# Incentive Compatible Overlay D2D System: A Group-Based Framework without CQI Feedback

Yi Zhang, *Student Member, IEEE,* Chih-Yu Wang, *Member, IEEE,* and Hung-Yu Wei, *Senior Member, IEEE*

**Abstract**—With the large expected demand of wireless communication, Device-to-Device (D2D) communication has been proposed as a promising technology to enhance network performance. Nevertheless, the selfish nature of potential D2D users may impale the performance of D2D-enabled network. In this paper, we propose a D2D-enabled cellular network framework, which support a novel group D2D mode under overlay D2D communication. The group-based design is derived from the discussions of two common D2D modes, divided and shared D2D modes, regarded as special cases. The proposed framework provides a pricing-based dynamic Stackelberg game for optimal mode selection and spectrum partitioning. We propose the incentive compatible pricing strategy to provide proper incentive for these selfish potential D2D pairs to make optimal choices in mode selection. Our results show that the pricing and spectrum partition strategy effectively prevents selfish potential D2D users from harming the system performance while fully exploits the potential of D2D communication.

**Index Terms**—Overlay, Device-to-Device communication, mode selection, spectrum partitioning, pricing, Stackelberg game.

## 1 Introduction

Device-to-Device (D2D) communication is a promising solution [1] to help next generation cellular communication system meet the challenging requirements in 5G standard such as Gbs-scale throughput and millions-scale device number [2]. D2D communication utilizes the opportunistic proximity between devices when occurs by allowing them to communicate directly instead of transmitting through conventional cellular links with base stations at a much further distance. This approach improves the spectrum utilization efficiency, energy efficiency, and offloading from base station. New types of services such as peer-to-peer services or mobile social network can also be realized in a more efficient way with the assistance of D2D communication.

One of the most challenging issues in integrating D2D communication into conventional cellular system is the spectrum assignment on both D2D links and cellular links. Two main approaches, the *underlay* and the *overlay* spectrum access, are proposed in the literature [3]. In underlay spectrum access, D2D users reuse the same spectrum / carriers as the conventional cellular spectrum subject to tolerable interference to the cellular users. Interference mitigation and QoE guarantees are the main challenges in this approach [4]. On the other hand, in the overlay approach the D2D users utilize a dedicated spectrum reserved for D2D communications. The interference to existing cellular users is not a threat here.

Under overlay D2D communication, there are two D2D modes frequently discussed, *divided D2D mode* and *shared D2D mode* [5–8]. For divided D2D mode, the dedicated D2D spectrum is equally split into multiple orthogonal resource units, which are then allocated to each D2D user. For shared D2D mode, on the other hand, all D2D users share the whole dedicated D2D spectrum. In brief, divided D2D mode guarantees the transmission quality when more users are accessing, while shared D2D mode provides higher spectrum utilization efficiency. In this paper, we propose a novel D2D mode, called **group D2D mode**, to extend these two common D2D modes into a more general form. For group D2D mode, the dedicated D2D spectrum is equally split into in a certain amount of spectrum resource units. By adopting proper group number, potential D2D pairs in D2D mode can enjoy good transmission quality meanwhile the framework can ensure high spectrum utilization efficiency.

The main challenge in overlay approach is the spectrum utilization efficiency. The service provider or base station reserve a dedicated spectrum for all D2D communications, which we refer to as *spectrum partitioning* [5]. The partition strategy should be aware of the loading and requirements from both potential D2D users and conventional cellular users. Furthermore, the spectrum partitioning within the dedicated D2D spectrum is also an important issue under the proposed group D2D mode.

Additionally, D2D users in D2D-enabled network usually are assumed to have the freedom to choose between D2D mode and cellular mode [5–7, 9–11]. Nevertheless, the choices of rational D2D users are more likely to be based on their self-interests instead of overall performance. That is, a rational user will choose D2D mode if and only if this mode is offering more benefits to him/her, such as higher transmission quality, lower service payment, or both. In such a case, these users may select the modes which are not favored by the base station, and the system performance will be degraded due to their selfish choices. Therefore, the *incentive* of these potential D2D users' *mode*

*Yi Zhang is with the Graduate Institute of Communication Engineering, National Taiwan University, Taiwan. Email: yzhang.cn@outlook.com.*

*Chih-Yu Wang is with the Research Center for Information Technology Innovation, Academia Sinica, Taiwan. Email: cywang@citi.sinica.edu.tw.*

*Hung-Yu Wei is with the Graduate Institute of Communication Engineering, Graduate Institute of Electrical Engineering, and Department of Electrical Engineering, National Taiwan University, Taiwan. Email: hywei@cc.ee.ntu.edu.tw.*

*selection* strategies in any given scenario should be studied in advance and be considered when integrating the overlay D2D framework into existing cellular system.

Our goal is to propose an incentive compatible mode selection and spectrum partitioning to maximize the overall system utility of D2D-enabled network, with service quality of existing conventional cellular users and incentive for selfish D2D users in mode selection in mind. The main idea of the proposed framework is to regulate mode selection by pricing, where the pricing strategy is purely based on the mode selection decisions applied by potential D2D pairs in previous rounds. We show that the system can reach the optimal configuration through the proposed pricing rule and self-regulation of potential D2D users without feedback of other information such as channel quality indicator (CQI). The D2D-enabled system benefits from such a design in two aspects: 1) the required feedback from potential D2D pairs is minimized, and 2) the room for undesired selfish behaviors is minimized since no additional information reporting process except mode selections is required. We show that the proposed framework achieve optimal performance under both divided and shared D2D modes through theoretical analysis and the performance under general group mode is also promising as we demonstrated through simulations. Besides, we prove that the proposed framework is incentive compatible for all three D2D modes.

### 1.1 Related Work

To address efficient and high performance D2D communication, resource management has been widely studied [12–17]. Several possible research directions are investigated, including channel assignment [13], power allocation [14, 15], relay selection [16, 17] and so on.

Mode selection in D2D communication has been studied in the literature with different game-theoretic approaches. Most previous works study D2D mode selection in underlay spectrum access. A two-armed Levy-bandit game [9] is proposed for D2D mode selection. Users consider cellular mode as a safe arm since it provides a fixed reward, while treat D2D mode as a risky arm since it provides a stochastic reward following compound Poisson distribution. In this case, users have their own belief to make decision and update it after each playing. Diaz *et al.* [10] introduce a distributed best response-based approach to D2D mode selection. The additionally imposed interference and backhaul constraints help guarantee that the final choices of users will reach the Nash equilibrium. To improve energy efficiency of network users, a coalition formation game [11] is proposed into the joint D2D mode selection and spectrum sharing. A coalition is formed when multiple D2D users would like to share the same spectrum as existing cellular user. The convergence of proposed coalition formation algorithm is guaranteed. Another coalition formation game [18] is proposed but with a different scenario, that is the spectrum of an existing traditional cellular user can only be reused by one D2D pair. In this work, mode selection and radio resource allocation are jointly considered for the first time. A matching game is formulated in [19] to tackle combinatorial problems and achieve a distributed solution in underlay D2D resource allocation. It performs a learning framework based on Markov approximation and the performance is not guaranteed, while our framework can guarantee the performance with the help of the proposed primal-dual pricing method.

In overlay spectrum access, on the other hand, divided D2D mode has been widely discussed. The proposed algorithm in [6] provide a contract-based mechanism for D2D mode selection. The objective is to improve resource efficiency under the threat that users may report their D2D channel information untruthfully. Nevertheless, the proposed algorithm handles one user at a time, while our approach can handle multiple users at once. A distance-based D2D mode selection strategy [5] is proposed with an optimal threshold affected by the BS density. The selfish nature of potential D2D pair in mode selection, on the other hand, is not discussed in this work. Stochastic geometry is adopted in [7] to estimate average spectrum efficiency in different modes, but the user distribution has to be known in advance. The cheat-proof is also not considered in [7]. A multi-hop multi-channel overlay D2D network is presented in [20] to optimize the network performance. However, the number of channels allocated for both cellular and D2D transmissions is fixed, while our approach is more flexible that the number of channels can be adjusted dynamically according to the network condition.

The shared D2D mode, on the other hand, is not discussed in most existing works such as [5–7, 20]. We will show that the shared D2D mode can greatly improve the overall system performance comparing to divided D2D mode due to higher spectrum utilization efficiency. A similar concept like shared D2D mode is proposed in [8]. One difference is that cellular users and potential D2D users in cellular mode are scheduled in a round-robin fashion. Another is that potential D2D users use a carrier sensing threshold to determine their transmission modes. In addition, the optimization problem in [8] is to maximize the total rate of D2D users under target rate constraint for cellular users, while the rate of cellular users under our proposed algorithm will remain unaffected regardless of the choices of potential D2D users. Moreover, a CSMA-like random access is introduced in [21] to address spatial reuse of D2D users in overlay D2D network. A D2D link will refrain from transmitting if any of its neighbors is transmitting.

Spectrum partitioning for D2D communication is another important issue in overlay approach. In [5], it is lack of adaptability and robustness that the optimal spectrum partitioning depends on fixed D2D mode selection threshold, which is purely due to the density of BS. Similar to [5], the spectrum partitioning in [8] can only be calculated by a joint function between target rate constraint of cellular users and carrier sensing threshold. However, no suggestions are presented in this paper on how to assign values for these two parameters. The authors in [7] suggest that the control of spectrum partitioning needs to be dynamically adjusted by considering D2D mode selection behaviors responded by users. We share a similar concept but our framework utilizes the dynamic pricing strategy as an additional tool to regulate the selfish behaviors of users in D2D mode selection.

In our previous work [22], mode selection and spectrum partitioning under both of divided and shared D2D modes

framework has been discussed. However, we propose group D2D mode in this paper to extend these two D2D modes into a more general form. Furthermore, we propose an approach to find target group number and near-optimal mode selection under group D2D mode framework. The cheat-proofness of the proposed framework is formally proved. And we also provide a discussion about its convergence.

### 1.2 Contributions and Organization

The main contributions of this paper are as follows.

- We propose a pricing-based D2D mode selection and spectrum partitioning framework. This framework utilizes the primal-dual method to retrieve the optimal model selection and spectrum partition configuration in both divided and shared modes. For the group D2D mode, a reinforcement learning approach is proposed to help potential D2D pairs perform mode selection, including choosing optimal D2D groups if in D2D mode. We further propose a dynamic approach to derive the target group number when integrating into the proposed dynamic Stackelberg game.
- Unlike traditional primal-dual method in existing works where the dual variable acts as a control parameter in optimization problem but not necessarily compatible to real incentive of users in regarding their real utility function. We propose a rationalized process to transform the virtual pricing strategy derived from the proposed primal-dual pricing method into real payment that satisfies the incentive compatibility of potential D2D users in mode selection.
- The proposed framework does not need potential D2D pairs to report their CQI, which reduces overhead in signal exchange between potential D2D pairs and the BS. A formal proof of the cheat-proofness of the proposed framework in general cases is provided.

The rest of this paper is organized as below. In Section 2, we introduce the proposed D2D-enabled framework and the proposed group D2D mode. A two-stage dynamic Stackelberg game is presented to manage the proposed framework. In Section 3, we formulate general form of the two-stage Stackelberg game by backward induction and propose a primal-dual method to handle the optimization problem. In Section 4, we analyze two special cases, divided and shared D2D modes. Major propositions, which contribute to the solution of the proposed D2D mode, are proposed here. Subsequently, the complete solution of group D2D mode is provided in Section 5. We evaluate and analyze the performance of the proposed pricing-based mode selection and spectrum partitioning framework in Section 6. Finally, we conclude our work in Section 7.

## 2 System Overview

We consider a D2D-enabled cellular system with one cell and multiple UEs. All UEs may communicate through conventional cellular communications, while some of them, denoted as potential D2D UEs, are D2D enabled. These UEs have formed transmitter-receiver pairs in advance. A part of UE pairs, of which both transmitter and receiver are potential D2D UEs, are denoted as potential D2D pairs.

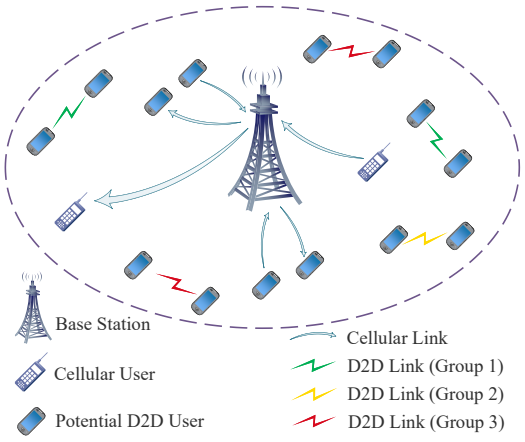



Fig. 1. System Overview

Other pairs, which can only communicate through BS in conventional way, are denoted as cellular pairs.

Formally speaking, we have a set of UE pairs $\mathcal{U}$ with total number of pairs $N = |\mathcal{U}|$, with a subset of cellular pairs $\mathcal{U}_c$ and potential D2D pairs $\mathcal{U}_d$, respectively. The key notations of this paper are listed in TABLE 1.

### 2.1 Spectrum Allocation

We proposed a spectrum partition design for this D2D-enabled cellular system working on overlay spectrum access. In the proposed system, all D2D connections are established on a dedicated spectrum without interference to/from the cellular connections. The BS partitions the spectrum for D2D and cellular connections following a partition strategy. We let $W$ be the total available bandwidth of the spectrum and $p$ be the proportion of total bandwidth reserved for D2D communication. Therefore, we have $Wp$ bandwidth for D2D communication and $W(1-p)$ for cellular communication.

