Learning for Matching Game in Cooperative D2D Communication with Incomplete Information

Yiling Yuan, Tao Yang Member, IEEE, Hui Feng, Member, IEEE, and Bo Hu, Member, IEEE

Abstract—This paper considers a cooperative device-to-device (D2D) communication system, where the D2D transmitters (DTs) act as relays to assist cellular users (CUs) in exchange for the opportunities to use licensed spectrum. Based on the interaction of each D2D pair and each CU, we formulate the pairing problem between multiple CUs and multiple D2D pairs as a one-to-one matching game. Unlike most existing works, we consider a realistic scenario with incomplete channel information. Thus, each CU lacks enough information to establish its preference over D2D pairs. Therefore, traditional matching algorithms are not suitable for our scenario. To this end, we convert the matching game to an equivalent non-cooperative game, and then propose a novel learning algorithm, which converges to a stable matching.

Index Terms—Cooperative D2D communication, matching game, incomplete information.

I. INTRODUCTION

RECENTLY, D2D communication has been extensively studied to provide better user experience. To implement this technology, one of the key issues is how to share licensed spectrum efficiently without degrading CUs’ performance greatly. We consider a cooperative D2D communication scheme, which exploits the advantages of cooperative relay and D2D communication [1]. The basic idea is that DTs act as relays for CUs in exchange for the transmission opportunities on the CUs’ channels. Thus, a win-win situation is achieved, which motivates CUs to share their spectrum with D2D pairs even if they have no surplus resource.

Most existing works [1–4] assume complete information, such as channel state information (CSI). However, collecting global information incurs heavy overhead, and thus may be not practical in large-scale networks. Besides, some information may be difficult to acquire, such as the CSI between CUs and DTs. Moreover, the latency requirement of some applications is stringent, such as D2D-based vehicle-to-vehicle communications. These facts motivate us to study distributed resource allocation scheme with incomplete information, where agents make decisions independently based on local information.

Game theory provides a framework to study the interactions of autonomous agents. There have been many game theoretical solutions in D2D networks [5]. In our context, CUs have preferences over D2D pairs and vice versa. Matching theory offers a suitable tool to study the cooperation between competitive CUs and competitive D2D pairs. There have been some matching-based resource allocation schemes for D2D communication [6, 8]. In this paper, we formulate the problem of pairing CUs with D2D pairs as a one-to-one matching game to seek a stable matching.

In the literature, authors of [9] have considered the incomplete information scenario, but do not investigate the pairing problem. Besides, similar cooperative scheme has been studied in cognitive radio networks recently [10–15], where secondary users (SUs) relay primary users’ (PUs) traffic for rewards of the transmission opportunities. Some works adopt auction [10], dynamic Bayesian game [11], and Stackelberg game [12] to tackle the incomplete information. Moreover, the authors of [13] consider the incomplete information in the matching game model. However, above works [10–13] assume PU has the knowledge of the relay rates, which depend on the SUs’ local information. In practice, such information is usually not known globally. In this paper, we consider a stronger incomplete information scenario, where CUs have no knowledge of the relay rates provided by the D2D pairs. The authors of [14], [15] consider the similar information assumption, but only consider single PU case. Instead, we consider the case with multiple CUs and multiple D2D pairs.

This paper focuses on the uplink resource sharing with incomplete information, because mobile devices are more likely to need help due to limited power budget. We formulate the pairing problem as a one-to-one matching game, based on the interaction between each CU and each D2D pair. Such interaction is described by Nash bargaining solution (NBS). Because the relay rates are unknown, CUs cannot establish preferences over D2D pairs. Thus, traditional matching algorithms, such as Gale-Shapley (GS) algorithm, are not suitable for our scenario. To the best of our knowledge, it is the first attempt to address the matching game with unknown preference. To overcome the difficulty, we convert the matching game to an equivalent non-cooperative game. At each period, each CU selects a D2D pair and a corresponding time allocation, and obtains a payoff as feedback. Based on the feedback, we propose a learning algorithm, which is proven to converge to a stable matching in probability. Moreover, the corresponding time allocation converges to the result of NBS with probability 1.
II. SYSTEM MODEL AND PROBLEM FORMULATION

A. System Model

We consider uplink resource sharing of a single cell with a base station (BS) denoted by $b$ and $M$ CUs. The set of CUs is denoted by $M$. Besides, there are $N$ D2D pairs, and the set of them is denoted by $N$. Each D2D pair contains one DT and one D2D receiver (DR). In this paper, we assume $M \leq N$. However, the proposed algorithm can be applied to the case where $M > N$. CU $m$ has been assigned to one cellular channel, namely channel $m$. There is no dedicated channel for D2D pairs. Therefore, D2D pairs relay the uplink traffic in exchange for access to the cellular channels.

