A Framework for Foresighted Resource Reciprocation in P2P Networks

Hyunggon Park, *Student Member, IEEE*, and Mihaela van der Schaar, *Senior Member, IEEE*

***Abstract*—We consider peer-to-peer (P2P) networks, where multiple peers are interested in sharing multimedia content. In such P2P networks, the shared resources are the peers' contributed content and their upload bandwidth. While sharing resources, autonomous and self-interested peers need to make decisions on the amount of their resource reciprocation (i.e., representing their actions) such that their individual utilities are maximized. We model the resource reciprocation among the peers as a stochastic game and show how the peers can determine optimal strategies for resource reciprocation using a Markov Decision Process (MDP) framework. Unlike existing resource reciprocation strategies, which focus on myopic decisions of peers, the optimal strategies determined based on MDP enable the peers to make foresighted decisions about resource reciprocation, such that they can explicitly consider both their immediate as well as future expected utilities. To successfully formulate the MDP framework, we propose a novel algorithm that identifies the state transition probabilities using representative resource reciprocation models of peers. These models express the peers' different attitudes toward resource reciprocation. We analytically investigate how the error between the true and estimated state transition probability impacts each peer's decisions for selecting its actions as well as the resulting utilities. Moreover, we also analytically study how bounded rationality (e.g., limited memory for reciprocation history and the limited number of state descriptions) can impact the interactions among the peers and the resulting resource reciprocation. Simulation results show that the proposed approach based on reciprocation models can effectively cope with a dynamically changing environment such as peers' joining or leaving P2P networks. Moreover, we show that the proposed foresighted decisions lead to the best performance in terms of the cumulative expected utilities.**

***Index Terms*—Bounded rationality, foresighted decision, Markov decision process, peer-to-peer (P2P) network, resource reciprocation game.**

I. Introduction

Peer-to-peer (P2P) applications (e.g., [1]–[3]) have recently become increasingly popular and represent a large majority of the traffic currently transmitted over the Internet. One of the unique aspects of P2P networks stems from their flexible and distributed nature, in which each peer can act as both server and client [4]. Hence, P2P networks provide a cost effective and easily deployable framework for disseminating large files without relying on a centralized infrastructure [5]. Due to these characteristics, it has been recently proposed to use P2P networks for general file sharing [2], [3], [6], [7] as well as multimedia streaming [5], [8]–[10]. Moreover, several media streaming systems have been successfully developed for P2P networks using different approaches such as tree-based or data-driven approaches (e.g., [11], [12]). In this paper, we focus on data-driven P2P systems such as CoolStreaming [8] and Chainsaw [9] for multimedia streaming, or BitTorrent systems [6], [7] for general file sharing, which adopt pull-based techniques [8], [9]. In these systems, data (i.e., multimedia stream or general files) are divided into chunks of uniform length, which are distributed over the P2P network. Each peer possesses several chunks which are shared among interested[1] peers. Information about the availability of the chunks is also periodically exchanged among the associated peers. Using this information, peers continuously associate themselves with other peers and exchange their chunks. While this approach has been successfully deployed in real-time multimedia streaming and file distributions over P2P networks, key challenges such as determining optimal resource reciprocation among self-interested peers still remain largely unaddressed. Specifically, pull-based techniques assume that the peers in the P2P network are altruistic and they provide their available chunks whenever requested. However, such a reciprocation strategy is undesirable from the perspective of a self-interested peer, who is aiming at maximizing its utility.

The resource reciprocation strategy deployed in BitTorrent is based on the equal upload bandwidth distribution. A peer in BitTorrent systems thus equally divides its available upload bandwidth among multiple leechers [6], [7]. However, for heterogeneous content and diverse peers (with different upload/download requirement), such reciprocation strategies are not optimal. The resource reciprocation in [8] is based on a heuristic scheduling algorithm, which enables the peers to determine the suppliers of required chunks and select the peer with the highest bandwidth. Alternatively, the resource reciprocation can be based on the random chunk selection algorithm as in [9]. As discussed, the solutions in [8] or [9] are implemented assuming that the associated peers are altruistic, such that they provide the chunks and bandwidth whenever requested. Hence, the resource reciprocation methods in these solutions do not

Manuscript received March 16, 2008; revised September 16, 2008. First published December 16, 2008; current version published January 09, 2009. This work was supported by NSF CAREER Award CCF-0541867, and grants from Microsoft Research. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ling Guan.

The authors are with the Electrical Engineering Department, University of California, Los Angeles CA 90095, USA (e-mail: hgpark@ee.ucla.edu; mihaela@ee.ucla.edu).

Digital Object Identifier 10.1109/TMM.2008.2008925

[1]In [6], [7], it is said that peer A is *interested* in peer B when B has chunks of the content that A would like to possess.

consider the strategic interactions of the heterogeneous and self-interested peers.

To take into account the interactions of heterogeneous and self-interested peers in P2P networks, game theoretic approaches have been proposed. In [13], a micropayment mechanism is used to model the rational peers' interactions, and the resulting equilibria emerging when different payment mechanisms are imposed. In general, a key assumption is that peers will follow the prescribed P2P protocols. It has been found, however, that self-interested peers will deviate from the prescribed protocols or free-ride unless preemptive solutions exist in the network. For example, in [14], mechanism design solutions are proposed in order to compel the peers to adhere to their reciprocation promises. In [15], an incentive scheme for compelling peers to contribute resources is proposed, which provides differential services based on the peer's past contributions. The interactions for different types of peers (e.g., homogeneous or heterogeneous) are analyzed using the notion of Nash equilibrium. In the above approaches, however, the peers determine their decisions (i.e., actions) to maximize their utilities myopically, without explicitly considering the future impact of the actions on their long-term utilities. In [16], the repeated interactions among peers are modeled as an evolutionary instantiation of the Prisoner's Dilemma and the Generalized Prisoner's Dilemma, and incentive techniques are proposed for peers in order to compel them to contribute their resources. However, this research only considers the case where peers have a limited set of simple actions, i.e., allowing download or ignoring download requests, but does not address how to divide each peer's available resources. Hence, they do not provide solutions for maximizing the foresighted utilities of peers, which is essential in P2P systems, where peers have long-term interactions.

To address these challenges, in this paper, we model the resource reciprocation among the interested peers as a stochastic game [17], where peers determine their resource distributions by explicitly considering the probabilistic behaviors (reciprocation) of their associated peers. Unlike existing resource reciprocation strategies, which focus on myopic decisions, we formalize the resource reciprocation game as a Markov Decision Process (MDP) [18] to enable peers to make foresighted decisions on their resource distribution in a way that maximizes their cumulative utilities, i.e., their immediate as well as future utilities.

To successfully formulate the resource reciprocation game as an MDP problem, the peers need to identify the associated peers' probabilistic behaviors for resource reciprocation. The probabilistic behaviors of the associated peers can be estimated using the past history of resource reciprocation and are represented by *state transition probabilities* in the MDP framework. In this paper, the state of a peer is defined as the set of received resources from each of the associated peers. Hence, the actions of the associated peers determine a peer's state. We propose a novel algorithm that can efficiently identify the state transition probabilities using peers' *reciprocation models*. The reciprocation models of the peers are motivated by [19], which classify the rational attitudes of players in a game towards their strategies as optimistic, pessimistic, and neutral archetypes. We construct reciprocation matrices to capture the reciprocation behaviors of peers. Then, the state transition probabilities are identified by linear combinations of weighted reciprocation matrices. Note that the decisions made by peers based on the estimated state transition probabilities can lead to different resource reciprocation strategies than those based on the true state transition probabilities, thereby possibly deviating from the actual derived utility. This impact on the accuracy of the estimated utility is analytically quantified.

Unlike the implicit assumptions on players' rationality in conventional game theory, where players have the abilities to collect and process relevant information, and select alternative actions among all possible actions [20], [19], we consider the *bounded rationality* [20] of peers. This is because perfectly rational decisions are often infeasible in practice due to memory and computational constraints. To illustrate the effects of bounded rationality, we consider cases where the peers have limited memory for storing the resource reciprocation history, and have a limited number of states based on which they make their decisions. We also quantify the impact of the bounded rationality on the peers' interactions and their utilities.

This paper is organized as follows. In Section II, we model the resource reciprocation among peers as a resource reciprocation game. In Section III, the types of peers in the considered P2P networks are discussed. They are classified based on their objectives in terms of utilities and their resource reciprocation attitudes. In Section IV, we analytically investigate the interactions among different types of peers with different constraints. In Section V, an algorithm that determines the state transition probabilities based on the reciprocation models is proposed. We analytically quantify the impact of this approach on the derived utility. Simulation results are provided in Section VI and conclusions are drawn in Section VII. For reader's convenience, we summarize several notations frequently used in this paper in Table I.

