Operational Research (2022) 22:2605–2630
https://doi.org/10.1007/s12351-021-00622-9

ORIGINAL PAPER

Optimal resource allocation for multiclass services in peer-to-peer networks via successive approximation

Shiyong Li[1] · **Wei Sun**[1] · **Huan Liu**[1]

Received: 13 November 2019 / Revised: 30 November 2020 / Accepted: 9 January 2021 /
Published online: 25 January 2021

Abstract

Peer-to-peer (P2P) networks support a wide variety of network services including elastic services such as file-sharing and downloading and inelastic services such as real-time multiparty conferencing. Each peer who acquires a service will receive a certain level of satisfaction if the service is provided with a certain amount of resource. The utility function is used to describe the satisfaction of a peer when acquiring a service. In this paper we consider optimal resource allocation for elastic and inelastic services and formulate a utility maximization model which is an intractable and difficult non-convex optimization problem. In order to resolve it, we apply the successive approximation method and approximate the non-convex problem to a serial of equivalent convex optimization problems. Then we develop a gradient-based resource allocation scheme to achieve the optimal solutions of the approximations. After a serial of approximations, the proposed scheme can finally converge to an optimal solution of the primal utility maximization model for resource allocation which satisfies the Karush–Kuhn–Tucker conditions.

Keywords Nonlinear programming · P2P networks · Resource allocation · Elastic and inelastic services · Successive approximation

Mathematics Subject Classification 68M10 · 68M20 · 90C30

✉ Wei Sun
wsun@ysu.edu.cn

Shiyong Li
shiyongli@ysu.edu.cn

Huan Liu
liuhuan5011@foxmail.com

[1] School of Economics and Management, Yanshan University, Qinhuangdao 066004, China

1 Introduction

Nowadays, peer-to-peer (P2P) networks have become an important network architecture supporting file sharing and video distribution over the Internet. P2Ps are different from traditional client/server data networks, where the performance mainly depends on a small number of powerful servers. In P2Ps the role of each peer is treated as the same. Each peer can not only serve as a server to provide network services for other peers, but also serve as a client to obtain services from others, thus avoiding the impact of server failure in the traditional networks. They can generate multiple routes between service providers and customers, thus improving the network throughput and routing optimization Wang et al. (2019). In recent years, many research scholars have carried out studies on P2Ps and applied them into various scenarios, e.g., distributed storage Yan et al. (2017), edge-cloud computing Song et al. (2020), multipath industrial networks Song et al. (2020), and vehicular networks Wang et al. (2018).

P2Ps can support a wide variety of network services such as file-sharing and video conferencing, e.g., Tencent meeting, Zoom, and are known to cause much network traffic over the Internet through various P2P protocols, e.g., BitTorrent, EDonkey, VoIP. In fact, each peer has a certain level of satisfaction when it requests a network service, which can be described as a form of utility function. Based on the shapes of utility functions, network services can be classified into two categories (Lee et al. 2005; Hande et al. 2007; Li et al. 2015, 2019). One type is the traditional data services, such as file download and upload, which are not very sensitive to the bandwidth requirements. These services are elastic in their bandwidth requirement intervals. Usually a concave utility function is used to describe peer's satisfaction for acquiring an elastic service. The other type is related to delay or bandwidth sensitive multimedia services, such as real-time streaming video service. These services are known as to be inelastic in their bandwidth requirement intervals. They usually have high requirements for bandwidth resource so as to guarantee certain level of QoS. The utility function for an inelastic service is often nonconcave, e.g., sigmoidal or general.

In recent years, many scholars have investigated the resource allocation for both elastic and inelastic services in P2Ps and presented some interesting resource allocation schemes. Firstly, resource allocation for elastic services mainly concentrates on resource pricing strategies, e.g., Eger and Killat (2007a, b), Kumar et al. (2011), Koutsopoulos and Iosifidis (2010), Li and Sun (2016) and Li et al. (2019). In this type of research, a resource pricing scheme is proposed to balance the resource requests and provisions, thus service providers adjust their bandwidth allocation according to the difference between the prices provided by service requesters and the prices charged by service providers, and finally achieve a weighted fairness among service requesters. Since the utility functions of elastic services are usually concave, then the resource allocation problem for elastic services is a convex optimization problem, which can be solved through the first order Lagrangian method. Besides resource pricing mechanisms, reputation-based methods are also used to encourage cooperation amongst selfish peers so as

to promote each peer to provide resource for others and achieve efficient resource allocation, e.g., Satsiou and Tassiulas (2010), Gupta et al. (2016) and Goswami et al. (2017).

P2Ps also provide a variety of inelastic services such as VoIP and real-time conferencing (e.g., Tencent meeting), however, the resource allocation problem of inelastic services is much more difficult than that of elastic services because the utility functions of inelastic services are often non-concave. Thus, how to achieve effective resource allocation for inelastic services becomes very important, which is also a crucial challenge and difficult problem. For example, P2P multiparty conferencing applications are considered in Chen et al. (2012), where a problem of utility maximization for resource allocation is formulated. In order to solve the nonstrictly concave optimization problem, a primal-dual distributed algorithm is presented and proven to converge to the global optimum under the proposed sufficient conditions. Resource allocation for inelastic services are also investigated in Li et al. (2017), and a heuristic algorithm using particle swarm optimization (PSO) is proposed to solve the difficult nonconvex optimization problem. As for the scenario with both elastic and inelastic services, the utility maximization model for resource allocation is also an intricate and difficult problem.

In this paper we consider resource allocation problem for P2Ps where both elastic and inelastic services are coexisting. We formulate the utility maximization (social welfare) model for resource allocation, i.e., the total satisfaction of all peers in the networks when they acquire these services. However, the utility maximization model is a non-convex problem which is hard to resolve. By applying the successive approximation method, we transform the non-convex problem into an equivalent convex optimization problem and develop a gradient-based resource allocation scheme to achieve the optimal solution of the approximations. After a serial of approximations, the proposed scheme can finally converge to an optimal solution of the primal utility maximization model for resource allocation which also satisfies the Karush–Kuhn–Tucker (KKT) conditions.

The rest of this paper is summarized as follows: We review research work on resource allocation for elastic and inelastic service in P2P networks in Sect. 1. Then we introduce the utility maximization model for resource allocation of both elastic and inelastic services and formulate a serial of approximations in Sect. 2. In Sect. 3 we develop a gradient-based resource allocation scheme to converge to the optimum of the primal resource allocation problem. Then we give some numerical examples to illustrate the performance of the proposed scheme in Sect. 4. Finally we conclude this paper in Sect. 5.

2 Related work

Realizing effective resource allocation in P2P networks has become an important research field and received extensive attention in recent years. P2P networks can provide elastic services such as file sharing and downloading. For the resource allocation of this type of services, in order to reduce the "free-riding" due to the peers' selfish nature, many scholars proposed incentive mechanisms to encourage

each peer to provide its own upload bandwidth for others, such as resource pricing, reputation-based schemes, and so on. Eger and Killat (2007a) presented a resource pricing scheme to achieve a fair bandwidth allocation among service requesters. They further expanded the scheme to achieve a weighted fairness among service requesters, such that service providers could adjust their service rates and offered prices (Eger and Killat 2007b). Later, Koutsopoulos and Iosifidis (2010) considered bandwidth allocation in a star topology P2P network where the peer access link to the backbone is the capacity bottleneck. They developed the problem of maximizing total network utility through controlling the bandwidth allocation for download and upload of each peer, respectively. Then Kumar et al. (2011) considered computing resources shared by users in P2Ps and proposed a resource pricing and allocation scheme. Pacifici et al. (2016) considered cache bandwidth allocation for P2P file-sharing networks which is formulated as a Markov decision process, and proposed three approximations to the optimal cache bandwidth allocation policy, so as to minimize inter-ISP traffic. Recently, Li and Sun (2016) and Li et al. (2019) applied the first order Lagrangian method and low-pass filtering scheme to design a novel scheme of resource pricing and allocation. Antal and Vinkó (2016) investigated max–min fair bandwidth allocation in BitTorrent communities and proposed an algorithm to realize max-min fairness bandwidth allocation in multi-swarm P2P content sharing community. Li et al. (2020) considered fair bandwidth allocation of access links in P2P file-sharing networks and developed a coupled network-wide utility maximization model which aims at achieving several kinds of fairness among requesting peers.

Besides resource pricing schemes, reputation-based mechanisms are also used to promote cooperation between selfish peers and achieve reasonable resource allocation for peers. For example, Satsiou and Tassiulas (2010) assumed each peer could earn its reputation analogous to its contributions and presented a reputation-based resource allocation scheme. Later, Gupta et al. (2016) described a reputation-based probabilistic resource allocation mechanism for avoiding free riding in unstructured P2P networks. Kang and Wu (2015) proposed a credit-based incentive mechanism to encourage peers to cooperate with each other in a heterogeneous network consisting of wired and wireless peers. The mechanism can provide differentiated service to peers with different credits through biased resource allocation. Goswami et al. (2017) investigated a reputation-based resource allocation mechanism through two non-cooperative games in P2P networks and proved evolutionary stability of reputation-based resource allocation (Goswami et al. 2018).

