hand, and adjust its own data plan fee F_1 and advertising price p on the other hand so that its average reward increases slowly and finally stabilizes. Because there are more action variables that directly affect the VNO's reward, i.e., the dimension of the VNO's action space is larger, the VNO's reward curve fluctuates a little more as compared with the NO.

Fig. 11 shows the price of mobile data sold by the NO, data plan fee, and price of ad sold by the VNO during the training process. It can be seen that as the number of training rounds increases, the price of the NO's mobile data also increases slowly. This is because the revenue that the NO obtains from selling mobile data to the VNO is higher than the data plan fee that they can charge by directly providing data plan to these MUs. So, the reward can be continuously increased until the price of data stabilizes at about 0.74 CNY/GB. It can be seen that the data plan fee of the VNO is also gradually increasing. This is because the price of mobile data sold by the NO to the VNO is increasing. In order to maintain its own income, the VNO must increase the data plan fee accordingly. Of course, the fee cannot be increased indefinitely, because an excessively high data plan fee may cause it to lose all MUs, thereby making its own income zero. So, there is a balance point in its data plan fee. It can be seen that the price of ads of the VNO is gradually decreasing. This is because the data plan fee of the VNO is gradually increasing. In order to obtain more mobile data rewards to compensate for the negative impact of high fees, MUs who choose the VNO choose to watch more ads. According to the formula of Proposition 5 (24b), an increase in the number of ads watched by all MUs will reduce the price of ads.

6.6 Utility Result of the DQN-based Algorithm

Fig. 12 shows the utilities of the NO and VNO with the DQN-based algorithm and other two benchmarks. Specifically, one benchmark is the SE solution under a static game solved with optimization theory in Section IV, and the other benchmark is a general random strategy, i.e., each agent

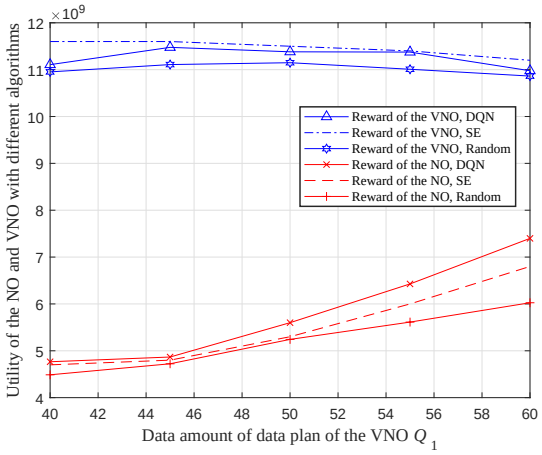


Fig. 12: Reward of the NO and VNO vs. data amount of the VNO's data plan under different algorithms.

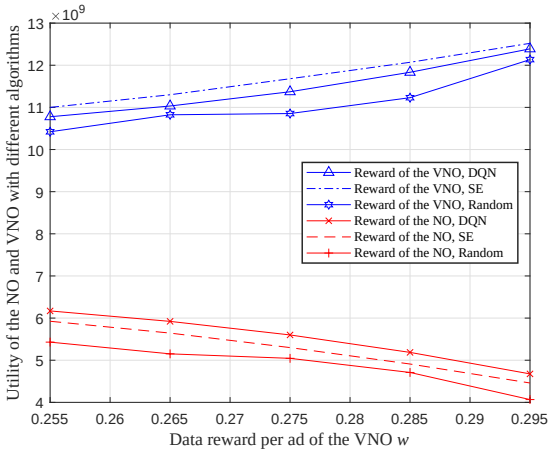


Fig. 13: Reward of the NO and VNO vs. data reward per ad under different algorithms.

