Future Generation Computer Systems 95 (2019) 277–291

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Stochastic game-based dynamic information delivery system for wireless cooperative networks☆

Li Feng [a], Amjad Ali [b,c], Hannan Bin Liaqat [d], Muhammad Aksam Iftikhar [c], Ali Kashif Bashir [e], Sangheon Pack [b,*]

[a] School of Computer Science and Communication Engineering, JiangSu University, China
[b] School of Electrical Engineering, Korea University, South Korea
[c] Department of Computer Science, COMSATS University Islamabad, Lahore Campus, Pakistan
[d] Department of Information Technology, University of Gujrat, Pakistan
[e] Department of Computing, Mathematics, and Digital Technology, Manchester Metropolitan University, UK

HIGHLIGHTS

- We develop a virtual selfishness queue (VSQ), to model the relay's selfish character after having analyzed the relationship between relay's selfishness and its time varying factors in WCNs.
- In harmony with the time-varying network state information, a novel dynamic data transmission scheme is proposed to coordinate the source and the relays to achieve their objectives, respectively.
- We employ the stochastic game to model the strategic interactions among selfish relays and prove the existence of Nash equilibrium (NE). Moreover, a combined Q-learning algorithm is raised for the relay to obtain the equilibrium strategy.
- We present a dynamic FA algorithm for the source to maximize the average network throughput whilst keeping the network stability and bounding relay's selfishness. In the proposed algorithm, the source executes the dynamic FA based solely on relays' current VSQ information and QSI.

ARTICLE INFO

Article history:
Received 31 July 2018
Received in revised form 28 December 2018
Accepted 9 January 2019
Available online 14 January 2019

Keywords:
Tactile internet
5G
Wireless resource allocation
Virtual selfish queue
Stochastic game
Incentive mechanism

ABSTRACT

The haptic communications is considered as the prime application running on the Tactile Internet. Therefore, Tactile Internet required to be highly reliable, provide a very low latencies, and required sufficient capacities at intermediate nodes to allow a large number of devices to communicate with each other simultaneously and autonomously. Moreover, the wireless cooperative network (WCN), is considered as one of the major component of the 5G technologies due to it promising advantages, such as improving wireless transmission capacity and reliability. However, the selfish nature of relay nodes may depress such enhancement and is not favored by the source node. In this paper, we propose an incentive-based dynamic flow allocation (FA) and forwarding strategy selection (FSS) scheme under time-varying selfishness. In the proposed scheme, the source node determines the FA to maximize the average network throughput under the constraints of network stability and selfishness boundaries, while each selfish relay executes the FSS to optimize its own profit with regard to the dynamic network state. Moreover, to cope with the conflicting interests between selfish relays a stochastic game model is employed to design a competition for haptic information forwarding and Nash equilibrium is proven also a combined Q-learning-based algorithm is proposed to guide the relays' forwarding strategies. Furthermore, by considering the stochastic property of the network state, the FA for the source is formulated as a stochastic optimization problem. Finally, by exploiting the concept of virtual selfishness queue, the problem is solved by using the Lyapunov optimization theory. Performance of the proposed scheme is evaluated with traditional FA approach and data queue-based FA approach. Numerical results exhibit that our scheme not only sustains a large network throughput but also achieves low latency and avoids the occurrence of a completely selfish relay in the long term.

☆ This research was supported by National Research Foundation (NRF) of Korea Grant funded by the Korean Government (MSIP) (No. 2017R1E1A1A01073742)

* Corresponding author.

E-mail addresses: fenglixidian@gmail.com (L. Feng), amjadali@korea.ac.kr (A. Ali), hannan.liaqat@uog.edu.pk (H.B. Liaqat), aksamiftikhar@cuilahore.edu.pk (M.A. Iftikhar), dr.alikashif.b@ieee.org (A.K. Bashir), shpack@korea.ac.kr (S. Pack).

https://doi.org/10.1016/j.future.2019.01.011
0167-739X/

1. Introduction

Recently, the Tactile Internet has attracted significant attention of both industrial and academic communities [1]. Because of its ability to transport real-time control and physical tactile experiences remotely, the Tactile Internet requires low latency and sufficient capacity to perform critical routine life tasks remotely over the Internet. The haptic communications will be the prime application running on the Tactile Internet [2] and the 5G communication systems are expected to play an integral part in underpinning the Tactile Internet at the wireless edge [3]. The cooperative diversity is considered as the one of the promising technique for enhancing the network throughput in 5G communication systems [4,5]. In wireless cooperative networks (WCNs), the participation of more relay nodes can significantly increases the reception reliability, improves the data rate, enhances the connectivity, and supports the low latencies. However, owing to the proliferation of the intelligence agents, the relay nodes that participate in cooperative transmission in WCNs are endowed with smart autonomic functions and pursue self-interest [6]. Such autonomous relays may present selfishness when delivering the intended information, which is the unavoidable result of many practical factors, e.g., finite energy and limited computing capacity. The selfish behavior of the relay significantly degrades the performance of WCNs, which have high requirements on delay time and data rates [7,8], and may further hinders the intersection of the larger Tactile Internet and emerging 5G systems.

In order to enhance the performance of WCNs under the scenario of selfish nodes, information delivery-related issues need to be addressed. First, the incentive-based mechanism can play a key role in stimulating the cooperation among relay nodes [9]. Second, as each relay updates its forwarding strategy based on its own profit and the interference causes the coupling among relays, an effective technique should be proposed to coordinate the forwarding strategies among relays for effective information delivery in WNCs. Third, as the relay's selfishness changes with its time-varying factors, a comprehensive mathematical model should be developed to portray the dynamics of the relay's selfishness. Besides, the flow allocation (FA) of the Tactile traffic must be completed under the time-varying selfishness and data queue state information (QSI). A dynamic program should be developed for the source node to perform Tactile information delivery in WCNs. Hence, it is important to introduce a dynamic information delivery scheme to coordinate both the FA for the source and the forwarding strategy selection (FSS) for each selfish relay, based on the time-varying network state (data queues and selfishness at relays).

In this study, we examine the dynamic delivery of Tactile information for WCNs with dynamic selfishness. Considering the conflicts of the selfish relays' benefits due to their mutual interferences, a stochastic game is presented to model the competition among relays for optimizing their own profits. Meanwhile, a Lyapunov optimization framework is employed for the source to maximize the average network throughput under the constraints of the relays' selfishness boundaries and network stability. Because the network stability also indicates the limitations of the information delivery delay [10], our scheme also meets the low-latency requirement for delivering real-time control in WCNs.

The main contributions of this study are listed as follows:

- Consistent with the time-varying network state, we propose a novel dynamic information delivery scheme to execute the remote real-time control in WCNs by coordinating the source and the relays to achieve their objectives.
- We develop a virtual selfishness queue (VSQ), to model the relay's selfish behavior after analyzing the relationship between the relay's selfishness and its time-varying factors in WCNs.

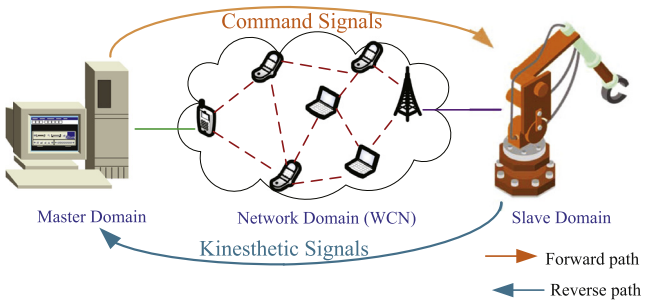


Fig. 1. Haptic information delivery over WCN..

- We employ the stochastic game to model the strategic interactions among selfish relays and prove the existence of Nash equilibrium. Moreover, a combined Q-learning algorithm is developed for the relay to obtain an equilibrium strategy.
- We present a dynamic FA algorithm for the source to maximize the average network throughput while keeping the network stability and bounding the relay's selfishness. In the proposed algorithm, the source executes the dynamic FA based solely on the relays' current VSQ information and QSI.

The rest of the paper is organized as follows. Section 2 presents the related works. The system model and incentive mechanism are presented in Section 3. In Section 4, we provide the problem formulation and build the virtual selfishness queue. In Section 5, we design a stochastic game for relays to update their forwarding strategies. A dynamic FA algorithm is proposed for the source in Section 6. Section 7 shows the performance of the scheme while Section 8 presents the simulation results to evaluate the proposed scheme. The conclusions are presented in Section 9.

2. Related works

Resource allocation including FA has received significant research attention in recent years [11]. A gradient-based FA algorithm was proposed in [12] regarding the fairness among flows; however, the nodes' selfishness in wireless networks was neglected. In order to study the strategic interactions among selfish nodes, game theory is considered a powerful tool in wireless networks [13] because it is the inherent nature for selfish nodes to reduce their costs whilst simultaneously maximizing their payoffs, e.g., energy resource. An evolutionary game-based packet forwarding scheme was introduced in [14]. However, they did not provide the specific decision-making process of autonomous relay and ignores the potential conflicting interests among the source and the relays. Stackelberg game is always used to manage the different objectives of the source and relays in wireless networks [15,16]. An incentive mechanism was proposed in [15] via Stackelberg game to obtain the optimal pricing and purchasing strategy, whilst a socially optimal approach routing was presented in [16] based on Stackelberg game. To ensure the cooperation among autonomous nodes over wireless cooperative networks, a core game was proposed in [17] by considering multi-sources and multi-relays. However, these schemes are mainly based on the static situations, e.g., the node's energy resource is assumed to be constant, which may not be valid for practical wireless networks.

3. System model

Considering the node's dynamic selfishness, which changes with its time-varying factors, an FA optimization problem was proposed in [18] based on the social and behavioral trust of the

Table 1
Key notations.

Notation	Interpretation
$\mathcal{G}, \mathcal{V}$	directed graph, node set
$\mathcal{L}, h^{\tau}_{nn}$	directed links, channel gain
$\mathcal{N}^{\mathcal{I}}, n^{\mathcal{I}}$	relay set, relay
$\mathcal{N}^{\mathcal{U}}, n^{\mathcal{U}}$	TD set, TD
$\mathcal{F}, r_n$	data session, delivery data
$\mathbf{F}^{\mathcal{I}}, f^{\mathcal{I}}_n$	relays' session set, $n^{\mathcal{I}}$' session
$\mathbf{F}^{\mathcal{U}}, f^{\mathcal{U}}_n$	TDs' session set, $n^{\mathcal{U}}$' session
$\mathcal{R}^g, \mathcal{R}^p$	reward, token consumption
t, τ	time frame, timeslot
ρ	slots number within one frame
$\mathcal{A}_n$	strategy set
$a_n, \mathbf{a}_{-n}$	$n^{\mathcal{I}}$' strategy, others' strategies
f, d	cooperative action, dropping action
P^T_n, P^R_n	transmission power, reception power
$\gamma^{\tau}_n(a_n, \mathbf{a}_{-n})$	SINR
W, D^t_n	bandwidth, queue backlog
γ^*_n, o^{τ}_n	SINR threshold, interference state
$\mathcal{R}^{\tau}_n(a_n, \mathbf{a}_{-n}, o^{\tau}_n)$	$n^{\mathcal{I}}$' utility
$U^t(\mathcal{F}^t)$	network throughput
w_n, ϖ_n	weights of $n^{\mathcal{I}}$
$\tilde{U}, S^t_n$	average throughput, $n^{\mathcal{I}}$' VSQ
$P^{Total,t}_n$	total energy consumption
$G^t, \mathcal{N}^{\mathcal{L}}$	stochastic game, player set
$\mathcal{S}, C^{tr}_n$	strategy set, forwarding cost
$s_n(o_n), \mathbf{s}_{-n}(o_n)$	mixed strategies of $n^{\mathcal{I}}$' and others
$V_n(s_n(o_n), \mathbf{s}_{-n}(o_n))$	discounted payoff
$s^*_n(o_n), \mathbf{s}^*_{-n}(o_n)$	equilibrium strategies of $n^{\mathcal{I}}$' and others
$V^*_n(s^*_n(o_n), \mathbf{s}^*_{-n}(o_n))$	equilibrium utility
T_{o_n, o'_n}	transition probability
$Q^*_n(o_n, a_n)$	optimal Q-value
$s^{\tau}_n(o_n, a_n)$	probability of $n^{\mathcal{I}}$' choosing action

nodes, while a routing application maximization was developed in [19] depending on the dynamic trust management. However, these studies executed the FA based on the individual node's trust aggregation and excluded the selfish nodes to the network by simply setting a trust threshold, which greatly wasted the potential relays and does not effectively avoid the existence of completely selfish relay. They did not jointly schedule the FA for the source and FSSs for relays in terms of dynamic network parameters. A flexible resource management scheme was built in [20] for Tactile communication in 5G-enabled Tactile Internet. To meet the stringent latency requirements for Tactile Internet capable networks, a predictive resource allocation algorithm including dynamic wavelength and bandwidth allocation was proposed in [21]. Nevertheless, existing works neglected the influence of the dynamic network state on system performance, which may lead to an unacceptable degradation of network stability [22].