P2P networks can also support inelastic services such as live streaming services. Therefore, to achieve reasonable resource allocation for inelastic services is also very important. Liang et al. (2011) investigated optimal bandwidth sharing in multiswarm multiparty P2P video-conferencing systems. Chen et al. (2012) considered resource allocation for P2P multiparty conferencing applications where quality of service (QoS) guarantee is a crucial challenge, and formulated the resource allocation problem as utility maximization. Liu et al. (2015) considered resource allocation in underprovisioned multioverlay peer-to-peer live video sharing services. They designed various objective functions for the upload bandwidth allocation problem and showed how optimal solutions could be computed using a bipartite flow

network. Rohmer et al. (2015) investigated the problem of maximizing the P2P streaming system capacity by effectively alternating between different resource allocation strategies, and combined different, even potentially conflicting, performance objectives when deciding which resource allocation strategy to use for the current time period. Mostafavi and Dehghan (2016) considered bandwidth sharing and allocation in helper-assisted P2P streaming using non-cooperative game theory and double auction, and proposed an algorithm to reach significant performance improvements in terms of utility per selling helpers. They Mostafavi and Dehghan (2017) also considered resource allocation for HD live streaming and proposed a decentralized, stochastic approximation helper selection mechanism which is adaptable to supply and demand pattern of various video channels. Recently, Li et al. (2017) also developed a utility maximization model for resource allocation of P2P inelastic services. In order to resolve the intrinsically difficult nonconvex optimization, they presented a heuristic algorithm using PSO to achieve the optimal resource allocation.

In this paper we consider resource allocation for multiclass services and introduce a utility maximization model for resource allocation, which is also a non-convex optimization problem. We apply successive approximation method to resolve the non-convex optimization, and approximate the primal problem into an equivalent convex optimization problem. The approximate approach is very useful in dealing with intricate difficult optimization problems (Rismanchian and Lee 2018). The successive approximation method is first introduced in Marks and Wright (1978). It has been applied into power control problems (Tran and Hong 2010; Vo et al. 2011), wireless random access problem (Vo et al. 2012), bandwidth allocation problems in single-path networks (Vo et al. 2013) and multipath networks (Vo et al. 2014), and cloud resource allocation problem for enterprise applications migration (Li and Sun 2020). This method is generally composed of two parts: inner-iterations and outer-iterations, and requires inner-iterations to converge to the optimum of the approximation problem. We prove the convergence of the resource allocation scheme derived from successive approximation method within a certain number of inner-iterations and illustrate the performance with some numerical examples.

3 Resource allocation model

3.1 Services and utility functions

P2P networks support a wide variety of network services. Each peer who acquires a service will receive a certain level of satisfaction if the service is provided with a certain amount of resource. The utility function derived from economics is found useful to describe the satisfaction of a peer when acquiring a service. According to the different shapes of utility functions, they can be divided into two categories: elastic services and inelastic services (Lee et al. 2005; Hande et al. 2007; Li et al. 2015). The former one mainly refers to traditional data services such as file uploading and downloading services in P2P file-sharing networks. The utility functions for this type of services are generally concave. The latter one is mainly related to multimedia video and audio services such as P2P multiparty conferencing services and

VoIP over P2P networks. This type of services are usually very sensitive to granted resources and the QoS will drop drastically if resources are below a certain threshold. They usually have non-concave utility functions, such as sigmoidal functions. The shapes of utility functions of elastic and inelastic services are generally illustrated in Fig. 1. We adopt the utility functions proposed for resource allocation of services in IP networks (Lee et al. 2005; Hande et al. 2007; Li et al. 2015). If a peer acquires an elastic service s, it will have a concave utility as follows

$$U^s(y) = w \log(y + 1) \tag{1}$$

and if it acquires an inelastic service r, then it will have a sigmoidal utility as follows

$$U^r(y) = \frac{w}{1 + e^{-a(y-b)}} - \frac{w}{1 + e^{ab}} \tag{2}$$

where y is the service rate, w is the willing-to-pay of the peer, a and b are parameters of the inelastic service. We find that for inelastic service r, there is a demarcation point y_0 that divides the utility function into a convex part and a concave part, that is, $d^2U^r(y)/dy^2 > 0$ if $y < y_0$, and $d^2U^r(y)/dy^2 < 0$ if $y > y_0$.

3.2 Model description

Consider a P2P network which is composed of a set of peers, a set S of elastic services and a set R of inelastic services. Each peer in the network acquires at least one service, elastic or inelastic. For example, a peer who is downloading files from other peers is also interested in taking part in a P2P multiparty conferencing. On the other hand each peer can also provide one or several services for others. Therefore, each peer acts as both a service customer and a service provider. In P2P networks, each peer uses its access link not only to obtain services from other peers, such as downloading files, but also to provide services for other peers, such as uploading files. Therefore, the upload bandwidth of a peer becomes a rare resource in the network, and other peers will compete for the upload bandwidth so as to obtain services. Therefore, the P2P networks are

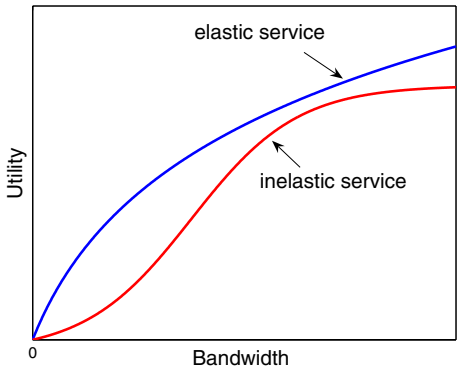


Fig. 1 Utility functions for elastic and inelastic services

faced with the problem of how to allocate the peers' upload bandwidth reasonably and effectively among service requesters, which is the main aim of this work.

Let the set P be peers acting as service providers that offer upload bandwidth to requesters. Also define the sets C^s and C^r of peers acting as service customers that request elastic services and inelastic services, respectively. Introduce x^s_{pc} and x^r_{pc} as the service rates offered by service provider p for customer c who requests elastic service s and inelastic service r, respectively. Then, each peer $c \in C^s$ receives a total bandwidth y^s_c offered by its providers $P^s(c)$ when it requests elastic service s, and peer $c \in C^r$ obtains a total bandwidth y^r_c offered by its providers $P^r(c)$ when it requests inelastic service r. Finally, the total bandwidth allocation of service provider p does not exceed its access link capacity C_p.

Then we formulate the resource allocation for multiclass services in P2P networks as the following optimization problem

$$\begin{aligned} \mathbf{P1}: \ \max \quad & \sum_{c:c\in C^s} U^s_c(y^s_c) + \sum_{c:c\in C^r} U^r_c(y^r_c) \\ \text{subject to} \quad & \sum_{p:p\in P^s(c)} x^s_{pc} = y^s_c, \ \sum_{p:p\in P^r(c)} x^r_{pc} = y^r_c, \\ & \sum_{c:c\in C^s(p)} x^s_{pc} + \sum_{c:c\in C^r(p)} x^r_{pc} \le C_p, \\ \text{over} \quad & x^{s\min}_{pc} \le x^s_{pc} \le x^{s\max}_{pc}, \\ & x^{r\min}_{pc} \le x^r_{pc} \le x^{r\max}_{pc}. \end{aligned} \tag{3}$$

Here, in the resource allocation problem **P1**, the objective is to maximize the aggregated utility of obtained bandwidth y^s_c and y^r_c over all service customers under the constraints that each service provider offers no more than its own access link capacity. As described by the equality of the resource allocation model, the aggregated bandwidth provision y^s_c (y^r_c) for elastic service s (inelastic service r) of service customer c is the sum of the rates x^s_{pc} (x^r_{pc}) that its service providers grant. Here, $x^{s\min}_{pc} \ge 0$ ($x^{r\min}_{pc} \ge 0$) is the minimal resource requirement of customer c from provider p for elastic service s (inelastic service r). Similarly, $x^{s\max}_{pc}$ ($x^{r\max}_{pc}$) is the maximal resource requirement of customer c from provider p for elastic service s (inelastic service r). Also, as described by the inequality in the optimization problem above, the bandwidth provision of provider p is constrained by its own upload capacity of access link, i.e. C_p.

3.3 Model analysis

In this part we give an analysis on the resource allocation problem **P1** for multiclass services in P2P networks. Firstly we obtain the Lagrangian

$$
\begin{aligned}
L_{P1}(\mathbf{x},\mathbf{y};\boldsymbol{\phi},\boldsymbol{\varphi}) = & \sum_{c:c\in C^s} U_c^s(y_c^s) + \sum_{c:c\in C^r} U_c^r(y_c^r) \\
& + \sum_{c:c\in C^s} \phi_c^s \left(\sum_{p:p\in P^s(c)} x_{pc}^s - y_c^s \right) + \sum_{c:c\in C^r} \phi_c^r \left(\sum_{p:p\in P^r(c)} x_{pc}^r - y_c^r \right) \\
& + \sum_{p:p\in P} \varphi_p \left(C_p - \sum_{c:c\in C^s(p)} x_{pc}^s - \sum_{c:c\in C^r(p)} x_{pc}^r \right),
\end{aligned} \tag{4}
$$

where $\boldsymbol{\phi}$ is the price vector with elements ϕ_c^s and ϕ_c^r, which can be interpreted as the prices per unit bandwidth paid by customer c when acquiring elastic service s and inelastic service r, respectively; $\boldsymbol{\varphi}$ is the price vector with element φ_p, which can be thought as the price per unit bandwidth charged by provider p when offering bandwidth allocation for customers; $\mathbf{x}$ is the service rate matrix with elements x_{pc}^s and x_{pc}^r for elastic service s and inelastic service r of customer c, respectively; $\mathbf{y}$ is the rate vector with elements y_c^s for elastic service s and y_c^r for inelastic service r of customer c.

