SOCIAL NORM AND LONG-RUN LEARNING IN PEER-TO-PEER NETWORKS

Yu Zhang, Mihaela van der Schaar
Department of Electrical Engineering, UCLA

ABSTRACT

We start by formulating the resource sharing in peer-to-peer (P2P) networks as a random-matching gift-giving game, where self-interested peers aim at maximizing their own long-term utilities. In order to provide incentives for the peers to voluntarily share their resources, we propose to design protocols that operate according to pre-determined social norms. To optimize their long-term performance when playing such a game, peers can learn to play the best response by solving individual stochastic control problems. We first show that when a peer learns in an environment in which its opponents play a fixed strategy, learning will provide an advantage for this peer (i.e. it will lead to an increased utility for the learning peer). If all the peers in the network learn, we prove that learning remains beneficial for the peers. Moreover, we prove that the network will converge to the “fully-cooperative state” (where a socially optimal outcome is attained) if the update error $\gamma$ of the peers’ reputations is sufficiently small and the benefit of participating in the stage game is sufficiently larger than the incurred cost.

Index Terms—Peer-to-Peer Networks, Social Norm, Stochastic Control, Markov Decision Process

1. INTRODUCTION

P2P systems rely on voluntary contribution of resources from the individual participants. However, P2P networks are vulnerable to intrinsic incentive problems since the upload service incurs costs to both the uploader and the downloader, but benefits only the downloader. Therefore, individual rationality results in free-riding behaviors among peers, at the expense of the collective social welfare of the network. Various incentive mechanisms have been proposed to encourage cooperation in P2P systems [2], with a large body of them relying on the idea of reciprocity, in which peers are rated based on their past behaviors. Differential service schemes which provide services based on the rating score are deployed to encourage peers to contribute their resources in order to receive in return such resources from others. Reciprocity-based mechanisms can be further classified into direct reciprocity and indirect reciprocity depending on how the rating score is generated. Existing direct reciprocity protocols do not scale well as the number of peers increases, since frequent interactions between two peers are required in order to build up accurate mutual ratings in this bilateral reciprocity paradigm. As a result, in current P2P networks, where a large population of peers is interacting and peers are randomly matched with each other [1], a new paradigm is necessary to prevent peers from free-riding.

In this paper, we develop a social norm based framework to analyze the performance of indirect reciprocity mechanisms in P2P networks. As pointed out in [3], in such highly interconnected networks there is a natural tendency for nodes to learn based on their interactions with the environment and adapt their strategies to maximize their long-term performances. For example, a peer might provide upload services voluntarily if there are many peers who also cooperate; whereas, in a network where free-riders are the majority, the peer will choose to change its sharing behavior and adapt to a more selfish strategy. Hence, it becomes essential to understand how the peers’ learning abilities influences their own performances as well as the evolution of the network in the long run. We provide a systematic methodology to model and analyze the dynamic learning and adaptive behaviors of peers operating in a P2P network governed by social norms. In particular, depending on its observation on the network state, each peer dynamically adapts its best response to maximize its long-term utility by solving an individual Markov decision problem. We first show that when a peer learn against a group of peers playing a fixed social strategy, it can always adapt to a strategy which leads to a performance that is at least as good as the social strategy. When all peers in the network adapt their strategies simultaneously, we further show that our social norm based framework also provides peers with the incentives to play cooperatively in the long run under certain conditions, and thereby the P2P network will converge to a socially optimal outcome.

The remainder of this paper is organized as follows. In Section 2, we present our proposed social norm based framework for indirect reciprocity in P2P networks. In Section 3, we discuss the case when there is only one peer learning in the network while all the peers play the fixed social strategy. In Section 4, we extend the analysis to the case when all peers are learning in the network. In Section 5, we present our experimental results and we draw our conclusion in Section 6.

2. SYSTEM MODEL

We consider a P2P network consisting of $N$ peers which is fully connected. The network is modeled as a discrete-time system. At the beginning of a period, each peer randomly selects another peer who possesses desirable content in order to request services [4]. The stage game played by such a pair of peers is modeled as an asymmetric gift-giving game in order to characterize the asymmetry of interests among peers. A peer who requests resources is called a client and the peer who is being requested is called a server. In the stage game, only the server is strategic: it determines how to select its actions in order to maximize its long-term utility. The server’s action space can be quantified by $z \in \mathcal{Z} = \{0,1\}$, where $z = 1$ represents the case when the server provides service to the client and $z = 0$ represents the case when the server does not provide service. When the server provides a service, it consumes a cost of $c$ and its client gains a benefit of $r$. The social welfare of the P2P network is quantified by the social utility $U$ that is defined as the average utility received by all peers in the network. It is obvious that a self-interested server who has the incentive to free-ride will always choose $z = 0$ if it expects to maximize its stage-game utility myopically, which gives rise to an undesirable outcome of $U = 0$ (i.e. no peers contribute their resources to the network) [2][3].