selects an action from its action space completely randomly in each epoch. It can be seen that the utility of NO under all three algorithms increases monotonically as the data amount Q_1 of the VNO's data plan increases. This is because the utility that an MU can get from the VNO's data plan increases as the amount of data in the VNO's data plan increases. So, more MUs are willing to choose the VNO's data plan, making the VNO willing to buy more data from the NO at a high price. When the data amount Q_1 of VNO's data plan increases, the revenue of VNO under the DQN-based algorithm decreases except for the point with $Q_1 = 40$ GB. The trend of decreasing monotonicity is because the VNO needs to pay more data cost to the NO as its data amount of the data plan increases. With different data amounts Q_1's of VNO's data plan, the performance of the VNO under the DQN-based algorithm is worse than the SE, while the performance of the NO is the opposite. This is because, with this parameter setting, the NO intentionally improves its mobile data price c in the training phase, further increasing its revenue, while the VNO can only be forced to accept and respond accordingly to obtain a conditionally optimal solution. Finally, the strategies of the NO and VNO reach a

new steady state, in which the NO benefits more from the increment of data price c while the VNO needs to pay more for its data cost. Moreover, the DQN-based algorithm makes both the NO and VNO achieve a higher reward than the random strategy, just because the random strategy can only obtain an expected reward for the whole action space, which is obviously lower than the solutions of other counterparts.

Fig. 13 compares the utilities of the NO and VNO with different algorithms given different data rewards per ad. As can be seen, the utility of the NO consistently decreases monotonically as data reward per ad w increases. This is because a larger data reward per ad make MUs watch more ads to get free data reward, resulting in higher data costs and lower profits for the VNO, which in turn increases their data plan costs, making their subscribers fewer and less willing to buy data, ultimately making revenue of the NO in the data market decreases. In contrast, as data reward per ad w increases, the VNO's revenue increases monotonically. This is because the NO's mobile data price c decreases monotonically, the VNO's mobile data cost continues to decrease and profit continues to increase. It can be seen that given different w's, the utility of the NO with the DQN-based algorithm is always higher than itself with the SE, while the utility of the VNO is the opposite. This is because different from the SE, with the DQN-based algorithm the NO intentionally raises its mobile data price c, which makes its reward further increase, and the VNO can only be forced to accept and respond accordingly, finally leading to a new stable state for the NO and VNO. Moreover, the DQN-based algorithm makes both the NO and VNO gain a higher reward than the random strategy.

7 Conclusion

In this paper, we have proposed a novel incentive mechanism for advertising via mobile data reward and modeled it as a three-stage Stackelberg game, for the scenario with a single NO and a single VNO which also is an EP serving multiple e-commerce merchants. We have obtained the closed-form optimal solution of the Nash equilibrium by backward induction. Further, for the scenario lack of knowledge on the interaction between the NO and VNO in a dynamic game, we have modeled it as a two-stage Stackelberg game in each cycle. Then, we have proposed a DQN-based algorithm to derive the optimal strategies of the NO and VNO in this scenario. The simulation results present the impact of the system parameters on the utilities of game players and social welfare. We also shed some light on the impact of system parameters on the proposed algorithm and other benchmarks and demomnstrate that the proposed DQN-based algorithm can learn a good strategy compared with the SE solution. For the future work, we will consider a more complex business model in the VNO combined with EP scenario, for example, multiple NOs, multiple VNOs, and heterogeneous EMs. In addition, we will consider using the policy-based DDPG algorithm to formulate and solve the problems for continuous actions.

Reference

[1] Q. Cheng, H. Shan, W. Zhuang, T. Q. S. Quek, and Z. Zhang, "When virtual network operator meets e-commerce platform: Advertising via data reward," in *Proc. IEEE/ACM IWQoS*, 2021, pp. 1–11.

[2] S. Singh and S. Jang, "Search, purchase, and satisfaction in a multiple-channel environment: How have mobile devices changed consumer behaviors?" *J. Retailing Consum. Serv.*, Jun. 2020.

[3] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2017-2022," [Online], https://hollandfintech.com/wp-content/uploads/2019/02/white-paper-c11-738429.pdf, Accessed Feb. 8 2021.

[4] F. Sun, F. Hou, H. Zhou, B. Liu, J. Chen, and L. Gui, "Equilibriums in the mobile-virtual-network-operator-oriented data offloading," *IEEE Trans. Veh. Technol.*, vol. 67, no. 2, pp. 1622–1634, Feb. 2018.

[5] Y. Zhang, S. Bi, and Y. A. Zhang, "Joint spectrum reservation and on-demand request for mobile virtual network operators," *IEEE Trans. Commun.*, vol. 66, no. 7, pp. 2966–2977, Jul. 2018.