Haptic communications will be the prime application running on the Tactile Internet [2]. The relationship between the haptic communications and Tactile Internet will be that of service and medium, respectively (e.g., the relationship between VoIP and the Internet). The functional architecture of the Tactile Internet for remote and real-time human-to-machine interaction, such as haptic communication is shown in Fig. 1. The master domain, which consists of a human and a human system interface, sends the command signals to control the operation of a remote robot (or tele-operator) in a slave domain through the network domain (the forward path). Moreover, the robot in the slave domain will feedback the kinesthetic signals to the master domain via the network domain for the human to learn the remote environment (the reverse path). We express the command signals on the forward path and the kinesthetic signals on the reverse path as Tactile information. The network domain, which is the 5G communication system containing wireless cooperative network in this study, provides a medium for Tactile information delivery between the master and slave domains. Therefore, the human can touch, feel, and manipulate the remote robot in an interactive environment. To extend the transmission convergence and improve the network capacity, we employ cooperative transmission in a 5G communication system and study the cooperative haptic information delivery in WCN (network domain) for real-time control. Table 1 lists notation used in this study. To describe the our system model in detail, we divided it into a number of sub-models and discuss each sub-model separately in following subsections.

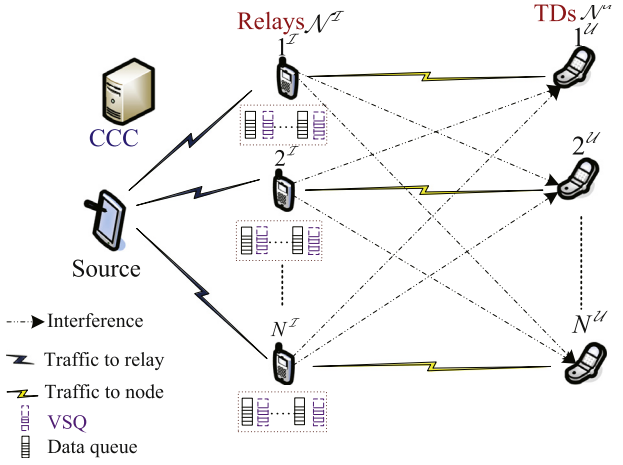


Fig. 2. WCNs with selfish relays.

3.1. Network model

In our network model, the potential selfish relay nodes are represented as a directed graph $\mathcal{G} = \{\mathcal{V}, \mathcal{L}\}$. The set of nodes $\mathcal{V}$ includes the source node s, relay nodes $\mathcal{N}^{\mathcal{I}} = \{1^{\mathcal{I}}, 2^{\mathcal{I}}, \ldots, N^{\mathcal{I}}\}$ within transmission range of source node, and other terminal devices (TDs) $\mathcal{N}^{\mathcal{U}} = \{1^{\mathcal{U}}, 2^{\mathcal{U}}, \ldots, N^{\mathcal{U}}\}$ that are out of the source's transmission range. The set of link $\mathcal{L}$ contains directional links which exists if and only if two nodes are within a each other transmission range.

We assume that the Tactile-based haptic information is delivered in the form of data packets. The source node receives the data packets about the haptic information from the master domain on a forward path (or slave domain on the reverse path) and sends the packets to the network nodes (relays and TDs). When the packets reached at their intended destination in the WCN (a *TD* or a relay), they further forwarded to the slave domain on the forward path (or master domain on the reverse path). Due to the limited delivery range, when the source wants to transmit TDs' packets, it requests the relays to forward the packets to its intended *TDs*. We also assume that each relay can forward the received packets to a specific *TD* and has an infinite buffer for backlogging the packets. There exist N communication pairs that share the same frequency resource as shown in Fig. 2. We further assume that both relay $n^{\mathcal{I}}$ and the *TD* $n^{\mathcal{U}}$ can receive data packets from the source and further send to the next domain. There exists a set of $\mathcal{F}$ active unicast data sessions in the considered network scenario. Let $f_{n^{\mathcal{I}}}$ and $f_{n^{\mathcal{U}}}$ denote the sessions whose source is s and the destination nodes are relay $n^{\mathcal{I}}$ and *TD* $n^{\mathcal{U}}$, respectively.

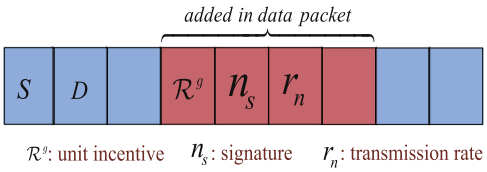


Fig. 3. Structure of data packet.

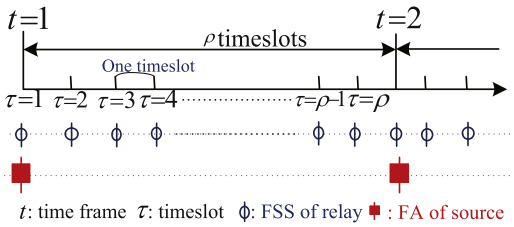


Fig. 4. Joint FA and FSS.

3.2. Incentive model

As the forwarding actions may consume a certain amount of energy; therefore, relay nodes may be unwilling to forward the received packets to save their limited energy for receiving their own packets such a selfish behavior of the relay may greatly reduce the network throughput. An incentive mechanism [23] is exploited to overcome selfishness of relay nodes, in which to stimulate the relays to participate in the cooperative packet forwarding, the source will pay tokens to the relays when it wants to deliver the packets to its intended *TD* via a relay nodes. If the relay node forwards the intended packets to the corresponding *TD* successfully, it will receive the tokens and may use them for receiving data from the source. Tokens that a relay node obtains from the source node as an incentive are related to its contribution to improve the network capacity. Specifically, when the relay $n^{\mathcal{I}}$ successfully participates in the cooperative packet forwarding to the *TD* $n^{\mathcal{U}}$ with a transmission rate of r_n, it will receive $\mathcal{R}^g r_n$ tokens as a reward, where $\mathcal{R}^g$ is the number of received tokens for unit packet. The specific expression of the delivery rate r_n for the communication pair n will be provided in the following subsection. When the relay has no tokens, it loses the chance to receive its own data packets from the source.

In order to enable the nodes to pay and earn tokens, a credit clearance center (CCC) [24] is employed to manage the tokens for the relay. Each network node can use wireless link to communicate with the CCC for the verification and the payment of virtual currency during the rewarding process. Furthermore, each relay holds a digitally signature to the CCC. Then, the relay can register itself to the CCC and know the number of its virtual currency. When the source sends a data packet whose destination is TD $n^{\mathcal{U}}$, there exists a space at the head of the packet which contains unit incentive $\mathcal{R}^g$ based on the packet type. During the transmission process, others can also add some forwarding information onto the packet. Specifically, if relay $n^{\mathcal{I}}$ receives and forwards this data packet, it will add its digitally signature to the packet. After the packet reaches the destination (*TD* $n^{\mathcal{U}}$) via $n^{\mathcal{I}}$, the destination will add link rate r_n to the packet. Then, the destination $n^{\mathcal{U}}$ extracts the forwarding information from the data packet and submits it as a report message to the CCC via wireless link, as shown in Fig. 3. The CCC will calculate and charge rewarding currency $\mathcal{R}^g r_n$ to relay $n^{\mathcal{I}}$ from source's account via a virtual bank, after it has verified the receipt. If relay $n^{\mathcal{I}}$ does not participate in the cooperative transmission, it will get nothing. Note that, to keep track the forwarding packet path truthfully, the currency exchange between the relay and source is virtual, and the currency is not actually sent by the source.

3.3. Joint FA and FSS model

Noting that the selfishness of relay $n^{\mathcal{I}}$ may varies with its residual energy and holding tokens[1] and QSI of $n^{\mathcal{I}}$ which evolves with the allocated rates of *TD* $n^{\mathcal{U}}$. The source aims to optimize the average network throughput by allocating the flow rates regarding the dynamic network state, i.e., the relays' data queues and selfishness. Each relay's objective is to take autonomous forwarding strategies for maximizing its own profit based on the interference. The specific illustration for the joint FA and FSS model is showed in Fig. 4. As there may exist conflicts of interests between the source node and the relay nodes. Moreover, collecting the relevant information of the relays incurs a non-negligible cost to the source; therefore, we consider a time-division system to dynamically coordinate the FA for the source and FSSs for the relays.

The time horizon is divided into time frames where each time frame contains ρ short timeslots indexed by τ. In this time-division system, the FA $\mathbf{F}^t = [\mathbf{F}^t_{\mathcal{I}}, \mathbf{F}^t_{\mathcal{U}}] = [f^t_{n^{\mathcal{I}}}, f^t_{n^{\mathcal{U}}}]_{n\in\mathcal{N}}$ is operated by the source node at the beginning of the time frame according to the dynamic network state. However, at timeslot τ within a specific time frame the forwarding strategy $a_n \in \mathcal{A}_n = \{f, d\}$, is updated by relay $n^{\mathcal{I}}$ to maximize its profit based on the interference, where f and d denote the relay's cooperative forwarding action and dropping action, respectively. When the network state changes and the next time frame comes, a new coordination period for the FA of the source and FSSs for each relay start. The variations in the relay's available resources (energy, tokens, or data queue) are much lower than those in its forwarding strategy, such as the timescale of the relay's energy variation is the hour [25] and the timescale of the forwarding strategy update is the second [26]. Various timescales of these factors lead to the different changing rates for the interference and the network state. Therefore, we assume that the network state is constant during one time frame and may vary over the frame boundaries, while the interference changes at each timeslot.

3.4. Data queuing model

The channel state information (CSI) of the link for communication pair n where $n \in \mathcal{N}$ at timeslot τ is defined as the channel gain h^τ_{nn}. We assume that h^τ_{nn} is independently and identically distributed over different timeslots. Moreover, we use P^T_n to designate the transmission power of relay $n^{\mathcal{I}}$. Then, by considering the interference introduced by the other communication pairs, the received signal-to-interference-plus-noise ratio (SINR) of communication pair n, $\gamma^\tau_n(a_n, \mathbf{a}_{-n})$ at timeslot t is given as follows:

$$\gamma^\tau_n(a_n, \mathbf{a}_{-n}) = \frac{h^\tau_{nn} P^T_n \mathbf{1}^{f,\tau}_n}{\sigma^2 + \sum_{m\in\mathcal{N}\setminus\{n\}} h^\tau_{mn} P^T_m \mathbf{1}^{f,\tau}_m}, \tag{1}$$

where σ^2 is the variance of the additive white Gaussian noise, $\mathbf{a}_{-n} = \{a_1, \ldots, a_{n-1}, a_{n+1}, \ldots, a_N\}$, h^τ_{mn} is the channel gain of the link between relay $m^{\mathcal{I}} \in \mathcal{N}^{\mathcal{I}}$ and node $n^{\mathcal{U}}$. $\sum_{m\in\mathcal{N}\setminus\{n\}} h^\tau_{mn} P^T_m \mathbf{1}^{f,\tau}_m$ denotes the interference which is introduced by others' information deliveries. P^T_m is the transmission power of relay $m^{\mathcal{I}}$ (e.g., $m \neq n$). The term $h^\tau_{nn} P^T_n \mathbf{1}^{f,\tau}_n$ denotes the practical received signal at node $n^{\mathcal{U}}$. The indicator function $\mathbf{1}^{f,\tau}_n$ is defined as follows:

$$\mathbf{1}^{f,\tau}_n = \begin{cases} 1 & \text{if } a_n = f, \\ 0 & \text{if } a_n = d. \end{cases} \tag{2}$$

[1] More details will be discussed in Section 4.

a_n is the forwarding strategy, such that $a_n \in \mathcal{A}_n = \{f, d\}$, and is updated by relay $n^{\mathcal{I}}$ to maximize its profit based on the interference. Without the loss of generality and under the framework of the Shannon formula, the transmission rate for communication pair n at timeslot t is $r_n^\tau = W \log\left(1 + \gamma_n^\tau(a_n, \mathbf{a}_{-n})\right), \forall t > 0, n \in \mathcal{N}$, where W is the bandwidth and $\gamma_n^\tau(a_n, \mathbf{a}_{-n})$ denotes the SINR of communication pair n, at timeslot t. The dynamic of the queue backlog D_n^t for communication link n ($\forall t > 0, n \in \mathcal{N}$) over different time frames denotes the data queuing model used in our study is given as

$$D_n^{t+1} = \max[D_n^t - \sum_{\tau=0}^{\rho-1} r_n^\tau, 0] + f_{n^{\mathcal{I}}}^t \rho, \tag{3}$$

where $\sum_{\tau=0}^{\rho-1} r_n^\tau$ and $f_{n^{\mathcal{I}}}^t \rho$ are the accumulative transmission rate and flow rate for communication pair n within time frame t, respectively. The first term in (3) corresponds to the departure process and the second term corresponds to the arrival process. Due to the time-varying variables (transmission rate r_n^τ and flow rate $f_{n^{\mathcal{I}}}^t$), both the departure and arrival processes are stochastic; therefore, the data queue backlogs change over time.

4. Problem formulation and virtual selfishness queue

In this section we will discuss our problem formulation and virtual selfishness queue procedure.

4.1. Problem formulation

In the WCNs with selfish relays, there naturally exist different objectives for the source and the selfish relays. In other words, the source allocates the flow rates to the network nodes to meet the requirement of a large network throughput for delivering sufficient Tactile information in WCNs, while the selfish relays select the forwarding strategies for maximizing their own profits. To coordinate the FA for the source and FSSs for relays, we consider a time-division system in Section 3. The network state is assumed to be constant during one time frame while the interference changes across different timeslots. Then, the source executes the FA according to the network state for each time frame that contains ρ timeslots, while the selfish relay determines the forwarding strategy based on its utility, which is related to its current interference, at each timeslot. The relay's forwarding strategies for various timeslots within a time frame impact the network state and further the source's FA at the next time frame. The source's FA at the current time frame influences the relays' selfishness and the forwarding strategies at the next time frame. The source's FA and each relay's FSSs affect each other over time via the network state.