To encourage the peers to cooperate, we design a new set of protocols in which peers are rewarded or punished by other peers based on the interaction rules specified by a pre-selected social norm. More specifically, each peer is tagged with a reputation $\theta \in \Theta = \{0,1,\ldots,L\}$ representing its social status. For notational convenience, a peer holding reputation $\theta$ is referred to as a $\theta$-peer. The reputation is maintained and updated by a trustworthy third-party device – e.g. the tracker – at the end of each period.

Formally, a social norm $\kappa$ is composed of a social strategy $\sigma$ and a reputation scheme $\tau$. The reputation-based behavioral strategy...
\( \sigma : \Theta \times \Theta \rightarrow \mathcal{A} \) imposes the action that a server should play depending on its own reputation as well as the reputation of the client. In this paper, we consider a set of simple strategies \( \Sigma = \{ \sigma_l, 0 \leq l \leq L + 1 \} \). When a peer adopts \( \sigma_l \), it provides services if and only if the reputation of the client is at least \( l \). The parameter \( l \) thus denotes the service threshold of \( \sigma_l \). If \( l = 0 \), the peer is fully cooperative; whereas if \( l = L + 1 \), the peer does not provide services to anyone. The reputation scheme \( \tau \) serves as the reward and punishment system in the social norm and it specifies how a peer’s reputation is updated according to its behavior. Specifically, \( \tau \) maps the server-client pair’s reputations and the server’s action to the server’s new reputation: \( \tau : \Theta \times \Theta \times \mathcal{A} \rightarrow \Theta \), and the update is subject to a small error probability \( \gamma \), which is referred to as the update error.

The working procedure of \( \tau \) can be briefly specified as follows: if a peer always plays with \( \sigma_l \), \( \tau \) assigns a reputation \( L \) to it; once a peer deviates from \( \sigma_l \), it gives it a reputation 0 and starts an L-period punishment for breaking the social norm; during the punishment phase, the peer has to follow the social norm in order to improve its reputation, and any time it deviates from \( \sigma_l \) again, the punishment is immediately restarted. Under the update rule of \( \tau \), a peer with a better social status faces larger threat of punishment from breaking the social norm and thus has less incentive to do so, which is in accordance to the social reality. Formally, \( \tau \) can be represented as

\[
\tau(\theta, \bar{\theta}, \sigma) = \begin{cases} 
\min[L, \theta + 1] & \text{if } z = \sigma(\theta, \bar{\theta}) \\
0 & \text{if } z = \sigma(\theta, \bar{\theta}) \end{cases}.
\]

As we assume that peers in the network try to maximize their individual utilities in their decision making, a peer’s decision on its action is thus determined by its belief on other peers’ actions. In this paper, we assume that each peer maintains a simple belief that all the peers except itself will follow the social strategy \( \sigma_l \). Denote the number of \( \theta \)-peers at period \( t \) as \( n(\theta) \), the expected stage-game utility that a \( \theta \)-peer can receive in period \( t \) is thus

\[
u^t(\theta) = \frac{1}{N - 1} \sum_{\bar{\theta}} n^t(\bar{\theta}) \nu^t(\theta, \bar{\theta}, \sigma(\theta, \bar{\theta})) - \sum_{a \in \mathcal{A}} \nu^t(\theta, a) \nu^t(\theta, \bar{\theta}, \sigma(a, \bar{\theta})) \] \tag{2}

where \( r(\theta, \bar{\theta}, \sigma(\theta, \bar{\theta})) \) is the stage-game benefit the peer can receive when the server has a reputation \( \theta \) and follows \( \sigma \); \( c(\theta, \bar{\theta}, \sigma(\theta, \bar{\theta})) \) is the stage-game cost of the peer if it follows a strategy \( \sigma \in \Sigma \) and the client’s reputation is \( \bar{\theta} \).