[6] C. Li, J. Li, Y. Li, and Z. Han, "Pricing game with complete or incomplete information about spectrum inventories for mobile virtual network operators," *IEEE Trans. Veh. Technol.*, vol. 68, no. 11, pp. 11 118–11 131, Nov. 2019.

[7] F. Lobo and T. Schauff, "Sponsored data e mobile marketing," [Online], https://goadmedia.com.br/, Accessed Feb. 8 2021.

[8] H. Yu, E. Wei, and R. A. Berry, "Monetizing mobile data via data rewards," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 4, pp. 782–792, Apr. 2020.

[9] M. Asghari, S. Yousefi, and D. Niyato, "An analysis of service bundles of mobile network operators with free services included," *IEEE Trans. Mobile Comput.*, vol. 19, no. 8, pp. 1789–1803, Aug. 2020.

[10] D. Brake and L. Belli, "Mobile zero rating: The economics and innovation behind free data," *Net Neutrality Reloaded: Zero Rating, Specialised Serv., Ad Blocking and Traffic Manage.*, pp. 132–154, 2016.

[11] R. Sen, S. Ahmad, A. Phokeer, Z. A. Farooq, I. A. Qazi, D. Choffnes, and K. P. Gummadi, "Inside the walled garden: Deconstructing facebook's free basics program," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 47, no. 5, pp. 12–24, Jul 2017.

[12] Y. Zhu, H. Yu, R. A. Berry, and C. Liu, "Cross-network prioritized sharing: An added value mvnos perspective," in *Proc. IEEE INFOCOM*, 2019, pp. 1549–1557.

[13] H. Yu, M. H. Cheung, L. Gao, and J. Huang, "Public Wi-Fi monetization via advertising," *IEEE/ACM Trans. Netw.*, vol. 25, no. 4, pp. 2110–2121, Aug. 2017.

[14] H. Yu, G. Iosifidis, B. Shou, and J. Huang, "Pricing for collaboration between online apps and offline venues," *IEEE Trans. Mobile Comput.*, vol. 19, no. 6, pp. 1420–1433, Jun. 2020.

[15] C. Joe-Wong, S. Sen, and S. Ha, "Sponsoring mobile data: Analyzing the impact on internet stakeholders," *IEEE/ACM Trans. Netw.*, vol. 26, no. 3, pp. 1179–1192, Jun. 2018.

[16] Y. Wu, H. Kim, P. H. Hande, M. Chiang, and D. H. K. Tsang, "Revenue sharing among ISPs in two-sided markets," in *Proc. IEEE INFOCOM*, 2011, pp. 596–600.

[17] Z. Shi, G. Yang, X. Gong, S. He, and J. Chen, "Quality-aware incentive mechanisms under social influences in data crowdsourcing," *IEEE/ACM Trans. Netw.*, pp. 1–14, 2021.

[18] G. Yang, S. He, Z. Shi, and J. Chen, "Promoting cooperation by the social incentive mechanism in mobile crowdsensing," *IEEE Commun. Mag.*, vol. 55, no. 3, pp. 86–92, Mar. 2017.

[19] B. Gu, X. Yang, Z. Lin, W. Hu, M. Alazab, and R. Kharel, "Multi-agent actor-critic network-based incentive mechanism for mobile crowdsensing in industrial systems," *IEEE Trans. Ind. Inf.*, vol. 17, no. 9, pp. 6182–6191, Sep. 2021.

[20] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," *arXiv preprint arXiv:1706.02275*, 2017.

[21] L. Xiao, Y. Li, G. Han, H. Dai, and H. V. Poor, "A secure mobile crowdsensing game with deep reinforcement learning," *IEEE Trans. Inf. Forensics Secur.*, vol. 13, no. 1, pp. 35–47, Aug. 2018.

[22] Q. Xu, Z. Su, and R. Lu, "Game theory and reinforcement learning based secure edge caching in mobile social networks," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 3415–3429, Mar. 2020.

[23] L. Duan, J. Huang, and B. Shou, "Pricing for local and global Wi-Fi markets," *IEEE Trans. Mobile Comput.*, vol. 14, no. 5, pp. 1056–1070, May. 2015.