Additionally, we propose a novel D2D mode, called **group D2D mode**, to balance D2D transmission quality and spectrum utilization efficiency. Under this mode, the dedicated spectrum is equally split into $K$ orthogonal resource units. The implementation of $K$ resource units is flexible that they can stand for either standard time-frequency resource blocks or subcarriers in LTE system. These units constitute a spectrum set $\mathcal{F} = \{\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_K\}$. Potential D2D pairs will make decisions to either communicate in cellular mode or choose one spectrum resource from $\mathcal{F}$ for D2D communications. We define $\mathcal{K} = \{\mathcal{K}_1, \mathcal{K}_2, ..., \mathcal{K}_K\}$ as the set of D2D groups and call $\mathcal{K}_k$ as D2D group $k$. The group $\mathcal{K}_k$ is formed by potential D2D pairs who simultaneously utilize spectrum resource $\mathcal{F}_k$. Notice that potential D2D pairs in the same D2D group will generate intra-group interference to each other.

For potential D2D pair $i \in \mathcal{U}_d$, its decision $x_{i,k} \in \{0, 1\}$ denotes whether or not pair $i$ selects group $k$ for D2D communication. The rule is that one pair can join only one D2D group, denoted by $\sum_{k\in\mathcal{K}} x_{i,k} \leq 1$. Here we let $\mathbf{X} = [x_{i,k}]$ denote the mode selection matrix. We further call potential D2D pairs selecting D2D mode as D2D pairs and denote the number of them by $m$. It can be seen that $m = \sum_{i\in\mathcal{U}_d} \sum_{k\in\mathcal{K}} x_{i,k}$, which indicates the loading of D2D

TABLE 1
List of Key Notation

| Notation | Definition |
|---|---|
| $\mathcal{U}, \mathcal{U}_c, \mathcal{U}_d$ | {total, cellular, potential D2D} UE pair set |
| $N$, $N_c$, $N_d$ | The number of UE pairs in $\mathcal{U}, \mathcal{U}_c, \mathcal{U}_d$ |
| $W$ | Total bandwidth in one macro D2D-enabled network |
| $p$ | Proportion of bandwidth reserved for D2D communication |
| $w_c$ | Bandwidth allocated to one cellular pair for either downlink or uplink |
| $w_d$ | Bandwidth allocated to one D2D pair |
| $K$ | The number of orthogonal resource units / D2D groups |
| $\mathcal{F}$ | D2D spectrum set, $\mathcal{F} = \{\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_K\}$ |
| $\mathcal{K}$ | D2D group set, $\mathcal{K} = \{\mathcal{K}_1, \mathcal{K}_2, ..., \mathcal{K}_K\}$ |
| $\text{SINR}_{i,k}$ | SINR of potential D2D pair $i$ in D2D group $k$ |
| $\text{SINR}_{i,up}$, $\text{SINR}_{i,down}$ | {uplink, downlink } SINR of potential D2D pair $i$ in cellular mode |
| $\text{SINR}_{i,c}$ | SINR of potential D2D pair $i$ in cellular mode, $\text{SINR}_{i,c} = \min\{\text{SINR}_{i,up}, \text{SINR}_{i,down}\}$ |
| $P_{ji}$ | Received power from the transmitter of pair $j$ to the receiver of pair $i$ |
| $N_0$ | Terminal noise at the receiver |
| $\Gamma$ | SNR gap |
| $\Pi, \Pi_c, \Pi_d$ | Network utility of {overall, cellular pairs, potential D2D pairs} |
| $r_i$ | Network utility of cellular pair $i$ |
| $r_{i,k}, r_{i,c}$ | Network utility of potential D2D pair $i$ in {D2D group $k$, cellular mode } |
| $a_i, a_{i,c}$, $a_{i,j}, a_{i,d}$ | Simplified parameters |
| $D(\boldsymbol{\mu})$ | Dual optimization problem of objective function $\Pi$, $\min_{\boldsymbol{\mu}} D(\boldsymbol{\mu}) = f_{\mathbf{X}}(\boldsymbol{\mu}) + g_m(\boldsymbol{\mu})$ |
| $\mathbf{X}$ | Mode selection matrix, $\mathbf{X} = [x_{i,k}]$; binary variable (1 or 0) $x_{i,k}$ denotes whether or not potential D2D pair $i$ choose D2D mode in group $k$ |
| $x_{i,d}$ | Mode selection indicator of potential D2D pair $i$ under two special cases |
| $m, m_k$ | The number of potential D2D pairs allocated in {D2D mode, D2D group $k$ } |
| $\boldsymbol{\pi}$ | Payment vector, $\boldsymbol{\pi} = (\pi_1, ..., \pi_K)$; potential D2D pairs pay additional $\pi_k$ for D2D communication in group $k$ |
| $\boldsymbol{\mu}$ | Dual variable vector, $\boldsymbol{\mu} = (\mu_1, ..., \mu_K)$ |
| $\pi, \mu$ | Real payment and virtual payment |
| $\delta_t$ | Dynamical stepsize sequence of subgradient method of $\mu$ |
| $T_i$ | SINR threshold of potential D2D pair $i$ |
| $Q_{i,k}(t)$ | Q-value of potential D2D $i$ in D2D group $k$ at stage $t$ |
| $\alpha_{i,k}, \beta$ | Learning rate and discount factor |
| $n_{i,k}$ | Visiting times of potential D2D $i$ in D2D group $k$ |
| $K_{min}$, $K_{max}$ | {minimal, maximal} number of groups that the system supports in group D2D mode |
| $gap(\mathcal{X}, N_d)$ | System-dependent gap with UE distribution $\mathcal{X}$ and number $N_d$ |

communication in the dedicated spectrum. For D2D group $k$, we introduce $m_k$ as the number of D2D pairs in it so that $m_k = \sum_{i \in \mathcal{U}_d} x_{i,k}$ and $m = \sum_{k \in \mathcal{K}} m_k$. Notice that divided and shared D2D modes are two special cases under the proposed group D2D mode:

- $K = m$: When $K = m$, that is the number of spectrum resources equals to the number of D2D pairs, the proposed group D2D mode reduces to divide D2D mode.
- $K = 1$: When $K = 1$, all D2D pairs form a grand group in D2D mode and they share the whole dedicated D2D spectrum. The group D2D mode reduces to shared D2D mode.

For the rest part of this paper, we denote divided, shared and group D2D modes as **MD**, **MS** and **MG**, respectively.

### 2.2 Link Quality

In the proposed framework, the SC-FDMA technique and OFDMA technique are adopted to uplink communication and downlink communication, respectively. And SC-FDMA technique will be adopted to D2D communication according to 3GPP TR 36.843. Notice that the choice of multiple access techniques does not influence the framework as long as it provides orthogonal access.

If pair $i$ chooses D2D mode in group $k$, which means $x_{i,k} = 1$ and $\sum_{k' \in \mathcal{K} \setminus \{k\}} x_{i,k'} = 0$, we denote its signal-to-noise-ratio (SINR) in D2D group $k$ as

$$\text{SINR}_{i,k} = \frac{P_{ii}}{\sum_{j \in \mathcal{U}_d \setminus \{i\}} x_{j,k} P_{ji} + N_0}, \tag{1}$$

where $P_{ji}$ is the received power from the transmitter of pair $j$ to the receiver of pair $i$ and $N_0$ is the terminal noise at the receiver. The corresponding network utility function is denoted by $r_{i,k}$ [23], in which a logarithmic function to achievable rate under certain SINR is considered, as follows.

$$r_{i,k} = \log(w_d \log(1 + \frac{\text{SINR}_{i,k}}{\Gamma})), \tag{2}$$

where $w_d$ is the bandwidth allocated to pair $i$ for D2D communication, $\Gamma$ is the SNR gap [24] according to applied modulation and coding schemes. The reasons we applied a logarithmic utility function are multiple folds: 1) it has been know that a logarithmic utility function provides better fairness than linear ones in overall utility maximization problem [23], 2) the concavity and strictly increasing propertiy capture the user's interests in higher throughput, and 3) the diminishing returns capture the fact that users will feel great differences if its rate doubles at lower rate but no significant differences when the achievable rate is already high enough. Moreover, it is a common practice to define the network utility as the logarithmic function of the achievable data rate in the literature [23, 25].

In contrast, if pair $i$ chooses cellular mode, which means $\sum_{k \in \mathcal{K}} x_{i,k} = 0$, we let $\text{SINR}_{i,up}$ be the uplink SINR value from transmitter $i$ to BS, and $\text{SINR}_{i,down}$ be the downlink SINR value from BS to receiver $i$. Similarly, we denote the network utility under cellular mode as

$$\begin{aligned} r_{i,c} &= \log(w_c \log(1 + \frac{\text{SINR}_{i,c}}{\Gamma})) \\ &= \log(w_c \log(1 + \frac{\min\{\text{SINR}_{i,up}, \text{SINR}_{i,down}\}}{\Gamma})), \end{aligned} \tag{3}$$

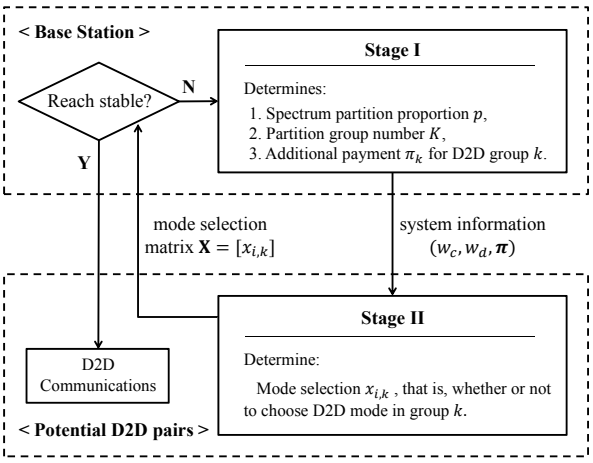



Fig. 2. A Two-Stage Dynamic Stackelberg Game

where $w_c$ is the bandwidth allocated to pair $i$ for either downlink or uplink. We select minimum value between uplink SINR and downlink SINR as cellular connection SINR [10, 11]. Notice that the same pair may achieve different network utilities under group D2D mode and cellular mode due to differences in SINR and/or allocated bandwidth.

The interference in D2D mode should be measured by the potential D2D pairs so that they can calculate the SINR and learn the transmission quality of the group. The interference in D2D mode mainly comes from other D2D pairs selecting the same group. Inspired by the measurement method proposed in [26], we propose a simple D2D interference measurement method for our framework. In our framework, a total of $K$ measurement periods is predefined at the beginning of each mode selection round. In a measurement period $k$, the transmitters of D2D pairs who select group $k$ will transmit reference signals simultaneously, while the receivers of all potential D2D pairs measure the interference level of that group. The measurement process ends after all D2D pairs have transmitted the reference signals in the defined periods according to the selected group. With this design, all potential D2D pairs can measure the interference they may experience in each group, given the current selection of other D2D pairs. Therefore, all potential D2D receivers can calculate the SINR accordingly.

### 2.3 Dynamic Stackelberg Game Model

Given a certain group number $K$, there exists a complex and interactive relationship between spectrum partition strategy for the BS and mode selection strategy for potential D2D pairs. The spectrum partition strategy should be aware of the D2D mode requests, which is reflected by $m$ and mode selection matrix $\mathbf{X}$. Nevertheless, the spectrum partition strategy also has a significant impact on the quality of D2D connections due to its influence on the allocated bandwidth to each connection, and therefore has an impact on the mode selection of each potential D2D pair. Additionally, due to the characteristic of D2D communications, selfish potential D2D pairs may be unwilling to select D2D mode when the network utility under the achievable rate is less than in cellular mode, even if it is better regarding the overall system performance and load balancing. Hence, we use game theory to analyze the proposed system.

Stackelberg game is known as a strategic game in economies in which the market leader takes action first and then the follower firms select a strategy to cope with sequentially. A dynamic Stackelberg game reaches equilibrium with dynamic strategic interactions. In this paper, we consider a two-stage dynamic Stackelberg game [27] in which the BS and potential D2D pairs dynamically interact in the leader-follower relationship, as shown in Fig. 2. We consider an one-shot game, that is, all pairs will start to communicate after the proposed dynamic Stackelberg game reaches stable state.

At Stage I, the BS acts as a leader. It determines spectrum partition proportion $p$, group number $K$ and additional D2D payment $\pi_k$ according to the previous feedback from potential D2D pairs, as shown in Fig. 2. We assume that all UEs have prepaid a fixed entrance fee to access the network. Therefore, $\pi_k$ here is additional payment for D2D communication in group $k$. Specifically, the payment vector $\boldsymbol{\pi} = (\pi_1, ..., \pi_K)$ is served as a tool to balance the loading between D2D and cellular modes while maintaining the incentive compatibility of potential D2D pairs in mode selection. At the end of Stage I, the BS announces system information $(w_c, w_d, \boldsymbol{\pi})$ to all potential D2D pairs.

The objective of the BS is to maximize the overall network utility, that is,

$$\max_{X,m} \Pi = \Pi_c + \Pi_d \tag{4}$$

$$\Pi_c = \sum_{i \in \mathcal{U}_c} r_i,\ \Pi_d = \sum_{i \in \mathcal{U}_d} [\sum_{k \in \mathcal{K}} x_{i,k} r_{i,k} + (1 - \sum_{k \in \mathcal{K}} x_{i,k}) r_{i,c}].$$

where $r_i$ is the network utility of cellular pair $i$, $\Pi_c$ and $\Pi_d$ are the network utility of all cellular pairs and potential D2D pairs, respectively.

Each potential D2D pair's objective is to maximize their own utility. We define the utility of each potential D2D pair as its network utility under selected mode minus the additional D2D payment (if exists), that is,

$$u_i = \begin{cases} r_{i,k} - \pi_k, & x_{i,k} = 1 \text{ and } \sum_{k' \in \mathcal{K} \setminus \{k\}} x_{i,k'} = 0, \\ r_{i,c}, & \sum_{j \in \mathcal{K}} x_{i,j} = 0. \end{cases} \tag{5}$$

At Stage II, potential D2D pairs are followers in the proposed Stackelberg game. They make mode selections by taking account of the system information announced by the BS and observing the expected utility under the selected mode. At the end of Stage II, potential D2D pairs report their mode selection matrix $\mathbf{X}$ to the BS.