TABLE I
SUMMARY OF NOTATIONS

Notation	Description
C_i	group of peer i (with N_{C_i} associated peers)
s_i	state of peer i, $s_i = (s_{i1}, \ldots, s_{iN_{C_i}})$
$\mathbf{S}_i$	state space of peer i
$\mathbf{a}_i$	action of peer i, $\mathbf{a}_i = (a_{i1}, \ldots, a_{iN_{C_i}})$
$\mathbf{A}_i$	action space of peer i
ψ_i	state mapping function of peer i
$R_i(s_i)$	reward of peer i in s_i
x_{ik}	allocated resources to peer k from peer i
n_{ik}	number of state descriptions
$P_{\mathbf{a}_i}(s_i, s_i')$	state transition probability from s_i to s_i' given $\mathbf{a}_i$
π_i	reciprocation policy of peer i
L_i	available maximum upload bandwidth of peer i
Δ_{ik}^O	degree of optimism of peer i to peer k
Δ_{ik}^P	degree of pessimism of peer i to peer k
$M_{il}^k(s_{ik})$	a reciprocation matrix of peer i in s_{ik}
D_{J_l}	distance metric (see Proposition 4)

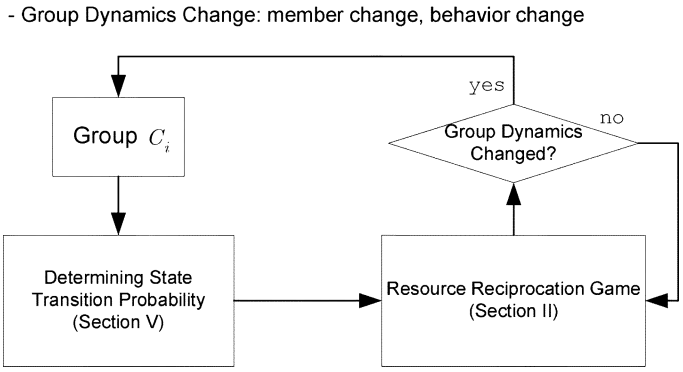

Fig. 1. Group update process and related processes.

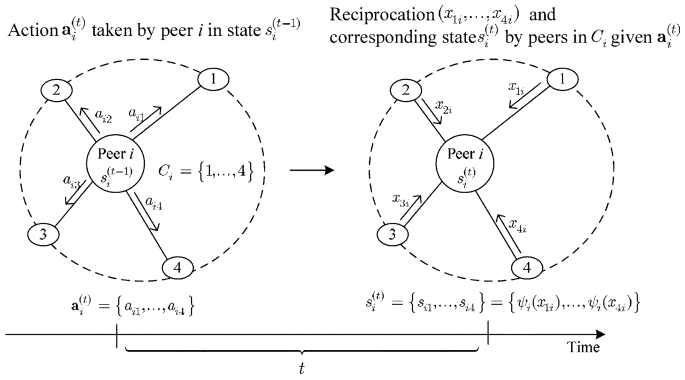

Fig. 2. Illustrative example for a resource reciprocation in C_i with four associated peers at t. The resource reciprocation at t is denoted by $(\mathbf{a}_i^{(t)}, s_i^{(t)})$ in this example.

II. A New Framework For Resource Reciprocation

In P2P networks, peers would like to associate themselves with other peers that possess multimedia content in which they are interested. When peers agree to share content with each other, they negotiate the amount of resources which they will provide to each other. We model the resource reciprocation among the peers as a resource reciprocation game. We begin with a motivating example describing why the foresighted decisions on the actions are important and how they can be beneficial to peers.

A. A Simple Motivating Example for Myopic and Foresighted Reciprocation

In this illustrative example, we consider a simple resource reciprocation game, where two self-interested peers interact with each other to negotiate what resources they will provide to each other. In this example, the peers' actions are the divisions of their available resources (e.g., percentage of available upload bandwidth) among their associated peers, and the states of the peers are determined based on their received resources. Hence, one peer's action can determine the other peer's state and reward[2]. We assume that the resource reciprocation behaviors are perfectly known by both peers. Thus, both peers know the probabilities with which the other peer takes a certain action given their own actions. In this illustrative example, we assume that the available actions of peer 1 and peer 2 are $A_1 = \{a_{12}^1, a_{12}^2\}$ and $A_2 = \{a_{21}^1, a_{21}^2\}$, respectively. Suppose that peer 1 and peer 2 currently take their actions a_{12}^1 and a_{21}^1, and hence, their current state is given by $s = (s_1, s_2) = (a_{21}^1, a_{12}^1)$. For example, peer 1 in state s_1 can take action a_{12}^1, while expecting that peer 2 will take action a_{21}^1 with probability $\Pr(a_{21}^1|a_{12}^1, s_1)$ and action a_{21}^2 with probability $\Pr(a_{21}^2|a_{12}^1, s_1)$. Hence, the expected reward $\bar{R}(a_{12}^1, s_1)$ for peer 1 that takes a_{12}^1 in state s_1 becomes

$$\bar{R}\left(a_{12}^1, s_1\right) = a_{21}^1 \Pr\left(a_{21}^1|a_{12}^1, s_1\right) + a_{21}^2 \Pr\left(a_{21}^2|a_{12}^1, s_1\right). \quad (1)$$

Similarly, the expected reward $\bar{R}(a_{12}^2, s_1)$ for peer 1 that takes a_{12}^2 in state s_1 can be expressed as $\bar{R}(a_{12}^2, s_1)$. Therefore, peer 1 makes decision on its action $a_{12}^{j^*}$, such that it maximizes its expected reward from state s_1, i.e.,

$$j^* = \arg\max_{j\in\{1,2\}} \left\{\bar{R}(a_{12}^j, s_1)\right\}. \quad (2)$$

[2]The reward can be defined as the total received resources or the resulting utilities. More detailed definition of the reward in this paper will be discussed in Section II-B.

Note that the decision of peer 1 given in (2) can be interpreted as *myopic* since it does not consider the future rewards by taking action $a_{12}^{j^*}$ in state s_1, but rather it focuses only on maximizing its immediate expected reward. Hence, this decision may not be optimal if the future rewards are considered.

Let us now consider the *foresighted* decisions of the peers on their actions, which can maximize the cumulative rewards including the immediate expected reward and the future (discounted) expected rewards. In this illustration, we assume that the peers make foresighted decisions considering the one step future reward. Hence, a foresighted peer 1 needs to consider the future reward $\bar{R}(a_{12}^{k^*}, s_1')$ that can be derived in a next state s_1' with the corresponding optimal action $a_{12}^{k^*} \in A_1$. Therefore, peer 1 in state s_1 determines its action considering the cumulative discounted expected reward, i.e.,

$$j^* = \arg\max_{j\in\{1,2\}} \left\{ \bar{R}\left(a_{12}^j, s_1\right) + \gamma \sum_{s_1'} \bar{R}\left(a_{12}^{k^*}, s_1'\right) \Pr\left(s_1'|a_{12}^j, s_1\right) \right\} \quad (3)$$

where γ is a constant referred to as the discount factor and k^* is determined by $k^* = \arg\max_{k\in\{1,2\}}\{\bar{R}(a_{12}^k, s_1')\}$. Note that the decisions in (2) are a subset of the decision in (3) (i.e., (3) is identical to (2) if $\gamma = 0$). Hence, if the decisions based on (2) and (3) are different, it can be inferred that an optimal action that maximizes the immediate expected reward cannot be the optimal action that maximizes the cumulative rewards.

Summarizing, as shown in the above example, peers need to take foresighted decisions when engaging in resource reciprocation games.

B. Resource Reciprocation Games in P2P Networks

Resource reciprocation games in P2P networks are played by the peers interested in each other's multimedia content. A resource reciprocation game is played in a *group*, where a group consists of a peer and its associated peers. A group can be swarms in [6], [7], partnerships in [8], or neighbors in [9]. We denote the associated group members of a peer i by C_i. Note that C_i does not include peer i but represents the associated peers with peer i. The peers in C_i are indexed by $1,\ldots,N_{C_i}$, i.e., $C_i = \{1,\ldots,N_{C_i}\}$. For a peer k in group C_i,

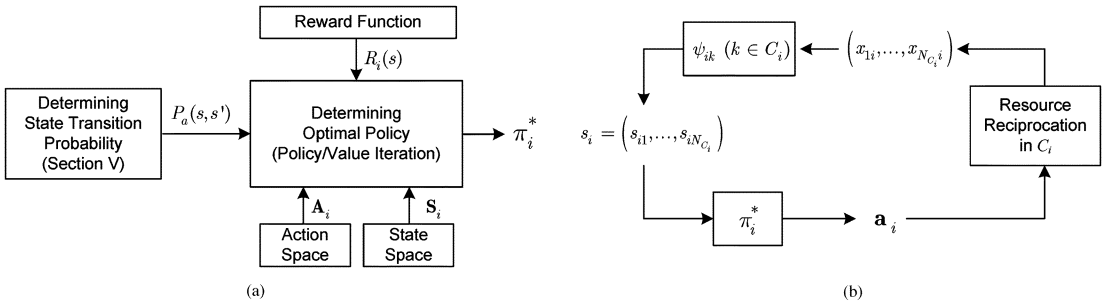

Fig. 3. Resource reciprocation game played by peer i based on the MDP. (a) Determining optimal policy (b) Determining optimal actions.

peer k also has its own group C_k which includes peer i. Due to the dynamics introduced by peers joining, leaving, or switching P2P networks, information about groups needs to be regularly (periodically) updated or it needs to be updated when group members change [6], [7]. This is shown in Fig. 1.

The resource reciprocation game in a group C_i is a stochastic game [17], which consists of

- a finite set of players (i.e., peers): $C_i \cup \{i\}$;
- for each peer $l \in C_i \cup \{i\}$, a nonempty set of actions: $\mathbf{A}_l$;
- for each peer $l \in C_i \cup \{i\}$, a preference relation (i.e., utility function) of peer l : $U_l(\cdot)$.

To play the resource reciprocation game, a peer can deploy an MDP, as discussed as follows.

For a peer i, an MDP is a tuple $< \mathbf{S}_i, \mathbf{A}_i, P_i, R_i >$, where $\mathbf{S}_i$ is the state space, $\mathbf{A}_i$ is the action space, $P_i : \mathbf{S}_i \times \mathbf{A}_i \times \mathbf{S}_i \rightarrow [0,1]$ is a state transition probability function that maps the state $s_i \in \mathbf{S}_i$ at time t, corresponding action $\mathbf{a}_i \in \mathbf{A}_i$ and the next state $s_i' \in \mathbf{S}_i$ at time $t+1$ to a real number between 0 and 1, and $R_i : \mathbf{S}_i \rightarrow \mathbb{R}$ is a reward function, where $R_i(s_i)$ is a reward derived in state $s_i \in \mathbf{S}_i$. The details are explained as follows.