We rewrite the Lagrangian (4) as following

$$
\begin{aligned}
L_{P1}(\mathbf{x},\mathbf{y};\boldsymbol{\phi},\boldsymbol{\varphi}) = & \sum_{c:c\in C^s} \left(U_c^s(y_c^s) - \phi_c^s y_c^s \right) + \sum_{c:c\in C^r} \left(U_c^r(y_c^r) - \phi_c^r y_c^r \right) \\
& + \sum_{c:c\in C^s} \sum_{p:p\in P^s(c)} x_{pc}^s \left(\phi_c^s - \varphi_p \right) + \sum_{c:c\in C^r} \sum_{p:p\in P^r(c)} x_{pc}^r \left(\phi_c^r - \varphi_p \right) + \sum_{p:p\in P} \varphi_p C_p.
\end{aligned} \tag{5}
$$

We find the first part in (5) is separable in variables y_c^s and y_c^r, and the second part is separable in variables x_{pc}^s and x_{pc}^r. Thus the objective function of the dual problem is described as

$$
\begin{aligned}
D(\boldsymbol{\phi},\boldsymbol{\varphi}) = & \max_{\mathbf{x},\mathbf{y}} L_{P1}(\mathbf{x},\mathbf{y};\boldsymbol{\phi},\boldsymbol{\varphi}) \\
= & \sum_{c:c\in C^s} \mathcal{P}^s(\phi_c^s) + \sum_{c:c\in C^r} \mathcal{P}^r(\phi_c^r) \\
& + \sum_{c:c\in C^s} \sum_{p:p\in P^s(c)} \mathcal{R}_{pc}^s(\phi_c^s,\varphi_p) + \sum_{c:c\in C^r} \sum_{p:p\in P^r(c)} \mathcal{R}_{pc}^r(\phi_c^r,\varphi_p) + \sum_{p:p\in P} \varphi_p C_p,
\end{aligned} \tag{6}
$$

where

$$
\mathcal{P}^s(\phi_c^s) = \max_{y_c^s} U_c^s(y_c^s) - \phi_c^s y_c^s, \tag{7}
$$

$$
\mathcal{P}^r(\phi_c^r) = \max_{y_c^r} U_c^r(y_c^r) - \phi_c^r y_c^r, \tag{8}
$$

$$
\mathcal{R}_{pc}^s(\phi_c^s,\varphi_p) = \max_{x_{pc}^{s\min} \le x_{pc}^s \le x_{pc}^{s\max}} x_{pc}^s(\phi_c^s - \varphi_p), \tag{9}
$$

$$
\mathcal{R}_{pc}^r(\phi_c^r,\varphi_p) = \max_{x_{pc}^{r\min} \le x_{pc}^r \le x_{pc}^{r\max}} x_{pc}^r(\phi_c^r - \varphi_p). \tag{10}
$$

We give an economic interpretation for Eqs. (7)–(10) as follows. In (7), service customer c acquires elastic service s with total bandwidth provision y_c^s and wants to maximize its own utility $U_c^s(y_c^s)$. Meanwhile, it needs to pay a fee for its obtained bandwidth to support the elastic service. Recall that ϕ_c^s is the price per unit bandwidth paid by customer c for service s, then $\phi_c^s y_c^s$ is regarded as the total fee paid by customer c. Therefore, the economic meaning of (7) is that each peer as *service customer* aims at achieving the objective of maximizing its own *profit*. The interpretation for (8) is similar to that of (7). As for (9), φ_p is considered as the price per unit bandwidth charged by service provider p. The product $x_{pc}^s \phi_c^s$ is the fee paid by customer c to provider p for elastic service s, and $x_{pc}^s \varphi_p$ is the expense demanded by provider p for its granting bandwidth x_{pc}^s. Then, the economic meaning of (9) is that each peer as *service provider* wants to achieve the objective of maximizing its own *revenue*.

At the optimum of sub-problems (7)–(10), the KKT conditions are satisfied, which are listed as following

$$\nabla_x L_{P1}(\mathbf{x}^*, \mathbf{y}^*; \boldsymbol{\phi}^*, \boldsymbol{\varphi}^*) = 0, \quad \text{and} \quad \nabla_y L_{P1}(\mathbf{x}^*, \mathbf{y}^*; \boldsymbol{\phi}^*, \boldsymbol{\varphi}^*) = 0 \tag{11}$$

$$\phi_c^{s*}\left(\sum_{p:p\in P^s(c)} x_{pc}^{s*} - y_c^{s*}\right) = 0, \quad \text{and} \quad \phi_c^{r*}\left(\sum_{p:p\in P^r(c)} x_{pc}^{r*} - y_c^{r*}\right) = 0 \tag{12}$$

$$\varphi_p^*\left(C_p - \sum_{c:c\in C^s(p)} x_{pc}^{s*} - \sum_{c:c\in C^r(p)} x_{pc}^{r*}\right) = 0 \tag{13}$$

Then, we can obtain the **dual problem** of resource allocation model **P1**

$$\begin{aligned} \min \quad & D(\boldsymbol{\phi}, \boldsymbol{\varphi}) \\ \text{over} \quad & \phi_c^s \geq 0, \phi_c^r \geq 0, \varphi_p \geq 0. \end{aligned} \tag{14}$$

The objective of the dual problem (14) is to minimize the total price charged by all service providers under the constraints that service customers are guaranteed with certain levels of satisfaction. In order to obtain the optimal price and bandwidth allocation, distributed algorithm should be developed to resolve the resource allocation model (3) and its dual problem (14). Traditional subgradient-based schemes can converge to the global optimum when only considering elastic services since their utility functions are all concave. However, these schemes do not work well when considering both elastic and inelastic services since the resource allocation problem becomes an intractable and difficult non-convex problem. They may produce suboptimal or even infeasible bandwidth allocation for each peer.

3.4 Approximation problem

In this part we will approximate the utility maximization problem to an equivalent convex optimization problem by applying the successive approximation method. The resource allocation problem **P1** equals to the following problem

$$\begin{aligned} \mathbf{P2}: \quad \max \quad & \log\left(\sum_{c:c\in C^s} U_c^s(y_c^s) + \sum_{c:c\in C^r} U_c^r(y_c^r)\right) \\ \text{subject to} \quad & \sum_{p:p\in P^s(c)} x_{pc}^s = y_c^s, \ \sum_{p:p\in P^r(c)} x_{pc}^r = y_c^r, \\ & \sum_{c:c\in C^s(p)} x_{pc}^s + \sum_{c:c\in C^r(p)} x_{pc}^r \le C_p, \\ \text{over} \quad & x_{pc}^{s\min} \le x_{pc}^s \le x_{pc}^{s\max}, \\ & x_{pc}^{r\min} \le x_{pc}^r \le x_{pc}^{r\max}. \end{aligned} \tag{15}$$

Lemma 1 *The primal problem* **P1** *and its extended problem* **P2** *share the same optimal or suboptimal solutions. Furthermore, if* $(\mathbf{x}^*, \mathbf{y}^*, \phi^*, \varphi^*)$ *is a KKT point of* **P1**, *then* $(\mathbf{x}^*, \mathbf{y}^*, \phi^*/V^*, \varphi^*/V^*)$ *is the KKT point of* **P2** *where* $V^* = \sum_{c:c\in C^s} U_c^s(y_c^{s*}) + \sum_{c:c\in C^r} U_c^r(y_c^{r*})$.

Proof The first part is obvious since the objective of **P2** is a monotonically increasing logarithm function. For the second part, we can obtain the Lagrangian of problem **P2** as follows □

$$\begin{aligned} L_{P2}(\mathbf{x},\mathbf{y};\zeta,\psi) = & \log\left(\sum_{c:c\in C^s} U_c^s(y_c^s) + \sum_{c:c\in C^r} U_c^r(y_c^r)\right) \\ & + \sum_{c:c\in C^s} \zeta_c^s\left(\sum_{p:p\in P^s(c)} x_{pc}^s - y_c^s\right) + \sum_{c:c\in C^r} \zeta_c^r\left(\sum_{p:p\in P^r(c)} x_{pc}^r - y_c^r\right) \\ & + \sum_{p:p\in P} \psi_p\left(C_p - \sum_{c:c\in C^s(p)} x_{pc}^s - \sum_{c:c\in C^r(p)} x_{pc}^r\right). \end{aligned} \tag{16}$$

Then the KKT point of **P2** satisfies the following equations

$$\nabla_x L_{P2}(\mathbf{x}^*, \mathbf{y}^*; \zeta^*, \psi^*) = 0 \quad \text{and} \quad \nabla_y L_{P2}(\mathbf{x}^*, \mathbf{y}^*; \zeta^*, \psi^*) = 0 \tag{17}$$

$$\zeta_c^{s*}\left(\sum_{p:p\in P^s(c)} x_{pc}^{s*} - y_c^{*s}\right) = 0 \quad \text{and} \quad \zeta_c^{r*}\left(\sum_{p:p\in P^r(c)} x_{pc}^{r*} - y_c^{r*}\right) = 0 \tag{18}$$

$$\psi_p^*\left(C_p - \sum_{c:c\in C^s(p)} x_{pc}^{s*} - \sum_{c:c\in C^r(p)} x_{pc}^{r*}\right) = 0 \tag{19}$$

Then the second statement can be easily verified by comparing the KKT conditions of **P1** and **P2** pair-by-pair. Thus this result can be obtained.

Now the problem **P2** is still a nonconvex optimization problem. Then, based on the transformations for an optimization problem (p.4.2.4 in Boyd and Vandenberghe 2004), we transform the equivalent problem **P2** into the following epigraph-form problem.

$$\begin{aligned}
\mathbf{P3}: \quad & \max \quad \mathbb{U} \\
& \text{subject to} \quad \mathbb{U} \le \log\left(\sum_{c:c\in C^s} U_c^s(y_c^s) + \sum_{c:c\in C^r} U_c^r(y_c^r)\right), \\
& \qquad \sum_{p:p\in P^s(c)} x_{pc}^s = y_c^s, \quad \sum_{p:p\in P^r(c)} x_{pc}^r = y_c^r, \\
& \qquad \sum_{c:c\in C^s(p)} x_{pc}^s + \sum_{c:c\in C^r(p)} x_{pc}^r \le C_p, \\
& \text{over} \qquad x_{pc}^{s\min} \le x_{pc}^s \le x_{pc}^{s\max}, \\
& \qquad\quad x_{pc}^{r\min} \le x_{pc}^r \le x_{pc}^{r\max}.
\end{aligned} \tag{20}$$

We find that the aforementioned model **P3** is still a nonconvex optimization problem, since the first constraint is still nonconvex. In order to obtain a convex approximate problem, we derive an inequality to replace the nonconvex constraint with a convex one. Following Jensen's inequality, we introduce an important result as follows.