We define a stable state as the state that no potential D2D pair has an incentive to deviate from its current mode selection anymore. This is formally defined as Nash equilibrium in game theory. The BS judges whether or not the dynamic Stackelberg game has reached stable state. If so, the proposed dynamic game is over and each pair starts to communicate. Otherwise, the BS and potential D2D pairs play another round of the proposed two-stage Stackelberg game. Here we define $t$ as the number of rounds of playing this Stackelberg game. We will testify that the proposed dynamic Stackelberg game can reach stable state, or Nash equilibrium by adopting the proposed algorithms.

## 3 Problem Formulation

In this section, we would like to analyze the two-stage Stackelberg game in general form. Following the standard backward induction process, we first check the incentive compatible conditions for potential D2D pairs when the partition proportion $p$, group number $K$, target D2D loading $m^*$, and D2D payment vector $\boldsymbol{\pi}$ are given.

Given a certain group number $K$, the bandwidth allocated for D2D communication is equally divided into $K$ slices. The BS allocated exactly one slice to each D2D group, that is,

$$w_d = \frac{Wp}{K}. \tag{6}$$

The potential D2D pairs in D2D mode within the same D2D group should share the same spectrum resource.

Each cellular pair or each potential D2D pair in cellular mode needs some bandwidth for both uplink and downlink communication. The bandwidth $w_c$ allocated to each pair for either downlink or uplink is

$$w_c = \frac{W(1-p)}{2(N-m)}. \tag{7}$$

### 3.1 Potential D2D Pairs: Mode Selection Game in Stage II

Potential D2D pairs are followers in the Stackelberg game. They should make decisions not only whether or not to choose D2D mode but also select the D2D groups they want to stay if in D2D mode. For potential D2D pair $i$, its utility described in (5) depends on whether it chooses cellular or D2D mode. A rational potential D2D pair will select the mode that maximizes its utility. Thus, potential D2D pair $i$ will select D2D mode of group $k$ when D2D link utility in D2D group $k$ is larger than both the utility in other D2D groups and the utility in cellular mode, which can be expressed as

$$\begin{aligned} x_{i,k} = 1 \text{ if and only if } k = \arg\max_j (r_{i,j} - r_{i,c} - \pi_j) \\ \text{and } r_{i,k} - r_{i,c} - \pi_k > 0, \end{aligned} \tag{8}$$

where $r_{i,j} = \log(\frac{Wp}{K}\log(1 + \frac{\text{SINR}_{i,j}}{\Gamma}))$ and $r_{i,c} = \log(\frac{W(1-p)}{2(N-m)}\log(1 + \frac{\text{SINR}_{i,c}}{\Gamma}))$, according to equations (2)(3)(6)(7). $r_{i,k}$ refers to the form of $r_{i,j}$.

To simplify the notations, we introduce two parameters

$$\begin{aligned} a_{i,j} &= \log(W\log(1 + \frac{\text{SINR}_{i,j}}{\Gamma})) \\ \text{and } a_{i,c} &= \log(W\log(1 + \frac{\text{SINR}_{i,c}}{\Gamma})). \end{aligned}$$

Substituting $a_{i,j}$ and $a_{i,c}$ back into (8), we have

$$x^*_{i,k} = \begin{cases} 1, & \text{if } k = \arg\max_j (a_{i,j} - a_{i,c} + \log\frac{2p(N-m)}{K(1-p)} - \pi_j) \\ & \quad \text{and } a_{i,k} - a_{i,c} - [\pi_k - \log\frac{2p(N-m)}{K(1-p)}] > 0. \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

The (9) are regarded as the incentive compatible conditions for potential D2D pairs to follow the specific mode selection strategy.

### 3.2 BS: Spectrum Allocation and Pricing Strategy in Stage I

The goal of the BS, who is the leader of the Stackelberg game, is to maximize total network utility of network. We first assume that the partition proportion $p$ and group number $K$ are fixed. In such a case, the BS should deal with spectrum allocation and pricing problems. The network utility of all cellular pairs in the system is given as

$$\begin{aligned} \Pi_c = \sum_{i\in\mathcal{U}_c} r_i &= \sum_{i\in\mathcal{U}_c} \log[w_c \log(1 + \frac{\text{SINR}_i}{\Gamma})] \\ &= \sum_{i\in\mathcal{U}_c} [a_i + \log\frac{(1-p)}{2(N-m)}], \end{aligned} \tag{10}$$

where $a_i = \log(W\log(1 + \frac{\text{SINR}_i}{\Gamma}))$. For cellular pair $i$, $r_i$ is the network utility and $\text{SINR}_i$ is the minimum between uplink SINR and downlink SINR.

Similarly, the network utility of all potential D2D pairs is

$$\begin{aligned} \Pi_d &= \sum_{i\in\mathcal{U}_d} [\sum_{k\in\mathcal{K}} x_{i,k} r_{i,k} + (1 - \sum_{k\in\mathcal{K}} x_{i,k}) r_{i,c}] \\ &= \sum_{i\in\mathcal{U}_d} r_{i,c} + \sum_{i\in\mathcal{U}_d}\sum_{k\in\mathcal{K}} x_{i,k}(r_{i,k} - r_{i,c}). \end{aligned} \tag{11}$$

Furthermore, the utility of the BS has been given in (4). According to (2)(3)(6)(7)(10)(11) and the definitions of $a_{i,k}$, $a_{i,j}$ and $a_{i,c}$, we simplify the utility function as

$$\begin{aligned} \Pi = &\sum_{i\in\mathcal{U}_c} [a_i + \log\frac{(1-p)}{2(N-m)}] + \sum_{i\in\mathcal{U}_d} [a_{i,c} + \log\frac{(1-p)}{2(N-m)}] \\ &+ \sum_{i\in\mathcal{U}_d}\sum_{k\in\mathcal{K}} x_{i,k}[a_{i,k} + \log\frac{p}{K} - a_{i,c} - \log\frac{(1-p)}{2(N-m)}]. \end{aligned} \tag{12}$$

The goal of the BS is to maximize (12), denoted as $\max_{X,m} \Pi$, under the incentive compatible constraint (9).

### 3.3 Incentive-aware Optimization: Primal-Dual Method

The incentive compatible constraints from potential D2D pairs in (9) in Stage II make the utility maximization problem facing by the BS in Stage I hard to be handled. Nevertheless, we observe that (9) can be integrated into Stage II to formulate as an optimization problem through primal-dual method [23, 25]. The primal formulation in (12) can be expressed in an equivalent form by introducing a set of three new variables as load metrics $N_c = |\mathcal{U}_c|$, $N_d = |\mathcal{U}_d|$ and $m = \sum_{i\in\mathcal{U}_d}\sum_{k\in\mathcal{K}} x_{i,k}$.

$$\begin{aligned} \Pi = &\sum_{i\in\mathcal{U}_d}\sum_{k\in\mathcal{K}} x_{i,k}(a_{i,k} - a_{i,c}) + \sum_{i\in\mathcal{U}_c} a_i + \sum_{i\in\mathcal{U}_d} a_{i,c} \\ &+ m\log\frac{p}{K} + (N-m)\log\frac{(1-p)}{2(N-m)}. \end{aligned} \tag{13}$$

The coupling constraint $m = \sum_{i\in\mathcal{U}_d}\sum_{k\in\mathcal{K}} x_{i,k}$ motivates us to turn to the Lagrangian dual decomposition method whereby a dual variable vector $\boldsymbol{\mu} = (\mu_1, ..., \mu_K)$ introduced for our utility function $\Pi$. The dual problem is

$$\min_{\boldsymbol{\mu}} D(\boldsymbol{\mu}) = \min_{\boldsymbol{\mu}} \{\Pi + \sum_{k\in\mathcal{K}} \mu_k (m_k - \sum_{i\in\mathcal{U}_d} x_{i,k})\}. \tag{14}$$

To solve the dual optimization problem of (14), we decouple it into two sub-problems:

$$\mathbf{D}: \min_{\boldsymbol{\mu}} D(\boldsymbol{\mu}) = f_{\mathbf{X}}(\boldsymbol{\mu}) + g_m(\boldsymbol{\mu}), \tag{15}$$

$$f(\boldsymbol{\mu}) = \max_{\mathbf{X}} \{ \sum_{i \in \mathcal{U}_d} \sum_{k \in \mathcal{K}} x_{i,k}(a_{i,k} - a_{i,c} - \mu_k) \}, \ \mathbf{X} = [x_{i,k}], \tag{16}$$

$$g(\boldsymbol{\mu}) = \max_{m} \{ m \log \frac{p}{K} + (N-m) \log \frac{(1-p)}{2(N-m)} + \sum_{k \in \mathcal{K}} \mu_k m_k \}. \tag{17}$$

Here we ignore the value of $(\sum_{i \in \mathcal{U}_c} a_i + \sum_{i \in \mathcal{U}_d} a_{i,c})$, which is only affected by background noise.

### 3.4 Spectrum Partition Strategy

Next, we relax the assumption that $p$ is given in Stage I. An efficient spectrum partition strategy should address the requirements from both cellular and D2D pairs and the loading of both modes in the system.

**Proposition 1.** *The optimal spectrum partition proportion is* $p^* = \frac{m}{N}$.

*Proof.* Recall that the original dual optimization function (14) is a differentiable concave function of $p$ given $m$ and $\mathbf{X}$ are fixed. Thus, the optimal $p^*$ can be found by checking its first order condition as

$$\frac{m}{p^*} - \frac{N-m}{1-p^*} = 0 \ \Rightarrow \ p^* = \frac{m}{N}. \tag{18}$$

□

In the proposed system, we propose a partition strategy for spectrum partition proportion by applying (18). It should be noted that $m$ is the number of potential D2D pairs who select D2D mode. For these potential D2D pairs, the total bandwidth $W\frac{m}{N}$ will be allocated to them. Consider conventional cellular system, the total bandwidth is equally divided $N$ into slices, which means that every UE pair is allocated $\frac{W}{N}$ bandwidth. In this way, the total bandwidth allocated for these potential D2D pairs is the same as $W\frac{m}{N}$. That is to say, the same size bandwidth will be allocated to them regardless of any mode selections they made. What's more, the rest bandwidth allocated for cellular pairs and potential D2D pairs in cellular mode, which have total number $N_c + N_d - m = N - m$, is $W(1-p^*) = W\frac{N-m}{N}$. Then each pair of them owns $\frac{W}{N}$ bandwidth. In such a case, the bandwidth allocated to existing cellular user will remain unaffected regardless of the choice of potential D2D pairs. Therefore, the service quality of cellular users and potential D2D pairs who stay in cellular mode will not be affected under the proposed strategy when we introduce D2D mode into conventional cellular system.

## 4 Divided D2D Mode & Shared D2D Mode

In this section, we will analyze the two special cases, divided and shared D2D modes, which provide some helpful propositions and contribute to the solution of the proposed group D2D mode. The optimal configuration of the proposed D2D-enabled cellular system in these two cases will be derived in this section.

### 4.1 Divided D2D Mode (MD)

The proposed group D2D mode reduces to divided D2D mode when $K = m$. Here we assume that the payment $\pi_k$ for any $k \in \mathcal{K}$ should be equal, that is,

$$\forall k \in \mathcal{K}, \ \pi_k = \pi. \tag{19}$$

Here we call $\pi$ as **real payment**, since potential D2D pairs need to actually pay $\pi$ to access D2D communication. We will show later in Section 5 that such a pricing scheme in fact is the necessary condition for optimal solution in general case when the group number is larger than 1, including the case that k = m here.

**Proposition 2.** *D2D pairs prefer to own an individual spectrum resource when* $K = m$.

*Proof.* The number of spectrum resources is exactly equal to the number of D2D pairs when $K = m$. We assume that D2D pair $i$ can observe the group selections of other D2D pairs. In other words, D2D pair $i$ can select its D2D group by considering the selections of other pairs. We know that total $(m-1)$ pairs can form at most $(m-1)$ groups, that is, there must exist at least one spectrum resource, denoted by resource $k^{'}$, owned by no one.

Under the proposed framework, a D2D pair in any D2D group pays the same additional D2D payment $\pi$ and shares D2D spectrum with the same bandwidth $w_d = Wp/K$. To achieve maximal utility as (5), D2D pairs would like to select a D2D group with best SINR quality. According to the definition of SINR in (1), we have

$$\begin{aligned} \text{SINR}_{i,k} &= \frac{P_{ii}}{\sum_{j \in \mathcal{U}_d \setminus \{i\}} x_{j,k} P_{ji} + N_0} \\ &\leq \frac{P_{ii}}{N_0} = \text{SINR}_{i,k'}, \quad (\forall k \in \mathcal{K} \backslash k^{'}). \end{aligned}$$

In this case, pair $i$ prefer to select D2D group $k^{'}$ rather than other groups. Hence, D2D pairs prefer to own an individual spectrum resource to avoid intra-group interference. □

By Proposition 2, we directly assume that the BS allocates spectrum resources to D2D pairs individually. Furthermore, potential D2D pairs only choose whether or not to select D2D mode.

Due to the fact that no group selection under divided D2D mode, with the coupling constraint $m = \sum_{i \in \mathcal{U}_d} x_{i,d}$, the dual optimization of (14) can be simplified as

$$\min_{\mu} D(\mu) = \min_{\mu} \{ \Pi + \mu(m - \sum_{i \in \mathcal{U}_d} x_{i,d}) \}, \tag{20}$$

where $\mu$ is a dual variable and $x_{i,d}$ is mode selection indicator of potential D2D pair $i$ under divide mode. Accordingly, two sub-problems are rewritten as

$$f(\mu) = \max_{\mathbf{X}} \{ \sum_{i \in \mathcal{U}_d} x_{i,d}(a_{i,d} - a_{i,c} - \mu) \}, \ \mathbf{X} = [x_{i,d}], \tag{21}$$

$$g(\mu) = \max_{m} \{ m \log \frac{p}{K} + (N-m) \log \frac{(1-p)}{2(N-m)} + \mu m \}. \tag{22}$$

Here $a_{i,d} = \log(W \log(1 + \frac{\text{SINR}_{i,d}}{\Gamma}))$, where $\text{SINR}_{i,d}$ is the SINR of potential D2D pair $i$ in divide D2D mode.