A peer’s long-term utility is evaluated using the infinite-horizon discounted sum criterion, and a peer’s expected overall utility in the repeated game starting from any period \( t_0 \) can be expressed as

\[
\nu^{\infty}(\theta) = \sum_{t = t_0}^{\infty} \beta^t \nu^t(\theta),
\]

where \( \beta \in [0, 1] \) is a peer-defined discount factor.

3. LEARNING AGAINST A FIXED STRATEGY

While adjusting its strategy during the repeated game, it should be noted that which strategy is the best for a peer depends on the peer’s own reputation as well as the strategies and the reputations of other peers. We start with a simple scenario in this section by analyzing how a peer learns to play in a network where all the other peers play a fixed strategy. For notational convenience, the learning peer is denoted as \( u \). At the beginning of each period \( t \), \( u \) observes its current reputation \( \theta^u_t \) and the reputation distribution \( \{n^t(\theta), 0 \leq \theta \leq L\} \)\(^1\) to determines its optimal strategy \( a^u \in \Sigma \). This decision problem can thus be solved under the framework of a Markov Decision Process (MDP) [5]. An MDP is a tuple \((\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \delta)\), where \( \mathcal{S} \) is a set of states, \( \mathcal{A} \) is a set of actions, \( p(a | s, a) \) is the probability of transition from state \( s \) to state \( s' \) after performing action \( a \), \( R(s, a) \) is the expected immediate reward for performing action \( a \) in state \( s \), and \( \delta \in [0, 1] \) is the discount factor.

In this paper, the action of \( u \) is a sharing strategy \( a \in \mathcal{A} \). The state \( s = \{\theta^u, (n(\theta), 0 \leq \theta \leq L)\} \). The state transition probability is

\[
p(s'|s,a) = \delta \left[ \frac{1}{n(t)} \sum \left\{ ((n(\theta)), a) | (n(\theta)) \right\} \right]
\]

Since all the peers other than \( u \) follow the social strategy \( \sigma_l \), it is easy to show that the transition of \( n(\theta) \) is Markovian and thus

\[
V(s) = \sum_{t=0}^{\infty} \delta^t \mathcal{E}(s', a') = R(s, a) + \delta \sum_{s'} p(s'|s)V(s').
\]

Using value iteration, the above MDP converges to the optimal policy for \( u \) as \( \pi^* : \mathcal{S} \rightarrow \mathcal{A} \), which maximizes the value function from each initial state \( s^0 \in \mathcal{S} \).

It should be noted that \( u \) tries to optimize its long-term utility by selecting a strategy from the set \( \mathcal{S} = \{\sigma_k, 0 \leq k \leq L + 1\} \) at any initial state \( s^0 \). Since the adopted social strategy \( \sigma_k \) is also included in \( \Sigma \), the optimal strategy chosen by \( u \) should deliver a long-term utility which is at least as high as that can be delivered by \( \sigma_k \). As a result, proposition 1 below concludes that \( \pi^* \) always outperforms the fixed social strategy \( \sigma_k \) in terms of the long-term utility, which is rigorously proved in [6] using the theory of stochastic control. The proof is omitted here due to the space limitation.

**Proposition 1:** With the optimal policy \( \pi^* \), the learning peer \( u \) can always achieve a higher long-term utility than the other peers adopting the social strategy \( \sigma_k \).

4. MULTI-PEER LEARNING AND THE LONG-RUN EVOLUTION

Section 3 analyzed the learning of a single peer against a group of peers playing the fixed social strategy \( \sigma_k \), and showed that such learning behavior is always beneficial for the learner in terms of its long-term utility. Here we extend the single-peer learning to multi-peer learning in this section, where all the peers in the network learn to adapt their strategies simultaneously, and analyze the evolution of the peers’ as well as the network’s behaviors in the long run when time goes to infinity.

For notational convenience, denote \( \mu = (n(\theta)) \) as the network state. At the beginning of period \( t \), each peer observes the current network state \( \mu^t = (n(\theta)) \) and solves an individual MDP problem to determine its best response against \( \mu^t \). Although none of the peers is following the social strategy \( \sigma_k \) unconditionally now, we still maintain the assumption that each peer holds the same belief that any peer other than itself plays \( \sigma_k \) at any stage of the game. To reduce the complexity.