We assume that each CU is assisted by at most one D2D pair, and each D2D pair can relay at most one CU due to limited battery capacity [2]. Similar to [13]–[15], we take the decode-and-forward protocol with parallel channel coding [10] as an example. When CU $m$ cooperates with D2D pair $n$, the normalized frame consists of three phases, as shown in Fig. 1. The first two phases both last $\frac{1-\alpha_m}{2}$ and are used for the relay transmission for the CU. Specifically, CU $m$ broadcasts its data with power $p_c$ to the BS and DT $n$ at first. Then, DT $n$ forwards received signal to the BS with power $p_d$. The third phase lasts $\alpha_m$, and is used by DT $n$ to transmit its data with power $p_d$ to DR $n$. We refer to $\alpha_m \in \mathcal{A} \triangleq [\alpha_L, \alpha_U]$ as time allocation.

The expected rate of CU $m$ in direct link is

$$R^C_m = \mathbb{E} \left[ \ln \left( 1 + \frac{p_c h_{mb}}{n_0} \right) \right],$$

where $h_{mb}$ is the channel gain from CU $m$ to the BS and $n_0$ denotes the noise power.

For simplicity, we assume every DT can decode all the CUs’ data in the first phase. Thus, cooperating with D2D pair $n$, the rate of CU $m$ in the first two phases is

$$r^C_{mn}(1-\alpha_m)/2 + \frac{\alpha_m}{2} \ln \left( 1 + \frac{p_c h_{mb}^m}{n_0} \right) \geq \alpha_m R^D_{mn},$$

where $h_{mb}^m$ is the channel gain from DT $n$ to the BS on channel $m$. Let $R^C_{mn} = \mathbb{E}[r^C_{mn}]$, and thus with time allocation $\alpha_m$, the expected rate of CU $m$ during the entire frame is $R^C_{mn} = (1-\alpha_m)R^C_{mn}$. Moreover, the expected rate of D2D pair $n$ during the entire frame is given by

$$R^D_{mn}(\alpha_m) = \alpha_m \mathbb{E} \left[ \ln \left( 1 + \frac{p_c h_{mb}^m}{n_0} \right) \right] \geq \alpha_m R^D_{mn},$$

where $h_{mb}^m$ is the channel gain of D2D pair $n$ on channel $m$. Assume that for each D2D link, the channel gains are i.i.d. across all the channels. Thus, we have $R^D_{mn} = R^D_{m'n'}, \forall m, \forall m' \in M$, and the value of $R^D_{mn}$ is denoted by $R^D_{mn}$.

Information Assumption: CU $m$ only knows $R^C_{mn}$ and has no knowledge of $R^D_{mn}$ and $R^m_n$, and D2D pair $n$ only knows $R^D_{mn}$. After cooperating with D2D pair $n$ at time $t$, CU $m$ gets a sample $r^C_{mn}(t)$ following a fixed unknown distribution.

B. Matching Based Framework

1) Bargaining Game for CU-D2D Pair $(m,n)$: To incentivize CU and D2D pair to cooperate mutually, a bargaining game is used to characterize the interaction between them. If CU $m$ cooperates with D2D pair $n$, the CU’s utility $U^C_{mn}$ and the D2D pair’s utility $U^D_{mn}$ are defined as

$$U^C_{mn}(\alpha_m) = R^C_{mn}(\alpha_m) - R^D_{mn},$$

$$U^D_{mn}(\alpha_m) = R^D_{mn}(\alpha_m).$$

We use NBS as the bargaining outcome to determine the time allocation, and thus the cooperation satisfies some useful properties and is beneficial for both sides. Hence, based on the concept of NBS [17], the time allocation is given by the following problem