**Proposition 3.** *Dual variable $\mu$ can be regarded as virtual payment and transformed into the real payment $\pi$ by*

$$\pi = \mu + \log \frac{2p(N-m)}{m(1-p)}. \quad (23)$$

*Proof.* Recall the mode selection of potential D2D pair $i$ by (9), we rewrite the incentive compatible constraint under divided mode as

$$x^*_{i,d} = \begin{cases} 1, & \text{if } a_{i,d} - a_{i,c} - [\pi - \log \frac{2p(N-m)}{m(1-p)}] > 0, \\ 0, & \text{otherwise.} \end{cases} \quad (24)$$

Back to the dual problem when $\mu$ is fixed, from (21) we have

$$x^*_{i,d} = \begin{cases} 1, & \text{if } a_{i,d} - a_{i,c} - \mu > 0, \\ 0, & \text{otherwise.} \end{cases} \quad (25)$$

We observe that (25) closely (but not exactly) resembles constraint (24). Dual variable $\mu$ can be considered as **virtual payment** in (25). Thus, the virtual payment $\mu$ can be transformed into the real payment $\pi$ announced by the BS with the pricing strategy in (23). □

Notice that the real D2D loading $m$ is not necessary equal to the target D2D loading $m^*$ defined by the BS. The realized loading depends on the selections of potential D2D pairs, which follow their own rationality. The virtual payment $\mu$ can regulate the number of D2D pairs to fit the target. According to subgradient method [28], $\mu$ is updated by

$$\mu_{t+1} = \mu_t - \delta_t(m^* - m) = \mu_t - \delta_t(m^* - \sum_{i \in \mathcal{U}_d} x^*_{i,d}), \quad (26)$$

where $\delta_t > 0$ is a dynamical stepsize sequence.

**Proposition 4.** *$g(\mu)$ is a concave function and the optimal solution $m^*$ is solvable.*

*Proof.* Substituting $K = m$ into function $g(\mu)$ from (22), the first and second derivatives of $g(\mu)$ of $m$ will be

$$\frac{\partial g(\mu)}{\partial m} = \log \frac{2p}{(1-p)} + \log \frac{(N-m)}{m} + \mu,$$
$$\frac{\partial^2 g(\mu)}{\partial m^2} = -[\frac{1}{(N-m)} + \frac{1}{m}] < 0.$$

Obviously, $g(\mu)$ is a concave function of $m$ because its second derivative is always negative. Notice that $m$ must be integer and $m \in [0, N_d]$. Given $\mu$ and $p$ are fixed, the optimal solution $m^*$ can be solved by checking the first order condition of $g(\mu)$ as

$$\hat{m} = \arg\max_m g(u), \quad m \in \{\left\lfloor \frac{\eta}{\eta+1} N \right\rfloor, \left\lceil \frac{\eta}{\eta+1} N \right\rceil\},$$
$$\eta = \frac{2pe^{\mu}}{1-p}, \quad m^* = \begin{cases} 0, & \hat{m} \le 0, \\ \hat{m}, & 0 < \hat{m} < N_d, \\ N_d, & N_d \le \hat{m}. \end{cases} \quad (27)$$

□

When the BS adopts the pricing strategy by (23), the actions of all potential D2D pairs will be absolutely governed by the BS, that is, the incentive compatible conditions in (24) are satisfied.

### 4.2 Shared D2D Mode (MS)

The proposed group D2D mode reduces to shared D2D mode when $K = 1$. The shared D2D mode framework has a similar structure to the divided D2D mode except that the D2D spectrum partitioned by the BS is shared by all D2D pairs, that is,

$$w_d = Wp. \quad (28)$$

The main difference of shared D2D mode from divided D2D mode is that potential D2D pairs in shared D2D mode will interfere each other. When D2D loading $m$ increases, the SINR of each D2D pair will generally decrease due to interference from other D2D pairs. Again, a rational potential D2D pair will select D2D mode if and only if the network utility minus the payment in D2D mode maximizes its utility. Additionally, the SINR in D2D mode should be higher than a minimum threshold under the interference so that the link can be established. Without loss of generality, we impose a minimum SINR constraint for potential D2D pair $i$ in shared D2D mode: $\text{SINR}_{i,d} \ge T_i$, where $T_i$ is a threshold determined by the communication system. On the other hand, potential D2D pair $i$ determines its own SINR threshold $T_i$ to satisfy its quality requirement of D2D link. Accordingly, the original incentive compatible conditions (9) will be modified as below.

$$x^*_{i,d} = \begin{cases} 1, & \text{if } \text{SINR}_{i,d} \ge T_i \text{ and} \\ & a_{i,d} - a_{i,c} - [\pi - \log \frac{2p(N-m)}{(1-p)}] > 0, \\ 0, & \text{otherwise.} \end{cases} \quad (29)$$

In shared D2D mode framework, each potential D2D pair $i$ can adopt mode selection strategy by (29). Theoretically, the total number of user pairs in D2D mode will be under a certain threshold.

Similar to the analysis of divided D2D mode, we can get the dual problem the same as (20)(21)(22) with $K = 1$ under shared D2D mode. The BS announces payment $\pi$ and target D2D loading $m^*$ under shared D2D mode as

$$\pi = \mu + \log \frac{2p(N-m)}{(1-p)}, \quad (30)$$

$$\hat{m} = \arg\max_m g(u), \quad m \in \{\left\lfloor N - \frac{(1-p)}{2pe^{\mu+1}} \right\rfloor, \left\lceil N - \frac{(1-p)}{2pe^{\mu+1}} \right\rceil\},$$
$$m^* = \begin{cases} 0, & \hat{m} \le 0, \\ \hat{m}, & 0 < \hat{m} < N_d, \\ N_d, & N_d \le \hat{m}. \end{cases} \quad (31)$$

In addition, the virtual payment $\mu$ is updated the same as (26).

### 4.3 Dynamic Stackelberg Game Under MD or MS

The complete dynamic Stackelberg game, adopting primal-dual pricing and partitioning update algorithm under divided D2D mode or shared D2D mode, is described in Algorithm 1.

**Proposition 5.** *The convergence of the proposed primal-dual algorithm can be guaranteed.*

*Proof.* Please refer to Appendix A for the proof. □

**Proposition 6.** *The framework is cheat-proof, that is, potential D2D pairs would like to truthfully report their actual mode selections to the BS.*

*Proof.* The proposed framework requires mode selection results instead of CQI from potential D2D pairs. We assume that only pair $i$ intends to cheat to the BS. According to

**Algorithm 1** The Dynamic Stackelberg Game Under MD or MS

**Initialization**: Set $m = randi([0, N_d])$. Set $p = \frac{m}{N}$. Set $\mathbf{X} = [x_{i,d}] = 0$. Set $\mu_0 = 0 - \delta_0(m - \sum_{i \in \mathcal{U}_d} x_{i,d})$.
**repeat**
  The BS updates $m$ according to MD-(27) / MS-(31).
  The BS updates $p$ according to (18).
  The BS updates $\mu$ according to (26).
  The BS computes $\pi$ according to MD-(23) / MS-(30).
  The BS announces system information $(w_c, w_d, \pi)$.
  **for** each $i \in \mathcal{U}_d$ **do**
    Potential D2D pair $i$ determines $x_{i,d}$ according to MD-(24) / MS-(29), and then reports its mode selection to the BS.
  **end for**
**until** $m = \sum_{i \in \mathcal{U}_d} x_{i,d}$.

the description of Algorithm 1, update value of $m$ and $p$ for next stage $t+1$ will be the same regardless mode selection results the BS has at stage $t$. While pair $i$ can guess the virtual payment $\mu_{t+1}$ of stage $t+1$. That is, $\mu_{t+1} = \mu_d = \mu_t - \delta(m - \sum_{j \neq i} x_{j,d} - 1)$ if pair $i$ chooses D2D mode, Otherwise $\mu_{t+1} = \mu_c = \mu_t - \delta(m - \sum_{j \neq i} x_{j,d})$. Obviously, $\mu_d - \delta = \mu_c$ and pair $i$ can predict all information the BS updates for next stage $t+1$. Unfortunately, the CQI of other pairs is private information to pair $i$, which means that pair $i$ could not predict $\sum_{i \in \mathcal{U}_d} x^*_{i,d}$ at stage $t+1$. If a potential D2D pair first misreports its mode selection as cellular mode at stage $t$ and then reports its real mode (D2D mode) at stage $t+1$, the D2D payment will be reduced at stage $t+1$ by the BS. However, the change in reports will be recognized by the BS and identifying that the configuration is not stable yet. A payment adjustment according to the real mode of this UE will be performed. Eventually, the D2D payment will be adjusted to the same value as the one if this UE reports real mode at first place. Moreover, misreporting may affect the convergence of our framework. However, the proposed framework is an one-shot game, that is, all pairs will start to communicate after the proposed dynamic Stackelberg game reaches stable state. This means that the user who misreports its mode selection will experience an extra delay in the communications. Given that the D2D payment will be the same eventually and there will be an additional delay in communication, it will truthfully report its mode selection. In a word, pair $i$ cannot manipulate the final result of the proposed game and has to truthfully report its mode selection. □

## 5 Group D2D Mode

After discussing two special cases, we relax $K$ for any integer with $K \in [1, N_d]$, that is the group D2D mode we have proposed. By relaxing the numerical characteristic of $m_k$ that $m_k$ should be integer only, we can have a proposition as below.

**Proposition 7.** *The necessary condition for optimal solution to* $\mathbf{X}$ *is that all dual variables* $\mu_k$ $(k \in \mathcal{K})$ *adopt the same value.*

*Proof.* Please refer to Appendix B for the proof. □

Due to proposition 7, we introduce $\mu$ as the value of $\mu_k$, that is,

$$\forall k \in \mathcal{K},\ \mu_k = \mu. \tag{32}$$

**Algorithm 2** Reinforcement Learning Approach

**Initialization**: Set $Q_{i,k}(t) = 0, n_{i,k} = 0$. Set $\beta$.
**repeat**
  **if** $\mathcal{S}_i = \{j \mid \text{SINR}_{i,j} \geq T_i \wedge r_{i,j} - r_{i,c} - \pi > 0\} \neq \emptyset$ **then**
    **if** $rand() \leq \gamma$ **then**
      Randomly choose D2D mode of group $k \in \mathcal{S}_i$ {Exploration}.
    **else**
      Choose D2D mode of group $k = \arg\max_j Q_{i,j}(t)$ {Exploitation}.
    **end if**
    $n_{i,k} = n_{i,k} + 1$, $\alpha_{i,k} = 1/n_{i,k}$.
    Update $Q_{i,k}(t+1) = (1 - \alpha_{i,k}) Q_{i,k}(t) + \alpha_{i,k}(r_{i,k} + \beta \max_j Q_{i,j}(t))$.
    Set $x^*_{i,k} = 1$ and $\sum_{k' \in \mathcal{K} \setminus \{k\}} x^*_{i,k'} = 0$.
  **else**
    Choose cellular mode. Set $\sum_{j \in \mathcal{K}} x^*_{i,j} = 0$.
  **end if**
**until** The dynamic Stackelberg game is end.

We will show later that the proposed framework can achieve better performance with the help of Proposition 7.

Here we denote $D(\boldsymbol{\mu})$ of (14) as $D(\mu)$ and $g(\boldsymbol{\mu})$ of (17) can also be simplified the same as (22). Furthermore, we apply (32) to (16) and then have

$$f(\boldsymbol{\mu}) = \max_{\mathbf{X}} \{ \sum_{i \in \mathcal{U}_d} \sum_{k \in \mathcal{K}} x_{i,k}(a_{i,k} - a_{i,c} - \mu) \},\ \mathbf{X} = [x_{i,k}]. \tag{33}$$

Similar to the derivation in Section 4, we obtain the relationship between (9) and (33), that is,

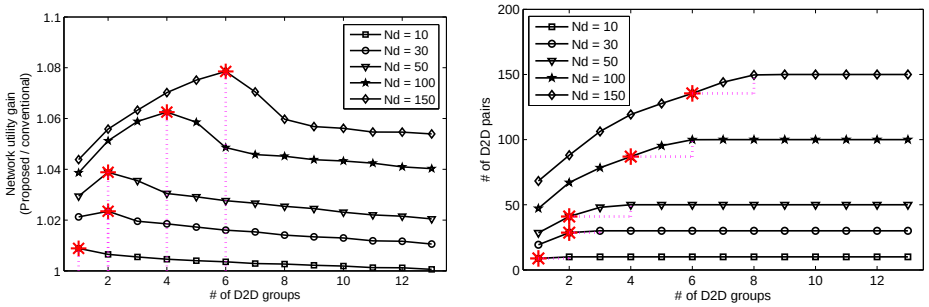
$$\pi_k = \log \frac{2p(N-m)}{K(1-p)} + \mu,\ \forall k \in \mathcal{K}. \tag{34}$$

Hence, $\pi_k$ should be equal for any $k \in \mathcal{K}$, which means that the proposed framework applies single additional payment under group D2D mode and is consistent with divided and shared D2D modes. We denote $\pi_k$ by $\pi$ for any $k \in \mathcal{K}$. For potential D2D pairs, they pay the same entrance $\pi$ to access overlay D2D communication and have the freedom to select a D2D group themselves by considering SINR quality.