Lemma 2 *For any vector* $\xi = (\xi_c^s, c \in C^s; \xi_c^r, c \in C^r)$ *where* $\xi_c^s > 0$, $\xi_c^r > 0$ *and* $\sum_{c\in C^s} \xi_c^s + \sum_{c\in C^r} \xi_c^r = 1$, *the following inequality holds*

$$\log\left(\sum_{c:c\in C^s} U_c^s(y_c^s) + \sum_{c:c\in C^r} U_c^r(y_c^r)\right) \ge \sum_{c:c\in C^s} \xi_c^s \log\left(\frac{U_c^s(y_c^s)}{\xi_c^s}\right) + \sum_{c:c\in C^r} \xi_c^r \log\left(\frac{U_c^r(y_c^r)}{\xi_c^r}\right). \tag{21}$$

The equality (21) holds if and only if

$$\xi_c^s = \frac{U_c^s(y_c^s)}{\sum_{c:c\in C^s} U_c^s(y_c^s) + \sum_{c:c\in C^r} U_c^r(y_c^r)}, \xi_c^r = \frac{U_c^r(y_c^r)}{\sum_{c:c\in C^s} U_c^s(y_c^s) + \sum_{c:c\in C^r} U_c^r(y_c^r)}. \tag{22}$$

Then we obtain the following approximation problem based on Lemma 2

$$\begin{aligned}
\mathbf{P4}: \max \quad & \mathbb{U} \\
\text{subject to} \quad & \mathbb{U} \le \sum_{c:c\in C^s} \xi_c^s \log\left(\frac{U_c^s(y_c^s)}{\xi_c^s}\right) + \sum_{c:c\in C^r} \xi_c^r \log\left(\frac{U_c^r(y_c^r)}{\xi_c^r}\right), \\
& \sum_{p:p\in P^s(c)} x_{pc}^s = y_c^s, \quad \sum_{p:p\in P^r(c)} x_{pc}^r = y_c^r, \\
& \sum_{c:c\in C^s(p)} x_{pc}^s + \sum_{c:c\in C^r(p)} x_{pc}^r \le C_p, \\
\text{over} \quad & x_{pc}^{s\min} \le x_{pc}^s \le x_{pc}^{s\max}, \\
& x_{pc}^{r\min} \le x_{pc}^r \le x_{pc}^{r\max}.
\end{aligned} \tag{23}$$

Then, we consider the canonical form for optimization problem, and deduce the following equivalent approximation problem

$$\begin{aligned}
\mathbf{P5}: \max \quad & \sum_{c:c\in C^s} \mathcal{U}_c^s(y_c^s, \xi_c^s) + \sum_{c:c\in C^r} \mathcal{U}_c^r(y_c^r, \xi_c^r) \\
\text{subject to} \quad & \sum_{p:p\in P^s(c)} x_{pc}^s = y_c^s, \quad \sum_{p:p\in P^r(c)} x_{pc}^r = y_c^r, \\
& \sum_{c:c\in C^s(p)} x_{pc}^s + \sum_{c:c\in C^r(p)} x_{pc}^r \le C_p, \\
\text{over} \quad & x_{pc}^{s\min} \le x_{pc}^s \le x_{pc}^{s\max}, \\
& x_{pc}^{r\min} \le x_{pc}^r \le x_{pc}^{r\max},
\end{aligned} \tag{24}$$

where $\mathcal{U}_c^s(y_c^s, \xi_c^s) = \xi_c^s \log\left(\frac{U_c^s(y_c^s)}{\xi_c^s}\right)$, $\mathcal{U}_c^r(y_c^r, \xi_c^r) = \xi_c^r \log\left(\frac{U_c^r(y_c^r)}{\xi_c^r}\right)$. Now the approximation problem **P5** indeed includes a series of approximations when we choose different values of ξ. Given an initial value of ξ, the solution to **P5** is a suboptimal solution to **P1**. After substituting this suboptimal solution, a new value ξ is deduced by the update rule (22). With this new value ξ, the corresponding new **P5** is solved. After a sequence of approximations, the solution to **P5** will finally converge to the global solution to **P1**. We will provide a resource allocation scheme for convergence to the stationary point of **P5**, which satisfies the KKT conditions for an optimization problem. At the stationary point the problem **P5** is equivalent to **P1**, thus the point is exactly the optimal resource allocation of **P1**.

It is not difficult to find that the extended utility $\mathcal{U}_c^s(y_c^s, \xi_c^s)$ for elastic service s is still a concave function since $U_c^s(y_c^s)$ is concave. For inelastic service r, we analyze the extended utility function $\mathcal{U}_c^r(y_c^r, \xi_c^r)$, and obtain the following result.

Lemma 3 *The extended utility functions $\mathcal{U}_c^r(y_c^r, \xi_c^r)$ are continuously differentiable and strictly concave for inelastic services.*

Proof We prove the result by verifying the second derivative of $\mathcal{U}_c^r(y_c^r, \xi_c^r)$ with respect to variable y_c^r. □

$$\frac{d^2\mathcal{U}_c^r(y_c^r, \xi_c^r)}{dy_c^{r2}} = \frac{\xi_c^{r2}}{U_c^{r2}(y_c^r)}\left(U_c^r(y_c^r)\frac{d^2U_c^r(y_c^r)}{dy_c^{r2}} - \left(\frac{dU_c^r(y_c^r)}{dy_c^r}\right)^2\right). \tag{25}$$

Thus, utility functions $\mathcal{U}(y_c^r, \xi_c^r)$ are strictly concave if the second derivatives are negative, that is, $U_c^r(y_c^r)d^2U_c^r(y_c^r)/dy_c^{r2} - \left(dU_c^r(y_c^r)/dy_c^r\right)^2 < 0$.

For elastic service with utility function (1), we can obtain

$$U_c^s(y_c^s)\frac{d^2U_c^s(y_c^s)}{dy_c^{s2}} - \left(\frac{dU_c^s(y_c^s)}{dy_c^s}\right)^2 = -\frac{w^2}{(y_c^s+1)^2}(\log(y_c^s+1)+1) < 0.$$

For inelastic service with utility function (2), we obtain

$$U_c^r(y_c^r)\frac{d^2U_c^r(y_c^r)}{dy_c^{r2}} - \left(\frac{dU_c^r(y_c^r)}{dy_c^r}\right)^2 = -\frac{a^2w^2e^{-a(y_c^r-b)}}{(1+e^{-a(y_c^r-b)})^4}\frac{e^{-2a(y_c^r-b)}+e^{ab}}{1+e^{ab}} < 0.$$

Thus the result is obtained.

Now the objective of approximation problem **P5** is strictly concave with respect to variables y_c^s and y_c^r, but is not strictly concave with respect to variables x_{pc}^s and x_{pc}^r. Meanwhile, the constraint conditions are linear, thus the constraint set of this optimization problem is convex. Thus, based on the convex optimization theory Bertsekas et al. (2003), we can obtain the following result.

Theorem 1 *For the approximation problem* **P5** *of resource allocation for multiclass services in P2P networks, there exists unique optimal resource allocation* y_c^{s*} *and* y_c^{r*} *for customer c when requesting elastic service s and inelastic service r. However, the optimal resource provision* x_{pc}^{s*} *and* x_{pc}^{r*} *from provider p to customer c is not necessarily unique.*

4 Resource allocation scheme

4.1 Algorithm description

In order to obtain the optimum of approximation problem **P5**, we firstly introduce the Lagrangian of problem **P5**

$$\begin{aligned} L_{P5}(\mathbf{x},\mathbf{y};\lambda,\mu;\xi) \;=\; & \sum_{c:c\in C^s}\mathcal{U}_c^s(y_c^s,\xi_c^s) + \sum_{c:c\in C^r}\mathcal{U}_c^r(y_c^r,\xi_c^r) \\ & + \sum_{c:c\in C^s}\lambda_c^s\left(\sum_{p:p\in P^s(c)}x_{pc}^s - y_c^s\right) + \sum_{c:c\in C^r}\lambda_c^r\left(\sum_{p:p\in P^r(c)}x_{pc}^r - y_c^r\right) \\ & + \sum_{p:p\in P}\mu_p\left(C_p - \sum_{c:c\in C^s(p)}x_{pc}^s - \sum_{c:c\in C^r(p)}x_{pc}^r\right). \end{aligned} \tag{26}$$

We apply the gradient projection method to solve approximate problem **P5** and present the following resource allocation algorithm to achieve the optimum. The algorithm only depends on locally available information of each service provider that supports elastic and/or inelastic services.