1) State Space $\mathbf{S}_i$: A state of peer i represents the set of received resources from the peers in C_i, expressed as

$$\{(x_{1i},\ldots,x_{N_{C_i}i}) \big| \, 0 \le x_{ki} \le L_k, \forall k \in C_i\} \tag{4}$$

where x_{ki} denotes the provided resources (i.e., rate) by peer k in C_i and L_k represents the available maximum upload bandwidth of peer k[3]. The total received rates of peer i in C_i is thus $\sum_{k \in C_i} x_{ki}$. Due to the continuity of x_{ki}, the cardinality of the set defined in (4) can be infinite. Hence, we assume that peer i has a function ψ_{ik} for peer k, which maps the received resource x_{ki} into one of n_{ik} discrete values[4], i.e., $\psi_{ik}(x_{ki}) = s_{ik} \in \{s_{ik}^1,\ldots,s_{ik}^{n_{ik}}\}$. These values are referred to as *state descriptions* in this paper. Hence, the state space can be considered to be finite. The number of state descriptions will impact its performance and this will be discussed in Section IV. The state space of peer i can be expressed as

$$\mathbf{S}_i = \{s_i = (s_{i1},\ldots,s_{iN_{C_i}}) | s_{ik} = \psi_{ik}(x_{ki}), k \in C_i\} \tag{5}$$

[3]Note that the available maximum upload bandwidth L_k can be time-varying, because it depends on the physical maximum upload bandwidth (L_k^{phy}) as well as the available data (i.e., chunks) (L_k^{data}) that can be transmitted. Hence, L_k can be determined by $L_k = \min\{L_k^{\text{phy}}, L_k^{\text{data}}\}$. For example, in the initial stage of file sharing, $L_k = L_k^{\text{data}}$ because peer k may not have enough chunks to transmit. However, L_k increases as peer k receives more chunks. While we assume that $L_k = L_k^{\text{phy}}$ in this paper, $L_k = L_k^{\text{data}}$ can be explicitly addressed by selecting discount factors, which will be discussed in Section III-A.

[4]A continuous value of x_{ik} can be discretized by peer i based on its quantization policy, as the bandwidth of each peer can be decomposed into several "units" of bandwidth by the client software, e.g., [21].

where s_{ik}^l denotes the lth segment among n_{ik} segments that corresponds to the lth state description of peer i. For simplicity, we assume that each segment represents the uniformly divided total bandwidth, i.e., $\psi_{ik}(x_{ki}) = s_{ik}^l$ if $(l-1)\cdot(L_k)/(n_{ik}) \le x_{ki} < l \cdot (L_k)/(n_{ik})$ for $1 \le l \le n_{ik}$.

2) Action Space $\mathbf{A}_i$: An action of peer i is its resource allocation to the peers in C_i. Hence, the action space of peer i in C_i can be expressed as

$$\mathbf{A}_i = \Bigg\{ \mathbf{a}_i = (a_{i1},\ldots,a_{iN_{C_i}}) | 0 \le a_{ik} \le L_i, \; 1 \le k \le N_{C_i}, \sum_{k \in C_i} a_{ik} \le L_i \Bigg\} \tag{6}$$

where $a_{ik} \in A_i$ denotes the allocated resources to peer k by peer i in C_i. Hence, peer i's action a_{ik} to peer k becomes peer k's received resources from peer i, i.e., $a_{ik} = x_{ik}$. To consider a finite action space, we assume that the available resources (i.e., upload bandwidth) of peers are decomposed into "units" of bandwidth [21]. Thus, the actions represent the number of allocated units of bandwidth to the associated peers in their groups. We define the *resource reciprocation* as a pair $(\mathbf{a}_i, s_i) = ((a_{i1},\ldots,a_{iN_{C_i}}),(s_{i1},\ldots,s_{iN_{C_i}}))$ comprising the peer i's action, a_{ik}, and the corresponding modeled resource reciprocation s_{ik}, which is determine as $s_{ik} = \psi_{ik}(x_{ki})$ for all $k \in C_i$. An illustrative resource reciprocation at t is shown in Fig. 2.

Note that various scheduling schemes can be used in conjunction with the resource allocation (i.e., actions) deployed by peers in order to consider the different priorities of the different data segments (chunks). We assume that the chunks that have higher quality impact on average multimedia quality have higher priority and are transmitted first when each peer takes its actions. However, other scheduling algorithms, such as the *rarest first* [6], [7] method for general file sharing applications or several scheduling methods proposed in e.g., [8] for multimedia streaming applications, can also be adopted. It is important to note that appropriate scheduling schemes need to be deployed in conjunction with our proposed resource reciprocation strategies, depending on the objectives of multimedia applications (e.g., maximizing achieved quality, minimizing the playback delay etc.). However, the selection of scheduling strategies was already investigated in several existing papers and it is not the focus of this paper, as existing scheduling solutions can be easily incorporated into the proposed framework.

3) State Transition Probability $P_{\mathbf{a}_i}(s_i, s_i')$: A state transition probability represents the probability that by taking an action, a peer will transit into a new state. We assume that the state transition probability depends on the current state and the action

taken by the peer, as peers decide their actions based on their currently received resources (i.e., state). Hence, given a state $s_i \in \mathbf{S}_i$ at time t, an action $\mathbf{a}_i \in \mathbf{A}_i$ of peer i can lead to another state $s_i' \in \mathbf{S}_i$ at t' $(t' > t)$ with probability $P_{\mathbf{a}_i}(s_i, s_i') = \Pr(s_i'|s_i, \mathbf{a}_i)$. Hence, for a state $s_i = (s_{i1}, \ldots, s_{iN_{C_i}})$ of peer i in C_i, the probability that an action $\mathbf{a}_i$ leads to a state transition from s_i to s_i' can be expressed as

$$P_{\mathbf{a}_i}(s_i, s_i') = \prod_{l=1}^{N_{C_i}} P_{a_{il}}(s_{il}, s_{il}'), \tag{7}$$

where $P_{a_{il}}(s_{il}, s_{il}') = \Pr(s_{il}'|s_{il}, a_{il})$. In this paper, the state transition probabilities of peers are identified based on the past resource reciprocation history. The details of how to build the state transition probability functions will be discussed in Section V.

4) Reward R_i: The utility of peer i downloading its desired multimedia content from its peers at rate x_i can be defined as

$$U_i(x_i) = \begin{cases} 0, & \text{if } x_i < R_i^{\text{req}}, \\ \rho_i \cdot Q_i(x_i), & \text{otherwise}, \end{cases} \tag{8}$$

where R_i^{req} is the minimum resource that corresponds to the minimum required utility and ρ_i is a constant representing the preference of peer i for the content. The minimum resources (rates) are explicitly considered in the utility definition in (8) in order to provide support the quality of service (QoS) required by delay-sensitive and bandwidth-intensive multimedia applications [22]. The derived quality $Q_i(x_i)$ with downloading rate x_i can be represented by a widely used quality measure, peak signal-to-noise ratio (PSNR), which is a non-decreasing function of x_i for multimedia applications [22]. Thus, we consider that the reward $R_i(s_i)$ for a peer i in state s_i is the total received resources in C_i. Since the state of a peer i is defined as the quantized received resources from the peers in C_i, the reward in each state can be represented by a random variable, i.e.,

$$R_i(s_i) = R_i(s_{i1}, \ldots, s_{iN_{C_i}}) = \sum_{k \in C_i} r_i(s_{ik}) \tag{9}$$

where $r_i(s_{ik})$ is a random variable that represents the received resource in s_{ik}. Thus, the resulting utility of peer i in state s_i is $U_i(\sum_{k \in C_i} r_i(s_{ik}))$.

5) Reciprocation Policy π_i^*: The solution to the MDP is represented by peer i's optimal policy π_i^*, which is a mapping from the states to optimal actions. The optimal policy can be obtained using well-known methods such as value iteration and policy iteration [18]. Hence, peer i can decide its actions based on the optimal policy π_i^*, i.e., $\pi_i^*(s_i) = \mathbf{a}_i$ for all $s_i \in \mathbf{S}_i$.

Hence, the resource reciprocation games in the P2P networks that consist of N total peers can be described by a tuple $(\mathcal{I}, \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$, where $\mathcal{I}$ is the set of N peers, $\mathcal{S}$ is the set of state profiles of all peers, i.e., $\mathcal{S} = \mathbf{S}_1 \times \cdots \times \mathbf{S}_N$, and $\mathcal{A} = \mathbf{A}_1 \times \cdots \times \mathbf{A}_N$ denotes the set of action profiles. $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is a state transition probability function that maps from the current state profile $s \in \mathcal{S}$, corresponding joint action $a \in \mathcal{A}$ and the next state profile $s' \in \mathcal{S}$, into a real number between 0 and 1, and $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^N$ is a reward function that maps an action profile $a \in \mathcal{A}$ and a state profile $s \in \mathcal{S}$ into the derived reward. Thus, in this paper, our focus is on the resource reciprocation game in a group, as this resource game can be extended to the resource reciprocation game in a P2P network. The resource reciprocation game, which includes the processes of determining an optimal policy and optimal actions, played by peer i based on the MDP is shown in Fig. 3.