### 5.1 Reinforcement Learning Approach

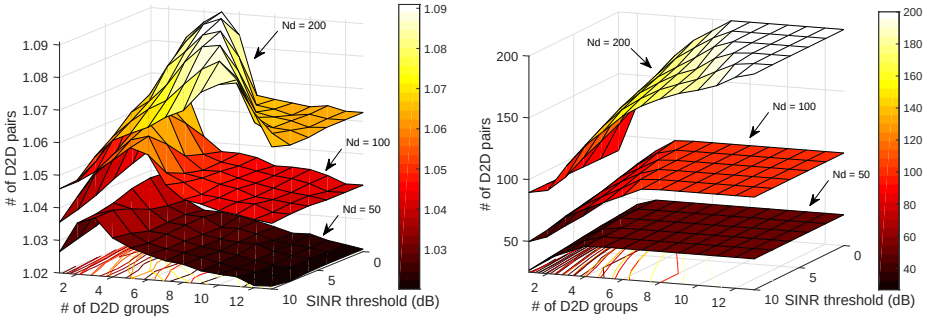
Recall that the BS announces the same additional payment $\pi$ for entering any D2D groups and the same bandwidth $w_d$ for each pair to communicate in D2D mode, potential D2D pairs need pay more attention to the SINR quality in each D2D group. Unlike divided D2D mode, under which the BS allocates spectrum resources to D2D pairs, they should determine their D2D groups by themselves under group D2D mode. Nevertheless, the optimality of their selections heavily depends on the choices of other pairs due to the intra-group interference. Naive solutions such as dynamic best response may fail to reach a stable state as a D2D group, which has best SINR quality in current stage, may be worse in the next stage, and vice versa. Therefore, we propose a reinforcement learning approach [29, 30] to guarantee that the proposed system can reach stable and potential D2D pairs can also achieve optimal long-term network utility. In addition, the smoother response in reinforcement learning helps reach the stable state by avoiding the ping-pong effects in dynamic best response approach.

Reinforcement learning [31] is one of the most popular learning algorithms by interacting with an environment and




(a) Network utility gain (b) # of D2D pairs

Fig. 3. Simulation with the increase of group number under different $N_d$: $N_c = 100$; $T_i = 6$ dB.




(a) Network utility gain (b) # of D2D pairs

Fig. 4. Simulation with the increase of group number and SINR threshold under different $N_d$ = 50, 100 and 200, respectively: $N_c = 100$.

has been widely used in the networking and communication areas, such as the dynamic provider selection problem in wireless resource management[32]. Here we focus on Q-learning, which is a model-free reinforcement learning technique. Q-learning algorithm works by learning optimal action-selection policy with the environment. For details, an agent $i$ learns a Q-function that maps the current stage $t$ and action $k$ to a utility value $Q_{i,k}(t)$, which can predict the total future discounted reward. At stage $(t+1)$, the agent plays an action $k$ due to its current action-selection policy, updates Q-value $Q_{i,k}(t+1)$ and then observes a feedback from the environment. There are two steps, exploitation and exploration [33]. An agent in exploitation step exploits the current learning knowledge by selecting one of the actions that has maximal Q-value. On the contrary, the agent in exploration step randomly selects an action to build its knowledge about the environment. In this approach, potential D2D pairs are not required to have the knowledge of the exact state of the entire network system, which is a desired property from the implementation perspective.

The proposed reinforcement learning approach [1] is described in Algorithm 2. $Q_{i,k}(t)$ denotes expected discounted Q-value of potential D2D $i$ in D2D group $k$ at time $t$, which is used to maintain its knowledge about D2D group $k$. The $r_{i,k}$ here is regarded as the current reward by selecting D2D group $k$. $\gamma$ is known as temperature of exploration, which means the agent has probability $\gamma$ to perform exploration step. There exists dynamic temperature adjustment mechanism, like $\varepsilon$-greedy [34], to harmonize the trade-off between exploration and exploitation. $\alpha_{i,k}$ denotes the learning rate of potential D2D $i$ in D2D group $k$, which controls the speed of adjustment of Q-value. We define $\alpha_{i,k} = \frac{1}{n_{i,k}}$, where $n_{i,k}$ is the number of times that potential D2D $i$ visits D2D group $k$ [35]. $\beta$ denotes the discount factor that determines the importance of future rewards. For next iteration, potential D2D pairs need update new Q-value with $\beta$ discount of previous Q-value. Each potential D2D pair learns and adapts its mode decision by adopting the proposed approach independently. Notice each D2D group may have more than one D2D pair, that is, the D2D pairs in the same D2D group interfere each other just like under shared D2D mode. Therefore, the SINR constraint derived from shared D2D mode is still applicable to group D2D mode. Here it is modified as $\text{SINR}_{i,j} \geq T_i$.

### 5.2 Dynamic Stackelberg Game Under Fixed MG

Given the BS has adopted fixed group number $K$ before playing the proposed Stackelberg game, we can derive payment $\pi$ and target D2D loading $m^*$ from previous analysis as

$$\pi = \mu + \log \frac{2p(N-m)}{K(1-p)}, \tag{35}$$

$$\hat{m} = \arg\max_m g(u), \quad m \in \{\left\lfloor N - \frac{K(1-p)}{2pe^{\mu+1}} \right\rfloor, \left\lceil N - \frac{K(1-p)}{2pe^{\mu+1}} \right\rceil\},$$
$$m^* = \begin{cases} 0, & \hat{m} \leq 0, \\ \hat{m}, & 0 < \hat{m} < N_d, \\ N_d, & N_d \leq \hat{m}. \end{cases} \tag{36}$$

1. An additional state (D2D pairs who simultaneously utilize group $k$ when pair $i$ selected group $k$) could be useful when the network is more dynamic and a rapid response to loading changes is necessary. Nevertheless, it is not necessary in this work since our goal is to reach the stable state, or Nash equilibrium. This could be considered as the future extension of this framework.

**Algorithm 3** The Dynamic Stackelberg Game Under MG with Fixed Group Number $K$

1: **Initialization**: Set $m = randi([0, N_d])$. Set $p = \frac{m}{N}$. Set $\mathbf{X} = [x_{i,k}] = 0$. Set $\mu_0 = 0 - \delta_0(m - \sum_{i\in\mathcal{U}_d}\sum_{k\in\mathcal{K}} x_{i,k})$. Set $Q_{i,k}(t) = 0$. Set $\beta, K$.
2: **repeat**
3: The BS updates $m$ according to (36).
4: The BS updates $p$ according to (18).
5: The BS updates $\mu$ according to (37).
6: The BS computes $\pi$ according to (35).
7: The BS announces system information $(w_c, w_d, \pi)$.
8: **for** each $i \in \mathcal{U}_d$ **do**
9: Potential D2D pair $i$ determines $x_{i,k}$ by **Algorithm 2** from step 3 to step 14, and then reports its mode selection to the BS.
10: **end for**
11: **until** $m = \sum_{i\in\mathcal{U}_d}\sum_{k\in\mathcal{K}} x_{i,k}$.

According to subgradient method, $\mu$ is updated by

$$\mu_{t+1} = \mu_t - \delta_t(m^* - \sum_{i\in\mathcal{U}_d}\sum_{k\in\mathcal{K}} x^*_{i,k}). \quad (37)$$

Given fixed $K$, the dynamic Stackelberg game under group D2D mode is described in Algorithm 3. Obviously, the mode selection strategy of potential D2D pairs is reinforcement learning instead of directly selecting the D2D group with best SINR quality.

Notice that with the same number of potential D2D pairs selecting D2D mode, divided D2D mode can guarantee transmission quality for each D2D pair, while shared D2D mode can support larger spectrum bandwidth and achieve higher spectrum utilization efficiency. These two modes exactly represent extreme cases of spectrum partitioning of D2D group number. There may exist a trade-off between transmission quality and spectrum utilization efficiency, that is, we can find an **optimal group number** so that the proposed system reaches maximal total network utility. We would like to verify our conjecture through simulations. Fig. 3 shows the number of D2D pairs and network utility gain with the increase of group number. In Fig. 3(a), we observe that each curve follows the definition of quasi-concave function [36] and has one and only one peak value. This confirms our assumption on the trade-off and suggests that an optimal group number can be found by iterative-update approaches. To further support our claim that quasi-concavity is satisfied in general, we adopt a more comprehensive set of numerical simulations. We observe that the property of quasi-concavity always exists with different number of D2D pairs ($10 \sim 200$) and different SINR threshold (0 dB $\sim$ 10 dB). A part of results are shown in Fig. 4 under different number of potential D2D pairs $N_d$ = 50, 100 and 200, respectively. It shows the quasi-concavity is more significant when the number of D2D users increases.

### 5.3 Dynamic Stackelberg Game Under Dynamic MG

Since the relation of group number to network utility is quasi-concave, the optimal group number can be found by iterative-based search approaches. A naive solution is MG-Search (**MG-S**) that the dynamic Stackelberg game starts with one group and increases group number by one at each convergence state. At each convergence state, all potential D2D pairs are required to report their CQI to the BS so that the total network utility can be known. MG-S approach can obtain the optimal group number when the total network utility starts to decrease. Finally, MG-S approach plays the Stackelberg game by optimal group number. Obviously, MG-S approach works slow and needs quite a number of signal exchanges between the BS and UEs. Additionally, it requires feedback regarding utility experiences by UEs and therefore is incompatible to our framework. Nevertheless, we treat it as a performance upper bound in the D2D-enabled system.

**Algorithm 4** The Dynamic Stackelberg Game Under MG with Dynamic Group Number $K$

1: **Initialization**: Set $m = randi([0, N_d])$. Set $p = \frac{m}{N}$. Set $\mathbf{X} = [x_{i,k}] = 0$. Set $\mu_0 = 0 - \delta_0(m - \sum_{i\in\mathcal{U}_d}\sum_{k\in\mathcal{K}} x_{i,k})$. Set $Q_{i,k}(t) = 0$. Set $\beta, K_{min}, K_{max}$. Set $K_0 = K_{min}$. Set $K = K_{max}$.
2: **repeat**
3: Play **Algorithm 3** from step 2 to step 10.
4: **if** $m == N_d$ **then**
5: $K_{max} = K$.
6: **else**
7: $K_{min} = K$.
8: **end if**
9: $K_0 = K$. $K = \lceil (K_{min} + K_{max})/2 \rceil$.
10: **until** $K_0 = K$.
11: Target group number is $K = K_0 - gap(\mathcal{X}, N_d)$.
12: Play **Algorithm 3** from step 2 to step 10 under group number $K$.

Here we recall the effects with the increase of group number shown in Fig. 3. In Fig. 3(b), the number of D2D pairs continually increases until the number of groups is larger than a certain value, which called **saturated group number**. All potential D2D pairs will select D2D mode if the number of groups is greater than or equal to the saturated group number. Specifically, the red asterisk on each curve shows the corresponding number of D2D pairs when the proposed system partitions the dedicated D2D spectrum by optimal group number. Obviously, there exists a gap between optimal group number and saturated group number. From the observations, the gap becomes wider with the increase of potential D2D pairs. For example, the optimal group number and saturated group number are 6 and 8, respectively, when $N_d = 150$, as shown in Fig. 3(b). Here the gap is 2, which is wider than the setting of $N_d = 30$.

By the above analysis, we propose a simple approach, called MG-Dynamic (**MG-D**), to play the Stackelberg game and meanwhile find target group number, which is defined as saturated group number minus system-dependent gap. The complete dynamic Stackelberg game by adopting MG-D approach is described in Algorithm 4. $K_{min}$ and $K_{max}$ are the minimal and maximal number of D2D group that the proposed framework desires to support under group D2D mode. $gap(\mathcal{X}, N_d)$ is the system-dependent gap we have discussed, which is related to both UE distribution $\mathcal{X}$ and number $N_d$ of potential D2D pairs. To reach system optimization, the service provider can adjust this gap variable according to the system information. MG-D has an advantage over MG-S that it doesn't need any CQI feedback from potential D2D pairs.

**Proposition 8.** *The framework under the proposed group D2D mode is also cheat-proof.*

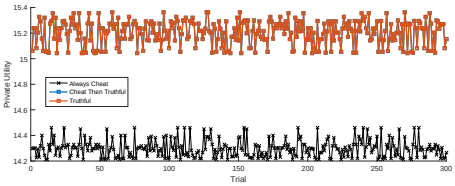
TABLE 2
Average private utility

| | Always Cheat | Cheat Then Truthful | Truthful |
|---|---|---|---|
| MD | 14.2978 | 15.2050 | 15.2050 |
| MS | 14.3045 | 16.6948 | 16.6959 |
| MG | 14.3949 | 16.0786 | 16.1405 |

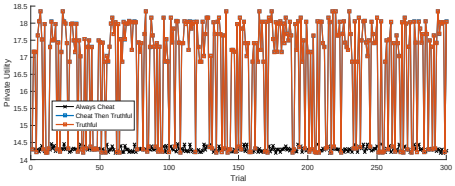
*Proof.* Since the proposed reinforcement learning has been shown to converge to the equilibrium of the game according to previous discussions, we only need to show that truth-telling strategy indeed is the equilibrium strategy in the proposed game model. Fortunately, we find that the proof of cheat-proofness under MD/MS also works under MG. We first define the cheating action in the proposed framework. If a potential D2D pair chooses to cheat, its choice will be: 1) select the mode which achieves lower private utility under MD/MS/MG, 2) specifically when selecting D2D mode under MG, select the D2D group which has minimal Q-value and update its Q-value with minimal Q-value of D2D groups.