Each service provider p updates its resource allocation for customer c who requests elastic service s with the following rule

$$x_{pc}^{s}(t+1) = \left((1-\theta)x_{pc}^{s}(t) + \theta\widetilde{x}_{pc}^{s}(t) + \theta\kappa(t)x_{pc}^{s}(t)(\lambda_{c}^{s}(t) - \mu_{p}(t))\right)_{x_{pc}^{s\min}}^{x_{pc}^{s\max}}, \tag{27}$$

$$\widetilde{x}_{pc}^{s}(t+1) = (1-\theta)\widetilde{x}_{pc}^{s}(t) + \theta x_{pc}^{s}(t), \tag{28}$$

$$\lambda_{c}^{s}(t) = \frac{\partial \mathcal{U}_{c}^{s}(y_{c}^{s}(t), \xi_{c}^{s})}{\partial y_{c}^{s}(t)}, \tag{29}$$

$$y_{c}^{s}(t) = \sum_{p:p\in P^{s}(c)} x_{pc}^{s}(t). \tag{30}$$

And each service provider p updates its resource allocation for customer c who requests inelastic service r with the following rule

$$x_{pc}^{r}(t+1) = \left((1-\theta)x_{pc}^{r}(t) + \theta\widetilde{x}_{pc}^{r}(t) + \theta\kappa(t)x_{pc}^{r}(t)(\lambda_{c}^{r}(t) - \mu_{p}(t))\right)_{x_{pc}^{r\min}}^{x_{pc}^{r\max}}, \tag{31}$$

$$\widetilde{x}_{pc}^{r}(t+1) = (1-\theta)\widetilde{x}_{pc}^{r}(t) + \theta x_{pc}^{r}(t), \tag{32}$$

$$\lambda_{c}^{r}(t) = \frac{\partial \mathcal{U}_{c}^{r}(y_{c}^{r}(t), \xi_{c}^{r})}{\partial y_{c}^{r}(t)}, \tag{33}$$

$$y_{c}^{r}(t) = \sum_{p:p\in P^{r}(c)} x_{pc}^{r}(t), \tag{34}$$

where $\kappa(t) > 0$ is the step size; $\theta > 0$ is the parameter for low-pass filtering; $a = (b)_{c}^{d}$ means $a = \min\{d, \max\{b, c\}\}$. Here, the augmented variables $\widetilde{x}_{pc}^{s}(t)$ and $\widetilde{x}_{pc}^{r}(t)$ are the optimal estimation of resource allocation $x_{pc}^{s}(t)$ and $x_{pc}^{r}(t)$, respectively, which can be assisted to remove the possible oscillation without changing the optimal solutions x_{pc}^{s*} and x_{pc}^{r*}.

Each service provider p updates its price $\mu_{p}(t)$ with the following rule

$$\mu_{p}(t+1) = \left(\mu_{p}(t) + \upsilon(t)\frac{z_{p}(t) - C_{p}}{C_{p}}\right)_{\mu_{p}(t)}^{+}, \tag{35}$$

$$z_{p}(t) = \sum_{c:c\in C^{s}(p)} x_{pc}^{s}(t) + \sum_{c:c\in C^{r}(p)} x_{pc}^{r}(t), \tag{36}$$

where $\upsilon(t) > 0$ is the step size; $a = (b)_{c}^{+}$ means $a = b$ if $c > 0$ and $a = \max\{0, b\}$ if $c = 0$.

Note that the approximation problem **P5** indeed includes a series of approximations, where each one is identified by a value ξ. If we select an appropriate value ξ, we can achieve the optimum of the corresponding approximation problem by applying the resource allocation scheme above. In order to guarantee that the approximations of problem **P5** finally become exact where the equality (21) always holds, each customer c updates its parameter ξ_c^s for elastic service s with the following rule

$$\xi_c^s = \frac{U_c^s(y_c^s(t))}{\sum_{c:c\in C^s} U_c^s(y_c^s(t)) + \sum_{c:c\in C^r} U_c^r(y_c^r(t))}, \tag{37}$$

and parameter ξ_c^r for inelastic service r with the following rule

$$\xi_c^r = \frac{U_c^r(y_c^r(t))}{\sum_{c:c\in C^s} U_c^s(y_c^s(t)) + \sum_{c:c\in C^r} U_c^r(y_c^r(t))}. \tag{38}$$

In the proposed resource allocation scheme above, if customer c requests elastic service s from provider p, it computes the price $\lambda_c^s(t)$ paid for provider p according to (29). And provider p calculates its charged price $\mu_p(t)$ according to (35), and updates its resource allocation $x_{pc}^s(t)$ for customer c with the rule of (27). We observe that resource allocation scheme (27) is a gradient-based fluid model which depends on the difference between the price $\lambda_c^s(t)$ paid by customer c and the price $\mu_p(t)$ charged by provider p. Meanwhile, if customer c requests inelastic service r from provider p, resource allocation scheme (31)–(34) will be executed to realize the optimum. On the other hand, provider p observes the aggregated load $z_p(t)$ from (36), and updates its charged price $\mu_p(t)$ according to (35). Thus the update rules for resource allocation and price are both a scaled gradient-based algorithm, which has been proven to be efficiently convergent to the optimum when choosing appropriate step sizes. However, each customer needs to learn the total utility values of all customers so as to update ξ_c^s with the law of (37) and ξ_c^r with the law of (38). Therefore, after each iteration each customer c communicates its utility value to all other customers in the network. In a new iteration the initial value is the stationary value of the previous iteration.

Therefore, the resource allocation algorithm can be described by the following pseudocode:

1: Initialization;
2: k:=0;
3: **repeat**
4: $\mathbf{x}^{(k)}(0) := \mathbf{x}^{(k-1)}(T)$ and $\mathbf{y}^{(k)}(0) := \mathbf{y}^{(k-1)}(T)$
5: Each customer c calculates $\xi_c^{s(k)}$ according to (37) and $\xi_c^{r(k)}$ according to (38);
6: **for** $t := 0$ to $T-1$ **do**
7: Each provider p updates its resource allocation $x_{pc}^{s(k)}(t)$ according to (27) and $x_{pc}^{r(k)}(t)$ according to (31);
8: Each customer c updates its prices $\lambda_c^{s(k)}(t)$ according to (29) and $\lambda_c^{r(k)}(t)$ according to (33);
9: Each provider p updates its price $\mu_p^{(k)}(t)$ according to (35)-(36);
10: **end for**
11: k:=k+1;
12: **until** convergence

The algorithm presents a modification of the successive approximation method where the number of inner-iterations is limited to a certain value T. In the proposed resource allocation scheme, the value in the last inner-iteration of the previous outer-iteration is the initial value in the next outer-iteration (Step 4). To solve the new approximate optimization, ξ is computed according to the resource allocation in the last iteration of the previous outer-iteration (Step 5). In next part we will investigate the performance of the proposed resource allocation scheme.

4.2 Performance analysis

In this part we first obtain the following result by applying convex optimization approach.

Lemma 4 *The KKT point of approximation problem* **P5** *is also the KKT point of extended problem* **P2**, *that is, if* $(\mathbf{x}^*, \mathbf{y}^*; \lambda^*, \mu^*; \xi^*)$ *is a KKT point of problem* **P5**, *then* $(\mathbf{x}^*, \mathbf{y}^*; \lambda^*, \mu^*)$ *is also a KKT point of problem* **P2**.

Proof From the Lagrangian (26), we can obtain the KKT conditions of **P5**

$$\nabla_x L_{P5}(\mathbf{x}^*, \mathbf{y}^*; \lambda^*, \mu^*; \xi^*) = 0 \quad \text{and} \quad \nabla_y L_{P5}(\mathbf{x}^*, \mathbf{y}^*; \lambda^*, \mu^*; \xi^*) = 0 \tag{39}$$

$$\lambda_c^{s*}\left(\sum_{p:p\in P^s(c)} x_{pc}^{s*} - y_c^{s*}\right) = 0 \quad \text{and} \quad \lambda_c^{r*}\left(\sum_{p:p\in P^r(c)} x_{pc}^{r*} - y_c^{r*}\right) = 0 \tag{40}$$

$$\mu_p\left(C_p^* - \sum_{c:c\in C^s(p)} x_{pc}^{s*} - \sum_{c:c\in C^r(p)} x_{pc}^{r*}\right) = 0 \tag{41}$$

□

We can verify that the point $(\mathbf{x}^*, \mathbf{y}^*; \lambda^*, \mu^*)$ also satisfies the Eqs. (17)–(19) which just are the KKT conditions of problem **P2** after substituting $\xi_c^{s*} = U_c^s(y_c^{s*})/V^*$ and $\xi_c^{r*} = U_c^r(y_c^{r*})/V^*$ into the equations above, where $V^* = \sum_{c:c\in C^s} U_c^s(y_c^{s*}) + \sum_{c:c\in C^r} U_c^r(y_c^{r*})$. Then this result is obtained.

We study the convergence of the proposed resource allocation scheme, and obtain the following theorem.

Theorem 2 *If the step sizes are sufficient small, then the proposed resource allocation scheme finally converges to a stationary point that satisfies the KKT conditions of primal problem* **P1**.

Proof Define $\mathbf{x}^{(k)}(0)$ to be the initial point of outer-iteration k, and $\mathbf{x}^{(k)*}$ to be the stationary point of outer-iteration k. We first prove that $\mathbf{x}^{(k)*}$ is obtainable in each outer-iteration k. Given a value ξ, the approximation problem **P5** is strictly concave with respect to variables $y_c^s(t)$ and $y_c^r(t)$, thus it has a unique optimum solution $\mathbf{y}^*$. However, the optimal resource allocation x_{pc}^{s*} and x_{pc}^{r*} from provider p to customer c for elastic service s and inelastic service r is not necessarily unique. With the assumptions on the small step sizes $\kappa(t) > 0$ and $\upsilon(t) > 0$, the proposed resource allocation scheme which is a gradient-based algorithm can converge to one of the optimums given $\xi_c^{s(k)}$ and $\xi_c^{r(k)}$ in each k outer-iteration Bertsekas et al. (2003). Furthermore, we apply low-pass filtering method in the proposed resource allocation scheme, which can remove the possible oscillation due to non-uniqueness of optimum. Thus the optimum $\mathbf{x}^{(k)*}$ as well as $\mathbf{y}^{(k)*}$ is obtainable. □

Now we prove the convergence of the proposed resource allocation scheme. Denote the objective of **P2** as $F(\mathbf{x}) \triangleq G(\mathbf{y}) = \log\left(\sum_{c:c\in C^s} U_c^s(y_c^s) + \sum_{c:c\in C^r} U_c^r(y_c^r)\right)$ where $y_c^s = \sum_{p:p\in P^s(c)} x_{pc}^s$ and $y_c^r = \sum_{p:p\in P^r(c)} x_{pc}^r$. The solution of **P5** indeed increases monotonically $G(\mathbf{y})$ in each outer-iteration.