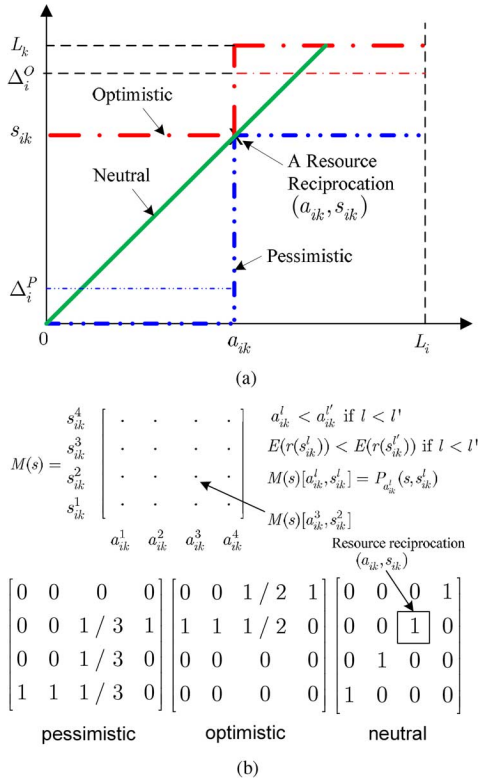

Fig. 4. Illustration of resource reciprocation based on peers' attitudes. (a) Resource reciprocation models (b) Examples of reciprocation matrices (four actions and four state descriptions).

We will discuss how the optimal policies can be determined based on different types of peers in the next section.

III. Categories of Peers in P2P Networks

The types of peers in the considered P2P networks can be characterized based on different criteria. In this paper, we categorize the peers as

- myopic or foresighted: depending on their objective utilities; and
- pessimistic, neutral, or optimistic: depending on their resource reciprocation attitudes.

These different types of peers affect how the resource reciprocation game is being played, thereby leading to various reciprocation policies π_i^*. More specifically, the decisions of myopic or foresighted peers directly influence their action selection. Moreover, various resource reciprocation attitudes lead to different state transition probability functions, which eventually impacts their actions and the resulting reciprocation policies.

A. Peer Types Depending on Their Adopted Utilities

In this paper, we consider two types of peers, myopic and foresighted peers, based on the utilities which they adopt. Myopic peers only focus on maximizing the immediate expected reward. Hence, the objective of a myopic peer i in state $s_i^{(t)} =$

$(s_{i1}, \ldots, s_{iN_{C_i}})$ at time t is to maximize its immediate expected reward, i.e.,

$$R_i^{\text{myo}}\left(s_i^{(t)}\right) \triangleq \sum_{s_i^{(t+1)} \in \mathbf{S}_i} P_{\mathbf{a}_i}\left(s_i^{(t)}, s_i^{(t+1)}\right) R_i\left(s_i^{(t+1)}\right). \tag{10}$$

Hence, the peer i takes its action $\mathbf{a}_i^*$ (i.e., upload bandwidth allocation) such that the action maximizes the immediate expected reward $R_i^{\text{myo}}(s_i^{(t)})$, i.e.,

$$\mathbf{a}_i^* = \arg\max_{\mathbf{a}_i \in \mathbf{A}_i} \sum_{s_i^{(t+1)} \in \mathbf{S}_i} P_{\mathbf{a}_i}\left(s_i^{(t)}, s_i^{(t+1)}\right) R_i\left(s_i^{(t+1)}\right) \quad \text{subject to} \sum_{k \in C_i} a_{ik} \le L_i. \tag{11}$$

As shown in (10), the immediate expected reward does not consider the future rewards.

Unlike the myopic peers, the foresighted peers take their actions considering the immediate expected reward as well as the future rewards. Since future rewards are generally considered to be worth less than the rewards received now [23], the foresighted peers try to maximize a cumulative discounted expected reward. The cumulative discounted expected reward for a foresighted peer i in state $s_i^{(t)} = (s_{i1}, \ldots, s_{iN_{C_i}})$ at time $t = t_c$ given a discount factor γ_i can be expressed as

$$R_i^{\text{fore}}\left(s_i^{(t_c)}\right) \triangleq \sum_{t=t_c+1}^{\infty} \gamma_i^{(t-(t_c+1))} \cdot E\left[R_i\left(s_i^{(t)}\right)\right]. \tag{12}$$

More precisely, the cumulative discounted expected reward $R_i^{\text{fore}}(s_i^{(t_c)})$ in (12) can be rewritten as (13), shown at the bottom of the page. Hence, peer i can determine a set of actions that maximizes $R_i^{\text{fore}}(s_i^{(t_c)})$ for every state in $\mathbf{S}_i$, which leads to an optimal policy π_i^*. The optimal policy π_i^* thus maps each state $s_i \in \mathbf{S}_i$ into a corresponding optimal action $\mathbf{a}_i^*$, i.e., $\pi_i^*(s_i) = \mathbf{a}_i^*$ for all $s_i \in \mathbf{S}_i$.

By comparing (10) and (13), we can observe that the myopic decisions are a special case of the foresighted decisions when $\gamma_i = 0$. Note that the discount factor γ_i in the considered P2P network can alternatively represent the belief of the peer i about the validity of the expected future rewards, since the state transition probability can be affected by system dynamics such as other peers' joining, switching, or leaving groups. Hence, for example, if the P2P network is in a transient regime, a small discount factor is desirable. However, a large discount factor can be used if the P2P network is in stationary regime [24]. Thus, we assume that the value of the discount factor can be determined by the peers using information based on their past experiences, reputation of their associated peers [25], etc.

B. Peer Types Based on Their Attitudes

Peers in the considered P2P networks can also be characterized based on their attitudes towards the resource reciprocation, which are pessimistic, neutral, or optimistic [19]. Let (a_{ik}, s_{ik}) be a resource reciprocation between peer i and peer k, i.e., a pair of peer i's action a_{ik} to peer k and the corresponding peer k's action x_{ki} that is mapped into s_{ik}. A peer i is *neutral* if it presumes that the resource reciprocation changes linearly[5] depending on its actions a'_{ik}. A peer i is *pessimistic* if it presumes that the resource reciprocation decreases fast for $a'_{ik} \le a_{ik}$ and increases slow for $a'_{ik} \ge a_{ik}$. On the other hand, an *optimistic* peer i presumes that the resource reciprocation decreases slow for $a'_{ik} \le a_{ik}$ and increases fast for $a'_{ik} \ge a_{ik}$. Illustrative examples of resource reciprocation shapes that correspond to peers' different attitudes are shown in Fig. 4. In this paper, we consider these resource reciprocation profiles, which will be referred to as *reciprocation models*. Note that these reciprocation models can be extended by considering different degrees of pessimism or optimism, which will be presented in Section V-B.

These types of peers discussed above obviously affect their resource reciprocation strategies. In the following sections, we discuss how the peers' attitudes can impact the way in which peers model the other peers' resource reciprocation behavior, and investigate several resulting properties that can be drawn for the various peer types.

IV. Analysis of Peers' Interactions Based on Their Attitudes

In this section, we investigate several properties of the interactions among peers that can have different memory sizes for maintaining their resource reciprocation history or different number of state descriptions. Moreover, we also study how different types of peers such as myopic/foresighted and their resource reciprocation models can impact their resource reciprocation. We analyze several interactions among the peers under particular conditions that allow us to capture how the peer's characteristics can influence their interactions.

A. Impact of History on Resource Reciprocation

In this section, we first investigate the impact of the memory size for the history of resource reciprocation and the reciprocation models of peers. We assume that a peer i has its own units of memory $m_i(l)$, where $l \ge 1$ denotes the index of each unit of memory, and one unit of memory is required to store a resource reciprocation, i.e., $m_i(l) = (\mathbf{a}_i^{(t-l+1)}, s_i^{(t-l+1)})$. We consider the interactions between a myopic peer i (i.e., it focuses on maximizing its immediate expected reward) that can only recall and process $m_i(1)$ (i.e., it identifies state transition probabilities based on its last resource reciprocation), and its

[5]Several levels of neutral attitudes can be represented by using different slopes [e.g., α_{ik} in (20)].

$$\underbrace{\sum_{s_i^{(t_c+1)} \in \mathbf{S}_i} P_{\mathbf{a}_i^{(t_c+1)}}\left(s_i^{(t)}, s_i^{(t_c+1)}\right) R_i\left(s_i^{(t_c+1)}\right)}_{\text{immediate expected reward}} + \underbrace{\sum_{t'=t_c+1}^{\infty} \gamma_i^{(t'-t_c)} \cdot \sum_{s_i^{(t'+1)} \in \mathbf{S}_i} P_{\mathbf{a}_i^{(t'+1)}}\left(s_i^{(t')}, s_i^{(t'+1)}\right) R_i\left(s_i^{(t'+1)}\right)}_{\text{cumulative discounted expected reward}}. \tag{13}$$

associated self-interested peers aiming at maximizing their utilities, in order to quantify how these constraints can affect the peers' decisions and the resulting utilities.

1) Resource Reciprocation of a Myopic and Pessimistic Peer for Self-Interested Peers: We first consider the resource reciprocation strategy that a myopic and pessimistic peer i who can recall and process only $m_i(1)$ will adopt, while interacting with its associated peers. We assume that the associated peers can identify the myopic and pessimistic peer and its reciprocation characteristics.

Proposition 1: When a myopic and pessimistic peer i that can recall and process only $m_i(1)$ interacts with self-interested peers aiming at maximizing their utilities in C_i, peer i will receive for download its minimum required resource R_i^{req} from each of its associated peers in C_i.