Furthermore, we define three different strategies, that is, *Always Cheat*, *Cheat Then Truthful* and *Truthful*. Under Always Cheat strategy, the pair will always choose to cheat; under Cheat Then Truthful strategy, the pair will play few rounds of cheating and then return to tell truth; under Truthful strategy, the pair will tell truth all the time. The key to guarantee the cheat-proofness is to check if the potential D2D pairs can derive more utility by choosing actions other then the truthful one. Nevertheless, the truthful one in any mode actually is the optimal action the pair can choose to maximize its own utility. One possibility of cheating is to influence service payment of D2D mode through suboptimal choices in previous rounds. This is the Cheat strategy we propose above. Nevertheless, this suggests that the cheating pair always receives a suboptimal utility unless it turns to tell the truth. Therefore, always cheating will not be the ration choice of a rational potential D2D pair. The last possibility is that the pair cheats in number of rounds but turn to tell the truth in latter rounds. This is the Cheat Then Truthful strategy we propose above. Nevertheless, the change in the action will be a signal to the BS that the system is still not in the stable state, and the update in payment and actions of other potential D2D pairs will follow up. The system will eventually converge back to the operation point that all potential D2D pairs play Truthful starting from the beginning. Notice that this argument applies in all proposed modes. Therefore, the cheat-proofness of the framework holds in all proposed modes. □

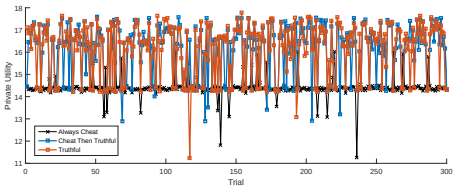
In order to strengthen our statement, we further evaluate the private utility of potential D2D pairs in three different strategies in all three modes through simulations. In our simulations, we randomly select a potential D2D pair that adopts three different strategies, respectively, and let other pairs always tell truth. The private utilities of the selected pair in different communication systems are shown in Fig. 5, and their average values are shown in Table 2. We observe that in most cases the pair achieves lower private utility under Always Cheat, which is reasonable since the pair always chooses suboptimal action and therefore receive lower utility. For Cheat Then Truthful strategy, on the other hand, it leads to the same results as the Truthful strategy. Notice that Truthful strategy under MG may not always achieve the best private utility is because the proposed algorithm is a learning-based approach with stochastic behaviors (exploration step and exploitation step). The average performance, as we illustrated in Table 2, shows that Truthful strategy and Cheat Then Truthful strategy actually perform identically. In conclusion, the proposed framework is cheat-proof since cheating will give no extra benefit to the potential D2D pair.




(a) Divided Mode (MD)




(b) Shared Mode (MS)




(c) Group Mode (MG) with two D2D groups ($K = 2$)

Fig. 5. Private utility of the selected pair in different communication systems: $N_d = 30$, $N_c = 100$.

### 5.4 Overhead & Convergence rate

As shown in Fig. 2, the BS announces system information $(w_c, w_d, \boldsymbol{\pi})$ to all potential D2D pairs at the end of Stage I. The signaling overhead is $3 \times 4$ bytes $= 96$ bits. At the end of Stage II, potential D2D pairs report their mode selection matrix $\mathbf{X} = [x_{i,k}]$ to the BS. The signaling overhead is $N_d K$ bits, where $N_d$ is the number of potential D2D pairs, $K$ is the number of D2D groups and $K \in [K_{min}, K_{max}]$. Therefore, the total signaling overhead in each iteration will be $(N_d K + 96)$ bits. Notice that it is reduced to $(N_d + 96)$ bits under MD or MS since potential D2D pairs only choose whether or not to select D2D mode under MD or MS.

The complexity analysis of Algorithm 1-4 are very difficult since Algorithm 1-4 are all in semi-distributed manner and their convergence rate is related to both UE distribution and the number of potential D2D pairs. Nevertheless, the convergence rate is centrally controlled by the BS, which means that the service provider can control the required iterations and signal overhead according to the network state. Specifically, the convergence rate of Algorithm 1 is related to the dynamic stepsize sequence $\{\delta_t\}$. The convergence rate of

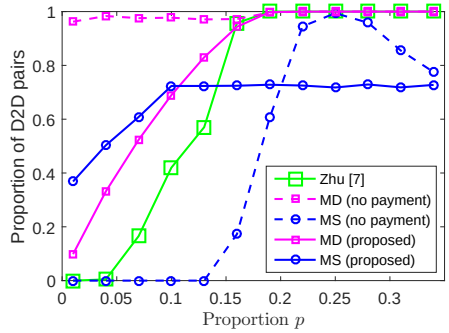
Algorithm 2 is also related to the temperature of exploration $\gamma$ and discount factor $\beta$. The convergence rates of both Algorithm 3 and algorithm 4 are also related to $\{\delta_t\}$, $\gamma$ and $\beta$. The service provider may adjust the step size accordingly in order to strike a balance between solution optimality and convergence rate.

### 5.5 Discussion about Convergence

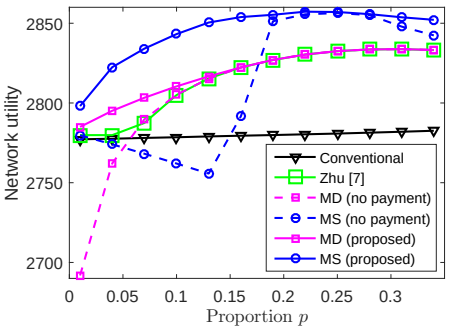
We find that the challenges of a general convergence proof for multi-agent Q-learning come from the dynamic environment. In general, the environment is non-stationary due to adaptation of other agents. Most discussions about the convergences of Q-learning in multi-agent systems are limited in two-agent game [37, 38]. In [37], the problem of policy learning in multi-agent environments using the stochastic game framework is investigated. Rationality and convergence are introduced as two properties for a learning agent when in the presence of other learning agents. The empirical results of the convergence of the proposed algorithm are presented in a number and variety of domains. Nevertheless, a theoretic proof is lacking.

In addition, most state-of-the-art multi-agent extensions of Q-learning require knowledge of other agents' actions, payoffs and Q-functions, or in other words, other agents' information is fully observable [39, 40]. This assumption limits their practicality as they require too much information which is usually unavailable or takes a significant cost to derive. In [39], a new Markov model, called multi-action replay process (MARP), is proposed for multi-agent coordination. A multi-agent Q-learning algorithm is then constructed as a cooperative reinforcement learning algorithm. However, when discussing the convergence of multi-agent Q-learning, it assumes that all of agents can share the information such as current states, actions and reward values. In [40], a distributed version of reinforcement Q-learning, $\mathcal{QR}$-learning, is developed for multi-agent Markov decision processes (MDPs). However, they assume that each network agent can observe the global state, including the states of other agents. All these algorithms are not practical in our system as they require too much information exchanges between agents, or D2D pairs in our system. The signaling overhead will be unacceptable. In addition, the exchanged information could be manipulated by selfish or malicious D2D pairs.

Nevertheless, there exists a series of literature discussing the relation of convergence to the existence of equilibrium in the corresponding game model. They show that a necessary condition for convergence is the existence of equilibrium [35, 41]. The authors in [42] applied results in evolutionary game theory to analyze the dynamic behavior of Q-learning. It appeared that for certain parameter settings, Q-learning is able to converge to a coordinated equilibrium in particular games. In other cases, unfortunately, it seems that Q-learners may exhibit cyclic behavior. Specifically, a Nash Q-learning [35] is proposed for non-cooperative multi-agent context using the framework of general-sum stochastic games. It concludes that the proposed Nash Q-learning consistently converges in the stochastic game, which has a unique equilibrium. In this framework, each agent's Nash Q-function is defined as the sum of its current reward plus its future rewards when all agents follow a joint Nash equilibrium strategy. Under this update rule, the actions of the agents will gradually converge to the Nash equilibrium. Note that, each agent must observe not only its own reward, but actions and rewards of other agents as well. We find that there is a clear link between the Nash Q-learning and the reinforcement learning we propose in this paper. Specifically, the prime-dual decomposition guarantees that the outcome will reach the optimal operating point, which has been proved to be an equilibrium in the proposed game model. This suggests that the proposed game model indeed has an unique equilibrium controlled by the BS through pricing. In addition, the effects of other agents to one agent is translated by the BS into the D2D service payment using equations (23),(30),(35) in the proposed framework. The payment can be considered as the global state the agent observed in the learning process. These links explain why the proposed algorithm always converges as we observed in the simulations.




(a) The proportion of potential D2D pairs which select in D2D mode with the increase of $p$

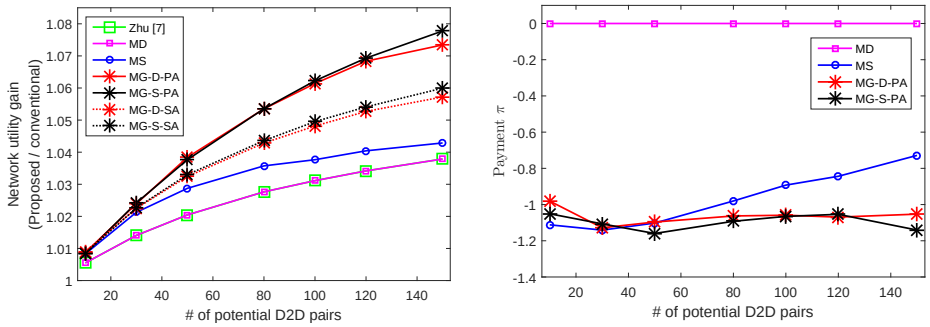



(b) The network utility in each communication system

Fig. 6. Comparison among different communication systems with increase of spectrum proportion $p$: $N = 200$; $N_d = 60$; $N_c = 140$; $T_i = 6$ dB.

## 6 Simulation Results

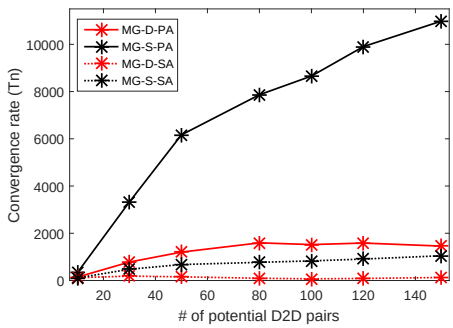
We evaluate the performance of the proposed system through simulations. We consider an urban macro hexagonal cell and $N$ UE pairs, including $N_d$ potential D2D pairs and $N_c$ cellular pairs, randomly and uniformly distributed within the cell. The key simulation parameters are list in TABLE. 3. To make it easier, we deploy all potential D2D pairs with the same SINR threshold. Specifically, the gap variable in MG-D approach is fixed by 1. Furthermore, all channel links follow the outdoor-to-outdoor channel model. The details about path loss, LOS probability, shadowing and fading strictly follow the description of 3GPP TR36.843 [43]. All simulation settings, if not mentioned, follow the



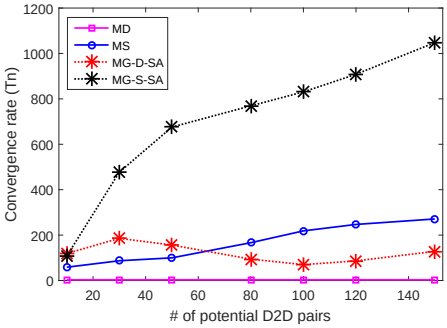

(a) Network utility gain




(b) D2D service payment




(c) Convergence rate - Part I




(d) Convergence rate - Part II

Fig. 7. Comparison among different communication systems with increase of potential D2D pairs $N_d$: **MG-D**: Dynamic algorithm; **MG-S**: Search algorithm; **PA**: Performance-aware; **SA**: Speed-aware. $N_c = 100$; $T_i = 6$ dB.

TABLE 3
List of Simulation Parameters

| | |
|---|---|
| ISD of urban macro (all UEs outdoor) | 500m |
| Carrier frequency | 2GHz |
| System bandwidth, $W$ | 20MHz |
| Transmit power of BS | 46dBm |
| Transmit power of UE | 23dBm |
| Minimum distance between transmitter and BS | $\geq$ 35m |
| Minimum distance between any two transmitters | $\geq$ 3m |
| Distribution range of potential D2D receiver to its corresponding transmitter | 50m |
| Discount factor $\beta$ | 0.2 |
| $K_{min}$, $K_{max}$ | 1, 10 |

suggested values in 3GPP TR36.843. We also present the numerical results of Zhu's algorithm in [7]. To fit Zhu's algorithm into our framework, we reduce its candidate communication modes of potential D2D pairs to cellular mode and dedicated mode only.

### 6.1 Effect of D2D Spectrum Partition Proportion

We first simulate the proposed system with the increase of partition proportion $p$ to understand the influence of D2D bandwidth to MD and MS systems. The performance comparison among the conventional cellular system, the D2D-enabled system with no payment and the proposed system under MD and MS are shown in Fig. 6. In conventional cellular system, all UEs, either conventional or potential D2D pairs, can only transmit in cellular mode. Under D2D-enabled system, no payment represents the case that each potential D2D pair selects the mode selfishly without any additional D2D payment for regulation. The proposed system, on the other hand, includes the optimal payment we derived in Section 4 under MD and MS, respectively.

Fig. 6(a) shows the average proportion of potential D2D pairs who will select D2D mode with the increase of D2D bandwidth proportion $p$. For the proposed system, we observe that MS attracts more potential D2D pairs than MD when $p$ is low. However, the proportion of D2D pairs saturates to around 0.72 under MS when $p$ increases to 0.1. This is due to the fact that the inter-pair interference significantly increases with the proportion of D2D pairs under MS. The interference reduces the incentive of potential D2D pairs to choose D2D mode, even with the benefit from the increasing bandwidth allocated to each potential D2D pair. On the contrary for MD, the number of D2D pairs $m$ grows steadily with the increase of $p$ since no inter-pair interference exists in D2D communication. For the D2D-enabled system with no payment, the network is not regulated by base station and thus pairs will select the mode purely based on the transmission quality. It can be seen that potential D2D pairs shows more interests to D2D transmission under MD while much less interests when under MS. These unregulated selections will degrade the overall system performance, as we will illustrate in Fig. 6(b). For Zhu's algorithm, the number of D2D pairs can be regulated because of its evolutionary game for adaptive mode selection. Nevertheless, our MD and MS modes with proposed payments outperform Zhu's in every settings

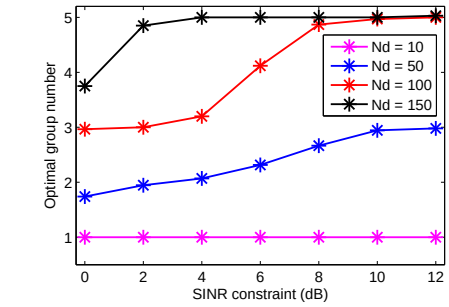
The network utility is shown in Fig. 6(b). The overall system achieves better performance with the increase of bandwidth reserved for D2D communication under both modes when the proposed pricing strategy is applied.