$$\alpha^*_m (R^C_{mn}) = \begin{cases} \frac{R^C_{mn} - R^D_{mn} \alpha_U}{2R^D_{mn}} & \alpha_m \in A \triangleq [\alpha_L, \alpha_U] \\ \alpha_U & \text{otherwise} \end{cases},$$

where $[x]^y = \max(a, \min(x, b))$. Based on (7), the D2D pair with higher relay rate can obtain larger transmission time. Moreover, it is easy to verify that $U^C_{mn}(\alpha^*_m(R^C_{mn}))$ is an increasing function of $R^C_{mn}$, which reflects the fact that the CU prefers to cooperate with the D2D pair offering higher relay rate. We will use $\alpha^*_m (R^C_{mn})$ and $\alpha^*_m$ interchangeably afterwards. When the problem (6) is infeasible, for convenience, we still let $\alpha^*_m$ be the associated time allocation, and thus have $U^C_{mn}(\alpha^*_m) \leq 0$ in this case.

2) Matching Game Model: CU and D2D pair can only be paired when they agree to cooperate mutually. Therefore, it is reasonable to model the pairing problem between the set of CUs and the set of D2D pairs as a one-to-one matching game under two-sided preferences. CU $m$ prefers D2D pair $n$ to D2D pair $n'$ (i.e., $m \succ_n m'$), if $U^C_{mn}(\alpha^*_m(R^C_{mn})) > U^C_{mn}(\alpha^*_m(R^C_{mn}))$. Similarly, D2D pair $n$ prefers CU $m$ to CU $m'$ (i.e., $m \succ_m m'$), if $U^D_{mn}(\alpha^*_m(R^C_{mn})) > U^D_{mn}(\alpha^*_m(R^C_{mn}))$, which is equivalent to $\alpha^*_m \succ \alpha^*_m$. Besides, if $U^C_{mn}(\alpha^*_m) > 0$, D2D pair $n$ is acceptable to CU $m$, which is denoted by $n \succ_m \emptyset$.

Mathematically, a matching is a function $\mu : M \cup N \rightarrow M \cup N \cup \{\emptyset\}$, such that $\mu(m) = n$ if and only if $\mu(n) = m$, $\mu \in \mathcal{M}$.
and \( \mu(m) \in N \cup \{0\} \), \( \mu(n) \in M \cup \{0\} \), for \( \forall m \in M, \forall n \in N \). Note that \( \mu(x) = 0 \) implies that user \( x \) is unmatched. We aim to seek a stable matching (SM), which is the major solution concept in matching game and defined as follows \cite{13}.

**Definition 1:** Let \( \mu \) be a matching. A CU-D2D pair \( (m, n) \) is a blocking pair if \( \mu(m) \neq n \), \( m > n \mu(n) \) and \( n > n \mu(m) \). \( \mu \) is individually rational if \( \mu(m) \geq n \), \( \forall m \in M \). Thus, \( \mu \) is stable if it is individually rational and there is no blocking pair.

SM captures the preferences of both sides and CUs will only be matched with acceptable D2D pairs in SM. The existence of SM is guaranteed \cite{13}. The challenge is that each CU cannot establish its preference due to the unavailability of \( R_{m}^{C} \). Thus, the traditional GS algorithm \cite{13} cannot be used to seek SMs.

III. LEARNING FOR MATCHING WITH INCOMPLETE INFORMATION

To overcome the difficulty, CU has to learn its preference from the interactions with D2D pairs. To this end, we convert the above matching game to an equivalent non-cooperative game, which enables us to exploit the rich learning techniques designed for non-cooperative game.

A. Equivalent Non-cooperative Game Model

We convert the matching game to a non-cooperative game \( G = (M, \{ B_{m} \}_{m \in M}, \{ C_{m} \}_{m \in M} \). Due to the priority of CUs on licensed spectrum, we let CUs be the players to propose to D2D pairs. The action of CU \( m \) is to select a D2D pair \( b_{m} \), which means CU \( m \) proposes to cooperate with D2D pair \( b_{m} \) with time allocation \( \alpha_{mb} \). Each CU can refuse to cooperate with any D2D pairs, which is denoted by action \( b_{0} \). Hence, the action set of CU \( m \) is \( B_{m} = N \cup \{ b_{0} \} \).