---

\(1\) The reputation distribution of all peers in the network except \( u \).
of the MDP problem, we further assume that each peer does not include \( \mu' \) in the formulation of its state, but uses it as the stationary reputation distribution in the computation of its long-term value function. Therefore, unlike Section 3, a peer’s state here is simply its own reputation, i.e., \( s = \theta \). Since \( \mu' \) is the public information which can be observed by all peers, \( \mu' \) is straightforward that all peers should learn the same best response \( \phi^*_{\mu'} \) in each period by solving the MDP problem. Based on \( \phi^*_{\mu'} \), the optimal strategy that a \( \theta \)-peer should play in this period is \( \phi^*_{\mu'}(\theta) \in \Sigma \), which only depends on its reputation \( \theta \).

We then analyze the evolution of the network under such learning dynamics. Denote \( p^\theta(i) \) as the probability of the event that there will be \( i \) \( \theta \)-peers moving toward the reputation \( 0 \) in period \( t \) and the rest of \( \theta \)-peers move toward the reputation \( 0 \leq \theta \leq L \). We can argue that the set \( \{ p^\theta(i) \}_{0 \leq \theta \leq L, 0 \leq i \leq n'(\theta) \} \) only depends on the network state \( \mu' = \{ n'(\theta) \} \), as well as the best response \( \phi^*_{\mu'} \). As \( \phi^*_{\mu'} \) only depends on \( \mu' \), it is straightforward that \( \{ p^\theta(i) \}_{0 \leq \theta \leq L, 0 \leq i \leq n'(\theta) \} \) also only depends on \( \mu' \).

Therefore, the state of the network \( \{ \mu', 0 \leq t < \infty \} \) evolves as a Markov chain on the finite space

\[
M = \left\{ \mu = (n(0), n(1), \ldots, n(L)) \mid n(0) \in \mathbb{N} \text{ for } 0 \leq \theta \leq L, \text{ and } \sum_{i=0}^{L} n(i) = N \right\}, \quad \text{with size } \left\{ \frac{N+1}{L} \right\}.
\]

Let \( P \) denote the transition matrix of this Markov chain, it is obvious that the transition probability \( p_{\mu\mu'} = p(\mu' \mid \mu) \in P \) for any two states \( \mu, \mu' \in M \) is positive when \( \gamma > 0 \). We can then conclude that \( \{ \mu' \} \) is an irreducible and aperiodic Markov chain. Indexing all elements in \( M \) from 1 to \( M = \left\{ \frac{N+1}{L} \right\} \) and using the row vector \( \eta' = (\eta_1, \eta_2, \ldots, \eta_M) \) to denote the state distribution of the Markov chain, it is well known that this Markov chain has a unique stationary distribution \( \eta^* \) satisfying

\[
\eta^* = \eta^* P,
\]

which also preserves the following properties:

1. Stability: Starting from an arbitrary initial reputation distribution \( \mu^0 \in M \) whose index is \( m \), we have that

\[
\lim_{t \to \infty} \eta^0 P^t \to \eta^*.
\]

2. Ergodicity: Starting from an arbitrary initial reputation distribution \( \mu^0 \in M \), the fraction of time that the network being in state \( m \) is \( \eta^*_{m} \) as \( t \to \infty \), for any \( 1 \leq m < M \).

Therefore, as \( t \) goes to infinity, the network will always converge to the stationary state distribution \( \eta^* \), where each term \( \eta^*_{m} \) denotes the average fraction of time that the network has the reputation distribution \( \mu^* \). Moreover, as all entries in \( P \) are positive, each term in \( \eta^* \), i.e., the fraction of time spent at each reputation distribution, takes positive values, and thus, we can draw the following conclusion.

**Lemma 1:** When \( \gamma > 0 \), the system never converges to a stationary reputation distribution \( \mu^* \), starting from any initial reputation distribution \( \mu^0 \).

Lemma 1 characterizes the long run behavior in the network when the network error is large. The network will always oscillate between different reputation distributions and will never converge. Therefore, learning is not beneficial in this case, neither to the long-term utility of an individual peer nor to the social utility of the network.

Since in real P2P networks, the error is usually small, we examine next the long-run behavior of the network with small \( \gamma \). To this end, we analyze the limiting distribution of the network when \( \gamma \) approaches 0, which is defined as follows.