4.1.1. FSS problem for selfish relay

Only if the relay forwards the data packet to the corresponding TD successfully, can it receive the tokens. The interference among the relays may influence the selfish relay's packet delivery and its revenue. Specifically, when relay $n^{\mathcal{I}}$ acts as a forwarder and if its SINR $\gamma_n^\tau(a_n, \mathbf{a}_{-n})$ is above the preset threshold γ_n^* at timeslot τ, it will gain the reward and consume its energy resource. If relay $n^{\mathcal{I}}$ acts as a forwarder but its SINR $\gamma_n^\tau(a_n, \mathbf{a}_{-n})$ is below the threshold γ_n^*, it cannot forward the packet successfully, then it receives a null reward but consumes its energy resource. Meanwhile, if the relay acts as a dropper, both its reward and energy consumption are zero. As the token number that the relay obtains is related to the transmission rate of its corresponding communication pair, relay $n^{\mathcal{I}}$'s payoff also depends on the others' forwarding strategies $\mathbf{a}_{-n}$. Then, we use $\mathcal{R}_n^\tau(a_n, \boldsymbol{a}_{-n}, o_n^\tau)$ to denote the payoff of $n^{\mathcal{I}}$ at timeslot τ, where o_n^τ is the current interference state of relay $n^{\mathcal{I}}$ and

$$o_n^\tau = \begin{cases} 1 & \text{if } \gamma_n^\tau(a_n, \mathbf{a}_{-n}) \geq \gamma_n^*, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Considering that the relay is selfish and updates its forwarding strategy to maximize its own payoff at each timeslot, we build the FSS problem for relay $n^{\mathcal{I}}$, $n^{\mathcal{I}} \in \mathcal{N}^{\mathcal{I}}$ as follows.

Problem 1. The FSS problem for relay $n^{\mathcal{I}}$ at timeslot τ is

$$\begin{aligned} &\max_{a_n} \ \mathcal{R}_n^\tau(a_n, \boldsymbol{a}_{-n}, o_n^\tau) && (5) \\ &\text{s.t. } a_n \in \{f, d\}, \forall n^{\mathcal{I}} \in \mathcal{N}^{\mathcal{I}}, && \text{(C1)} \\ &\quad o_n^\tau \in \{0, 1\}, \forall 0 < \tau \leq \rho, n^{\mathcal{I}} \in \mathcal{N}^{\mathcal{I}}, && \text{(C2)} \end{aligned}$$

where (C1) and (C2) are the constraints of the relay's forwarding strategy and interference state, respectively.

Due to the relays' discrete decision variables and severe coupling in the interference among relays, the traditional optimization method [27] cannot be applied to solve Problem 1. Using the advantage of game theory in overcoming the conflicting interests among selfish relays, we employ a stochastic game [28] to model the competition among relays for maximizing their utilities. Moreover, in the stochastic game, we propose a novel learning algorithm for the relay to select its forwarding strategy across the timeslots. The details are presented in Section 5.

4.1.2. FA problem for source node

Source node s controls the flow rates $\mathbf{F}^t$ at time frame t to maximize the network throughput for delivering Tactile information in WCNs. However, the existence of the selfish relays in WCNs always has a negative impact on the network throughput. Especially, when the relay is completely selfish, it denies to forward any data packets, which results in an enormous decrease in network throughput or even in the failure of system operation [6]. Thus, the source is required to reduce the relays' selfishness and prevent complete selfishness. Moreover, network stability is a natural constraint in the dynamic system for remote control [22]. Therefore, the objective of the source node in our study is to maximize the average network throughput while maintaining network stability and preventing the relay node from becoming completely selfish. The function $U^t(\mathbf{F}^t)$ defines the network throughput at time t as follows:

$$U^t(\mathbf{F}^t) = \sum_n \left[\omega_n \log(e + f_{n^{\mathcal{I}}}^t) + \varpi_n \log(e + f_{n^{\mathcal{U}}}^t)\right], \tag{6}$$

where log and e represent the logarithm and the constant, respectively. Additionally, ω_n and ϖ_n denote the weights. In (6), the first term represents the sum of the weighted data traffic for relay $n^{\mathcal{I}}$ where $n \in \mathcal{N}$, while the second term represents the sum of the weighted data traffic for node $n^{\mathcal{U}}$. The term $\log(e+f_{n^{\mathcal{I}}}^t)$ presents the traffic flow rate between the source and relay $n^{\mathcal{I}}$, and $\log(e + f_{n^{\mathcal{U}}}^t)$ reflects the flow rate between the source and the node $n^{\mathcal{U}}$. Next, we construct the FA problem for the source.

Problem 2. In order to maximize the long-term network throughput, the FA problem for the source is

$$\max \ \overline{U} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[U^t(\mathbf{F}^t)] \tag{7}$$

$$\begin{aligned} &\text{s.t. } f_{n^{\mathcal{I}}}^t \geq 0, f_{n^{\mathcal{U}}}^t \geq 0, \forall t \geq 0, n \in \mathcal{N}, && \text{(C3)} \\ &\quad 0 \leq f_{n^{\mathcal{I}}}^t + f_{n^{\mathcal{U}}}^t \leq A_n^t, \forall t \geq 0, n \in \mathcal{N}, && \text{(C4)} \\ &\quad \text{there exit no completely selfish nodes,} && \text{(C5)} \\ &\quad \text{the network is stable,} && \text{(C6)} \end{aligned}$$

where A_n^t is the flow rate budget, (C3) shows that the link rate cannot be negative, (C4) represents the constraint of the maximal rate for each link, (C5) is the constraint of relay's selfishness boundary, and (C6) is the network stability constraint.

It is impossible to directly find a solution to Problem 2 via the optimization method because there exists no comprehensive model to portray the relay's dynamic selfishness in terms of its time-varying factors. Moreover, there is no clear definition of the complete selfishness of the relay. In the following subsection, we employ a novel technique to model the relay's dynamic selfishness and determine the selfishness boundary. Then, Problem 2 can be transformed into a solvable stochastic optimization problem.

4.2. Virtual Selfishness Queue (VSQ)

During data transmission, the relay's selfishness is influenced by both the holding token and the residual energy of the relay. A relay with small battery power and large number of tokens may use all the available energy and then show no willingness for further data transmission. However, if a relay has sufficiently large battery power but very few tokens, then it will tend to help deliver data for earning tokens to pay for data acquisition. The relay's residual energy and holding tokens may vary with its receiving and forwarding actions; especially when the energy consumption of the relay includes traffic reception and transmission. Because relay $n^{\mathcal{I}}$ uses power P_n^T to achieve its transmission goal when it decides to forward the packets, the energy consumption of relay $n^{\mathcal{I}}$ for forwarding its received data traffic at timeslot τ is $P_n^T \mathbf{1}_n^{f,\tau}$. Inspired by [29], the energy consumption of its own data reception is $P^R f_{n^{\mathcal{I}}}^t$ and the energy consumption of data reception for node $n^{\mathcal{U}}$ is $P^R f_{n^{\mathcal{U}}}^t$, where P^R denotes the energy consumption of the unit data reception. Thus, the total energy consumption $P_n^{Total,t}$ of relay $n^{\mathcal{I}}$ at time frame t is as follows:

$$P_n^{Total,t} = P^R \rho (f_{n^{\mathcal{I}}}^t + f_{n^{\mathcal{U}}}^t) + \sum_{\tau=0}^{\rho-1} P^T \mathbf{1}_n^{f,\tau}. \tag{8}$$

The first term and the second term on the right hand side of (8) are the cumulative energy consumption of relay $n^{\mathcal{I}}$ for receiving and forwarding the data packets within time frame t, respectively. Moreover, based on the proposed incentive mechanism in the system model, each relay earns tokens by forwarding received packets and spends tokens to receive its own data traffic. Specifically, if the relay forwards the data packet to its corresponding TD successfully ($o_n^\tau = 1$) with transmission rate r_n^τ, it will receive $\mathcal{R}^g r_n^\tau o_n^\tau$ tokens at timeslot τ. Additionally, when the relay buys its own data traffic with a rate of $f_{n^{\mathcal{I}}}^t$, it will consume $\mathcal{R}^p f_{n^{\mathcal{I}}}^t$ tokens, where $\mathcal{R}^p$ is the spent tokens of the relay for unit data packet. A decrease in the number of holding tokens will decrease the selfishness of relay $n^{\mathcal{I}}$ and encourage $n^{\mathcal{I}}$ to forward more data packets. Meanwhile, the depletion of residual energy or an increase in the number of holding tokens will increase the selfishness of relay $n^{\mathcal{I}}$. Motivated by [30] that used a virtual energy queue to limit power consumption while at the same time sustaining the throughput performance, our objective is to construct a virtual queue, called virtual selfishness queue (VSQ), to model the relays' selfishness dynamics over time. In this study, we regard the decrease in relay's selfishness as the departure of VSQ and the increase in selfishness as the arrival of VSQ. Then, the backlog of VSQ S_n^{t+1} can be formulated as

$$S_n^{t+1} = \max[S_n^t - \mathcal{R}^p f_{n^{\mathcal{I}}}^t \rho, 0] + P_n^{Total,t} + \mathcal{R}^g \sum_{\tau=0}^{\rho-1} r_n^\tau o_n^\tau, \tag{9}$$

where $S_n^{t+1} \in [0, \infty)$ and o_n^τ is the interference state. $\mathcal{R}^g \sum_{\tau=0}^{\tau=\rho-1} r_n^\tau o_n^\tau$ and $\mathcal{R}^p f_{n^{\mathcal{I}}}^t \rho$ are the cumulative tokens that the relay earns and spends during time frame t, respectively. Similarly, because of the time-varying variables, the VSQ value changes over time. Moreover, based on the built VSQ in (9), if its value S_n^t becomes larger, relay $n^{\mathcal{I}}$ is more reluctant to forward the data packets of TD $n^{\mathcal{U}}$ and its selfishness is higher. Specifically, when its VSQ value approaches to infinity, i.e., $S_n^t \to \infty$, relay $n^{\mathcal{I}}$ is completely selfish and will refuse any forwarding service requests about TD $n^{\mathcal{U}}$. Additionally, $S_n^t = 0$ means that relay $n^{\mathcal{I}}$ is altruistic and may forward the received data packets.

4.3. Stochastic optimization problem for FA

In this subsection, we consider a non-cooperative system, in which each selfish relay behave as a learning agent that adjusts its forwarding strategy, i.e. whether forwards the data packets or not (f or d), over the timeslots based on its profit by assuming that each relay's VSQ and QSI keep constant. We employed the VSQ to model the dynamics of the relay's selfishness in terms of its time-varying factors. A larger VSQ value indicates higher selfishness of the relay. Therefore, we can use the bounded selfishness (VSQ value) to avoid complete selfishness of the relay. The following definition is to determine the network stability and relay's selfishness boundary.

Theorem 1. *A discrete time queue is called mean rate stable if [31]*

$$\lim_{t\to\infty} \frac{1}{t} \mathbb{E}[\mathcal{D}_n^t] = 0, \tag{10}$$

where $\mathcal{D}_n^t$ denotes a specific queue.

To ensure system stability and further limit the delivery latency [10] for Tactile information delivery, the data queues for all relays in the network should be mean rate stable [21]. Then, (10) can be used to determine whether the network is stable or not. Here, we also employ (10) to bound the relay's dynamic selfishness in the long term. If the relay's VSQ is mean rate stable, its selfishness is bounded. Then, **Problem 2** is translated as the following optimization problem.

Problem 3. The FA problem for the source is

$$\max\ \overline{U} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[U^t(\mathbf{F}^t)] \tag{11}$$

s.t. (C3) and (C4)

$$\lim_{t\to\infty} \frac{1}{t} \mathbb{E}[D_n^t] = 0, \forall t \geq 0, n \in \mathcal{N}, \tag{C7}$$

$$\lim_{t\to\infty} \frac{1}{t} \mathbb{E}[S_n^t] = 0, \forall t \geq 0, n \in \mathcal{N}, \tag{C8}$$

where (C7) is the data queue stability constraint and (C8) is the selfishness boundary constraint to prevent the relay from being completely selfish.

Problem 3 can be understood from the stochastic programming perspective. The objective is to design an algorithm that enables the data rates $\mathbf{F}^t$ passing through the source to satisfy all constraints and maximize the network utility as much as possible. We introduce our proposed dynamic FA algorithm in Section 6. The proposed algorithm achieves the optimal solution of Problem 3 in terms of average network throughput.

Remark 1. The source allocates the flow rates to the network nodes at the beginning of each time frame based on the relays' queue information including data queues and VSQs (Section 6). Meanwhile, each relay updates its forwarding strategy at each timeslot based on its own profit that is also related to its queue information (Section 5). The allocated flow rate and a series of forwarding strategies for the relay will cause variations in its queue

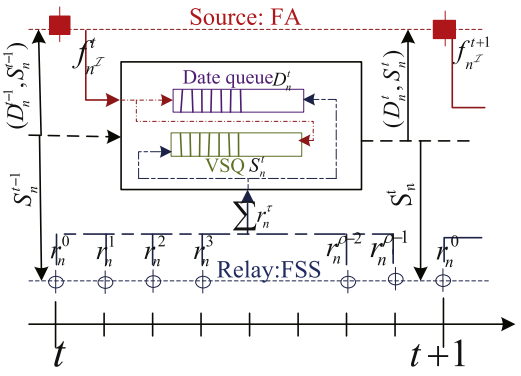


Fig. 5. Dynamics of queue information, FA and FSS.