Given an action profile \( b = (b_{1}, b_{2}, \cdots, b_{M}) \), each D2D pair selects the CU offering the maximal time allocation among the CUs proposing to it and rejects the others. If more than one CUs offers the maximal time allocation, the D2D pair will choose one of them based on a predefined rule. The CU chosen by D2D pair \( n \) is denoted by \( C_{m}(b_{n}) \), which can reflect the preference of D2D pair \( n \). Thus, the utility of CU \( m \) is:

\[
\pi_{m}(b_{m}, b_{-m}) = \begin{cases} 
\frac{U_{m}^{C}(a_{mb}) - \theta}{U_{mb}^{C} + w_{m}}, & \text{if } C_{m}(b_{n}) = m, b_{m} \neq b_{0}, \\
-\theta, & \text{if } C_{m}(b_{n}) \neq m, b_{m} \neq b_{0}, \\
0, & \text{if } b_{m} = b_{0},
\end{cases}
\]

where \( b_{-m} \) is the action profile of all the CUs except CU \( m \), and \( \theta > 0 \) is an arbitrarily small number and denotes the negotiation cost. Assume \( \theta \) is sufficiently small so that \( U_{m}^{C}(a_{mn}) - \theta > 0 \) if \( U_{mb}^{C}(a_{mb}) > 0 \). In the first case, \( \theta \) makes sure that CUs only select acceptable D2D pairs at equilibriums. The first two cases imply acceptance and rejection of the CU’s proposal, respectively. The third case means that the CU refuses to cooperate with any D2D pairs.

Given an action profile \( b \), its associated matching \( \mu_{b} \) is obtained as follows: for \( \forall m \in M, \forall n \in N \), \( \mu_{b}(m) = n \) and \( \mu_{b}(n) = m \) if and only if \( C_{m}(b) = m \). Hence, the relationship between the pure Nash equilibrium (PNE) of \( G \) and the SM can be stated as follows, which implies that an SM can be found via finding a PNE of \( G \).

**Theorem 1:** If action profile \( b \) is a PNE, \( \mu_{b} \) is an SM. Conversely, if \( \mu \) is an SM, there is a PNE \( b \) such that \( \mu_{b} = \mu \).

**Proof:** On the one hand, let \( b \) be a PNE. We will prove the stability of \( \mu_{b} \) by contradiction. The individual rationality is easy to verify. Suppose there is a blocking pair \( (m, n) \) in \( \mu_{b} \). Thus, CU \( m \) can take action \( b'_{m} = n \) to improve its utility, which violates our assumption. Therefore, \( \mu_{b} \) is stable.

On the other hand, let \( \mu \) be an SM. We construct an action profile \( b \) as follows: for CU \( m \), if \( \mu(m) = n \), it takes action \( b_{m} = n \) and action \( b_{0} \) otherwise. We will prove that \( b \) is a PNE by contradiction. Suppose \( b \) is not a PNE, so there exists a CU \( m \) deviating to take action \( b'_{m} \neq b_{m} \). If \( b'_{m} = b_{0} \), \( \mu \) is not individually rational. Besides, if \( b'_{m} \in N \), there is a blocking pair in \( \mu \). Thus, \( \mu \) is not stable, which violates our assumption. Therefore, \( b \) is a PNE.

To develop the learning algorithm, we show that \( G \) is a weakly acyclic under better-replies game (WABRG), which enables us to adopt better-reply with inertia (BRI) learning algorithm \cite{19} to find the PNE of \( G \). WABRG means that from any action profile, there is a better-reply path that terminates in a PNE in a finite number of steps. A better-reply path is a sequence of action profiles \( (b^{1}, b^{2}, \cdots, b^{t}, \cdots) \), where for each \( t \), there is a CU \( m \) such that \( b'_{m} \neq b_{m} \), \( b'_{m} = b_{m} \) and \( \pi_{m}(b^{t}) > \pi_{m}(b^{t-1}) \). In other words, in successive action profiles, only one CU changes its action to improve its utility.

**Theorem 2:** The proposed game \( G \) is a WABRG.

**Proof:** Suppose \( b^{0} \) is not a PNE. We will construct a better-reply path that ends at a PNE to prove the theorem.