$$\begin{aligned} F(\mathbf{x}^{(k-1)*}) &= G(\mathbf{y}^{(k-1)*}) = \sum_{c:c\in C^s} \mathcal{U}_c^s(y_c^{s(k)}(0), \xi_c^{s(k)}) + \sum_{c:c\in C^r} \mathcal{U}_c^r(y_c^{r(k)}(0), \xi_c^{r(k)}) \\ &\le \sum_{c:c\in C^s} \mathcal{U}_c^s(y_c^{s(k)*}, \xi_c^{s(k)}) + \sum_{c:c\in C^r} \mathcal{U}_c^r(y_c^{r(k)*}, \xi_c^{r(k)}) \\ &\le G(\mathbf{y}^{(k)*}) = F(\mathbf{x}^{(k)*}). \end{aligned} \tag{42}$$

The second equality is deduced by substituting $\mathbf{y}^{(k)}(0) = \mathbf{y}^{(k-1)*}$ and $\xi_c^{s(k)} = U_c^s(y_c^{s(k-1)*})/V^{(k-1)*}$, $\xi_c^{r(k)} = U_c^r(y_c^{r(k-1)*})/V^{(k-1)*}$, where $V^{(k-1)*} = \sum_{c:c\in C^s} U_c^s(y_c^{s(k-1)*}) + \sum_{c:c\in C^r} U_c^r(y_c^{r(k-1)*})$, into the right-hand size. The first inequality is obtained since $\mathbf{y}^{(k)*}$ is the optimum of **P5** given a value $\xi^{(k)}$. The second inequality is satisfied from (21). Meanwhile, $G(\mathbf{y})$ is a continuous function, then $G(\mathbf{y})$ is bounded as $\mathbf{y}$ is bounded (recall that $\mathbf{x}$ is bounded). Furthermore,

the sequence $\{G(\mathbf{y}^{(k)*}), k = 1, 2, \dots\}$ monotonically increases, thus it can converge from convex optimization (Bertsekas et al. 2003, Prop.A.3). Then the sequence $\{\sum_{c:c\in C^s} U_c^s(y_c^{s(k)*}) + \sum_{c:c\in C^r} U_c^r(y_c^{r(k)*}), k = 1, 2, \dots\}$ also converges.

From Lemma 4 we have known that the stationary point of the approximation problem **P5** is also the KKT point of **P2**. On the other hand, based on Lemma 1, we have obtained the KKT conditions of **P2** are equivalent to those of **P1**. Thus the proposed resource allocation scheme can eventually converge to the stationary point which satisfies the KKT conditions of primal problem **P1**. Then, this theorem is completed.

Recall that the utility maximization model of resource allocation for multiclass services in P2Ps is a non-convex optimization problem, since the utility functions for inelastic services are sigmoidal, then the final stationary points satisfying the KKT conditions may be suboptimal solutions to the primal problem. By substituting these KKT points into the utility maximization model and comparing the corresponding objective value, the final optimal solution can be derived, as well as the optimal objective.

5 Numerical examples and discussions

In this part we consider the performance of the proposed resource allocation mechanism and present some numerical examples to show its convergence in P2P networks. We firstly investigate the scheme in a simple network architecture and then analyze its performance in large scale networks.

5.1 Simple network architecture

Consider a simple network which is composed of two service providers and four service customers. The service providers have the access link capacity $C = (C_1, C_2) = (5, 10)$ Mbps. The minimal and maximal resource requirements for each elastic or inelastic service are 0.01 Mbps and 10 Mbps, respectively. The initial rates $x_{pc}^s(t)$ and $x_{pc}^r(t)$ are all set to 0.5 Mbps, and the initial parameters ξ_c^s and ξ_c^r are all chosen 0.25. The low-pass filtering parameter is chosen $\theta = 0.2$.

5.1.1 Resource allocation for elastic services

In this case we consider the four customers are only requesting elastic services and analyze the proposed resource allocation scheme. The willingness-to-pay of four customers is $w = (4, 3, 2, 1)$, then the utility functions when customers request these elastic services are given by: $U_1^s(y_1^s) = 4\log(y_1^s + 1)$, $U_2^s(y_2^s) = 3\log(y_2^s + 1)$, $U_3^s(y_3^s) = 2\log(y_3^s + 1)$, and $U_4^s(y_4^s) = \log(y_4^s + 1)$.

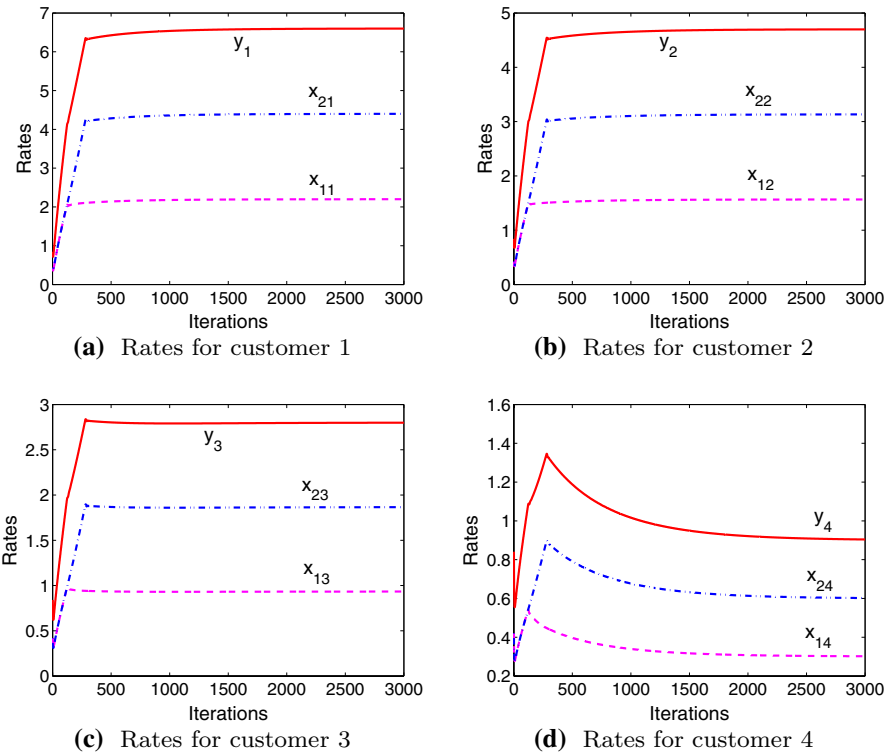
We observe that the proposed scheme has two levels of convergence. The outer-iterations update ξ, and the inner-iterations solve the convex approximation problem **P5**. In order to guarantee the convergence of every outer-iteration process, the

step sizes should be small enough and the number of inner-iterations should be large enough. Then we choose the step sizes $\kappa(t) = \upsilon(t) = 0.2/t$ and the number of inner-iterations as a fixed value $T = 500$.

We depict the simulation results obtained from the proposed resource allocation scheme in Fig. 2 and can observe its validity and performance. We find that the scheme gradually tends to a steady state and finally converges to an optimal resource allocation for each service customer within reasonable iterations.

We list the optimum derived from the proposed scheme in Table 1. Meanwhile, we also presented the optimal solution by using nonlinear programming software LINGO in this table. Since the objective function is not a strictly concave function with the variable $x = (x^s_{pc}, p \in P, c \in C^s)$, the optimal resource allocated by service providers for customers is not unique, which is also discussed in Theorem 1. However, the objective function is a strictly concave function with the variable $y = (y^s_c, c \in C^s)$, thus as we can observe in Table 1, the optimal total resource obtained by each customer is unique.

Next we analyze the convergence speed of the proposed resource allocation algorithm. In order to improve the convergence speed of the algorithm, we conduct several experiments to find an appropriate value for the number T of inner-iterations.



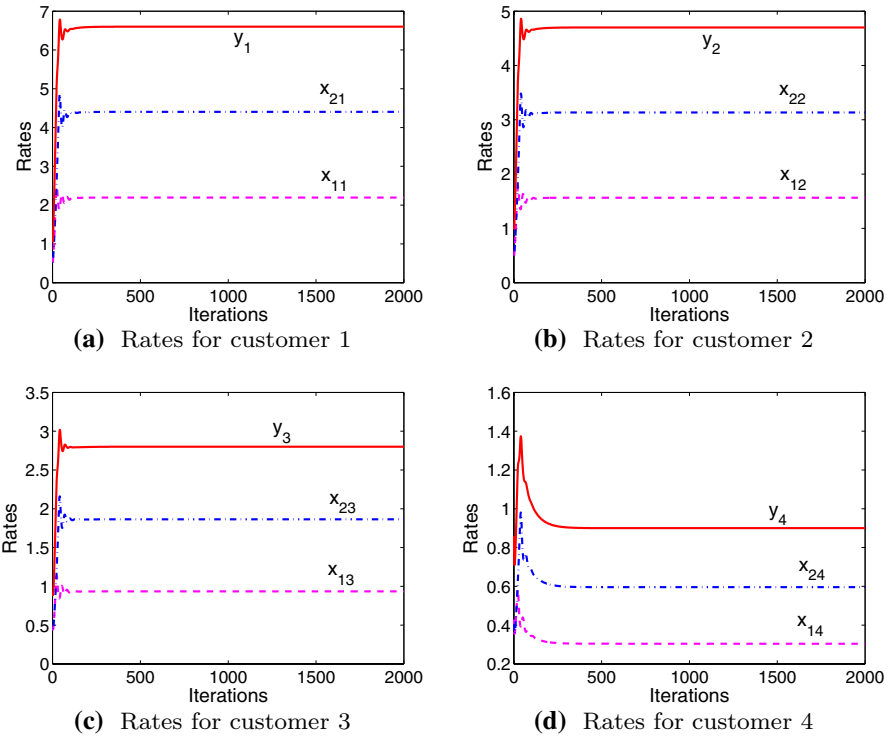
(a) Rates for customer 1 **(b)** Rates for customer 2