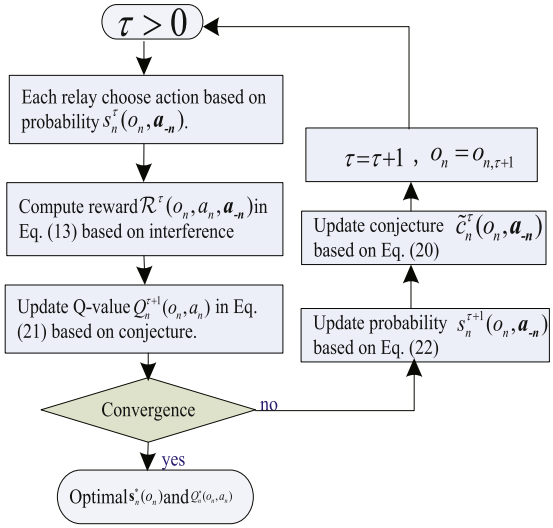


Fig. 6. Flow chart of **Algorithm 1**.

information including QSI and VSQ in the next time frame ((3) and (9)). Then, the FA for the source and FSSs for each relay will be executed depending on the new queue values, which influence each other, and the co-evolution across the time frames (Fig. 5).

5. Stochastic game among relays with different resource states

In this section, we consider a non-cooperative system, in which each selfish relay behaves as a learning agent that adjusts its forwarding strategy across the timeslots based on its profit, by assuming that each relay's VSQ and QSI are constant. Based on the relays' dynamic interference state in various timeslots, we formalize the relays' sequential forwarding decisions (Problem 1) within a specific time frame t through non-cooperative game playing because the proposed incentives engender a competition among relays that play strategies on their activation.

5.1. Stochastic game

Because of the complicated wireless environment, the relay does not know other relays' payoffs and possible actions. Therefore, we build a stochastic game to explore the relay's mixed forwarding strategy, i.e., the probability of the relay selecting f action to forward the received packets at a specific timeslot by considering the relay's selfishness. Cost parameters, payoff functions, and forwarding strategies are private knowledge among the game players in this study.

Let $G^t = [\mathcal{N}^{\mathcal{L}}, \mathcal{S}, \mathcal{R}(\cdot)]$ denote the stochastic game where $\mathcal{N}^{\mathcal{L}}$ is the index set for relays, $\mathcal{S}$ is the mixed strategy set, and $\mathcal{R}(\cdot)$ is the utility function of the player. The utility of each relay depends on its own behavior and also on the choice of other relays' actions. Formally, stochastic game G^t is expressed as

- *Players:* The players are relays $\mathcal{N}^{\mathcal{L}}$ in the network.
- *Strategies:* The strategy of a player is the probability of the relay selecting f action. Then, the stochastic mixed strategy for relay $n^{\mathcal{I}}$ is s_n where $s_n \in \mathcal{S}$.
- *Payoffs:* Player's payoff is determined by multiple factors, including the cost of forwarding the packet and the reward obtained by forwarding or dropping the packet.

When the relay acts as a forwarder and its SINR γ_n^τ is above the threshold γ_n^*, it will obtain $\mathcal{R}^g r_n^\tau$ tokens at timeslot τ. Moreover, the relays with different VSQ values will react differently to the consumption of their energy resources. For instance, the node with higher selfishness (more tokens and shorter residual energy) will appreciate its residual energy more and its forwarding cost will be relatively higher than that of the others. We define the forwarding cost of relay $n^{\mathcal{I}}$ considering its current VSQ as the relative cost:

$$C_n^{tr} = P_n^T S_n^t. \tag{12}$$

Then, by normalizing the costs to this expected benefit, we obtain the payoff of relay $n^{\mathcal{I}}$ at timeslot τ as

$$\mathcal{R}_n^\tau(a_n, \boldsymbol{a}_{-n}, o_n^\tau) = \begin{cases} \mathcal{R}^g r_n^\tau - C_n^{tr} & \text{if } a_n = f \text{ and } o_n^\tau = 1, \\ -C_n^{tr} & \text{if } a_n = f \text{ and } o_n^\tau = 0, \\ 0 & \text{if } a_n = d, \end{cases} \tag{13}$$

where o_n^τ is the interference state defined in Section 4 and $a_n = f$ means that relay $n^{\mathcal{I}}$ is willing to forward the packet at timeslot τ. The interference state transition from o_n^τ to $o_n^{\tau+1}$ for relay $n^{\mathcal{I}}$ across the timeslots is also determined by the stochastic mixed strategies of other relays. In the non-cooperative game, each relay chooses the mixed strategy independently to maximize its total expected discounted payoff. Then, we define the total expected discounted reward of relay $n^{\mathcal{I}}$ within the time frame t (ρ timeslots and $\rho \gg 1$) as

$$V_n(s_n(o_n), \boldsymbol{s}_{-n}(o_n)) = \mathbb{E}\left[\sum_{\tau=0}^{\rho-1} \beta^\tau \mathcal{R}_n(s_n^\tau(o_n), \boldsymbol{s}_{-n}^\tau(o_n)) \mid o_n^0 = o_n\right], \tag{14}$$

where $\beta^\tau \in [0, 1)$ is the discount factor, o_n is the interchangeable expression of o_n^τ, $s_n^\tau(o_n)$ is the mixed strategy of relay $n^{\mathcal{L}}$ at state o_n, $\boldsymbol{s}_{-n}^\tau(o_n) = (s_1^\tau(o_1), \ldots, s_{n-1}^\tau(o_{n-1}), s_{n+1}^\tau(o_{n+1}), \ldots, s_N^\tau(o_N))$ and $\mathcal{R}_n(s_n^\tau(o_n), \boldsymbol{s}_{-n}^\tau(o_n)) = \mathbb{E}[\mathcal{R}_n^\tau(o_n, a_n, \boldsymbol{a}_{-n})]$. In the stochastic game, each relay behaves as a learning agent whose task is to find the optimal mixed forwarding strategy $s_n^*(o_n)$ for each interference state o_n where $o_n \in \mathcal{O}$, within time frame t.

Theorem 2. *A tuple of N strategies $\{s_n^*(o_n), \boldsymbol{s}_{-n}^*(o_n)\}$ is a Nash equilibrium within time frame t, if a relay $n^{\mathcal{I}}$*

$$V_n^*(s_n^*(o_n), \boldsymbol{s}_{-n}^*(o_n)) \geq V_n(s_n(o_n), \boldsymbol{s}_{-n}^*(o_n)), \tag{15}$$

$\forall s_n^*(o_n) \in \mathcal{S}$, *for each state o_n.*

Every strategic-form game has a mixed strategy equilibrium [32], i.e., there always exists an Nash equilibrium in our game formulation of stochastic strategy adaptation. The equilibrium strategy satisfies the Bellman's optimal equation.

$$V_n(s_n^*(o_n), \boldsymbol{s}_{-n}^*(o_n)) = \max_{a_n \in \mathcal{A}_n} \mathbb{E}[\mathcal{R}_n(a_n, \boldsymbol{s}_{-n}^*(o_n))] + \beta \sum_{o' \in \mathcal{O}} T_{o_n, o'_n}(a_n, \boldsymbol{s}_{-n}^*(o_n)) V_n(s_n^*(o'_n), \boldsymbol{s}_{-n}^*(o'_n)), \tag{16}$$

where T_{o_n,o'_n} is the state transition probability, $\mathcal{A}_{-n} = \{\mathcal{A}_1, \ldots, \mathcal{A}_{n-1}, \mathcal{A}_{n+1}, \ldots, \mathcal{A}_N\}$ and $\mathbb{E}[\mathcal{R}_n(a_n, \boldsymbol{s}^*_{-n}(o_n))] = \sum_{\boldsymbol{a}^*_{-n}\in\mathcal{A}_{-n}}[\mathcal{R}^\tau_n(o_n, a_n, \boldsymbol{a}^*_{-n}) \prod_{m\in\mathcal{N}/\{n\}} s^*_m(o_n)]$.

5.2. Solution techniques for stochastic game

In this subsection, we consider the Q-learning algorithm as the benchmark to solve the stochastic game because it needs no prior knowledge about the interference state transition probabilities $T_{o_n,o'_n}(\cdot)$. We define the optimal Q-value Q^*_n of relay $n^{\mathcal{I}}$ as the current expected reward plus its future expected discounted rewards when all relays follow the Nash equilibrium strategies,

$$Q^*_n(o_n, a_n) = \mathbb{E}[\mathcal{R}_n(a_n, \boldsymbol{s}^*_{-n}(o_n))] + \beta \sum_{o'_n\in\mathcal{O}} T_{o_n,o'_n}(a_n, \boldsymbol{s}^*_{-n}(o_n)) V_n(s^*_n(o'_n), \boldsymbol{s}^*_{-n}(o'_n)). \quad (17)$$

Then, we have

$$Q^*_n(o_n, a_n) = \mathbb{E}[\mathcal{R}_n(a_n, \boldsymbol{s}^*_{-n}(o_n))] + \beta \sum_{o'_n\in\mathcal{O}} T_{o_n,o'_n}(a_n, \boldsymbol{s}^*_{-n}(o_n)) \max_{b_n\in\mathcal{A}_n} V_n(b^*_n(o'_n)). \quad (18)$$

The combined Q-learning process attempts to find the optimal Q-value $Q^*_n(o_n, a_n)$ in a recursive way using the relays' information set $\mathbb{I} = \{a_n, o_n, o'_n, s^\tau_n | \forall n \in \mathcal{N}\}$, where $o_n(= o^\tau_n)$ and $o'_n(= o^{\tau+1}_n)$ are the interference states observed by relay $n^{\mathcal{I}}$ at timeslot τ and $\tau+1$, respectively. Additionally, a_n is the forwarding action of relay $n^{\mathcal{I}}$ taken according to the current forwarding strategy policy s^τ_n at timeslot τ. Then, we derive the combined Q-learning updating rule (19) which is given in Box I. where $\alpha^\tau \in [0, 1)$ is the learning rate.

According to (19), to learn the optimal strategy, relay $n^{\mathcal{I}}$ needs to know not only its own strategy, but also other relays' interference states and available forwarding strategies, which are their private information. It is difficult for the relays to obtain the exact private information of their opponents in the non-cooperative scenario; the relay can only obtain its own local information, such as the interference state, forwarding strategy, and historical rewards. To avoid obtaining others' private strategy information, the relay $n^{\mathcal{I}}$ estimates how its competitors' strategic decisions vary in response to its own actions. Based on [33], the estimated probability that all other relays choose the action set $\boldsymbol{a}_{-n}$ at timeslot τ is

$$\tilde{c}^\tau_n(o_n, \boldsymbol{a}_{-n}) = \tilde{c}^{\tau-1}_n(o_n, \boldsymbol{a}_{-n}) - w^{o_n,\boldsymbol{a}_{-n}}_n [s^\tau_n(o_n, a_n) - \tilde{s}^{\tau-1}_n(o_n, a_n)], \quad (20)$$

for $n \in \mathcal{N}$. For each relay, any change in their current strategy will encourage other competing relays to make well-defined changes in the next timeslot. By incorporating (19) and (20), the combined Q-learning algorithm rule is modified as (21) which is given in Box II. An alternative solution is to vary the action probabilities as a graded function of the Q-value. The most common method is to use a Boltzmann distribution [34]. Relay $n^{\mathcal{I}}$ chooses action $a_n \in \{f, d\}$ in state o_n at timeslot τ with probability

$$s^\tau_n(o_n, a_n) = \frac{e^{Q^\tau_n(o_n,a_n)/\nu}}{\sum_{b_n\in\mathcal{A}_n} e^{Q^\tau_n(o_n,b_n)/\nu}}, \quad (22)$$

where ν is a positive parameter and $e^{Q^\tau_n(o_n,a_n)/\nu}$ is related to the Q-value for a specific action a_n of relay $n^{\mathcal{I}}$. Moreover, $\sum_{b_n\in\mathcal{A}_n} e^{Q^\tau_n(o_n,b_n)/\nu}$ denotes the sum of Q-values for relay's all potential actions. The entire combined Q-learning algorithm is summarized in **Algorithm 1**. The flowchart of **Algorithm 1** is shown in Fig. 6, which is also the running example of solution to stochastic game.