**Definition 1:** The limiting state distribution of the network is defined by

\[
\eta^* = \lim_{\gamma \to 0} \eta^*.
\]

Using the fact that each entry in \( P \) is a polynomial of \( \gamma \), it can be shown in [6] that \( \eta^* \) exists and is unique. However, it is rather complicated to derive the closed form of \( \eta^* \), which requires us to write down the close form of \( P \) for each \( \gamma \), and solve Eq. (6). Instead, we characterize \( \eta^* \) by analyzing its properties in the rest of this section.

To do this, we start by characterizing the properties of the best response \( \phi^*_{\mu'} \) when \( \gamma \) goes to 0.

When \( \gamma \to 0 \), the reputation transition of a peer becomes deterministic once its action is specified. If the learnt strategy of a \( \theta \)-peer is \( \sigma_{l} \), it is easy to show that this peer can always obtain a higher current reward as well as a higher future reward by adopting \( \sigma_{l} \). This argument is valid for any reputation \( \theta \), and hence, we can conclude that the optimal strategy \( \phi^*_{\mu'}(\theta) \) has a service threshold which is at least \( k_{l} \), regardless of its reputation \( \theta \) and the current state \( \mu \) of the network.

**Lemma 2:** When \( \gamma \to 0 \), \( \phi^*_{\mu'}(\theta) \) always has a service threshold which is at least \( k_{l} \), for any \( \theta \in \Theta \) and \( \mu \in M \).

**Proof:** See technical report [6].

Further, we analyze the ordinal relationship between \( \theta \) and \( \phi^*_{\mu'}(\theta) \).

In any period, the average service that has to be provided for a peer is constant regardless of its reputation, since it only depends on the current state of the network. On the other hand, the average service that a peer expects to receive monotonically increases with its reputation. Consequently, a peer’s incentive to follow the social strategy \( \sigma_{l} \) monotonically increases with its reputation, with the resulting optimal strategy more close to \( \sigma_{l} \). This is formalized in the following lemma as the service thresholds of \( \{ \phi^*_{\mu'}(\theta) \} \) always preserve the order which is non-increasing on \( \theta \) [6].

**Lemma 3:** When \( \gamma \to 0 \), for any two reputations \( \theta_{1} > \theta_{2} \) and any \( \mu \in M \), we always have the service threshold \( \phi^*_{\mu'}(\theta_{1}) \) equal to or smaller than \( \phi^*_{\mu'}(\theta_{2}) \).

Now we can use the above two properties of the best response to characterize the evolution of the network and peers’ behavior in the long run. With the monotonicity on service thresholds in the optimal strategies, it is obvious that the higher a peer’s reputation is, the more likely it will become cooperative and adopt the social strategy \( \sigma_{l} \).

Meanwhile, an \( L \)-peer will always choose to follow \( \sigma_{l} \), and maintains its reputation with \( L \) is an absorbing reputation in the network. Therefore, if a \( \theta \)-peer provides some services to others in its optimal strategy \( \phi^*_{\mu'}(\theta) \), the probability that its reputation moves towards \( L \) is positive and it will finally reach the reputation \( L \) in finite number of periods and remain to be a \( L \)-peer thereafter. Therefore, only two reputations can exist in the reputation distribution after sufficiently.
long time: a peer will either hold the reputation $L$ and follow $\sigma^L$, or hold the reputation $0$ and choose not to provide any service by following the strategy $\sigma^0$; all the reputations other than $L$ and $0$ are transient reputations. Since a peer’s reputation will not change once it reaches either $L$ or $0$, the network will converge to a stationary reputation distribution. This conclusion is formalized in the following proposition.

**Proposition 2:** When $\gamma \to 0$, there is a unique stationary reputation distribution $\mu^* = \{n^*(\theta)\}$ in the network, which preserves the following property

$$n^*(\theta) = 0, \text{ for any } 1 \leq \theta \leq L-1.$$

**Proof:** See technical report [6].

Since transient reputations do not exist in $\mu^*$, it will be desirable if $n^*(0)$ equals to 0 as well. In this case, there are only $L$-peers in the network, i.e. $n^*(L) = n^*$, and all peers will be mutually cooperative. The social utility is then maximized as $r-c$. This situation can be realized when the ratio of the stage-game benefit and cost $r/c$ and the peer’s discount factor $\delta$ are sufficiently large, if the future benefit of holding a good reputation outweighs the instant cost of providing services, and even a 0-peer will have sufficient incentive to follow $\sigma^L$ in order to increase its reputation. The above argument is formalized in the next proposition and its proof can be found in [6].