[24] D. Bergemann and A. Bonatti, "Targeting in advertising markets: implications for offline versus online media," *The RAND J. Econ.*, vol. 42, no. 3, pp. 417–443, Jun. 2011.

[25] C. Zhou, M. L. Honig, and S. Jordan, "Utility-based power control for a two-cell cdma data network," *IEEE Trans. Wireless Commun.*, vol. 4, no. 6, pp. 2764–2776, Nov. 2005.
[26] H. Guo, X. Zhao, L. Hao, and D. Liu, "Economic analysis of reward advertising," *Production and Operations Management*, vol. 28, no. 10, pp. 2413–2430, Feb. 2019.
[27] S. P. Anderson and B. Jullien, "The advertising-financed business model in two-sided media markets," *Handbook of Media Econ.*, 2015.
[28] B. N. Anand and R. Shachar, "Advertising, the matchmaker," *The RAND J. Econ.*, vol. 42, no. 2, pp. 205–245, May. 2011.
[29] Z. Han, D. Niyato, W. Saad, T. Basar, and A. Hjrungnes, *Game Theory in Wireless and Communication Networks: Theory, Models, and Applications*. Cambridge Univ. Press, 2012.
[30] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
[31] K. Yang, H. Shan, T. Sun, R. Hu, Y. Wu, L. Yu, Z. Zhang, and T. Q. S. Quek, "Reinforcement learning-based mobile edge computing and transmission scheduling for video surveillance," *IEEE Trans. Emerging Top. Comput.*, vol. 10, no. 2, pp. 1142–1156, April-Jane 2022.
[32] C. Zhu, Y. H. Chiang, Y. Xiao, and Y. Ji, "Flexsensing: A QoI and latency-aware task allocation scheme for vehicle-based visual crowdsourcing via deep Q-network," *IEEE Internet Things J.*, vol. 8, no. 9, pp. 7625–7637, Mar. 2021.
[33] K. He and J. Sun, "Convolutional neural networks at constrained time cost," in *Proc. IEEE CVPR*, 2015, pp. 5353–5360.
[34] Tecent, "Tecent Data card," https://wk.10010.com/index.html#/details/4G, Accessed May 8 2021.

Qi Cheng received the B.S. degree in information engineering from Zhejiang University, Hangzhou, China, in 2019, and is currently working toward the M.S. degree in electronics and communication engineering at Zhejiang University, Zhejiang, China.

His research interests include multimedia communication, network economy, and reinforcement learning.

Hangguan Shan (Member, IEEE) received the B.Sc. degree in electrical engineering from Zhejiang University, Hangzhou, China, in 2004, and the Ph.D. degree in electrical engineering from Fudan University, Shanghai, China, in 2009.

From 2009 to 2010, he was a Postdoctoral Research Fellow with the University of Waterloo, Waterloo, ON, Canada. Since 2011, he has been with the College of Information Science Electronic Engineering, Zhejiang University, where he is currently an Associate Professor. He is also with the Zhejiang Provincial Key Laboratory of Information Processing Communication Networks, Hangzhou and SUTD-ZJU IDEA, Hangzhou. His current research interests include cross-layer protocol design, resource allocation, and the quality-of-service provisioning in wireless networks.

Dr. Shan has served as a Technical Program Committee Member of various international conferences, including the IEEE Global Communications Conference, the IEEE International Conference on Communications, the IEEE Wireless Communications and Networking Conference, and the IEEE Vehicular Technology Conference (VTC). He has coreceived the Best Industry Paper Award from the 2011 IEEE WCNC held in Quintana Roo, Mexico. He was an Editor of the IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING.

Weihua Zhuang (Fellow, IEEE) is a University Professor and a Tier I Canada Research Chair in Wireless Communication Networks at University of Waterloo, Canada. Her research focuses on network architecture, algorithms and protocols, and service provisioning in future communication systems. She is the recipient of 2021 Women's Distinguished Career Award from IEEE Vehicular Technology Society, 2021 Technical Contribution Award in Cognitive Networks from IEEE Communications Society, 2021 R.A. Fessenden Award from IEEE Canada, and 2021 Award of Merit from the Federation of Chinese Canadian Professionals in Ontario. She was the Editor-in-Chief of the IEEE Transactions on Vehicular Technology from 2007 to 2013, General Co-Chair of 2021 IEEE/CIC International Conference on Communications in China (ICCC), Technical Program Chair/Co-Chair of 2017/2016 IEEE VTC Fall, Technical Program Symposia Chair of 2011 IEEE Globecom, and an IEEE Communications Society Distinguished Lecturer from 2008 to 2011. She is an elected member of the Board of Governors and the Executive Vice President of the IEEE Vehicular Technology Society. Dr. Zhuang is a Fellow of the IEEE, Royal Society of Canada, Canadian Academy of Engineering, and Engineering Institute of Canada.