(c) Rates for customer 3 **(d)** Rates for customer 4

Fig. 2 Performance of the resource allocation algorithm for elastic services with parameters $T = 500$ and $\kappa(t) = \upsilon(t) = 0.2/t$

Table 1 The optimum for the resource allocation model: elastic services

Variable	x_{11}^{s*}	x_{21}^{s*}	x_{12}^{s*}	x_{22}^{s*}	y_1^{s*}	y_2^{s*}
Algorithm	2.1987	4.3995	1.5660	3.1326	6.5982	4.6986
LINGO	0.6968	5.9031	1.6972	3.0027	6.5999	4.6999
Variable	x_{13}^{s*}	x_{23}^{s*}	x_{14}^{s*}	x_{24}^{s*}	y_3^{s*}	y_4^{s*}
Algorithm	0.9334	1.8657	0.3020	0.6021	2.7991	0.9041
LINGO	2.0556	0.7443	0.5503	0.3496	2.7999	0.8999



(a) Rates for customer 1 **(b)** Rates for customer 2

(c) Rates for customer 3 **(d)** Rates for customer 4

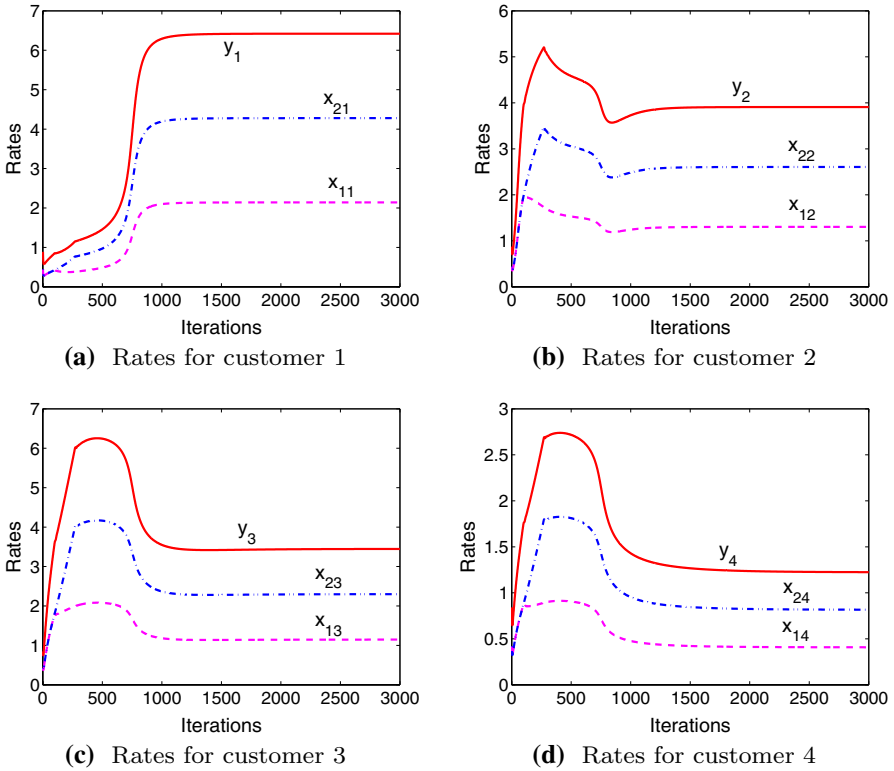
Fig. 3 Performance of the resource allocation algorithm for elastic services with parameters $T = 100$ and $\kappa(t) = \upsilon(t) = 0.2$

Also we apply a constant step size for the subgradient-based update law since it often has a faster convergence speed than the diminishing step size. In this simulation we choose the inner-iterations number $T = 100$ and step sizes $\kappa(t) = \upsilon(t) = 0.2$, and depict the performance of the scheme in Fig. 3. We find that the convergence speed of the algorithm is improved at this time, and the optimum can be achieved in fewer iterations.

5.1.2 Resource allocation for multiclass services

In this part we investigate the performance of the proposed resource allocation scheme for both elastic and inelastic services. Without loss of generality, we assume the first two customers request inelastic services with utility functions $U_1^r(y_1^r) = \frac{6}{1+e^{-(y_1^r-4)}} - \frac{6}{1+e^4}$ and $U_2^r(y_2^r) = \frac{4}{1+e^{-(y_2^r-2)}} - \frac{4}{1+e^2}$. The others request elastic services with utility functions $U_3^s(y_3^s) = 2\log(y_3^s + 1)$, and $U_4^s(y_4^s) = \log(y_4^s + 1)$. We firstly choose the number of inner-iterations $T = 500$ and the diminishing step sizes $\kappa(t) = \upsilon(t) = 0.2/t$, and present the behavior of the scheme in Fig. 4. We observe that the scheme finally converges to an optimal resource allocation $x^* =$ (2.1474, 4.2731, 1.3096, 2.5993, 1.1372, 2.3098, 0.4057, 0.8178)Mbps within reasonable iterations.

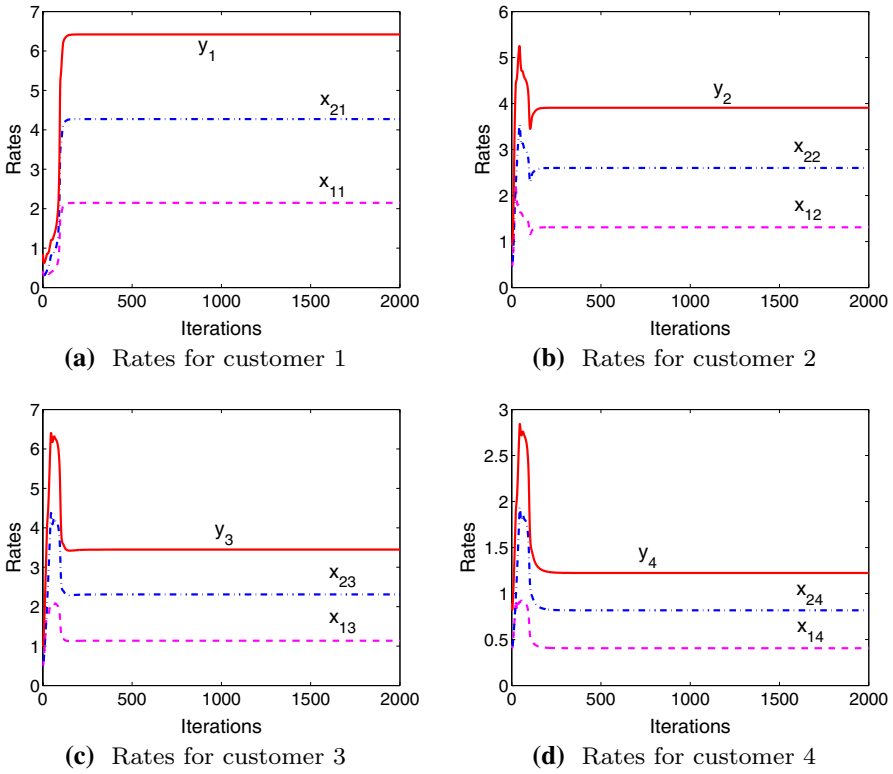
We also investigate the convergence speed of the resource allocation scheme in this scenario where elastic and inelastic services are coexisting. We choose the inner-iterations number $T = 100$ and constant step sizes $\kappa(t) = \upsilon(t) = 0.2$, and depict the scheme behavior in Fig. 5. We can observe that the scheme converges



(a) Rates for customer 1 **(b)** Rates for customer 2

(c) Rates for customer 3 **(d)** Rates for customer 4

Fig. 4 Performance of the resource allocation algorithm for elastic and inelastic services with parameters $T = 500$ and $\kappa(t) = \upsilon(t) = 0.2/t$



(a) Rates for customer 1 **(b)** Rates for customer 2

(c) Rates for customer 3 **(d)** Rates for customer 4

Fig. 5 Performance of the resource allocation algorithm for elastic and inelastic services with parameters $T = 100$ and $\kappa(t) = \upsilon(t) = 0.2$

much faster than that when choosing larger inner-iterations number and diminishing step sizes.

5.2 Large scale networks

Now we consider the performance of the resource allocation scheme in large scale networks with different number of peers. The access link capacity of service providers are all assumed to be 20 Mbps. The low-pass filtering parameter is $\theta = 0.2$ and the initial parameters are all $\xi_c^s = 0.25$. And the number of inner-iterations is $T = 100$ and the step sizes are $\kappa(t) = \upsilon(t) = 0.2$.

5.2.1 Resource allocation for elastic services

Considering the different willingness-to-pay of customers when they request services, we assume that there are four types of customers in the P2P networks. Each type of customers has the same number and is with one of the utility functions

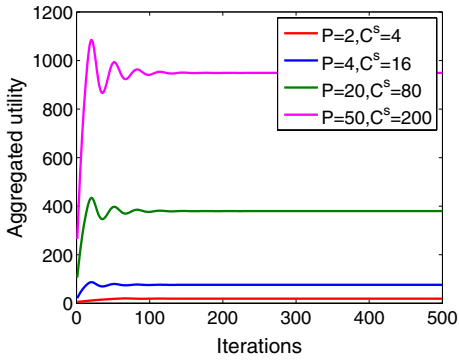


Fig. 6 Performance of the resource allocation algorithm for elastic services in large scale networks

discussed in Sect. 5.1.1. In Fig. 6, we depict the evolution of aggregated utility in large scale P2P networks with different number of peers. We find that the number of peers does not influence the convergence speed of the resource allocation scheme. The aggregated utility increases gradually with the number of peers but, in almost all cases, the optimal value is achieved within the same number of iterations (e.g., 200 iterations). In fact, it is the parameters of the resource allocation scheme such as inner-iterations number and step sizes that mainly affect the convergence speed, as we have discussed in Sect. 5.1.