Specifically, MS performs significant better due to higher spectrum utilization efficiency from sharing spectrum. For the D2D-enabled system with no payment applied, on the other hand, the system performance may be worse than the conventional cellular system. The network utility under MD is worse than conventional cellular system when $p$ is low. Nevertheless, it obtains the same overall utility as the one under the proposed pricing strategy when $p$ is larger than $0.19$, since the optimal choices (in terms of overall system performance) for all potential D2D pairs are D2D mode, which is exactly the same as the selfish choice of these D2D pairs even when no payment is applied. Additionally, the performance of MS without payment is strictly lower than the one under the proposed pricing strategy. The allocated D2D bandwidth is under-utilized with significantly lower number of pairs choosing D2D mode. In general, the proposed pricing strategy system impresses a performance gain by fully exploiting the advantage of D2D system while avoiding undesired selfish selection by a proper pricing regulation. The network utility of Zhu's algorithm is better than conventional cellular system and similar to our MD mode. Nevertheless, it never outperforms any of the proposed modes in all simulations.

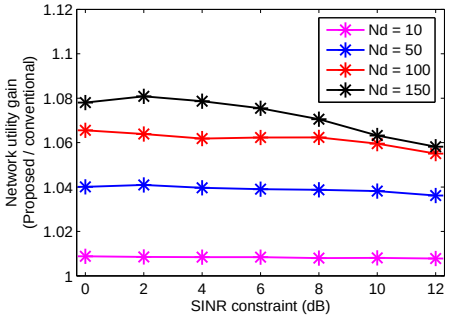
### 6.2 Dynamic Stackelberg Game under Different Proposed Modes

We compare the performance among different frameworks of D2D-enabled cellular system. In Fig. 7, we simulate two different approaches, MG-S and MG-D, under group D2D mode. Generally, reinforcement learning approach requires large learning scale to converge to near-global optimal state. We observe that the approach may run with small learning scale and converge to some local optimal state by adopting different initial parameters. Here we would like to show two typical cases, Performance-Aware (**PA**) and Speed-Aware (**SA**), by initializing different D2D payments. PA stresses the preference of system performance. In other words, the goal of PA is to reach the global optimal configuration regardless of convergence rate. On the contrary, SA focuses on improving convergence speed. The system, therefore, may reach some local optimization under SA.

We define network utility gain as the ratio of total network utility of proposed D2D-enabled system over conventional one. In general, the overall system achieves better network utility gain with the increased number of potential D2D pairs, as shown in Fig. 7(a). The performance of MS is better than MD because of higher spectrum utilization efficiency. By considering a trade-off between transmission quality and spectrum utilization efficiency, MG achieves better performance than both of MD and MS. Besides, the algorithm under PA has more network utility gain than under SA. For either PA or SA, we know that the search approach (MG-S) definitely can achieve the best network utility gain among the proposed MD, MS, MG-S and MG-D approaches. Fig. 7(a) shows that the proposed dynamic approach (MG-D) really can reach near-optimal solution. We also observe that the network utility gain of Zhu's algorithm is exactly the same as the gain of MD. It illustrates that the optimal choices for all potential D2D pairs here are D2D mode, which is also consistent with those discussed in [7].




(a) Optimal group number




(b) Network utility gain

Fig. 8. Comparison among different number of potential D2D pairs with the increase of SINR constraint $T_i$ under MG: $N_c = 140$.

Fig. 7(b) shows the D2D service payment with the increase of potential D2D pairs. We observe that there exists no additional payment for potential D2D pairs under MD but negative D2D payment under either MS or MG. The main reason is that it is interference-free to each potential D2D pairs regardless any mode it selects under MD. However, D2D pairs under either MS or MG should suffer from an intra-group interference. The goal of the BS is to optimize total system performance while each potential D2D pair only cares about its individual utility. Negative D2D payment announced by the BS is therefore provided to compensate the utility loss of potential D2D pairs in D2D mode in the optimal configuration.

For either PA or SA, MG-D converges quite faster than MG-S, as shown in Fig. 7(c) and Fig. 7(d). Here we calculate the convergence rate as the value of $t$, namely the number of rounds by playing the proposed two-stage Stackelberg game. The number of signals exchanges is highly related to the number of rounds required to converge to the stable state. Specifically, in each round of playing the two-stage Stackelberg game, all potential D2D pairs must report their mode selection decisions to the BS. Therefore, the required times of signal exchange for each D2D pair until the stable state is exactly the number of rounds required for convergence, or the convergence rate we defined in Fig. 7. Specifically, MG-D-PA needs extreme larger learning scale than MG-D-SA. PA is recommended when the system is relatively stable and therefore a lower convergence rate is acceptable. On the other hand, SA is more desired when the system is fast-changing and requires rapid reconfiguration.

### 6.3 Effect of SINR Constraint

Here we simulate with the increase of SINR constraint under MG, as shown in Fig. 8. In Fig. 8(a), the more

potential D2D pairs the system has, the more D2D group the system needs. With the increase of SINR constraint, optimal group number increases accordingly and stabilizes at some certain value. The reason is that higher SINR constraint means lower toleration of intra-group interference, which can be reduced by adding more D2D group. In Fig. 8(b), we observe that the performance of the proposed system with larger number of potential D2D pairs is likely to be affected by SINR constraint. Obviously, more potential D2D pairs mean more interference is generated in the system. In this situation, mode selections of some potential D2D pairs will be sensitive to SINR constraint. A larger SINR constraint can easily prevent them from choosing D2D mode even though they achieve more network utility in D2D mode.

## 7 Conclusion

We presented a pricing-based game-theoretic framework for optimal mode selection and spectrum partitioning for D2D communication. The proposed D2D-enabled system displays a significant performance improvement over conventional cellular system. The results show that the BS can manage D2D-enabled network through the simple price design. Besides, target group number can be found by adopting the proposed dynamic algorithm under group D2D mode. Furthermore, both performance-aware and speed-aware settings are mentioned with their own advantages. We also observe that spectrum partitioning and SINR constraint can affect the mode selections of potential D2D pairs in the proposed system.

## Acknowledgment

This work was supported by the Ministry of Science and Technology under Grant MOST 105-2221-E-001-003-MY3, 105-2221-E-002-014-MY3, 103-2221-E-002-086-MY3 and the Academia Sinica under Thematic Research Grant.

## References

[1] X. Lin, J. G. Andrews, A. Ghosh, and R. Ratasuk, "An overview of 3gpp device-to-device proximity services," *IEEE Communications Magazine*, vol. 52, no. 4, pp. 40–48, April 2014.

[2] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka, H. Tullberg, M. A. Uusitalo, B. Timus, and M. Fallgren, "Scenarios for 5g mobile and wireless communications: the vision of the metis project," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 26–35, May 2014.

[3] A. Asadi, Q. Wang, and V. Mancuso, "A survey on device-to-device communication in cellular networks," *IEEE Communications Surveys Tutorials*, vol. 16, no. 4, pp. 1801–1819, Fourthquarter 2014.

[4] J. Liu, N. Kato, J. Ma, and N. Kadowaki, "Device-to-device communication in lte-advanced networks: A survey," *IEEE Communications Surveys Tutorials*, vol. 17, no. 4, pp. 1923–1940, Fourthquarter 2015.

[5] X. Lin and J. G. Andrews, "Optimal spectrum partition and mode selection in device-to-device overlaid cellular networks," in *2013 IEEE Global Communications Conference (GLOBECOM)*, Dec 2013, pp. 1837–1842.

[6] S. T. Su, B. Y. Huang, C. Y. Wang, C. W. Yeh, and H. Y. Wei, "Protocol design and game theoretic solutions for device-to-device radio resource allocation," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 5, pp. 4271–4286, May 2017.

[7] K. Zhu and E. Hossain, "Joint mode selection and spectrum partitioning for device-to-device communication: A dynamic stackelberg game," *IEEE Transactions on Wireless Communications*, vol. 14, no. 3, pp. 1406–1420, March 2015.

[8] B. Cho, K. Koufos, and R. Jntti, "Spectrum allocation and mode selection for overlay d2d using carrier sensing threshold," in *2014 9th International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM)*, June 2014, pp. 26–31.

[9] S. Maghsudi and S. Staczak, "Transmission mode selection for network-assisted device to device communication: A levy-bandit approach," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 7009–7013.

[10] C. G. Diaz, W. Saad, B. Maham, D. Niyato, and A. S. Madhukumar, "Strategic device-to-device communications in backhaul-constrained wireless small cell networks," in *2014 IEEE Wireless Communications and Networking Conference (WCNC)*, April 2014, pp. 1661–1666.

[11] H. Chen, D. Wu, and Y. Cai, "Coalition formation game for green resource management in d2d communications," *IEEE Communications Letters*, vol. 18, no. 8, pp. 1395–1398, Aug 2014.

[12] P. K. Mishra, S. Pandey, and S. K. Biswash, "Efficient resource management by exploiting d2d communication for 5g networks," *IEEE Access*, vol. 4, pp. 9910–9922, 2016.

[13] G. Katsinis, E. E. Tsiropoulou, and S. Papavassiliou, "Joint resource block and power allocation for interference management in device to device underlay cellular networks: A game theoretic approach," *Mobile Networks and Applications*, pp. 1–13, 2016.

[14] A. Abrardo and M. Moretti, "Distributed power allocation for d2d communications underlaying/overlaying ofdma cellular networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1466–1479, March 2017.

[15] F. Yang and X. Zhang, "Distributed optimal green power allocation for d2d based cellular networks with long-term qos constraint," in *2016 IEEE Global Communications Conference (GLOBECOM)*, Dec 2016, pp. 1–6.

[16] O. Delgado and F. Labeau, "D2d relay selection and fairness on 5g wireless networks," in *2016 IEEE Globecom Workshops (GC Wkshps)*, Dec 2016, pp. 1–6.

[17] R. Ma, Y. J. Chang, H. H. Chen, and C. Y. Chiu, "On relay selection schemes for relay-assisted d2d communications in lte-a systems," *IEEE Transactions on Vehicular Technology*, vol. PP, no. 99, pp. 1–1, 2017.

[18] J. Zhao, K. K. Chai, Y. Chen, J. Schormans, and J. Alonso-Zarate, "Joint mode selection and radio resource allocation for d2d communications based on dynamic coalition formation game," in *Proceedings of European Wireless 2015; 21th European Wireless Conference*, May 2015, pp. 1–5.

[19] S. M. A. Kazmi, N. H. Tran, W. Saad, Z. Han, T. M. Ho, T. Z. Oo, and C. S. Hong, "Mode selection and resource allocation in device-to-device communications: A matching game approach," *IEEE Transactions on Mobile Computing*, vol. PP, no. 99, pp. 1–1, 2017.

[20] J. Dai, J. Liu, Y. Shi, S. Zhang, and J. Ma, "Analytical modeling of resource allocation in d2d overlaying multi-hop multi-channel uplink cellular networks," *IEEE Transactions on Vehicular Technology*, vol. PP, no. 99, pp. 1–1, 2017.

[21] J. Lyu, Y. H. Chew, and W. C. Wong, "A stackelberg game model for overlay d2d transmission with heterogeneous rate requirements," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 10, pp. 8461–8475, Oct 2016.

[22] Y. Zhang, C. Y. Wang, and H. Y. Wei, "Incentive compatible mode selection and spectrum partitioning in overlay d2d-enabled network," in *2015 IEEE Globecom Workshops (GC Wkshps)*, Dec 2015, pp. 1–6.

[23] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 6, pp. 2706–2716, June 2013.

[24] J. M. Cioffi, G. P. Dudevoir, M. V. Eyuboglu, and G. D. Forney, "Mmse decision-feedback equalizers and coding. ii. coding results," *IEEE Transactions on Communications*, vol. 43, no. 10, pp. 2595–2604, Oct 1995.

[25] K. Shen and W. Yu, "Distributed pricing-based user association for downlink heterogeneous cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1100–1113, June 2014.

[26] C. Y. Wang, G. Y. Lin, C. C. Chou, C. W. Yeh, and H. Y. Wei, "Device-to-device communication in lte-advanced system: A strategy-proof resource exchange framework," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 12, pp. 10 022–10 036, Dec 2016.

[27] R. Gibbons, "Stackelberg model of duopoly," in *A primer in game theory*. New York, Harvester Wheatsheaf, 1992, ch. 2.

[28] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and distributed computation: numerical methods*. Prentice hall Englewood Cliffs, NJ, 1989, vol. 23.

[29] D. Fudenberg and D. K. Levine, *The theory of learning in games*. MIT press, 1998, vol. 2.

[30] D. Niyato and E. Hossain, "Dynamics of network selection in heterogeneous wireless networks: An evolutionary game approach," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 4, pp. 2008–2017, May 2009.

[31] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press Cambridge, 1998, vol. 1, no. 1.

[32] P. Vamvakas, E. E. Tsiropoulou, and S. Papavassiliou, "Dynamic provider selection & power resource management in competitive wireless communication markets," *Mobile Networks and Applications*, pp. 1–14, 2017.

[33] J. G. March, "Exploration and exploitation in organizational learning," *Organization science*, vol. 2, no. 1, pp. 71–87, 1991.