Algorithm 1 Combined Q-learning algorithm

1: Given the timeslot index $\tau = 0$,
2: **for** each node $n^{\mathcal{L}} \in \mathcal{N}^{\mathcal{L}}$, $a_n \in \mathcal{A}$, $o_n \in \mathcal{O}$ **do**
3: Initialize strategy $s^t_n(o'_n, a_n)$ and $w^{o_n,\boldsymbol{a}_{-n}}_n > 0$;
4: **end for**
5: Evaluate the initial state $o_n = o^t_n$.
6: **for** $\tau > 0$ **do**
7: Choose action a_i according to the forwarding strategy $s^\tau_n(o_n, a_n)$;
8: Measure the received SINR with the feedback information of intended receiver;
9: Observe the interference state $o'_n = o^{\tau+1}_n$ by identifying the transmission strategy and comparing γ_n with the threshold γ^*_n;
10: A reward $\mathcal{R}^\tau(o_n, a_n, \boldsymbol{a}_{-n})$ can be achieved;
11: Update $Q^{\tau+1}_n(o_n, a_n)$-value based on $\tilde{c}^\tau_n(o_n, \boldsymbol{a}_{-n})$ according to (18);
12: Update strategy $s^{\tau+1}_n(o_n, \boldsymbol{a}_{-n})$ according to (22);
13: Update conjecture $c^{\tau+1}_n(o_n, \boldsymbol{a}_{-n})$ according to (20);
14: $o_n = o_{n,\tau+1}$, $\tau = \tau + 1$;
15: **end for**

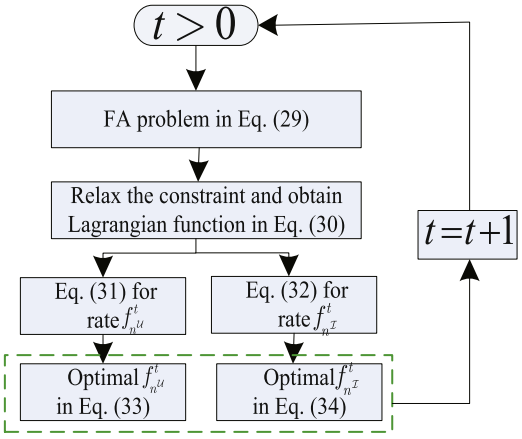


Fig. 7. Flowchart of **Algorithm 2**.

5.3. Convergence of the combined Q-learning Algorithm

In this section, we focus on the convergence of the combined Q-learning algorithm. By establishing the convergence of a general Q-learning process updated by a pseudo contraction operator [35], we can prove the convergence of the learning algorithm. Additionally, we use $\mathbf{Q}$ to denote the space of all Q-values.

Theorem 3. *Regardless of any initial values chosen for $Q^0_n(s^o_n, a_n)$, if τ is sufficiently large, the iteration that is defined by*

$$Q^{\tau+1} = (1 - \alpha^\tau)Q^\tau + \alpha^\tau(\mathcal{H}^\tau Q^\tau), \quad (23)$$

converges to Q^* w.p.1. and the algorithm converges, where $\mathcal{H}^\tau$ is the mapping function.

Proof. See Appendix A. □

6. Dynamic FA for source

To solve Problem 3, we propose a dynamic optimization method to design the FA framework for the source. The challenge behind Problem 3 is that we determine an FA decision $\mathbf{F}^t$ at time frame t for

$$Q_n^{\tau+1}(o_n, a_n) = (1-\alpha^\tau)Q_n^\tau(o_n, a_n) + \alpha^\tau \left\{ \mathbb{E}[\mathcal{R}_n^\tau(o_n, a_n, \boldsymbol{a}_{-n})] + \beta \max_{b_n \in \mathcal{A}_n} Q_n^\tau(o'_n, b_n) \right\}$$
$$= (1-\alpha^\tau)Q_n^\tau(o_n, a_n) + \alpha^\tau \left\{ \sum_{\boldsymbol{a}_{-n} \in \mathcal{A}_{-n}} \left[\prod_{m \in \mathcal{N}\setminus\{i\}} s^\tau(o_m, a_m) \mathcal{R}_i^\tau(o_n, a_n, \boldsymbol{a}_{-n}(o_n)) \right] + \beta \max_{b_n \in \mathcal{A}_n} Q_n^\tau(o'_n, b_n) \right\}, \quad (19)$$

Box I.

$$Q_n^{\tau+1}(o_n, a_n) = (1-\alpha^\tau)Q_n^\tau(o_n, a_n) + \alpha^\tau \left\{ \sum_{\boldsymbol{a}_{-n} \in \mathcal{A}_{-n}} \tilde{c}_n^\tau(o_n, \boldsymbol{a}_{-n}) \mathcal{R}_i^\tau(o_n, a_n, \boldsymbol{a}_{-n}(o_n)) + \beta \max_{b_n \in \mathcal{A}_n} Q_n^\tau(o'_n, b_n) \right\} \quad (21)$$

Box II.

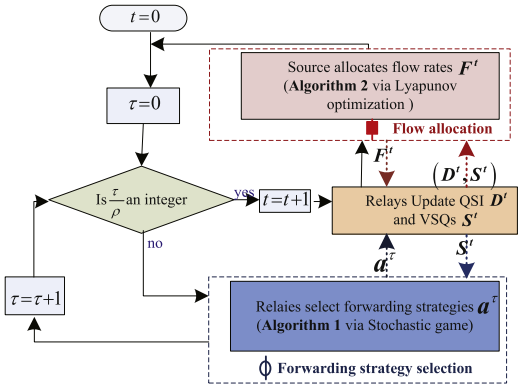


Fig. 8. Interrelationship between the stochastic game and the Lyapunov framework.

stabilizing the queues (both data queues and VSQs) while also maximizing the average network throughput. By controlling the arrival and departure processes of the queues appropriately via Lyapunov *drift-plus-penalty* method [36], the FA decisions can stabilize the queues, whilst maximizing the average network throughput.

Let $\boldsymbol{\Theta}^t$ denote the matrix containing the queues $\{D_n^t, S_n^t | n \in \mathcal{N}\}$. We define the quadratic Lyapunov function at time frame t as

$$L(\boldsymbol{\Theta}^t) = \frac{1}{2}\sum_n [D_n^t]^2 + \sum_n [S_n^t]^2. \quad (24)$$

The conditional expected Lyapunov drift at time frame t is defined as follows:

$$\Delta(\boldsymbol{\Theta}^t) :\triangleq \mathbb{E}[L(\boldsymbol{\Theta}^{t+1})|\boldsymbol{\Theta}^t] - \mathbb{E}[L(\boldsymbol{\Theta}^t)], \quad (25)$$

where the expectation is taken over by the randomness of departure and arrival processes of the queues. Following the Lyapunov optimization framework, we add the penalty term $-V\mathbb{E}[U^t(\mathbf{F}^t)|\Theta^t]$ to (25) to obtain the following *drift-plus-penalty* term,

$$\Delta_V(\boldsymbol{\Theta}^t) = \Delta(\boldsymbol{\Theta}^t) - V\mathbb{E}[U^t(\mathbf{F}^t)|\boldsymbol{\Theta}^t]. \quad (26)$$

Here, $V > 0$ is a control parameter. Through minimizing *drift-plus-penalty* term at each time frame, we can limit the increases of relay's selfishness and data queues, and also improve the network throughout. Then, the objective of FA problem for the source is further achieved. Then, we have the following theorem regarding the *drift-plus-penalty* term.

Theorem 4. *For any feasible FA decision that can be implemented at time frame t, we define (27) (given in Box III), where $\mathbf{R}^t$ denotes the matrix $\{\sum_{\tau=0}^{\rho-1} r_n^\tau | \forall n \in \mathcal{N}\}$ and B is an upper bound on the term $\frac{1}{2}[(\mathbf{R}^t)^H(\mathbf{R}^t) + (\mathbf{F}_{\mathcal{U}}^t)^H(\mathbf{F}_{\mathcal{U}}^t)] + \frac{1}{2}[(\mathcal{R}^p)^2(\mathbf{F}_{\mathcal{I}^t})^H\mathbf{F}_{\mathcal{I}}^t + (P^T)^2(\mathbf{I}^t)^H\mathbf{I}^t + (\mathcal{R}^g)^2(\mathbf{R}^t)^H\mathbf{R}^t + (P^R)^2(\mathbf{F}_{\mathcal{U}}^t)^H\mathbf{F}_{\mathcal{U}}^t + (P^R)^2(\mathbf{F}_{\mathcal{I}}^t)^H\mathbf{F}_{\mathcal{I}}^t]$, which holds under the fact that the flow rates satisfy the properties of boundedness.*

Proof. See Appendix B. □

Our dynamic rate allocation policy is designed to observe the data queues $\mathbf{D}^t = \{D_n^t | n \in \mathcal{N}\}$ and VSQ $\mathbf{S}^t = \{S_n^t, n \in \mathcal{N}\}$, as well as to make a flow rate allocation decision $\mathbf{F}^t$ for minimizing the right-hand-side (RHS) of (27) for the current time. The non-constant part of the RHS of (27) can be written as (28) (given in Box IV), where $\chi^t = \sum_n \sum_{\tau=0}^{\rho-1}[D_n^t r_n^\tau + S_n^t \mathcal{R}^g r_n^\tau o_n^\tau - T_n^t P^T \mathbf{1}_n^{I,\tau}]$ is a constant for time frame t. Next, Problem 3 is translated into a series of optimization problems for each time frame. Then, the FA problem for the source to relays at time frame t is

$$\max \sum_n [V w_n \log(f_{n^{\mathcal{I}}}^t + e) - S_n^t P^R f_{n^{\mathcal{I}}}^t + \mathcal{R}^p S_n^t f_{n^{\mathcal{I}}}^t$$
$$V\varpi_n \log(f_{n^{\mathcal{U}}}^t + e) - D_n^t f_{n^{\mathcal{U}}}^t - S_n^t P^R f_{n^{\mathcal{U}}}^t] \quad (29)$$
$$\text{s.t. } f_{n^{\mathcal{I}}}^t + f_{n^{\mathcal{U}}}^t \le A_n^t, f_{n^{\mathcal{I}}}^t \ge 0, f_{n^{\mathcal{U}}}^t \ge 0.$$

(29) has a strictly concave function that can be decomposed by the dual decomposition method [27]. Relaxing the constraint by introducing the Lagrangian multiplier γ_n^t associated with $f_{n^{\mathcal{I}}}^t + f_{n^{\mathcal{U}}}^t \le A_n^t$ for the source's FA to relay $n^{\mathcal{I}}$ at time frame t, the Lagrangian is formed as

$$L(\mathbf{F}^t) = \sum_n [V w_n \log(f_{n^{\mathcal{I}}}^t + e) - S_n^t P^R f_{n^{\mathcal{I}}}^t + \mathcal{R}^p S_n^t f_{n^{\mathcal{I}}}^t \quad (30)$$
$$+V\varpi_n \log(f_{n^{\mathcal{U}}}^t + e) - D_n^t f_{n^{\mathcal{U}}}^t - S_n^t P^R f_{n^{\mathcal{U}}}^t - \gamma_n^t (f_{n^{\mathcal{U}}}^t + f_{n^{\mathcal{I}}}^t - A_n^t)].$$

There is no coupling in the term $\sum_n$, suggesting that the source allocates the flow rates to different relays independently. Thus, the optimal flow rate $f_{n^{\mathcal{U}}}^t$ is obtained by solving

$$\max_{0 \le f_{n^{\mathcal{U}}}^t \le A_n^t} V\varpi_n \log(f_{n^{\mathcal{U}}}^t + e) - D_n^t f_{n^{\mathcal{U}}}^t - S_n^t P^R f_{n^{\mathcal{U}}}^t - \gamma_n^t f_{n^{\mathcal{U}}}^t. \quad (31)$$

Similarly, the optimal flow rate $f_{n^{\mathcal{I}}}^t$ allocated to relay $n^{\mathcal{I}}$ is obtained by solving

$$\max_{0 \le f_{n^{\mathcal{I}}}^t \le A_n^t} V w_n \log(f_{n^{\mathcal{I}}}^t + e) - S_n^t P^R f_{n^{\mathcal{I}}}^t + \mathcal{R}^p S_n^t f_{n^{\mathcal{I}}}^t - \gamma_n^t f_{n^{\mathcal{I}}}^t. \quad (32)$$

(31) and (32) are the concave optimization problems, which can be solved efficiently by the gradient descent method [27]. Given

$$\Delta_V(\boldsymbol{\Theta}^t) \le B - V\mathbb{E}[U^t(\mathbf{F}^t)|\boldsymbol{\Theta}^t] + \mathbb{E}[(\mathbf{F}^t_{\mathcal{U}} - \mathbf{R}^t)^H\mathbf{D}^t|\boldsymbol{\Theta}^t] + \mathbb{E}[(\mathbf{I}^t P^T - \mathcal{R}^p\mathbf{F}^t_{\mathcal{I}} + \mathcal{R}^g\mathbf{R}^t + \mathbf{F}^t_{\mathcal{U}}P^R + \mathbf{F}^t_{\mathcal{I}}P^R)^H\mathbf{S}^t|\boldsymbol{\Theta}^t], \tag{27}$$

Box III.

$$\begin{aligned} &VU^t(\mathbf{F}^t) - (\mathbf{F}^t_{\mathcal{U}})^H\mathbf{D}^t + (\mathbf{R}^t)^H\mathbf{D}^t - (\mathbf{I}^t)^H P^T\mathbf{S}^t + \mathcal{R}^p\mathbf{F}^t_{\mathcal{I}}\mathbf{S}^t - \mathbf{F}^t_{\mathcal{U}}P^R\mathbf{S}^t - \mathbf{F}^t_{\mathcal{I}}P^R\mathbf{S}^t - \mathcal{R}^g\mathbf{R}^t\mathbf{S}^t \\ &= V\sum_n w_n\log(f^t_{n^{\mathcal{I}}} + e) + V\sum_n \varpi_n\log(f^t_{n^{\mathcal{U}}} + e) - \sum_n D^t_n f^t_{n^{\mathcal{U}}} - \sum_n S^t_n P^R f^t_{n^{\mathcal{U}}} - \sum_n S^t_n P^R f^t_{n^{\mathcal{I}}} + \mathcal{R}^p\sum_n S^t_n f^t_{n^{\mathcal{I}}} + \chi^t. \end{aligned} \tag{28}$$

Box IV.