If there are any rejected CUs, we let them take action \( b_{0} \) successively to obtain \( b^{1}, b^{2}, \cdots, b^{i} \), such that the CUs unmatched in \( \mu_{b^{i}} \) take action \( b_{0} \). Furthermore, according to Theorem 2.33 in \cite{18}, there exists a finite sequence of matchings \( \mu_{1}, \mu_{2}, \cdots, \mu_{k}, \cdots, \mu_{k} \), where \( \mu_{i} = \mu_{i-1} \), \( \mu_{k} \) is stable, and there is a blocking pair \( (m_{k}, n_{k}) \) for \( \mu_{k} \) such that \( m_{k+1} \) is obtained from \( m_{k} \) by satisfying the blocking pair \( (m_{k}, n_{k}) \). Thus, we let CU \( m_{1} \) select D2D pair \( n_{1} \) to obtain \( b^{i+1} \). Similarly, we let rejected CUs take action \( b_{0} \) successively to obtain \( b^{i+2}, b^{i+2}, \cdots, b^{k} \). Note that the above process will not change the associated matching, i.e., \( \mu_{b^{i}} = \mu_{b} \). Repeating the above process, we can obtain an action profile \( b^{k} \) such that \( \mu_{b^{k}} = \mu_{k} \) and the unmatched CUs in \( \mu_{k} \) take action \( b_{0} \). Note that \( b^{k} \) is exactly the constructed action profile in the proof of Theorem 1, so it must be a PNE. Besides, it is easy to find that the sequence \( b^{0}, b^{1}, \cdots, b^{k} \) is a better-reply path. Hence, we can verify Theorem 2.

B. Learning Algorithm

Because each CU’s utility is related to \( R_{m}^{C} \), each CU has to learn its utility from the interactions with D2D pairs. Furthermore, the action of CU \( m \) can be redefined as a proposal \( \bar{b}_{m} = (b_{m}, a_{mb}) \), where the time allocation is unknown in the case of incomplete information. Therefore, CUs have to make proposals explicitly to help D2D pairs establish
Algorithm 1: Extended BRI with Q-learning (EBRI-Q)

1: Initialize $\hat{R}_{mn}^{C}(1),\alpha_{m}, \forall n \in N; \forall m' \in M$ and $\hat{R}_{mn}(b), \forall b \in M'(r_{mn}B_{mn})$.
2: for $t = 2, 3, \cdots, T$ do:
3: With probability $\epsilon(t)$, uniformly select D2D pairs $b_{m}'$:
   a) With probability $1 - \epsilon$, choose the time allocation as $\alpha_{m} = \alpha_{m}^{t}r_{mn}$.
   b) With probability $\epsilon$, choose the time allocation as $\alpha_{m} = \alpha_{m}^{t}b_{m}'$.
4: With probability $1 - \epsilon(t)$, choose D2D pairs $b_{m}'$ by following better reply with inertia and choose $\alpha_{m}' = \alpha_{m}^{t}b_{m}'$:
   a) With probability $\epsilon$, select D2D pair $b_{m}' = b_{m}^{'t}$.
   b) With probability $1 - \epsilon$, select D2D pair $b_{m}'$ according to the distribution, which is over the D2D pairs that are better replies to CU’s full memory of length $L$ than $b_{m}^{'t}$ with respect to $\hat{R}_{mn}$. 
5: Observe joint proposal $b_{m}'$ and choice of each D2P pair. Get achieved rate $r_{mn}'(t)$ if cooperating with D2P pair $n$.
6: Update the estimation of $R_{mn}^{C}$, $\forall n \in N$:

\[
\hat{R}_{mn}^{C}(t) = \hat{R}_{mn}^{C}(t-1) + \lambda(t)(\hat{C}_{mn}(b_{m}'^{t}) - \hat{R}_{mn}^{C}(t-1)),
\]

where $\lambda(t) = 1/(1 + \sum_{t=1}^{\infty}1(\hat{C}_{mn}(b_{m}'^{t})) = m)$ and $1(\cdot)$ is indicator function.
7: Update its time allocation: $\alpha_{m}' = \alpha_{m}^{t}r_{mn}'(t), \forall n \in N$.
8: Update time allocations of other CUs according to their proposals:

\[
\alpha_{m}' = \begin{cases} \alpha_{m}^{t}, & \text{if } b_{m}' = n \text{ and } \alpha_{m}^{t} \neq \alpha_{e} \\ \alpha_{m}^{t-1}, & \text{otherwise}. \end{cases}
\]
9: Update estimated utility with respect to joint D2P pair selection $b_{m}' = (n, b_{m})$, $\forall n \in N; \forall b_{m} \in B_{mn}$:

\[
\hat{R}_{mn}(b) = \begin{cases} (1 - \alpha_{m}')R_{mn}^{C}(t) - R_{mn}^{C} - \theta & \text{if } \hat{C}_{mn}(b_{m}') = m \\ -\theta & \text{otherwise}. \end{cases}
\]