[34] E. Rodrigues Gomes and R. Kowalczyk, "Dynamic analysis of multiagent q-learning with $\varepsilon$-greedy exploration," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 369–376.

[35] J. Hu and M. P. Wellman, "Nash q-learning for general-sum stochastic games," *Journal of machine learning research*, vol. 4, no. Nov, pp. 1039–1069, 2003.

[36] J. A. dos Santos Gromicho, *Quasiconvex optimization and location theory*. Springer Science & Business Media, 2013, vol. 9.

[37] M. Bowling and M. Veloso, "Rational and convergent learning in stochastic games," *International joint conference on artificial intelligence*, vol. 17, no. 1, pp. 1021–1026, 2001.

[38] ——, "Multiagent learning using a variable learning rate," *Artificial Intelligence*, vol. 136, no. 2, pp. 215–250, 2002.

[39] H.-E. Kim and H.-S. Ahn, "Convergence of multiagent q-learning: Multi action replay process approach," *Intelligent Control (ISIC), 2010 IEEE International Symposium on*, pp. 789–794, 2010.

[40] S. Kar, J. M. F. Moura, and H. V. Poor, "$\mathcal{QD}$-learning: A collaborative distributed strategy for multi-agent reinforcement learning through $\mathrm{Consensus} + \mathrm{Innovations}$," *IEEE Transactions on Signal Processing*, vol. 61, no. 7, pp. 1848–1862, April 2013.

[41] A. Greenwald, K. Hall, and R. Serrano, "Correlated q-learning," *ICML*, vol. 3, pp. 242–249, 2003.

[42] K. Tuyls, P. J. Hoen, and B. Vanschoenwinkel, "An evolutionary dynamical analysis of multi-agent learning in iterated games," *Autonomous Agents and Multi-Agent Systems*, vol. 12, no. 1, pp. 115–153, 2006.

[43] 3GPP, "Study on LTE device to device proximity services; Radio aspects," 3rd Generation Partnership Project (3GPP), Technical Report (TR) 36.843, 03 2014, version 12.0.1. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2544

[44] D. P. Bertsekas, *Convex optimization theory*. Athena Scientific Belmont, 2009.

[45] S. Boyd, L. Xiao, and A. Mutapcic, "Subgradient methods," *lecture notes of EE392o, Stanford University, Autumn Quarter*, vol. 2004, 2003.


## APPENDIX A
## PROOF OF PROPOSITION 5

Here we discuss the convergence issue and dynamic step size rule in details. Our proof for convergence issue mainly derives from the analysis of [44] and [45].

Recall that $\mu_k = \mu\,(\forall k \in \mathcal{K})$ and $m = \sum_{k\in\mathcal{K}} m_k$. Then from the original dual optimization function (14), we have $\frac{\partial D(\mu)}{\partial \mu} = m(\mu) - \sum_{k\in\mathcal{K}} \sum_{i\in\mathcal{U}_d} x_{i,k}(\mu)$. Furthermore, the constraint that $m = \sum_{k\in\mathcal{K}} \sum_{i\in\mathcal{U}_d} x_{i,k} \leq N_d$ should be satisfied, where $N_d$ is the total number of potential D2D pairs. The subgradient of dual optimization function $D(\mu)$, therefore, is bounded, that is

$$\sup_t \{\|\partial D(\mu_t)\|\} \leq N_d. \tag{38}$$

Here we denote $D(\mu)_t$ as the best objective value so far found in $t$ iterations, that is,

$$D(\mu)_t = \min_{\tau=1,\ldots,t} D(\mu_\tau) \text{ or } D(\mu)_t = \min\{D(\mu)_{t-1}, D(\mu_t)\},$$

where $D(\mu_\tau)$ is real value of dual optimization problem (14) at time $\tau$. Furthermore, we define $\mu_t \in \mathbb{M}$ and $\mathbb{M}$ is a subset of $\mathbb{R}$. Proposition 6.3.1 in [44] shows that for all $\nu \in \mathbb{M}$ and $t \geq 0$,

$$\|\mu_{t+1} - \nu\|^2 \leq \|\mu_t - \nu\|^2 - 2\delta_t(D(\mu_t) - D(\nu)) + \delta_t^2 \|\partial D(\mu_t)\|^2,$$

where $\delta_t$ is the step size in time $t$. We let $\mu^*$ be a point that minimizes $D(\mu)$, then $D(\mu^*)$ will be the optimal value of the problem. By replacing $\nu$ with $\mu^*$, we have

$$\begin{aligned}
&\|\mu_{t+1} - \mu^*\|^2 \\
&\leq \|\mu_t - \mu^*\|^2 - 2\delta_t(D(\mu_t) - D(\mu^*)) + \delta_t^2 \|\partial D(\mu_t)\|^2 \\
&\leq \|\mu_1 - \mu^*\|^2 - 2\sum_{\tau=1}^{t} \delta_\tau (D(\mu_\tau) - D(\mu^*)) \\
&\qquad + \sum_{\tau=1}^{t} \delta_\tau^2 \|\partial D(\mu_\tau)\|^2.
\end{aligned}$$

Using $\|\mu_{t+1} - \mu^*\|^2 \geq 0$, we have

$$2\sum_{\tau=1}^{t} \delta_\tau (D(\mu_\tau) - D(\mu^*)) \leq \|\mu_1 - \mu^*\|^2 + \sum_{\tau=1}^{t} \delta_\tau^2 \|\partial D(\mu_\tau)\|^2. \tag{39}$$

In addition,

$$\sum_{\tau=1}^{t} \delta_\tau (D(\mu_\tau) - D(\mu^*)) \geq (\sum_{\tau=1}^{t} \delta_\tau) \min_{\tau=1,\ldots,t} (D(\mu_\tau) - D(\mu^*)). \tag{40}$$

Combining (39) and (40), we have the inequality

$$\begin{aligned}
D(\mu)_t - D(\mu^*) &= \min_{\tau=1,\ldots,t} D(\mu_\tau) - D(\mu^*) \\
&\leq \frac{\|\mu_1 - \mu^*\|^2 + \sum_{\tau=1}^{t} \delta_\tau^2 \|\partial D(\mu_\tau)\|^2}{2(\sum_{\tau=1}^{t} \delta_\tau)}.
\end{aligned} \tag{41}$$

Finally, using the bound (38), we obtain the basic inequality

$$\begin{aligned}
D(\mu)_t - D(\mu^*) &= \min_{\tau=1,\ldots,t} D(\mu_\tau) - D(\mu^*) \\
&\leq \frac{\|\mu_1 - \mu^*\|^2 + N_d^2 \sum_{\tau=1}^{t} \delta_\tau^2}{2(\sum_{\tau=1}^{t} \delta_\tau)}.
\end{aligned}$$

We can state that

$$D(\mu)_t - D(\mu^*) \leq \frac{\mathbf{dist}(\mu_1, \mathcal{U}^*)^2 + N_d^2 \sum_{\tau=1}^{t} \delta_\tau^2}{2(\sum_{\tau=1}^{t} \delta_\tau)}, \quad (42)$$

Where $\mathcal{U}^*$ denotes the optimal set, and $\mathbf{dist}(\mu_1, \mathcal{U}^*)$ is the (Euclidean) distance of $\mu_1$ to the optimal set.

Several different types of step size rules can be used as follows.

- *Constant step size*: $\delta_\tau = h$ is a constant, independent of $\tau$. From (42), we have

  $$D(\mu)_t - D(\mu^*) \leq \frac{\mathbf{dist}(\mu_1, \mathcal{U}^*)^2 + N_d^2 h^2 t}{2ht}.$$

  The righthand side convergences to $N_d^2 h/2$ as $t \to \infty$. Thus, for the subgradient method with fixed step size $h$, $D(\mu)_t$ that converges to within $N_d^2 h/2$ of optimal.
- *Costant step length*: $\delta_\tau = h/\|\partial D(\mu_\tau)\|$. This means that $\|\mu_{\tau+1} - \mu_\tau\| = h$. From (41), we have

  $$\begin{aligned} D(\mu)_t - D(\mu^*) &\leq \frac{\mathbf{dist}(\mu_1, \mathcal{U}^*)^2 + \sum_{\tau=1}^{t} \delta_\tau^2 \|\partial D(\mu_\tau)\|^2}{2(\sum_{\tau=1}^{t} \delta_\tau)} \\ &\leq \frac{\mathbf{dist}(\mu_1, \mathcal{U}^*)^2 + h^2 t}{2(\sum_{\tau=1}^{t} \delta_\tau)}. \end{aligned}$$

  And we have $\delta_\tau = h/\|\partial D(\mu_\tau)\| \geq h/N_d$. Applying this to the denominator of the above inequality gives

  $$D(\mu)_t - D(\mu^*) \leq \frac{\mathbf{dist}(\mu_1, \mathcal{U}^*)^2 + h^2 t}{2ht/N_d}. \quad (43)$$

  The righthand side converges to $N_d h/2$ as $t \to \infty$, so in this case the subgradient method converges to within $N_d h/2$ of optimal.

In a word, the convergence of the proposed algorithm can be guaranteed.

## APPENDIX B
## PROOF OF PROPOSITION 7

By relaxing $m_k$ as a continuous variable within feasible range, we try to prove that adopting the same value for all dual variables $\mu_k$ ($k \in \mathcal{K}$) is the best solution.

Our discussion here bases on D2D group $k$ and its number of D2D pairs $m_k$. According to the definitions, we have $m = m_k + \sum_{k' \in \mathcal{K} \setminus \{k\}} m_{k'} = m_k + m_{-k}$. Substituting it into $g(\boldsymbol{\mu})$ from (17), the first and second derivatives of $g(\boldsymbol{\mu})$ of $m_k$ will be

$$\begin{aligned} \frac{\partial g(\boldsymbol{\mu})}{\partial m_k} &= \log \frac{2p}{K(1-p)} + \log(N - m_k - m_{-k}) + 1 + \mu_k, \\ \frac{\partial^2 g(\boldsymbol{\mu})}{\partial {m_k}^2} &= -\frac{1}{(N - m_k - m_{-k})} = -\frac{1}{(N-m)} < 0. \end{aligned}$$

Obviously, $g(\boldsymbol{\mu})$ is a concave function of $m_k$ because its second derivative is always negative. By checking its first order condition, optimal $m_k^*$, therefore, will be

$$\begin{aligned} &\frac{\partial g(\boldsymbol{\mu})}{\partial m_k} = 0 \\ \Rightarrow &\log \frac{2p}{K(1-p)} + \mu_k + 1 = -\log(N - m_k^* - m_{-k}) \\ \Rightarrow &\log \frac{2p}{K(1-p)} + \log e^{\mu_k + 1} = -\log(N - m_k^* - m_{-k}) \\ \Rightarrow &\frac{2p e^{\mu_k+1}}{K(1-p)} = \frac{1}{(N - m_k^* - m_{-k})} \\ \Rightarrow &m_k^* + m_{-k} = N - \frac{K(1-p)}{2p} e^{-\mu_k - 1} \end{aligned}$$

For any $k, j \in \mathcal{K}$ and $k \neq j$, due to the symmetry we have

$$\begin{aligned} &m^* = m_k^* + m_{-k}^* = m_j^* + m_{-j}^* \\ \Rightarrow &N - \frac{K(1-p)}{2p} e^{-\mu_k - 1} = N - \frac{K(1-p)}{2p} e^{-\mu_j - 1} \\ \Rightarrow &\mu_k = \mu_j \end{aligned}$$

That is to say, all dual variables $\mu_k$ ($k \in \mathcal{K}$) should be the same when $m$ reach optimal value.

**Yi Zhang** received the B.S. degree in software engineering from Software College, Xiamen University (XMU), China, in 2014. He received the M.S. degree from Graduate Institute of Communication Engineering (GICE), National Taiwan University (NTU), Taipei, Taiwan, in 2016. He has been an assistant engineer in Quanzhou Institute of Equipment Manufacturing, Haixi Institutes, Chinese Academy of Sciences (CAS), during 2016-2017. He is currently pursuing the Ph.D. degree in GICE at NTU. His primary research areas include wireless communication, game theory and optimization theory.

**Chih-Yu Wang** received the B.S. and Ph.D. degrees in electrical engineering and communication engineering from National Taiwan University (NTU), Taipei, Taiwan, in 2007 and 2013, respectively. He has been a visiting student in University of Maryland, College Park in 2011. He is currently an Assistant Research Fellow with the Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan. His research interests include game theory, wireless communications, social networks, and data science.

**Hung-Yu Wei** received his B.S. degree in electrical engineering from National Taiwan University (NTU) in 1999. He received his M.S. and Ph.D. degrees in electrical engineering from Columbia University, in 2001 and 2005, respectively. He was a summer intern at Telcordia Applied Research in 2000 and 2001. He was with NEC Labs America from 2003 to 2005. He joined Department of Electrical Engineering at the National Taiwan University in July 2005. He is currently a Professor with the Department of Electrical Engineering and Graduate Institute of Communication Engineering at National Taiwan University. His research interests include broadband wireless communications, fog computing, cross-layer design for wireless multimedia, IoT, and game theoretic models for networking. He has been actively participating in NGMN, IEEE 802.16, 3GPP, OpenFog Consortium, and IEEE P1934 standardization activities.

He was the recipient of the Recruiting Outstanding Young Scholar Award from the Foundation for the Advancement of Outstanding Scholarship in 2006, NTU Excellent Teaching Award in 2008, K. T. Li Young Researcher Award from ACM Taipei Chapter and IICM in 2012, CIEE Excellent Young Engineer Award in 2014, and Wu Ta You Memorial Award from Ministry of Science and Technology in 2015. He was the chair of IEEE Vehicular Technology Society Taipei Chapter during 2016-2017. He is currently the Secretary of IEEE P1934 Working Group. He also serves as an Associate Editor for IEEE IoT journal.