Algorithm 2 Dynamic FA for the source

1: At each time frame t, the source observes the relays' data queue states $\mathbf{D}^t$ and VSQ states $\mathbf{S}^t$.
2: The source obtains FA $\mathbf{F}^t$ by solving (31) and (32) via the gradient descent method.
3: The relays update $\mathbf{D}^t$ and $\mathbf{S}^t$ according to (3) and (9).
4: $t = t + 1$.

an optimal γ^t_n, the rate $f^t_{n^{\mathcal{U}}}$ allocated to node $n^{\mathcal{U}}$ can be calculated by applying the KarushKuhn–Tucker (KKT) method, which results in

$$f^t_{n^{\mathcal{U}}} = \left[\frac{Vw_n}{(\gamma^t_n + D^t_n + P^R S^t_n)} - e\right]^{[0,A^t_n]}, \tag{33}$$

where $[x]^{[0,A^t_n]}$ denotes the projection of x onto $[0, A^t_n]$. The flow rate $f^t_{n^{\mathcal{I}}}$ is

$$f^t_{n^{\mathcal{I}}} = \left[\frac{V\varpi_n}{(\gamma^t_n + S^t_n P^R - S^t_n\mathcal{R}^p)} - e\right]^{[0,A^t_n]}. \tag{34}$$

Then, the dynamic FA for source is given by **Algorithm 2**. The flowchart of **Algorithm 2** is shown in Fig. 7, which is also the running example of FA problem at each time frame.

Remark 2. The FA in **Algorithm 2** is performed in a centralized way, which may bring communication and computing overheads to the source. The utility function for each relay is a concave function of the flow rates. Distributed algorithms can therefore be found using dual decomposition methods. Choosing either a centralized way or a distributed way depends on the practical communication scene and the network node's properties. If the autonomous relay is willing to compute and submit the optimal flow rates for the source, the distributed way will be better for the scheme. Otherwise, the scheme can only use a centralized way.

7. Performance analysis

In this section, we will analyze the performance of the proposed scheme.

7.1. Summary of the overall solution

Fig. 8 summarizes the overall solution and the interrelationship of the stochastic game and the Lyapunov framework. The decision of FA is made by the source at each time frame based on the QSI and VSQ information of the relays while the decision of forwarding strategy is made by each relay at every timeslot depending on its profit (Fig. 8). Specifically, at the beginning of each time frame, we first employ the dynamic FA algorithm (**Algorithm 2**) for the source to determine the flow rates of network nodes such that the average network throughput is maximized under the constraints of network stability and selfishness boundary. Then, for ρ timeslots within a specific time frame, we employ the combined Q-learning algorithm (**Algorithm 1**) for the relay to select the forwarding strategy to maximize its own profit at each timeslot. By weaving the concepts from Lyapunov optimization and stochastic game, both the source and relays can achieve their objectives in an orderly manner.

7.2. Performance achieved in our scheme

For the dynamic FA of the source, the gap between the average network throughput under our algorithm and the optimal one[2] is bounded by the term $\frac{B}{V}$ based on the property of the Lyapunov optimization framework [37]. Similarly, the average data queue length of the network in this study employing Lyapunov optimization is also related to parameter V. Based on Little's theorem, the average data queue length can represent the information delivery delay [10]. Then, by setting an appropriate V, the average network throughput can be close to the optimal one and the relatively low delivery delay can also be achieved. Thus, our scheme can meet the requirements of the sufficient capacity and low latency for real-time control in WCNs. Regarding the FSS of each relay, aided by the incentive mechanism, our scheme can effectively manage the dynamic selfishness of relays and reduce the selfishness' negative influence on the network throughput. The autonomous relay will actively participate in the FSS based on **Algorithm 1** for maximizing its utility. As the relay's reward depends on the transmission rate of its communication pair (the incentive mechanism proposed), the instantaneous network throughput can also be maintained. In our scheme, the relay with the lower VSQ value (more residual energy and fewer holding tokens) will be allocated more data traffic of the TD ((33) for FA in Section 6) and then undertake more forwarding tasks owing to its lower forwarding cost ((13) for FSS in Section 5). The relay with the higher VSQ value (less energy resource and sufficient tokens) will be allocated less data traffic of TD and forward a smaller number of data packets. Because the relay's VSQ value indicates the count of its available resources including residual energy and holding tokens, the scheme also contributes to the resource balance for the relays.

7.3. Main motivations for dividing time horizon

Noting that, there naturally exist conflicting interests between the source and the relays for the information delivery in WCNs [16]. Moreover, in the practical communication system, the node's

[2] The optimal network throughput is obtained based on a dynamic algorithm, which only maximizes the network throughput and does not consider network stability or selfish boundary.

source states may change much slower than channel fading. To deal with the different objectives between the source and relays, we propose a two-timescale information delivery scheme based on the different timescale states. In this two-timescale scheme, the source only needs to collect the current QSI and VSQ information of relays at each time frame. It benefits from low signaling overhead and computational complexity. Motivated by the analysis in [38], for FA of the source at a large time frame, the signaling overhead is $O(2N)$ due to that the total number of collected data bits is $2N$. Meanwhile, the source has to calculate the flow rates for each relay and TD. By using primal–dual interior point methods, the computational complexity of FA problem with $2N$ variables is $O(8N^3)$. While, for FSS of the relay, there exists no communication overhead owing to that the relay executes the FSS just according to its local information at each timeslot. And also, since each relay chooses its own strategy based on the Q-value, the total computational complexity of FSS problems for relays is $O(N\rho)$. Thus, the signaling overhead of our two-timescale delivery scheme for one time frame is $O(2N)$, and the computational complexity is $O(8N^3 + N\rho)$. Compared with the one-timescale algorithm whose signaling overhead is $O((2N)\rho)$ and computational complexity is $O(\rho(8N^3 + N))$, the communication overhead imposed by the proposed algorithms in our scheme is trivial. However, the high prediction accuracy is required in our scheme owing to that it will significantly affect the outputs of **Algorithm 1**. In addition, the iteration complexity[3] of (**Algorithm 1**) is $O(1)$ due to the fact that it is iterated only once within each timeslot. Besides, by referring to Section 6 in [39], we know that the iteration complexity of **Algorithm 2** is $O(\frac{1}{\delta})$ when setting the tolerance deviation of the algorithm iterations as $\frac{1}{N}\|\mathbf{F}_{\mathcal{I}}^t - \mathbf{F}_{\mathcal{I}}^{t^*}\| + \frac{1}{N}\|\mathbf{F}_{\mathcal{U}}^t - \mathbf{F}_{\mathcal{U}}^{t^*}\| < \delta$ where δ stands for the accuracy index.[4]

8. Simulation results

This section presents the performance analysis results of the proposed scheme. First, we illustrate the convergence of the proposed combined Q-learning algorithm for stochastic games. Second, we present the dynamic processes of the FA for the source. Then, we present the queue length of the proposed scheme and demonstrate the characteristics of network throughput.

8.1. Simulation settings

In our simulation, we consider a cooperative 5G system with multi-hop cooperative delivery. We consider a total of 20 communication pairs that are uniformly distributed. For the sake of simplicity, we consider that the normalized duration of the timeslot is 1 and $B = 1$ denotes the normalized bandwidth spacing. Moreover, we model the channel process as Gaussian random variables as i.i.d over different timeslots. Our proposed scheme starts with the queue lengths including data queues $\mathbf{D}^0$ and VSQs $\mathbf{S}^0$: $S_1^0 = 1, S_2^0 = 5, S_3^0 = 3, D_1^0 = 6, D_2^0 = 4, D_3^0 = 1$. Other simulation parameters are listed in Table 2 in compliance with previous studies [4,40].

To demonstrate the effectiveness of the proposed scheme, we compare the performance of our proposed scheme with the following two schemes. **Traditional FA approach (TFA):** TFA divides the data rates equally among all the network nodes at each time frame and ignores the nodes' differences in the node-selfishness and QSIs. **Data queue-based FA approach (QBFA):** QBFA makes the dynamic FA to maximize the average utility when just considering the network stability. Additionally, for the sake of fairness, the incentive mechanisms in the compared algorithms are similar to our scheme.

[3] The iteration complexity is defined as the maximum number of algorithm iterations before the agent obtains the optimal strategy at a specific system state.

[4] $\mathbf{F}_{\mathcal{I}}^{t^*}$ and $\mathbf{F}_{\mathcal{U}}^{t^*}$ are the theoretical optimum values for flow rates at time frame t.

Table 2
Simulation settings.

Parameter	Value
Transmit power P^T	2
Receive power P^R	0.08
Maximum transmit power P^{max}	3
Spent tokens $\mathcal{R}^p$	0.42
Earned tokens $\mathcal{R}^g$	0.1
Maximum data rate for link between s and $n^{\mathcal{I}}$ A_n^t	40
Number of short timeslots with a frame ρ	200

8.2. Processes of dynamic strategy decisions and queues

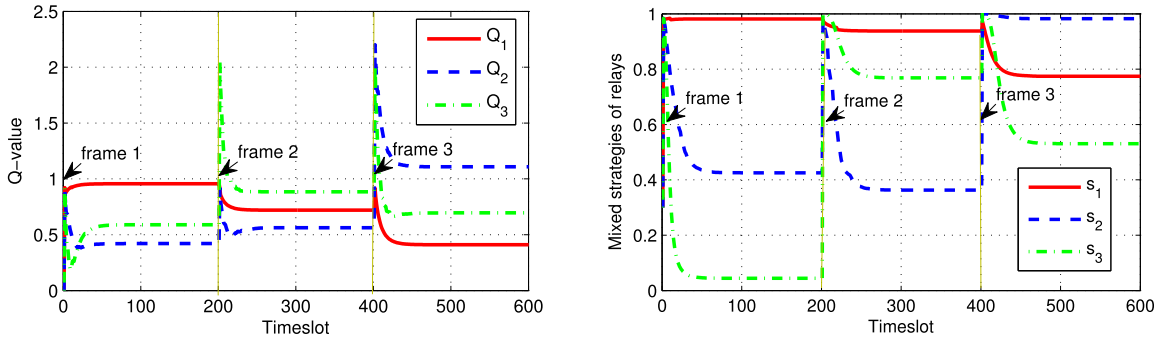
Fig. 9 shows the convergence of the proposed combined Q-learning algorithm by using stochastic game. Fig. 9(a) exhibits the expected payoffs of the relays, whereas Fig. 9(b) presents the optimal mixed forwarding strategies for relays across the timeslots. It proves the robust convergence of **Algorithm 1**, that verifies the result of Theorem 3. The happens because the relay nodes also considers others' strategies during the learning process.

Fig. 10 illustrates the basic ideal of the proposed scheme (the FA processes results of **Algorithm 2**) to deal with the selfishness of relay nodes and to improve the overall network throughout with the help of source's FAs and incentive mechanism. In our proposed scheme, when a relay exhibits selfishness for its *TD's* data traffic, it is assigned less traffic to decrease the packet-dropping rate and to obtains more traffic of its own to decrease its selfishness (based on the incentive mechanism). Fig. 10(a) demonstrates the dynamics of allocated flow rates to the relay nodes, whereas Fig. 10(b) shows the dynamics of allocated flow rates to nodes $\mathcal{N}^{\mathcal{U}}$ via a corresponding relays. Fig. 10 indicates that the FA decision is made dynamically from frame-to-frame and it is influenced by queues $\mathbf{D}^t$ and $\mathbf{S}^t$. From this figure, we can also observe that the larger node-selfishness results in the more flow traffic allocated to relay but the fewer traffic for the corresponding *TD* via a relay. Hence, these results meets the normal considerations for maximizing the overall network throughput when the source executes the FAs to the relay nodes.

Fig. 11 presents the evolution of queues (i.e., VSQs and data queues) of the proposed scheme across time frames. Fig. 11(a) plots the VSQs' lengths for relay nodes, whereas, Fig. 11(b) demonstrates the dynamic lengths of data queues. From Fig. 11, we can observe that the proposed scheme keeps the lengths of both queues below a positive value. This happens because the source considers both the boundaries of relay's selfishness and the data queues while performing the FAs. Since the relays' selfishness are bounded by the positive values which implies that the proposed scheme can also prevent the autonomous relay from being completely selfish, that may effectively depress the selfishness's negative influence on network throughput. Therefore, this figure also shows that the proposed scheme can ensures the network stability.