5.2.2 Resource allocation for multiclass services

In this case of resource allocation for multiclass services in large scale networks, we assume that both customers requesting elastic services and those requesting inelastic services are divided into two types. Each type of customers has the same one of the utility functions considered in Sect. 5.1.2. Then in Fig. 7, we present the evolution of aggregated utility of multiclass services in different scale networks. Similar to the results for only elastic services in Sect. 5.2.1, the optimal objective can be finally reached in almost the same number of iterations. However, it appears to converge slower than that in the case for only elastic services. This is rather expected. When considering only elastic services, the service customers have the same structure of logarithmic utility functions and run the separate update steps in parallel. We

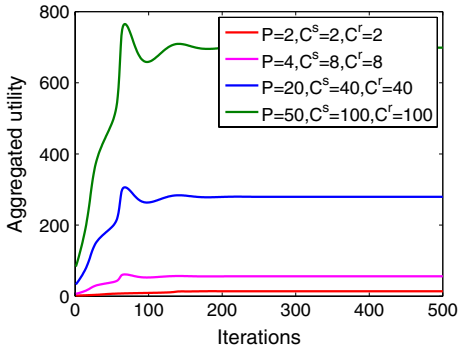


Fig. 7 Performance of the resource allocation algorithm for multiclass services in large scale networks

believe that the convergence speed can be also improved if we choose proper parameters such as more suitable inner-iterations number and step sizes.

6 Conclusions

In recent years P2P networks have played an important role in supporting elastic applications (e.g., file sharing) and inelastic applications (e.g., video distribution) over the Internet, and been applied into many scenarios, e.g., distributed storage, cloud computing, edge computing and vehicular networks. Especially, P2Ps have well supported online video conferencing applications (e.g., Tencent meeting, Zoom) in 2020 due to the epidemic of COVID-19. However, it is a crucial challenge and difficult problem to achieve reasonable and effective resource allocation for peers who acquire both elastic and inelastic services. In this paper we concentrate on resource allocation for both elastic and inelastic services in P2P networks, and formulate the utility maximization model for peers who request these services. The utility maximization model is an intractable and difficult non-convex optimization problem, since the inelastic services have non-concave utility functions. In order to obtain the optimal resource allocation, we approximate the utility maximization problem to an equivalent convex optimization problem by applying the successive approximation method, and design a gradient-based resource allocation scheme to achieve the optimal solution of the approximations. The proposed scheme is proven to converge to an optimal solution of the primal utility maximization model which also satisfies the KKT conditions. Numerical examples verify the convergence of resource allocation scheme for both elastic and inelastic services. For further research work, we will investigate the resource allocation of multiclass applications in edge computing which builds on P2Ps.

Acknowledgements The authors would like to thank the anonymous reviewers and Associate Editor for very detailed and helpful comments and suggestions to improve this work, and the support from the National Natural Science Foundation of China (Nos. 71671159 and 71971188), the Humanity and Social Science Foundation of Ministry of Education of China (No. 16YJC630106), and the Natural Science Foundation of Hebei Province (Nos. G2018203302, G2020203005).

References

Antal E, Vinkó T (2016) Modeling max-min fair bandwidth allocation in BitTorrent communities. Comput Optim Appl 66(2):383–400

Bertsekas DP, Nedic A, Ozdaglar AE (2003) Convex analysis and optimization. Athena Scientific, Belmont

Boyd S, Vandenberghe L (2004) Convex optimization. Cambridge University Press, New York

Chen M, Ponec M, Sengupta S, Li J, Chou PA (2012) Utility maximization in peer-to-peer systems with applications to video conferencing. IEEE/ACM Trans Netw 20(6):1681–1694

Eger K, Killat U (2007) Resource pricing in peer-to-peer networks. IEEE Commun Lett 11(1):82–84

Eger K, Killat U (2007) Fair resource allocation in peer-to-peer networks (extended version). Comput Commun 30(16):3046–3054

Goswami A, Gupta R, Parashari GS (2017) Reputation-based resource allocation in P2P systems: a game theoretic perspective. IEEE Commun Lett 21(6):1273–1276
Goswami A, Parashari GS, Gupta R (2018) Evolutionary stability of reputation-based incentive mechanisms in P2P Systems. IEEE Commun Lett 22(2):268–271
Gupta R, Singha N, Singh YN (2016) Reputation based probabilistic resource allocation for avoiding free riding and formation of common interest groups in unstructured P2P networks. Peer-to-Peer Netw Appl 9(6):1101–1113
Hande P, Zhang S, Chiang M (2007) Distributed rate allocation for inelastic flows. IEEE/ACM Trans Netw 15(6):1240–1253
Kang X, Wu Y (2015) Incentive mechanism design for heterogeneous peer-to-peer networks: a Stackelberg game approach. IEEE Trans Mob Comput 14(5):1018–1030
Koutsopoulos I, Iosifidis G (2010) A framework for distributed bandwidth allocation in peer-to-peer networks. Perform Eval 67(4):285–298
Kumar C, Altinkemer K, De P (2011) A mechanism for pricing and resource allocation in peer-to-peer networks. Electron Commer Res Appl 10(1):26–37
Lee JW, Mazumdar RR, Shroff NB (2005) Non-convex optimization and rate control for multi-class services in the Internet. IEEE/ACM Trans Netw 13(4):827–840
Li S, Sun W (2016) A mechanism for resource pricing and fairness in peer-to-peer networks. Electron Commerce Res 16(4):425–451
Li S, Sun W (2020) Utility maximisation for resource allocation of migrating enterprise applications into the cloud. Enterpr Inform Syst. https://doi.org/10.1080/17517575.2020.1730445
Li S, Sun W, Tian N (2015) Resource allocation for multi-class services in multipath networks. Perform Eval 92:1–23
Li S, Jiao L, Zhang Y, Wang Y, Sun W (2017) A scheme of resource allocation for heterogeneous services in peer-to-peer networks using particle swarm optimization. IAENG Int J Comput Sci 44(4):482–488
Li S, Zhang Y, Wang Y, Sun W (2019) Utility optimization-based bandwidth allocation for elastic and inelastic services in peer-to-peer networks. Int J Appl Math Comput Sci 29(1):111–123
Li S, Zhang Y, Sun W (2019) Optimal resource allocation model and algorithm for elastic enterprise applications migration to the cloud. Mathematics 7(10):1–20
Li S, Sun W, Li Q-L (2020) Utility maximization for bandwidth allocation in peer-to-peer file-sharing networks. J Ind Manag Optim 16(3):1099–1117
Liang C, Zhao M, Liu Y (2011) Optimal bandwidth sharing in multiswarm multiparty P2P videoconferencing systems. IEEE/ACM Trans Netw 19(6):1704–1716
Liu J, Ahmad S, Buyukkaya E et al (2015) Resource allocation in underprovisioned multioverlay peer-to-peer live video sharing services. Peer-to-Peer Netw Appl 8(3):399–413
Marks BR, Wright GP (1978) A general inner approximation algorithm for nonconvex mathematical programs. Oper Res 26(4):681–683
Mostafavi S, Dehghan M (2016) Game-theoretic auction design for bandwidth sharing in helper-assisted P2P streaming. Int J Commun Syst 29(6):1057–1072
Mostafavi S, Dehghan M (2017) A stochastic approximation resource allocation approach for HD live streaming. Telecommun Syst 64(1):87–101
Pacifici V, Lehrieder F, Dan G (2016) Cache bandwidth allocation for P2P file-sharing systems to minimize inter-ISP traffic. IEEE/ACM Trans Netw 24(1):437–448
Rismanchian F, Lee YH (2018) Moment-based approximations for first- and second-order transient performance measures of an unreliable workstation. Oper Res Int J 18(1):75–95
Rohmer T, Nakib A, Nafaa A (2015) A learning-based resource allocation approach for P2P streaming systems. IEEE Netw 29(1):4–11
Satsiou A, Tassiulas L (2010) Reputation-based resource allocation in P2P systems of rational users. IEEE Trans Parallel Distrib Syst 21(4):466–479
Song F, Zhu M, Zhou Y, You I, Zhang H (2020) Smart collaborative tracking for ubiquitous power IoT in edge-cloud interplay domain. IEEE Internet Things J 7(7):6046–6055
Song F, Ai Z, Zhou Y, You I, Choo R, Zhang H (2020) Smart collaborative automation for receive buffer control in multipath industrial networks. IEEE Trans Ind Inf 16(2):1385–1394
Tran NH, Hong CS (2010) Joint rate and power control in wireless network: a novel successive approximations method. IEEE Commun Lett 14(9):872–874

Vo PL, Tran NH, Hong CS (2011) Joint rate and power control for elastic and inelastic traffic in multihop wireless networks. In: The Procedings of the IEEE global telecommunications conference (GLOBECOM 2011), pp 1–5

Vo PL, Lee S, Hong CS (2012) The random access NUM with multiclass traffic. EURASIP J Wirel Commun Netw 242:1–12

Vo PL, Tran NH, Hong CS, Lee S (2013) Network utility maximisation framework with multiclass traffic. IET Netw 2(3):152–161

Vo PL, Le TA, Lee S, Hong CS, Kim B, Song H (2014) Multi-path utility maximization and multipath TCP design. J Parallel Distrib Comput 74(1):1848–1857

Wang K, Yin H, Quan W, Min G (2018) Enabling collaborative edge computing for software defined vehicular networks. IEEE Netw 32(5):112–117

Wang K, Quan W, Cheng N, Liu M, Liu Y, Anthony Chan H (2019) Betweenness centrality based software defined routing: observation from practical Internet datasets. ACM Trans Internet Technol 19(4):1–19

Yan H, Gao D, Su W, Foh CH, Zhang H, Vasilakos A (2017) Caching strategy based on hierarchical cluster for named data networking. IEEE Access 5:8433–8443