**Proposition 3:** When $\gamma \to 0$ and $\frac{r}{1-\delta} > c$, there is a unique stationary reputation distribution $\mu^*$ in the network, with $n^*(L) = n^*$.

5. EXPERIMENTS

In our experiments, we simulate the long-run behavior of 1000 peers. The punishment length $L = 5$ and a peer’s initial reputation is randomly selected from the set $\{0,1,\ldots,5\}$. For notational convenience, we also assume a normalized service cost $c = 1$.

In our previous analysis, each peer solves an MDP problem offline at the beginning of each period $t$ to obtain the optimal strategy for the current network state $\mu^t$. Due to the high computational complexity, this offline approach is difficult to be implemented in real P2P networks. Therefore, we use the online Q-Learning method instead [5]. It is a simple update step based on the experienced tuple $(s^{t-1}, a^{t-1}, r^{t-1}, s^t)$:

$$Q^t(s^{t-1}, a^{t-1}) \leftarrow Q^t(s^{t-1}, a^{t-1}) + \alpha [r^{t-1} + \delta \max_{a'} Q^t(s', a') - Q^t(s^{t-1}, a^{t-1})],$$

where $s^{t-1}, a^{t-1}, r^{t-1}$, and $s^t$ are the state, performed action, and corresponding utility for a peer in period $t-1$, respectively; $s'$ is the peer’s resulting state in period $t$; and $\alpha \in [0,1]$ is the time-varying learning rate parameter.

During the learning process, to judiciously trade off exploration and exploitation [5], we use the $\varepsilon$-greedy action selection method: with probability $1 - \varepsilon$, take the greedy action that maximizes the action-value function; and, with probability $\varepsilon$, take an action randomly and uniformly over the action set $\mathcal{A}$.

We run the network for $10^6$ periods and counting the average fraction of each reputation in the population over time. Figure 1 and Figure 2 illustrates the detailed results. Figure 1 shows the reputation distributions when $\gamma = 0.2$ and $0.001$, respectively. As we have proved in Section 4, when $\gamma$ is large, each reputation takes a positive fraction in the long run, as the network oscillates and cannot converge to a unique stationary reputation distribution. On the other hand, when $\gamma \to 0$, only 0-peers and $L$-peers remain in the long run, while 0-peers provides no service and $L$-peers are mutually cooperative to each other. Figure 2 shows the reputation distribution when the stage-game benefit $r$ varies. Here we adopt a $\gamma$ close to 0 and thus only 0-peers and $L$-peers remain. When $r$ is small, the instant saving of the service cost outweigh the future benefit of obtaining a high reputation and hence, a majority of peers will maintain the minimum reputation (the reputation of 0). When $r$ is large, it is more attractive to obtain a high reputation to receive higher future download benefit, and thus the network converges to the all-cooperative state with $n^*(L) = N$.

![Figure 1](image1.png)

**Figure 1.** The average fraction of each reputation after $10^6$ periods $(l_0 = 2, r = 3, \delta = 0.8)$: (a) $\gamma = 0.2$; (b) $\gamma = 0.001$

![Figure 2](image2.png)

**Figure 2.** The average fraction of each reputation after $10^6$ periods $(l_0 = 2, \gamma = 0.001, \delta = 0.8)$: (a) $r = 1.5$; (b) $r = 5$

6. CONCLUSION

We modeled the P2P resource sharing as a random-matching gift-giving game and used the concept of social norm to formalize the incentive mechanisms which are based on indirect reciprocity. With peers learning the optimal sharing strategy by solving individual Markov decision problems, we first proved that a learning peer can always outperform opponents deploying a fixed strategy and showed the fact that learning is advantageous in terms of the long-term utility in this case. We then extended the single-peer learning to the multi-peer learning where all peers learn to play using best response dynamics. Different from the single-peer case, learning is not necessarily beneficial here as the network will oscillate and never converge when $\gamma$ is large. However, we are able to show that when $\gamma$ is sufficiently small, and $r/c$ and $\delta$ are sufficiently large, such best response dynamics can lead the network to the “all-cooperative” state in which the social utility of the network is maximized. Our experiments illustrated that when using online Q-learning methods, peers can also learn to play cooperatively in the long run and achieve a socially optimal outcome under the same conditions.

7. REFERENCES