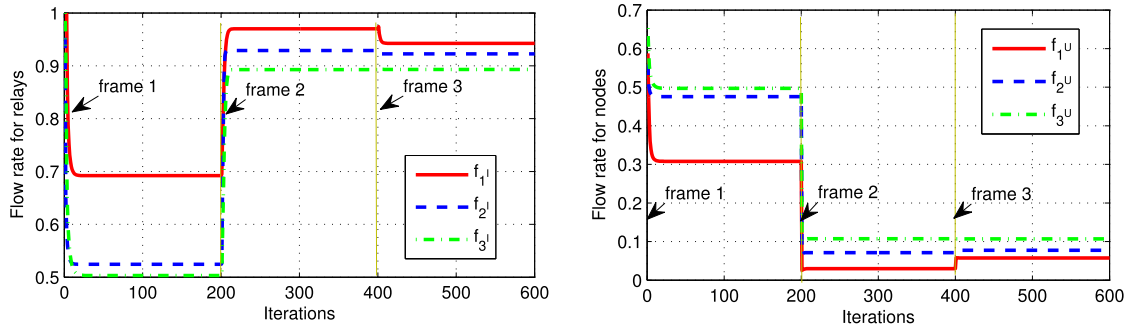
8.3. Average queue length and network throughput for different approaches

Fig. 12 presents the time-average lengths of the queues versus the control parameter V. Fig. 13(a) exhibits the length of data queue D_1, whereas Fig. 13(b) shows the length of VSQ S_1 for communication pair 1. From this figure, we can observe that the parameter V impacts the average length of the queues and changes the values. As the data queue length may represent the information delivery delay, we can adjust the appropriate parameter V to let the data queue length and further the delivery delay of our scheme in satisfactory states. Its clear from this figure that the proposed scheme performs better than the other approaches with smaller values of data queues and VSQs because our proposed scheme



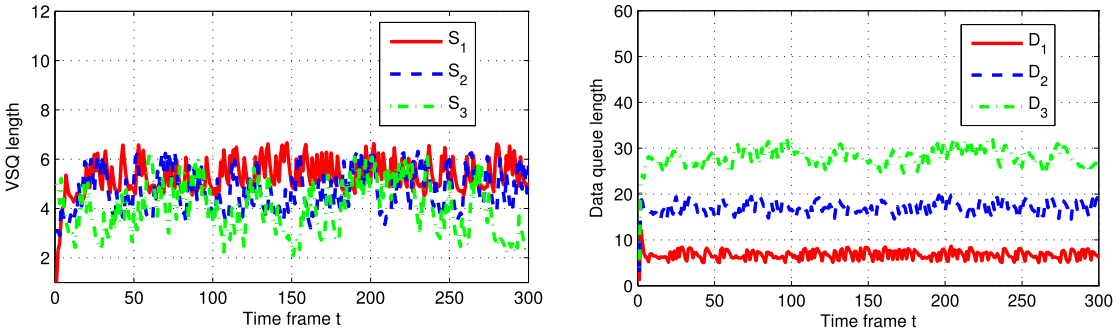
(a) Q-value vs. timeslot. (b) Forwarding strategy vs. timeslot.

Fig. 9. Convergence of the learning algorithm.



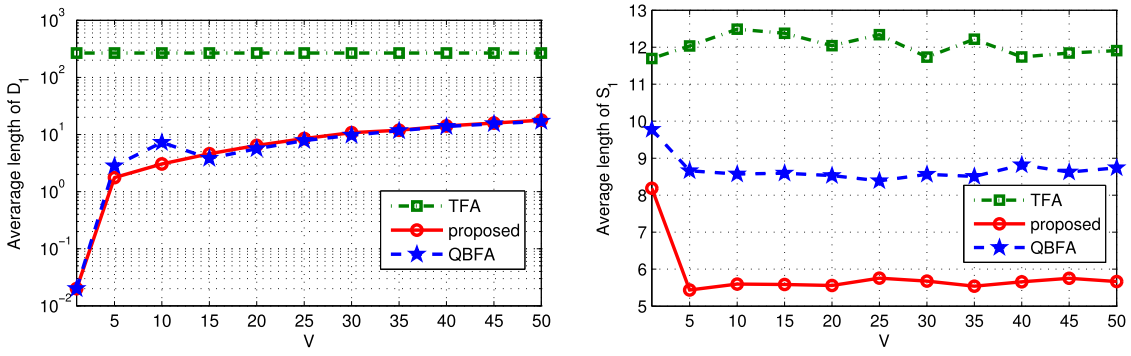
(a) Flow rate for relays. (b) Flow rate for nodes.

Fig. 10. Rate allocation across the time frames.



(a) VSQ process. (b) Data queue process.

Fig. 11. Processes for queues across the time frames..



(a) Data queue D_1. (b) VSQ S_1.

Fig. 12. Average queue lengths vs. V.

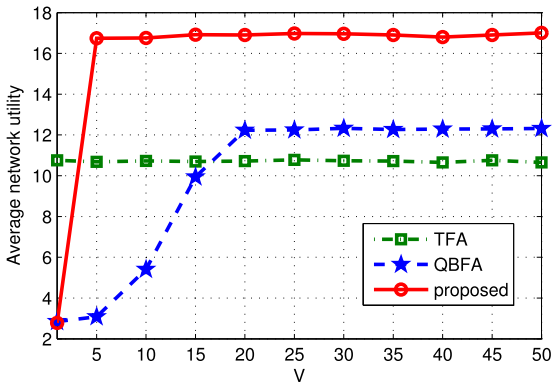


Fig. 13. Utility vs. V.

considers both the data queue state and the selfishness of the relays when executing the FAs.

Fig. 13 depicts the average network throughput (utility) versus the control parameter V for different approaches. The average throughput achieved by the proposed algorithm is larger than that of the other compared approaches because the source node in our proposed scheme considers the relays' selfishness while performing the FAs. Then, the relay with lower node-selfishness undertakes more forwarding tasks in our proposed scheme, which significantly reduces the packet-dropping.

9. Conclusions

In this paper, we employed a cooperative transmission in 5G communication systems and studied the Tactile information delivery via 5G systems to understand the real-time human-to-machine interactions remotely for the Tactile Internet. Specifically, we developed a dynamic information delivery scheme in WCNs to coordinate the FA for the source and FSS for each relay according to the relays' dynamic selfishness. Considering the interference among the relays, we employed the stochastic game to model the strategic interactions among selfish relays and prove the existence of Nash equilibrium. We proposed a combined Q-learning algorithm for the relay to obtain the equilibrium strategy and then a stochastic optimization model was presented for the source node in terms of the dynamic network state to allocate the flow rates to maximize the average network throughput under the constraints of the network stability and selfishness boundary. Aided by relay's VSQ and the Lyapunov optimization theory, an effective iteration algorithm was established to solve the optimization problem. Our scheme has a practical significance for managing the dynamic delivery of Tactile information and reducing the negative influence of selfishness on network throughput, contributing to the wide applications of the Tactile Internet and smart autonomous devices.

Appendix A. Proof of Theorem 3

Based on Theorem in [41], if the mapping $\mathcal{H}^\tau$ meets the following conditions: (1) there exists a number $0 < \beta < 1$ (2) a sequence $\xi^\tau \geq 0$ converging to zero w.p. 1, such that $\| \mathcal{H}^\tau Q^\tau - \mathcal{H}^\tau Q^* \| \leq \beta \| Q^\tau - Q^* \| + \xi^\tau$ for all $Q^\tau \in \mathbf{Q}$ and $Q^* \in E[\mathcal{H}^\tau Q^*]$, then the algorithm converges. Similar to the proof of theorem in [33], we can prove the convergence of our combined Q-learning algorithm.

Appendix B. Proof of Theorem 4

Followed from (24), we have

$$L(\boldsymbol{\Theta}^{t+1}) - L(\boldsymbol{\Theta}^t) = \frac{1}{2}(\boldsymbol{\Theta}^{t+1})^H \boldsymbol{\Theta}^{t+1} - \frac{1}{2}(\boldsymbol{\Theta}^t)^H \boldsymbol{\Theta}^t,$$

$$\begin{aligned}
&= \frac{1}{2}(\max[\mathbf{D}^t - \mathbf{R}^t, 0] + \mathbf{F}_{\mathcal{U}}^t)^H (\max[\mathbf{D}^t - \mathbf{R}^t, 0] + \mathbf{F}_{\mathcal{U}}^t) \\
&+ \frac{1}{2}(\max[\mathbf{T}^t - \mathcal{R}^p \mathbf{F}_{\mathcal{I}}^t, 0] + \mathbf{I}^t P^T + \mathcal{R}^g \mathbf{R}^t + \mathbf{F}_{\mathcal{U}}^t P^R + \mathbf{F}_{\mathcal{I}}^t P^R)^H \\
&\quad (\max[\mathbf{T}^t - \mathcal{R}^p \mathbf{F}_{\mathcal{I}}^t, 0] + \mathbf{I}^t P^T + \mathcal{R}^g \mathbf{R}^t + \mathbf{F}_{\mathcal{U}}^t P^R + \mathbf{F}_{\mathcal{I}}^t P^R) \\
&- \frac{1}{2}(\mathbf{D}^t)^H \mathbf{D}^t - \frac{1}{2}(\mathbf{S}^t)^H \mathbf{S}^t, \qquad (35)
\end{aligned}$$

where we used the queue evolution in (3) and (9). For any non-negative scalar quantities D, f, and r, the inequality

$$(\max[D - r, 0] + f)^2 \leq D^2 + r^2 + f^2 + 2D(f - r), \qquad (36)$$

holds. Similarly,

$$\begin{aligned}
&\left(\max[\mathbf{S}^t - \mathcal{R}^p \mathbf{F}_{\mathcal{I}}^t, 0] + \mathbf{I}^t P^T + \mathcal{R}^g \mathbf{R}^t + \mathbf{F}_{\mathcal{U}}^t P^R + \mathbf{F}_{\mathcal{I}}^t P^R\right)^2 \\
&\leq (\mathbf{S}^t)^H \mathbf{S}^t + (\mathcal{R}^p)^2 (\mathbf{F}_{\mathcal{I}^t})^H \mathbf{F}_{\mathcal{I}}^t + (P^T)^2 (\mathbf{I}^t)^H \mathbf{I}^t \\
&+ (\mathcal{R}^g)^2 (\mathbf{R}^t)^H \mathbf{R}^t + (P^R)^2 (\mathbf{F}_{\mathcal{U}}^t)^H \mathbf{F}_{\mathcal{U}}^t + (P^R)^2 (\mathbf{F}_{\mathcal{I}}^t)^H \mathbf{F}_{\mathcal{I}}^t \\
&+ 2(\mathbf{I}^t P^T + \mathcal{R}^g \mathbf{R}^t + \mathbf{F}_{\mathcal{U}}^t P^R + \mathbf{F}_{\mathcal{I}}^t P^R)^2 - (\mathcal{R}^p \mathbf{F}_{\mathcal{I}}^t)^H \mathbf{S}^t. \qquad (37)
\end{aligned}$$

Then, we have

$$\begin{aligned}
&L(\boldsymbol{\Theta}^{t+1}) - L(\boldsymbol{\Theta}^t) \\
&\leq \frac{1}{2}(\mathbf{R}^t)^H (\mathbf{R}^t) + \frac{1}{2}(\mathbf{F}_{\mathcal{U}}^t)^H (\mathbf{F}_{\mathcal{U}}^t) + (\mathbf{F}_{\mathcal{U}}^t - \mathbf{R}^t)^H \mathbf{D}^t \\
&+ \frac{1}{2}(\mathcal{R}^p)^2 (\mathbf{F}_{\mathcal{I}^t})^H \mathbf{F}_{\mathcal{I}}^t + (P^T)^2 (\mathbf{I}^t)^H \mathbf{I}^t + \frac{1}{2}(\mathcal{R}^g)^2 (\mathbf{R}^t)^H \mathbf{R}^t \\
&+ \frac{1}{2}(P^R)^2 (\mathbf{F}_{\mathcal{U}}^t)^H \mathbf{F}_{\mathcal{U}}^t + \frac{1}{2}(P^R)^2 (\mathbf{F}_{\mathcal{I}}^t)^H \mathbf{F}_{\mathcal{I}}^t \\
&+ (\mathbf{I}^t P^T + \mathcal{R}^g \mathbf{R}^t + \mathbf{F}_{\mathcal{U}}^t P^R + \mathbf{F}_{\mathcal{I}}^t P^R)^2 - (\mathcal{R}^p \mathbf{F}_{\mathcal{I}}^t)^H \mathbf{S}^t \\
&\leq B + (\mathbf{F}_{\mathcal{U}}^t - \mathbf{R}^t)^H \mathbf{D}^t \\
&+ (\mathbf{I}^t P^T - \mathcal{R}^p \mathbf{F}_{\mathcal{I}}^t + \mathcal{R}^g \mathbf{R}^t + \mathbf{F}_{\mathcal{U}}^t P^R + \mathbf{F}_{\mathcal{I}}^t P^R)^H \mathbf{S}^t, \qquad (38)
\end{aligned}$$

where B is an upper bound on the term $\frac{1}{2}[(\mathbf{R}^t)^H (\mathbf{R}^t) + (\mathbf{F}_{\mathcal{U}}^t)^H (\mathbf{F}_{\mathcal{U}}^t)] + \frac{1}{2}[(\mathcal{R}^p)^2 (\mathbf{F}_{\mathcal{I}^t})^H \mathbf{F}_{\mathcal{I}}^t + (P^T)^2 (\mathbf{I}^t)^H \mathbf{I}^t + (\mathcal{R}^g)^2 (\mathbf{R}^t)^H \mathbf{R}^t + (P^R)^2 (\mathbf{F}_{\mathcal{U}}^t)^H \mathbf{F}_{\mathcal{U}}^t + (P^R)^2 (\mathbf{F}_{\mathcal{I}}^t)^H \mathbf{F}_{\mathcal{I}}^t]$, which holds under the fact that the wireless link transmission rates satisfy the properties of boundedness.