10: End for

C. Implementation Issues

At the beginning of each frame, CUs will send their proposals to the BS. Then, the BS broadcasts a proposal list containing all the CUs’ proposals at a dedicated channel. Meanwhile, all the D2P pairs will listen to this channel. After receiving CUs’ proposals, each D2P pair will accept one of them. Then, each matched D2P pair will send a feedback to the BS using the channel occupied by its matched CU. Based on these feedback, the BS obtains the final matching and informs the result to the CUs. Thus, each CU knows its partner and can begin its data transmission.

Except the above hand-shaking procedure, no extra overhead is needed in the proposed algorithm. Thus, each iteration has low signaling overhead. Note that (9)-(11) can be calculated in constant time. Moreover, the estimated utility is only needed in BRI. Thus, according to BRI, the algorithm only needs to estimate $\sum_{t=1}^{\infty}1(\hat{C}_{mn}(b, b_{m}^{'t}), \forall b \in N$. Therefore, the computational complexity of each iteration is $O(LN)$.

IV. Simulation Results

Simulation results are presented to evaluate the performance of the proposed algorithm. The channel gain is $\eta D^{-K}$, where $D$ is the distance between receiver and transmitter, $K = 4$ is the path loss exponent, and $\eta$ is fast fading with exponential distribution. The cell radius is $400$ m. CUs are randomly...
distributed in an area of at least 300 m away from the BS. The distance between the DT and the BS is uniformly distributed between 150 and 250 m. The length of D2D link is uniformly distributed between 10 and 60 m. Besides, we set $n_0 = -100$ dBm, $p_c = p_d = 20$ mW, $\alpha_L = 0.1$, $\alpha_U = 0.5$, $\zeta = \epsilon_0 = 0.1$ and $\xi = 0.2$, and the length of memory in BRI is set to 4.

At first, we investigate the convergence behavior of the proposed learning algorithm. For illustration purposes, we consider a small network with 2 CUs and 2 D2D pairs. There is only one SM, where CU 1 is matched with D2D pair 2 and CU 2 is matched with D2D pair 1. The results are given in Fig. 2 and Fig. 3. The results are averaged over 1000 simulations with the same topology. Fig. 2 presents the convergence of the time allocation estimation, where the estimation $\alpha_{mn}$ is normalized by the true value $\alpha_{mn}^*$. It is observed that the sequence $\{\alpha_{mn}\}$ converges to $\alpha_{mn}^*$ asymptotically, which is consistent with Theorem 3. The convergence of CUs’ behaviors is given in Fig. 3. It can be found that CU 1 and CU 2 could acquire their correct partners. This result implies that PNE or SM will be achieved eventually.

Next, we compare the proposed algorithm with other distributed algorithms in a larger network with 4 CUs and 5 D2D pairs. Fig. 4 shows the achieved system throughput over time for different algorithms. The results are averaged over 1000 simulations with different topologies. In the classical exploration-exploitation $\epsilon$-greedy algorithm, at each period, every CU selects the best D2D pairs so far with probability $1 - \epsilon$, and some random D2D pair with probability $\epsilon$. Besides, the time allocation estimations are updated similarly to our algorithm. We take $\epsilon = 0.1$ in the simulation. In the random algorithm, each CU selects D2D pair randomly and proposes $\alpha_L$ as time allocation to guarantee its performance. We present the non-cooperative scheme as well, where every CU takes action $b_0$. It can be observed that our algorithm yields significant gain over other learning algorithms. Besides, the performance loss due to incomplete information is small. It is also worth mentioning that the cooperative scheme achieves much better performance than non-cooperative scheme, which verifies the efficiency of the cooperative scheme.