Proof: Let $m_i(1) = (\mathbf{a}_i, s_i) = ((a_{i1}, \ldots, a_{iN_{C_i}}), (s_{i1}, \ldots, s_{iN_{C_i}}))$ be a recent resource reciprocation for peer i, where $s_{ik} = \psi_{ik}(x_{ki})$. As shown in Fig. 4, a pessimistic peer i presumes that

$$\begin{cases} a'_{ik} < a_{ik} \Rightarrow s'_{ik} = \Delta^P_{ik}(< s_{ik}) \\ a'_{ik} > a_{ik} \Rightarrow s'_{ik} = s_{ik} \end{cases} \tag{14}$$

where a'_{ik} denotes the actions that peer i can take in the next resource reciprocation. Given the conditions in (14), peer i allocates its available resources to maximize its reward. If peer k reduces its current allocated resources x_{ki} to $x'_{ki}(< x_{ki})$ in the next resource reciprocation interaction, which leads to decrease of peer i's reward, peer i can adjust its actions in response to the change of peer k's resource allocation. Peer i can compensate the reward reduction only if there exists a peer $k'(\neq k) \in C_i$ with the resource reciprocation $(a_{ik'}, s_{ik'})$ such that

$$a'_{ik'} > a_{ik'} \Rightarrow s'_{ik'} > s_{ik'} \tag{15}$$

where $s'_{ik'} = \psi_{ik'}(x_{k'i})$. However, peer i cannot find such a peer k' which satisfies (15), since this contradicts (14). Therefore, if peers in C_i identify peer i's reciprocation characteristics, they will select actions which provide the minimum required resources, i.e., $a_{ki} = R_i^{\text{req}}$ for $k \in C_i$. ∎

As a result of Proposition 1, the total received download rates that peer i can achieve in C_i is at most $N_{C_i} \cdot R_i^{\text{req}}$. A similar conclusion can be drawn from the interactions between a myopic and optimistic peer and its associated self-interested peers.

2) Resource Reciprocation of a Myopic and Optimistic Peer for Self-Interested Peers: Let us now consider the resource reciprocation strategy that a myopic and optimistic peer only with a recent resource reciprocation will take.

Proposition 2: When a myopic and optimistic peer i that can recall and process only $m_i(1)$ interacts with self-interested peers aiming at maximizing their utilities in C_i, peer i will receive for download its minimum required resource R_i^{req} from each of its associated peers in C_i.

Proof: Let $m_i(1) = (\mathbf{a}_i, s_i) = ((a_{i1}, \ldots, a_{iN_{C_i}}), (s_{i1}, \ldots, s_{iN_{C_i}}))$ be a recent resource reciprocation for peer i, where $\sum_{k \in C_i} a_{ik} \leq L_i$. The current reward is $\sum_{k \in C_i} r(s_{ik})$. As shown in Fig. 4, an optimistic peer i presumes that

$$\begin{cases} a'_{ik} < a_{ik} \Rightarrow s'_{ik} = s_{ik} \\ a'_{ik} > a_{ik} \Rightarrow s'_{ik} = \Delta^O_{ik}(> s_{ik}) \end{cases} \tag{16}$$

for peer $k, k \in C_i$. Based on $m_i(1) = (\mathbf{a}_i, s_i)$ and the condition in (16), peer i can take its next action $\mathbf{a}'^*_i$ such that it maximizes the immediate expected reward, i.e.,

$$\mathbf{a}'^*_i = \arg\max_{\mathbf{a}'_i \in \mathbf{A}_i} \sum_{k \in C_i} r(s'_{ik}) \quad \text{subject to} \sum_{k \in C_i} a'_{ik} \leq L_i \tag{17}$$

where $s'_i = (s'_{i1}, \ldots, s'_{iN_{C_i}})$ is the resulting state for action $\mathbf{a}'^*_i$. Based on the condition in (16), it can be easily shown that a solution to the optimization problem in (17) is given by

$$a'^*_{ik^*} = R^{\text{req}}_{k^*} \text{ and } a'^*_{ik} = a_{ik} + \eta_k \text{ for all } k \in C_i \backslash \{k^*\} \tag{18}$$

where η_k is a positive constant satisfying $\sum_{k \in C_i \backslash \{k^*\}} \eta_k = L_i - R^{\text{req}}_{k^*}$ and a peer $k^* \in C_i$ is selected by

$$k^* = \arg\min_{k \in C_i} \left\{ \Delta^O_{ik} - s_{ik} \right\}. \tag{19}$$

Equation (18) and (19) imply that peer i selects peer k^* that currently provides it with the most resources. Then, the peer i allocates the minimum required resource $R^{\text{req}}_{k^*}$ to peer k^*. Hence, the associated peers in C_i prefer not to be selected by peer i, which will lead to the associated peers selecting their actions $a_{ki} = R_i^{\text{req}}$ for all $k \in C_i$. ∎

From the above proposition, we can conclude that the total received download rates that peer i can achieve in C_i are at most $N_{C_i} \cdot R_i^{\text{req}}$.

Based on the above two propositions (i.e., Proposition 1 and Proposition 2), it can be observed that myopic and pessimistic/optimistic peers, which base their resource reciprocation only on the observed recent reciprocation, will receive only the minimum resource reciprocation from their associated self-interested peers, since there are no utility benefits for these peers to adopt other reciprocation policies. These results can be explained based on the peer's pessimistic or optimistic attitudes for the resource reciprocation. Since these attitudes provide the peer an overly simplified perspective on resource reciprocation (i.e., minimum/unchanged or unchanged/maximum reciprocation, respectively), the peer cannot effectively adopt its policies, which leads to inefficient response to the resource reciprocation of self-interested peers.

3) Resource Reciprocation of a Myopic and Neutral Peer for Self-Interested Peers: We now consider the resource reciprocation among a myopic and neutral peer and its associated self-interested peers. In this analysis, we show that the best strategy that a myopic and neutral peer i that can recall and process only $m_i(1)$ can adopt is the tit-for-tat (TFT) strategy. The TFT strategy is currently deployed in BitTorrent system as a peer selection strategy [6], [7]. Specifically, a peer with the TFT strategy in BitTorrent systems selects a fixed number of peers that provides the highest upload rates (i.e., the most cooperative), and equally divides and allocates its resources to the selected peers.

Proposition 3: When a myopic and neutral peer i that can recall and process only $m_i(1)$ interacts with self-interested peers aiming at maximizing their utilities in C_i, the strategy that the peer i adopts is the TFT strategy.

TABLE II
COMPARISON OF RESOURCE RECIPROCATION STRATEGIES

Peer Type	No. of Reciprocation Models	$m_i(l)$	Rewards
Foresighted	> 1	$l > 1$	(a)
Myopic	> 1	$l > 1$	(b)
	1 (neutral)	$l = 1$	$\geq N_{C_i} \cdot R_i^{req}$ (c)
	1 (pessimistic, optimistic)	$l = 1$	$N_{C_i} \cdot R_i^{req}$

Proof: Let $m_i(1) = (\mathbf{a}_i, s_i) = ((a_{i1}, \ldots, a_{iN_{C_i}}), (s_{i1}, \ldots, s_{iN_{C_i}}))$ be a recent resource reciprocation for peer i, where $\sum_{k \in C_i} a_{ik} \leq L_i$. Fig. 4 shows that a neutral peer i given $m_i(1) = (\mathbf{a}_i, s_i)$ presumes that

$$a'_{ik} \neq a_{ik} \Rightarrow s'_{ik} = \psi_{ik}(x_{ki} = \alpha_{ik} \cdot a'_{ik}) \tag{20}$$

where $\alpha_{ik} = s_{ik}/a_{ik}$ for $k \in C_i$. Therefore, to maximize peer i's rewards, it allocates the minimum required resources R_k^{req} to peer $k(\neq k^*) \in C_i$ (i.e., $a_{ik} = R_k^{\text{req}}$) and the residual available resources $L_i - \sum_{k \in C_i \setminus \{k^*\}} R_k^{\text{req}}$ to peer k^* (i.e., $a_{ik^*} = L_i - \sum_{k \in C_i \setminus \{k^*\}} R_k^{\text{req}}$), where the peer k^* is selected by

$$k^* = \arg\max_{k \in C_i} \{\alpha_{ik} = s_{ik}/a_{ik}\}. \tag{21}$$

The peer selection rule in (21) is the TFT strategy, as peer i selects the peer with the highest α_{ik}. ■

Hence, the TFT strategy deployed in BitTorrent system is a simple extension (i.e., it allows a peer to select multiple peers rather than one) of the strategy that a myopic and neutral peer can take.

The conclusions from the propositions presented in this section are summarized in Table II (Note that the comparison of (a)–(c) in Table II will be discussed in Section VI-C). These propositions show that a peer who myopically determines its actions using a single reciprocation model and a single resource reciprocation history (i.e., $m_i(1)$) cannot adopt an efficient reciprocation policy. Although the TFT strategy enables a neutral peer to achieve higher download rates than a pessimistic or optimistic peer, actions based on this strategy result in lower expected rewards than myopic or foresighted actions determined considering well-estimated associated peers' behaviors, as presented in Section VI-C. The difference between these approaches can be easily understood by considering that methods such as TFT are based on feedback information rather than predictive information based on peer's models. This shows the importance of accurately modeling the associated peers' behavior. Hence, a peer should identify its associated peers' behavior (i.e., the state transition probabilities) using multiple reciprocation models and the history of several resource reciprocation. How to determine the state transition probabilities based on multiple reciprocation models and resource reciprocation will be discussed in Section V.

In the subsequent section, we determine the impact that different numbers of deployed state descriptions have on the policy selected by the peers, and hence, its rewards.