Adding $-V\mathbb{E}[U^t(\mathbf{F}^t, \mathbf{h}^t) | \boldsymbol{\Theta}^t]$ to both sides of (37) and taking a expectation, yields

$$\begin{aligned}
&\Delta_V(\boldsymbol{\Theta}^t) \leq B - V\mathbb{E}[U^t(\mathbf{F}^t) | \boldsymbol{\Theta}^t] + \mathbb{E}[(\mathbf{F}_{\mathcal{U}}^t - \mathbf{R}^t)^H \mathbf{D}^t | \boldsymbol{\Theta}^t] \\
&+ \mathbb{E}[(\mathbf{I}^t P^T - \mathcal{R}^p \mathbf{F}_{\mathcal{I}}^t + \mathcal{R}^g \mathbf{R}^t + \mathbf{F}_{\mathcal{U}}^t P^R + \mathbf{F}_{\mathcal{I}}^t P^R)^H \mathbf{S}^t | \boldsymbol{\Theta}^t].
\end{aligned}$$

This completes the proof of Theorem 4.

References

[1] M. Maier, M. Chowdhury, B. Rimal, D. Van, The Tactile Internet: Vision recent progress open challenges, IEEE Commun. Mag. 54 (5) (2016) 138–145.

[2] A. Aijaz, Towards 5G-enabled Tactile Internet: Radio resource allocation for haptic communications, in: IEEE Wireless Commu. and Networking Conf. Workshops (WCNCW), April 2016, pp. 145–150.

[3] A. Aijaz, M. Dohler, A. Aghvami, V. Friderikos, Realizing the Tactile Internet: Haptic communications over next generation 5G cellular networks, IEEE Wirel. Commun. 24 (2) (2017) 82–89.

[4] P. Li, S. Guo, On the multicast capacity in energy-constrained lossy wireless networks by exploiting intrabatch and interbatch network coding, IEEE Trans. Parallel Distrib. Syst. 24 (11) (2013) 2251–2260.

[5] A. Ali, K.S. Kwak, N.H. Tran, Z. Han, D. Niyato, F. Zeshan, M.T. Gul, D.Y. Suh, RaptorQ-Based efficient multimedia transmission over cooperative cellular cognitive radio networks, IEEE Trans. Veh. Technol. 67 (8) (2018) 7275–7289.

[6] Z. Kong, Y. Kwok, J. Wang, On the impact of selfish behaviors in wireless packet scheduling, in: Proc. of IEEE ICC, May 2008, pp. 3253–3257.

[7] B. Cao, Y. Li, C. Wang, G. Feng, Dynamic cooperative media access control for wireless networks, Proc. Wirel. Commun. Mob. Comput. 15 (13) (2015) 1759–1772.

[8] B. Cao, Q. Chen, G. Feng, Y. Li, C. Wang, Revisiting relay assignment in cooperative communications, Wirel. Netw. (2016) 1–15.

[9] Z. Li, H. Shen, Game-theoretic analysis of cooperation incentive strategies in mobile ad hoc networks, IEEE Trans. Mob. Comput. 11 (8) (2012) 1287–1303.

[10] E. Orallo, M. Olmos, J. Cano, C. Calafate, CoCoWa: A collaborative contact-based watchdog for detecting selfish nodes, IEEE Trans. Mob. Comput. 14 (6) (2015) 1162–1175.
[11] C. Luo, S. Guo, T. Yang, Green communication in energy renewable wireless mesh networks: Routing, rate control, and power allocation, IEEE Trans. Parallel Distrib. Syst. 25 (12) (2014) 3211–3220.
[12] K. Wu, W. Liao, Flow allocation in multi-hop wireless networks: A cross-layer approach, IEEE Trans. Wirel. Commun. 7 (1) (2008) 269–276.
[13] F. Wu, T. Chen, S. Zhong, A game-theoretic approach to stimulate cooperation for probabilistic routing in opportunistic networks, IEEE Trans. Wirel. Commun. 12 (4) (2013) 1573–1583.
[14] C. Tang, A. Li, X. Li, When reputation enforces evolutionary cooperation in unreliable MANETs, IEEE Trans. Syst. Man Cybern. 45 (10) (2015) 2190–2201.
[15] X. Kang, Y. Wu, Incentive mechanism design for heterogeneous peer-to-peer networks: A Stackelberg game approach, IEEE Trans. Mob. Comput. 14 (5) (2015) 1018–1030.
[16] W. Krichene, M. Castillo, A. Bayen, On social optimal routing under selfish learning, IEEE Trans. Control. Netw. Syst. PP (99) (2016) 479–488.
[17] D. Li, Y. Xu, J. Liu, Distributed relay selection over multi-source and multi-relay wireless cooperative networks with selfish nodes, Comput. Commun. 33 (17) (2010) 2145–2153.
[18] A. Clark, A. Pande, K. Govindan, R. Poovendran, Using social information for flow allocation in MANETs, in: Proc. of IEEE INFOCOM, Nov. 2013, pp. 1–10.
[19] F. Bao, I. Chen, M. Chang, J. Cho, Hierarchical trust management for wireless sensor networks and its applications to trust-based routing and intrusion detection, IEEE Trans. Netw. Serv. Manag. 9 (2) (2012) 169–183.
[20] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, 5G-enabled Tactile Internet, IEEE J. Sel. Areas Commun. 34 (3) (2016) 460–473.
[21] E. Wong, M. Dias, L. Ruan, Predictive resource allocation for Tactile Internet capable passive optical LANs, J. Lightwave Technol. 35 (13) (2017) 2629–2641.
[22] X. Ma, M. Sheng, Y. Zhang, Green communications with network cooperation: A concurrent transmission approach, IEEE Commun. Lett. 16 (12) (2012) 1952–1955.
[23] H. Zhu, X. Lu, Y. Fan, SMART: A secure multilayer credit-based incentive scheme for delay-tolerant networks, IEEE Trans. Veh. Technol. 58 (8) (2009) 4628–4639.
[24] Q. Xu, Z. Su, S. Guo, A game theoretical incentive scheme for relay selection services in mobile social networks, IEEE Trans. Veh. Technol. PP (99) (2015) 1–11.
[25] C. Adika, L. Wang, Autonomous appliance scheduling for household energy management, IEEE Trans. Smart Grid. 5 (2) (2014) 673–682.
[26] D. Brown, F. Fazel, A game theoretic study of energy efficient cooperative wireless networks, J. Commun. Netw. 13 (3) (2011) 266–276.
[27] D. Palomar, M. Chiang, A tutorial on decomposition methods for network utility maximization, IEEE J. Sel. Areas Commun. 24 (8) (2006) 1439–1451.
[28] F. Fu, M. Schaar, Learning to compete for resources in wireless stochastic games, IEEE Trans. Veh. Technol. 58 (4) (2009) 1904–1919.
[29] W. Xu, Y. Zhang, Q. Shi, X. Wang, Energy management and cross layer optimization for wireless sensor network powered by heterogeneous energy sources, IEEE Trans. Wirel. Commun. 14 (5) (2015) 2814–2826.
[30] C. Qiu, Y. Hu, Y. Chen, Lyapunov optimized cooperative communications with stochastic energy harvesting relay, IEEE Internet Things J. 5 (2) (2018) 1323–1333.
[31] L. Georgiadis, M. Neely, L. Tassiulas, Resource allocation and cross-layer control in wireless networks, Found. Trends Netw. 1 (1) (2006) 1–144.
[32] D. Fudenberg, J. Tirole, Game Theory, MIT, 1992.
[33] X. Chen, Z. Zhao, H. Zhang, Stochastic power adaptation with multiagent reinforcement learning for cognitive wireless mesh networks, IEEE Trans. Mob. Comput. 12 (11) (2013) 2155–2166.
[34] R. Sutton, A. Barto, Reinforcement Learning: An Introduction, MIT, 1998.
[35] C. Szepesvari, M. Littman, A unified analysis of value-function-based reinforcement learning algorithms, Neural Comput. 11 (8) (1999) 2017–2060.
[36] M. Neely, Stochastic Network Optimization with Application to Communication and Queueing Systems, Morgan & Claypool, 2010.
[37] M. Neely, Dynamic optimizaton and learning for renewal systems view document, IEEE Trans. Automat. Control 58 (1) (2012) 32–46.
[38] N. Omidvar, A. Liu, V. Lau, F. Zhang, Optimal hierarchical radio resource management for HetNets with flexible backhaul, IEEE Trans. Wirel. Commun. 17 (7) (2018) 4239–4255.
[39] W. Wu, Q. Yang, P. Gong, K. Kwak, Adaptive multi-homing resource allocation for time-varying heterogeneous wireless networks without timescale separation, IEEE Trans. Commun. 64 (9) (2016) 3794–3807.
[40] L. Feng, Q. Yang, K. Kim, Dynamic rate allocation and forwarding strategy adaption for wireless networks, IEEE Signal Process. Lett. 25 (7) (2018) 1034–1038.
[41] C. Watkins, P. Dayan, Q-learning algorithms, Mach. Learn. 8 (1992) 279–292.

Li Feng received the B.S. degree in electronic information engineering from Zhengzhou University of Aeronautics, China, in 2011, the M.E. degrees in telecommunications engineering from Xidian University, China, in 2014, the Ph.D. degree in communication and information systems from Xidian University, in 2017. She is currently a Post-Doctoral Fellow with the UWB Wireless Communications Research Center, Inha University, Korea. Her research interests include wireless resource allocation, stochastic network optimization, and node selfishness analysis.

Amjad Ali received his B.S. and M.S. degrees in Computer Science from COMSATS Institute of Information Technology, Pakistan in 2006 and 2008, respectively. He received the Ph.D. from Electronics and Radio Engineering Department at Kyung Hee University, South Korea in 2015. Currently he is Post-Doctoral Fellow at MNC Research Center in the Department of Electrical Engineering at Korea University, South Korea. He published various peer-reviewed journal papers and served as TPC for several IEEE related conferences. His main research interests include Internet of Things, cognitive radio networks 5G networks, Tactile Internet, stochastic network optimization, wireless resource management, and vehicular networks.

Hannan Bin Liaqat received the B.Sc. and M.Sc. degrees from COMSATS Institute of Information Technology, Lahore, Pakistan, in 2006 and 2009, respectively. He is currently working toward the Ph.D. degree with the Mobile and Social Computing Laboratory, School of Software, Dalian University of Technology, Dalian, China. He was a Lecturer with the Department of Computer Science, University of Gujrat, Gujrat, Pakistan. His research interests include ad-hoc social networks, mobile computing, and social computing.

Muhammad Aksam Iftikhar completed his B.S. degree in computer engineering from University of Engineering and Technology, Lahore in 2007. He received M.S. degree in computer science from the same university in 2010. In 2014, he completed his Ph.D. at Pakistan Institute of Engineering and Applied Sciences, Islamabad. His primary research areas include Image Processing (especially medical image processing and analysis), Computer Vision, and Pattern Recognition. He has published various research articles in international journals and conferences.

Ali Kashif Bashir is working as Senior Lecturer at Manchester Metropolitan University, United Kingdom. He received his Ph.D. in 2012 from computer science and engineering from Korea University, South Korea. In the past, he held appointments with Osaka University, Japan; Nara National College of Technology, Japan; the National Fusion Research Institute, South Korea; Southern Power Company Ltd., South Korea, and the Seoul Metropolitan Government, South Korea. His research interests include: cloud computing, NFV/SDN, network virtualization, network security, IoT, computer networks, RFID, sensor networks, wireless networks, and distributed computing. He is serving as the Editor-in-chief of the IEEE INTERNET TECHNOLOGY POLICY NEWSLETTER and the IEEE FUTURE DIRECTIONS NEWSLETTER. He is an Editorial Board Member of journals, such as the IEEE ACCESS, the Journal of Sensor Networks, and the Data Communications.

Sangheon Pack received the B.S. and Ph.D. degrees in computer engineering from Seoul National University, Seoul, South Korea, in 2000 and 2005, respectively. From 2005 to 2006, he was a Post-Doctoral Fellow with the Broadband Communications Research Group, University of Waterloo, Waterloo, ON, Canada. In 2007, he joined the Faculty of Korea University, where he is currently a Full Professor with the School of Electrical Engineering. His research interests include future Internet, soft warized networking (SDN/NFV), information centric networking/delay tolerant networking, and vehicular networks. He was a recipient of the IEEE/Institute of Electronics and Information Engineers Joint Award for IT Young Engineers Award in 2017, the Korean Institute of Information Scientists and Engineers Young Information Scientist Award in 2017, the Korean Institute of Communications and Information Sciences Haedong Young Scholar Award in 2013, the LG Yonam Foundation Overseas Research Professor Program in 2012, and the IEEE ComSoc APB Outstanding Young Researcher Award in 2009. He served as the TPC Chair for EAI Qshine 2016, a Publication Co-Chair for the IEEE INFOCOM 2014 and ACM MobiHoc 2015, a Co-Chair for IEEE VTC 2010-fall transportation track, a Co-Chair for IEEE WCSP 2013 Wireless Networking Symposium, and a Publicity Co-Chair for IEEE SECON 2012. He is an Editor of the Journal of Communications Networks, IET Communications, and he is a Guest Editor of the IEEE Transactions on Emerging Topics in Computing.