V. Conclusion

This paper considers a cooperative D2D communication system with incomplete information. We model the pairing
problem between multiple CUs and multiple D2D pairs as a one-to-one matching game and propose a novel learning algorithm, which converges to a stable matching. The simulation results verify our analysis and show that the proposed algorithm outperforms the classical $\epsilon$-greedy algorithm. In the future work, the location information will be considered to divide CUs and D2D pairs into small groups to speed up the learning process. Moreover, the learning algorithm with faster convergence rate will also be investigated.

REFERENCES

[1] Y. Cao, T. Jiang, and C. Wang, “Cooperative device-to-device communications in cellular networks,” *IEEE Wirel. Commun.*, vol. 22, no. 3, pp. 124–129, Jun. 2015.

[2] Q. Wu et al., “Energy-efficient D2D overlaying communications with spectrum-power trading,” *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4404–4419, Jul. 2017.

[3] S. Shalmashi and S. B. Slimane, “Cooperative device-to-device communications in the downlink of cellular networks,” in *Proc. IEEE Wireless Commun. Netw. Conf.* (WCNC), Apr. 2014, pp. 2265–2270.

[4] M. Seif et al., “Cooperative D2D communication in downlink cellular networks with energy harvesting capability,” in *Proc. IEEE Int. Wireless Commun. Mobile Comput. Conf.* (IWCMC), Jun. 2017, pp. 183–189.

[5] L. Song et al., “Game-theoretic resource allocation methods for device-to-device communication,” *IEEE Wirel. Commun.*, vol. 21, no. 3, pp. 136–144, Jun. 2014.

[6] Z. Zhou et al., “Energy-efficient matching for resource allocation in D2D-enabled cellular networks,” *IEEE Transactions on Vehicular Technology*, vol. 66, no. 6, pp. 5256–5268, June 2017.

[7] S. Bayat et al., “Matching theory: Applications in wireless communications,” *IEEE Signal Processing Magazine*, vol. 33, no. 6, pp. 103–122, Nov. 2016.

[8] Z. Zhou et al., “Social big-data-based content dissemination in internet of vehicles,” *IEEE Transactions on Industrial Informatics*, vol. 14, no. 2, pp. 768–777, Feb. 2018.

[9] C. Ma et al., “Cooperative spectrum sharing in D2D-enabled cellular networks,” *IEEE Trans. Commun.*, vol. 64, no. 10, pp. 4394–4408, Oct. 2016.

[10] S. K. Jayaweera, M. Bkassiny, and K. A. Avery, “Asymmetric cooperative communications based spectrum leasing via auctions in cognitive radio networks,” *IEEE Transactions on Wireless Communications*, vol. 10, no. 8, pp. 2716–2724, Aug. 2011.

[11] Y. Yan, J. Huang, and J. Wang, “Dynamic bargaining for relay-based cooperative spectrum sharing,” *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 8, pp. 1480–1493, Aug. 2013.

[12] X. Feng, H. Wang, and X. Wang, “A game approach for cooperative spectrum sharing in cognitive radio networks,” *Wireless Communications and Mobile Computing*, vol. 15, no. 3, pp. 538–551, 2015.

[13] X. Feng et al., “Cooperative spectrum sharing in cognitive radio networks: A distributed matching approach,” *IEEE Trans. Commun.*, vol. 62, no. 8, pp. 2651–2664, Aug. 2014.

[14] L. Duan, L. Gao, and J. Huang, “Cooperative spectrum sharing: A contract-based approach,” *IEEE Trans. Mob. Comput.*, vol. 13, no. 1, pp. 174–187, Jan. 2014.

[15] M. Lopez-Martinez et al., “A superprocess with upper confidence bounds for cooperative spectrum sharing,” *IEEE Trans. Mob. Comput.*, vol. 15, no. 12, pp. 2939–2953, Dec. 2016.

[16] J. N. Laneman and G. W. Wornell, “Distributed space-time-coded protocols for exploiting cooperative diversity in wireless networks,” *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2415–2425, Oct. 2003.

[17] Z. Han et al., *Game Theory in Wireless and Communication Networks: Theory, Models, and Applications*. Cambridge University Press, 2011.

[18] A. Roth and M. A. O. Sotomayor, *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*. Cambridge Univ. Press, 1992.

[19] A. C. Chapman et al., “Convergent learning algorithms for unknown reward games,” *SIAM Journal on Control and Optimization*, vol. 51, no. 4, pp. 3154–3180, 2013.