B. Impact of Number of State Descriptions on Rewards

As discussed previously, a peer's received resources from the associated peers are characterized by a state, and each state is represented by a set of finite state descriptions. Since a provided rate x_{ki} from peer k is mapped into $s_{ik} = \psi_{ik}(x_{ki})$ by peer i using finite state descriptions, there exists a quantization error $|s_{ik} - x_{ki}|$. Hence, there is an error between the expected rewards computed based on the state descriptions and the actual rewards. We use the variance of the expected rewards to quantify how accurately the computed expected rewards represent the actual rewards. Moreover, we will show that the variance of the expected rewards will decrease as finer state descriptions (i.e., more state descriptions) are used, providing more accurate modeling for the resource reciprocation of the associated peers. Hence, this is consistent to minimizing the mean square error (MMSE) between the actual rewards and the expected rewards computed based on the state descriptions.

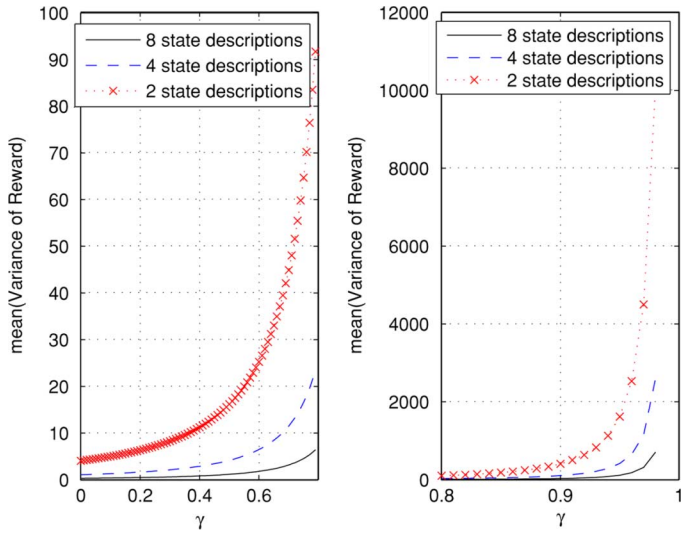

Fig. 5. Variances of the expected rewards given eight, four, and two state descriptions.

To compare the variances induced by different number of state descriptions, we assume that an optimal policy π_i^* given an $n_{ik} \times n_{ik}$ state transition probability matrix P is known to peer i. As will be shown, the expected rewards $\mathbf{J}^* = [J^*(s_{ik}^1), \ldots, J^*(s_{ik}^{n_{ik}})]^T$ of peer i from peer k based on the policy π_i^* can be obtained by (28). Since no prior information about the action of peer k is available, we assume that $r_i(s_{ik}^l)$ in (28) is a uniform random variable, i.e., the mean and the variance of $r_i(s_{ik}^l)$ for s_{ik}^l (i.e., $(l-1) \cdot \frac{L_k}{n_{ik}} \leq x_{ki} < l \cdot \frac{L_k}{n_{ik}}$) are given by $E(r(s_{ik}^l)) = ((2l-1)L_k)/(2n_{ik})$ and $V(r(s_{ik}^l)) = (1)/(12)(L_k/n_{ik})^2$.

The simulation results in Fig. 5 show several illustrative examples for the variances of the expected rewards given eight, four, and two state descriptions. These results clearly show that having a larger number of state descriptions can decrease the variance of the expected rewards regardless of the value of discount factors, thereby leading to a more accurate computation of the expected rewards.

V. DETERMINING THE STATE TRANSITION PROBABILITIES

A. State Transition Probability Computation Based on Empirical Frequency

A peer i can identify its state transition probabilities based on the frequency of the reciprocation. For this, we consider a table T_i^k that stores the history of resource reciprocation for peer k given actions of peer i. An element $T_i^k(s_{ik}^{l_1}, s_{ik}^{l_2}, a_{ik})$ of the table T_i^k denotes the number of state transitions from $s_{ik} = s_{ik}^{l_1}$ to $s'_{ik} = s_{ik}^{l_2}$, given an action a_{ik}. Hence, the state transition

probability $P_{a_{ik}}(s_{ik} = s_{ik}^{l_1}, s'_{ik} = s_{ik}^{l_2})$ based on the empirical frequency can be expressed as

$$P_{a_{ik}}(s_{ik} = s_{ik}^{l_1}, s'_{ik} = s_{ik}^{l_2}) = \frac{T_i^k\left(s_{ik}^{l_1}, s_{ik}^{l_2}, a_{ik}\right)}{\sum_{h=1}^{n_{ik}} T_i^k\left(s_{ik}^{l_1}, s_{ik}^{h}, a_{ik}\right)}. \quad (22)$$

Algorithm 1 shows the steps for determining the state transition probabilities based on the empirical frequency.

A disadvantage of this algorithm is that it may require a considerable amount of observations of the resource reciprocation over time to accurately identify the state transition probabilities. To reduce the number of required observations, we propose an alternative algorithm that can efficiently identify the state transition probabilities by modeling the peers' attitudes.

Algorithm 1: Determining State Transition Probability based on Empirical Frequency

Set: initial state $s_i = (s_{i1}, \ldots, s_{iN_{C_i}})$, initialize T_i^k with $T_i^k(s_{ik}^{l_1}, s_{ik}^{l_2}, a_{ik}) = 1$ for all $s_{ik}^{l_2}(1 \le l_2 \le n_{ik})$, for all $a_{ik} \in A_i$, and for all $k \in C_i$.

1: Observe resource reciprocation given an action $\mathbf{a}_i$; $(x_{1i}, \ldots, x_{N_{C_i}i})$
2: State Mapping; $(x_{1i}, \ldots, x_{N_{C_i}i}) \rightarrow s'_i = (s'_{i1}, \ldots, s'_{iN_{C_i}})$, where $s'_{il} = \psi_{il}(x_{li})$ for all $l \in C_i$
3: Update T_i^k for all $k \in C_i$; $T_i^k(s_{ik}^{l_1}, s_{ik}^{l_2}, a_{ik}) \leftarrow T_i^k(s_{ik}^{l_1}, s_{ik}^{l_2}, a_{ik}) + 1$ if $s_{ik} = s_{ik}^{l_1}$ and $s'_{ik} = s_{ik}^{l_2}$
4: Compute $P_{\mathbf{a}_i}(s_i, s'_i)$ using and (7) and (22)

B. State Transition Probabilities Based on Reciprocation Models

The resource reciprocation models of peers are discussed in Section III. A set of the state transition probability functions that correspond to the resource reciprocation models is called reciprocation matrix. The set of m available reciprocation matrices of peer i in s_{ik} for peer k is denoted by $\mathbf{M}_i^k(s_{ik}) = \{M_{i1}^k(s_{ik}), \ldots, M_{im}^k(s_{ik})\}$, where $M_{il}^k(s_{ik})$ is a matrix with its element $M_{il}^k(s_{ik})[a_{ik}, s'_{ik}] = P_{a_{ik}}(s_{ik}, s'_{ik})$ as shown in Fig. 4. Hence, a reciprocation matrix $M_{il}^k(s_{ik}) \in \mathbf{M}_i^k(s_{ik})$ for a pessimistic peer i taking action a_{ik} in s_{ik} (given its resource reciprocation $(\hat{a}_{ik}, s_{ik})$) shown in Fig. 4 can be expressed by (23), shown at the bottom of the previous page, where $W_P = |\{l | \Delta_{ik}^P \le s_{il} \le s_{ik}\}|$ is the number of state descriptions between s_{ik} and Δ_{ik}^P, and a_{ik} denotes the available actions that can be taken in the next state. Δ_{ik}^P represents the degree of pessimism for the resource reciprocation. Hence, $s_{ik}^1 \le \Delta_{ik}^P \le s_{ik}$.

Similarly, a reciprocation matrix of an optimistic peer i in s_{ik} for peer k shown in Fig. 4 can be represented by (24), shown at the bottom of the previous page, where $W_O = |\{l | s_{ik} \le s_{il} \le \Delta_{ik}^O\}|$ is the number of state descriptions between s_{ik} and Δ_{ik}^O. Δ_{ik}^O represents the degree of optimism for the resource reciprocation. Hence, $s_{ik} \le \Delta_{ik}^O \le s_{ik}^{n_{ik}}$.

The reciprocation matrix for a neutral peer can also be expressed as follows. A neutral peer i presumes that an action $a_{ik} \neq \hat{a}_{ik}$ will lead to linear changes in resource reciprocation from the current resource reciprocation $(\hat{a}_{ik}, s_{ik})$. Hence, the reciprocation matrix of a neutral peer i can be expressed as

$$M_{il}^k(s_{ik})[a_{ik}, s'_{ik}] = \begin{cases} 1, & \text{if } s'_{ik} = \psi_{ik}(x_{ki} = \alpha \cdot a_{ik}) \\ 0, & \text{otherwise} \end{cases} \quad (25)$$

where $\alpha = s_{ik}/\hat{a}_{ik}$ denotes the slope determined based on the current resource reciprocation. In the following subsection, we propose an algorithm that uses the discussed reciprocation matrices to efficiently identify the state transition probability functions.

C. Building State Transition Probability Functions Based on Reciprocation Matrices

We assume that a peer i has a predetermined initial action $\mathbf{a}_i^I = (a_{i1}^I, \ldots, a_{iN_{C_i}}^I)$ that is used for initializing the reciprocation matrices, i.e., a peer i has a predetermined action $a_{ik}^I \in A_i$ for peer k and the resulting s_{ik}. Based on the initial resource reciprocation between peer i and peer $k \in C_i$, (a_{ik}^I, s_{ik}), the reciprocation matrices of peer i can be initialized based on (23), (24), and (25). Note that peer i can have several reciprocation matrices, since it can select several levels of pessimism (or optimism) for resource reciprocation based on Δ_{ik}^P and Δ_{ik}^O. The next step is to determine and adjust the weights for each reciprocation matrices, such that the weighted sum of reciprocation matrices represents a set of state transition probability functions.