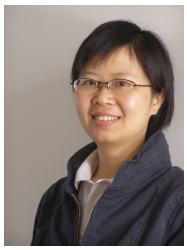
Tony Q.S. Quek (Fellow, IEEE) received the B.E. and M.E. degrees in electrical and electronics engineering from the Tokyo Institute of Technology in 1998 and 2000, respectively, and the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology in 2008.

He is currently the Cheng Tsang Man Chair Professor with the Singapore University of Technology and Design (SUTD). He is also the Head of the ISTD Pillar, the Sector Lead of the SUTD AI Program, and the Deputy Director of the SUTD-ZJU IDEA. His current research interests include wireless communications and networking, network intelligence, the Internet of Things, URLLC, and big data processing.

Dr. Quek serves as an Elected Member of the IEEE Signal Processing Society SPCOM Technical Committee. He was an Executive Editorial Committee Member of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. He was honored with the 2008 Philip Yeo Prize for Outstanding Achievement in Research, the 2012 IEEE William R. Bennett Prize, the 2015 SUTD Outstanding Education Awards Excellence in Research, the 2016 IEEE Signal Processing Society Young Author Best Paper Award, the 2017 CTTC Early Achievement Award, the 2017 IEEE ComSoc AP Outstanding Paper Award, the 2020 IEEE Communications Society Young Author Best Paper Award, the 2020 IEEE Stephen O. Rice Prize, and the 2016-2019 Clarivate Analytics Highly Cited Researcher. He has been actively involved in organizing and chairing sessions and has served as a member of the Technical Program Committee and symposium chairs in a number of international conferences. He is serving as an Editor for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and the Chair of the IEEE VTS Technical Committee on Deep Learning for Wireless Communications. He was an Editor of IEEE TRANSACTIONS ON COMMUNICATIONS and IEEE WIRELESS COMMUNICATIONS LETTERS. He is a Distinguished Lecturer of the IEEE Communications Society.

Zhaoyang Zhang (Member, IEEE) received the Ph.D. degree from Zhejiang University, Hangzhou, China, in 1998.

He is currently a Qiushi Distinguished Professor with Zhejiang University. He has coauthored more than 300 peer-reviewed international journal articles and conference papers. His current research interests include fundamental aspects of wireless communications and networking, such as information theory and coding, network signal processing and distributed learning, AI-empowered communications and networking, network intelligence with synergetic sensing, and computation and communication. He was a co-recipient of seven conference best paper awards including ICC 2019. He was awarded the National Natural Science Fund for Distinguished Young Scholars by NSFC in 2017. He is serving or has served as Editor for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON COMMUNICATIONS, and IET Communications. He served as a General Chair, a TPC Co-Chair or a Symposium Co-Chair of the WCSP 2013/2018, the Globecom 2014 Wireless Communications Symposium, and the VTC-Spring 2017 Workshop HMWC, etc.



Fen Hou (Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, Canada, in 2008.

She is currently an Associate Professor with the State Key Laboratory of IoT for Smart City, the Department of Electrical and Computer Engineering, and GuangdongHong Kong-Macau Joint Laboratory for Smart Cities, University of Macau. Her research interests include resource allocation intelligent computing networks, mechanism design and optimal user behavior in crowd sensing networks, etc.

She was a recipient of the IEEE Globecom Best Paper Award in 2010 and the Distinguished Service Award in the IEEE MMTC in 2011. She served as the TPC chair and co-chair for several IEEE conferences such as ICCS 2014, INFOCOM 2014, ICCC 2015, ICC 2016, ICCC 2021, etc. She currently serves as an Associate Editor of IET Communications.