Let $\mathbf{M}_i^k(s_{ik}) = \{M_{i1}^k(s_{ik}), \ldots, M_{im}^k(s_{ik})\}$ be the set of reciprocation matrices that are initialized by peer i with the initial resource reciprocation (a_{ik}^I, s_{ik}). The weights of peer i for the reciprocation matrices are denoted by $w_i^k(s_{ik}) = (w_{i1}^k(s_{ik}), \ldots, w_{im}^k(s_{ik}))$ for peer $k \in C_i$. We also define $H_i^k(s_{ik}) = (h_{i1}^k(s_{ik}), \ldots, h_{im}^k(s_{ik}))$ as the set of number of hits, where the resource reciprocation between peer i and peer k are matched to non-zero elements in the reciprocation matrices. Specifically, if a resource reciprocation (a_{ik}, s'_{ik}) is matched up to a non-zero element in $M_{il}^k(s_{ik})[a_{ik}, s'_{ik}]$, then

$$M_{il}^k(s_{ik})[a_{ik}, s'_{ik}] = \begin{cases} 1, & \text{if } a_{ik} < \hat{a}_{ik}, s'_{ik} = \Delta_{ik}^P \text{ or } a_{ik} > \hat{a}_{ik}, s'_{ik} = s_{ik} \\ 1/W_P, & \text{if } a_{ik} = \hat{a}_{ik}, \Delta_{ik}^P \le s'_{ik} \le s_{ik} \\ 0, & \text{otherwise} \end{cases} \quad (23)$$

$$M_{il}^k(s_{ik})[a_{ik}, s'_{ik}] = \begin{cases} 1, & \text{if } a_{ik} < \hat{a}_{ik}, s'_{ik} = s_{ik} \text{ or } a_{ik} > \hat{a}_{ik}, s'_{ik} = \Delta_{ik}^O \\ 1/W_O, & \text{if } a_{ik} = \hat{a}_{ik}, s_{ik} \le s'_{ik} \le \Delta_{ik}^O \\ 0, & \text{otherwise} \end{cases} \quad (24)$$

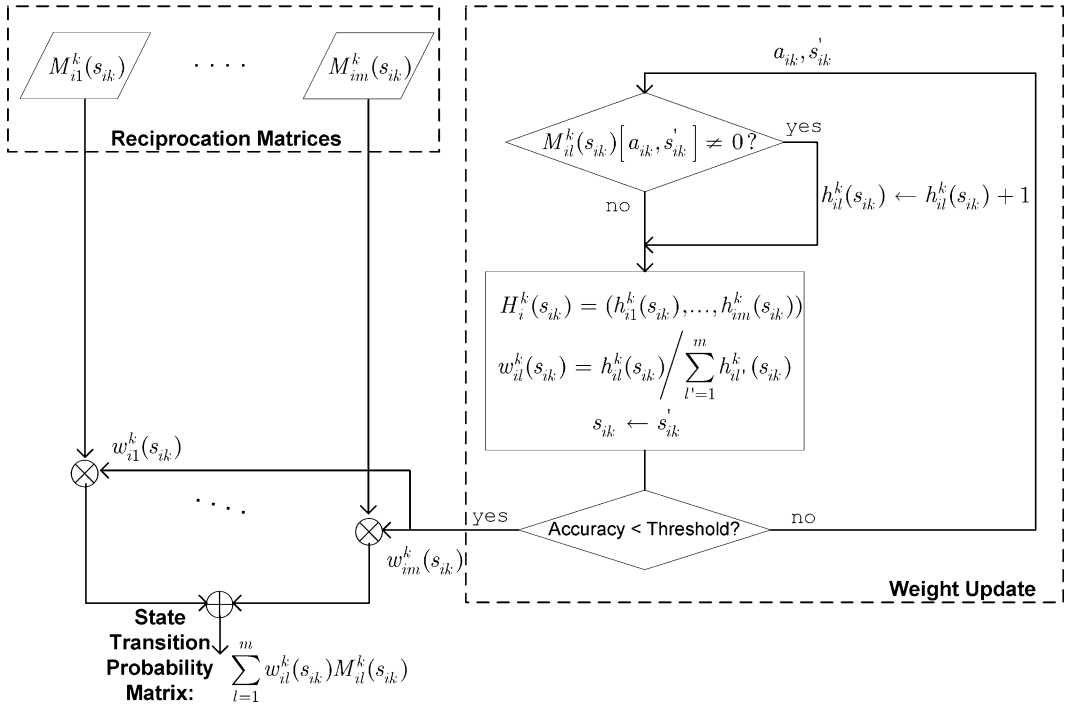

Fig. 6. Block diagram for updating the weights of the reciprocation matrices.

the number of hits increases, i.e., $h^k_{il}(s_{ik}) \leftarrow h^k_{il}(s_{ik}) + 1$. $H^k_i(s_{ik})$ is used to update the weights of reciprocation matrices for peer k, i.e.,

$$w^k_{il}(s_{ik}) = h^k_{il}(s_{ik}) / \sum_{l'=1}^{m} h^k_{il'}(s_{ik}). \tag{26}$$

$H^k_i(s_{ik})$ can be initialized by $H^k_i(s_{ik}) = (1,\ldots,1)$. Hence, the initial weights are $w^k_i(s_{ik}) = (1/m,\ldots,1/m)$, and thus, the initial state transition probability matrix for peer k is given by $\sum_{l=1}^{m} w^k_{il}(s_{ik}) M^k_{il}(s_{ik}) = (1)/(m)\sum_{l=1}^{m} M^k_{il}(s_{ik})$. This update process is depicted in Fig. 6 and the detailed algorithm is presented in Algorithm 2.

Algorithm 2: Algorithm for Updating Peer i's Weights of Reciprocation Matrices for Peer k

Set: initial resource reciprocation (a^I_{ik}, s_{ik}), and Δ^P_{ik}, Δ^Q_{ik}, a desired maximum reward deviation δ_i.

1: Initialize $\mathbf{M}^k_i(s_{ik}) = \{M^k_{i1}(s_{ik}),\ldots,M^k_{im}(s_{ik})\}$ with (a^I_{ik}, s_{ik}) based on (23), (24), and (25)
2: Initialize $H^k_i(s_{ik}) = (1,\ldots,1)$, $w^k_i(s_{ik}) = (w^k_{i1}(s_{ik}),\ldots,w^k_{im}(s_{ik})) = (1/m,\ldots,1/m)$
3: **repeat**
4: Take action $a_{ik} \in A_i$, observe reciprocation x_{ki}, and determine $s'_{ik} = \psi_{ik}(x_{ki})$
5: **if** $\mathbf{M}^k_i(s'_{ik})$ does not exist
6: Initialize $\mathbf{M}^k_i(s'_{ik}) = \{M^k_{i1}(s'_{ik}),\ldots,M^k_{im}(s'_{ik})\}$ with (a_{ik}, s'_{ik}) based on (23), (24), and (25)
7: Initialize $H^k_i(s'_{ik}) = (1,\ldots,1)$, $w^k_i(s'_{ik}) = (w^k_{i1}(s'_{ik}),\ldots,w^k_{im}(s'_{ik})) = (1/m,\ldots,1/m)$
8: **else**
9: **for all** l such that $1 \le l \le m$ **do**//`Update` $H^k_i(s_{ik})$
10: **if** $M^k_{il}(s_{ik})[a_{ik}, s'_{ik}] \neq 0$ **then**
11: $h^k_{il}(s_{ik}) \leftarrow h^k_{il}(s_{ik}) + 1$
12: Update w^k_i using (26)
13: $s_{ik} \leftarrow s'_{ik}$
14: **until** $D_{J_l} \le \delta_i$ //D_{J_l} `is a distance metric. See Proposition 4.`

Finally, based on the identified weights, $P_{a_{ik}}(s_{ik}, s'_{ik})$ can be obtained by

$$P_{a_{ik}}(s_{ik}, s'_{ik}) = \left\{\sum_{l=1}^{m} w^k_{il}(s_{ik}) M^k_{il}(s_{ik})\right\}[a_{ik}, s'_{ik}] \tag{27}$$

In the next section, we investigate how the error in estimating the state transition probabilities can affect the peers' decisions and the resulting utilities.

D. Impact of Estimation Error in State Transition Probability Matrices on Rewards

We study the impact of the state transition probability estimation error on the cumulative discounted expected rewards. As an illustration, we consider a case where peer i and peer k are in a group and reciprocate their resources.

Suppose that P is an $n_{ik} \times n_{ik}$ state transition probability matrix of peer i for peer k given peer i's optimal actions a^*_{ik}, which are determined by the optimal policy π^*_i, i.e., $a^*_{ik} = \pi^*_i(s_i = s_{ik})$ for $s_i \in \mathbf{S}_i$. Each element $[P]_{l_1 l_2} = P[l_1, l_2] = P_{a^*_{ik}}(s^{l_1}_{ik}, s^{l_2}_{ik})$ denotes the transition probability from $s^{l_1}_{ik}$ to $s^{l_2}_{ik}$ given the optimal action a^*_{ik}, and its lth row vector is denoted by $\mathbf{p}_l$. We assume that P is an irreducible matrix since different actions of a peer can induce different actions from its associated peers [26]. Therefore, there exists a steady state distribution $\boldsymbol{\nu}$, i.e., $\lim_{h\to\infty}[P]^h = \mathbf{1}\boldsymbol{\nu}$, where $\mathbf{1} = (1,\ldots,1)^T$. We use the L-norm ($L \ge 1$) to represent the distance between two vectors $\mathbf{p}_l = (p_{l1},\ldots,p_{ln})$ and $\mathbf{q}_l = (q_{l1},\ldots,q_{ln})$, denoted by $\|\mathbf{p}_l - \mathbf{q}_l\|_L$. The L-norm of a vector $\mathbf{x} \in \mathbb{R}^n$ is defined by $\|\mathbf{x}\|_L = (\sum_{h=1}^{n} |x_h|^L)^{1/L}$.

Proposition 4: For the true and estimated state transition probability matrices P and P', let $\mathbf{J}^*$ and $\mathbf{J}'^*$ be the cumulative discounted expected rewards of a peer i from a peer k based on an optimal policy π_i^* and $\pi_i'^*$, respectively. Then, given a discount factor γ, the discrepancy between $\mathbf{J}^*$ and $\mathbf{J}'^*$ from initial state s_{ik}^l is bounded by $D_{J_l} \triangleq \|(\mathbf{p}_l - \mathbf{p}_l') + (\gamma)/(1-\gamma)(\boldsymbol{\nu} - \boldsymbol{\nu}')\|_L \|\mathbf{r}\|_{L'}$.

Proof: The discounted expected rewards $\mathbf{J}^* = [J^*(s_{ik}^1), \ldots, J^*(s_{ik}^n)]^T$ of a peer i from a peer k based on an optimal policy π_i^* for P can be computed by

$$\begin{bmatrix} J^*\left(s_{ik}^1\right) \\ \vdots \\ J^*\left(s_{ik}^{n_{ik}}\right) \end{bmatrix} = P \begin{bmatrix} r\left(s_{ik}^1\right) \\ \vdots \\ r\left(s_{ik}^{n_{ik}}\right) \end{bmatrix} + \gamma P \begin{bmatrix} J^*\left(s_{ik}^1\right) \\ \vdots \\ J^*\left(s_{ik}^{n_{ki}}\right) \end{bmatrix}. \quad (28)$$

A compact expression for (28) is given by

$$\mathbf{J}^* = P\mathbf{r} + \gamma P \mathbf{J}^*, \quad (29)$$

and the solution to (29) is expressed as

$$\mathbf{J}^* = [I - \gamma P]^{-1} P\mathbf{r} = [P + \gamma P^2 + \gamma^2 P^3 + \cdots]\mathbf{r}. \quad (30)$$

Without loss of generality, we consider a cumulative discounted expected reward from s_{ik}^l, i.e., $\mathbf{J}^*(s_{ik}^l)$. Using the expression in (30), $\mathbf{J}^*(s_{ik}^l)$ can be expressed as

$$\mathbf{J}^*(s_{ik}^l) = ([P]_{l1}, \ldots, [P]_{ln_{ik}})\mathbf{r} + \gamma([P]_{l1}^2, \ldots, [P]_{ln_{ik}}^2)\mathbf{r} + \gamma^2([P]_{l1}^3, \ldots, [P]_{ln_{ik}}^3)\mathbf{r} + \cdots. \quad (31)$$

Since $\lim_{h\to\infty}[P]^h = \mathbf{1}\boldsymbol{\nu}$, (31) can be approximated by

$$\mathbf{J}^*(s_{ik}^l) \approx \mathbf{p}_l\mathbf{r} + \gamma\boldsymbol{\nu}\mathbf{r} + \gamma^2\boldsymbol{\nu}\mathbf{r} + \cdots = \mathbf{p}_l\mathbf{r} + \frac{\gamma}{1-\gamma}\boldsymbol{\nu}\mathbf{r}. \quad (32)$$

Similarly, $\mathbf{J}'^*(s_{ik}^l)$ can be computed based on an optimal policy $\pi_i'^*$ for P' as

$$\mathbf{J}'^*(s_{ik}^l) \approx \mathbf{p}_l'\mathbf{r} + \gamma\boldsymbol{\nu}'\mathbf{r} + \gamma^2\boldsymbol{\nu}'\mathbf{r} + \cdots = \mathbf{p}_l'\mathbf{r} + \frac{\gamma}{1-\gamma}\boldsymbol{\nu}'\mathbf{r}. \quad (33)$$

Using the approximations as in (32) and (33), the discrepancy between the two cumulative discounted expected rewards are expressed as

$$\begin{aligned} &|\mathbf{J}^*\left(s_{ik}^l\right) - \mathbf{J}'^*\left(s_{ik}^l\right)| \\ &\approx \left|\left[(\mathbf{p}_l - \mathbf{p}_l') + \frac{\gamma}{1-\gamma}(\boldsymbol{\nu} - \boldsymbol{\nu}')\right]\mathbf{r}\right| \\ &\le \left\|(\mathbf{p}_l - \mathbf{p}_l') + \frac{\gamma}{1-\gamma}(\boldsymbol{\nu} - \boldsymbol{\nu}')\right\|_L \|\mathbf{r}\|_{L'} \end{aligned}$$

where $1/L + 1/L' = 1$. Since $\|\mathbf{r}\|_{L'}$ is a non-negative constant, the discrepancy between the two cumulative discounted expected rewards from s_{ik}^l is bounded by $D_{J_l} = \|(\mathbf{p}_l - \mathbf{p}'_l) + (\gamma)/(1-\gamma)(\boldsymbol{\nu} - \boldsymbol{\nu}')\|_L \|\mathbf{r}\|_{L'}$. ■

Note that since $\|(\mathbf{p}_l - \mathbf{p}'_l) + (\gamma)/(1-\gamma)(\boldsymbol{\nu} - \boldsymbol{\nu}')\|_L \le \|(\mathbf{p}_l - \mathbf{p}_l')\|_L + (\gamma)/(1-\gamma)\|(\boldsymbol{\nu} - \boldsymbol{\nu}')\|_L$, the error of cumulative discounted expected rewards is bounded by both the distances of state transition probabilities and the stationary distributions. Hence, we can conclude that the discrepancy between the cumulative discounted expected rewards can be affected by the estimation error in state transition probability matrix and the stationary distribution, as well as each peer's discount factor. Since $\lim_{\gamma\to 1}(\gamma)/(1-\gamma) = \infty$, D_{J_l} is dominated by the stationary distribution error. As $\gamma \to 0$, however, it is dominated by the error in state transition probabilities. Hence, a peer can adjust the impact of the estimated state transition probability error on the cumulative discounted expected reward by changing γ. We remark that the approximation of $\mathbf{J}^*(s_{ik}^l)$ given by (32) is very accurate as shown in Fig. 7. Fig. 7 depicts the average discrepancy between $\mathbf{J}^*(s_{ik}^l)$ and its approximation for different number of states, i.e., $|(\mathbf{p}_l\mathbf{r} + \gamma\boldsymbol{\nu}\mathbf{r} + \gamma^2\boldsymbol{\nu}\mathbf{r} + \cdots) - (\mathbf{p}_l\mathbf{r} + (\gamma)/(1-\gamma)\boldsymbol{\nu}\mathbf{r})|$. These illustrative examples show that the discrepancy increases as γ approaches 0.5, while the discrepancy decreases as γ approaches 0 or 1. These observations are reasonable since the approximation becomes more accurate as either $\mathbf{p}_l\mathbf{r}$ or $\boldsymbol{\nu}\mathbf{r}$ dominates.

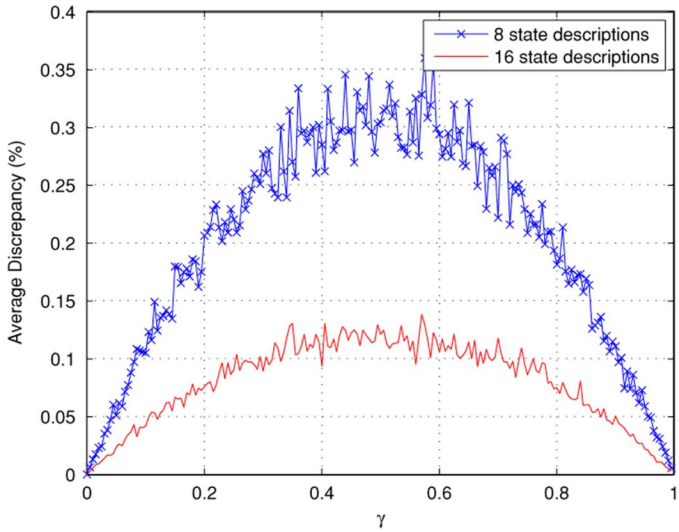

Fig. 7. Average discrepancy between true and approximated cumulative discounted expected rewards: 2-norm ($L = 2$) is used.

VI. Simulation Results

A. Comparison of Various Approaches for Identifying the State Transition Probabilities

We discuss two algorithms, one is based on the empirical frequency and the other one is based on the reciprocation models, to identify the state transition probability matrices in Section V. As discussed in Proposition 4, D_{J_l} is proposed to quantify the maximum discrepancy between the discounted expected rewards from a state, which is induced by the true and estimated state transition probability error. To illustrate these tradeoffs between the efficiency and the accuracy of the two proposed approaches, we deploy them to identify the true state transition probability. For the reciprocation model based approach, three reciprocation models that represent the pessimistic, optimistic, and neutral behaviors are used. In D_{J_l}, we assume that $\|\mathbf{r}\|_{L'}$ is normalized to 1, and the 2-norm (i.e., $L = L' = 2$) is used. The results are shown in Fig. 8.

Since the observations are generated based on a stationary state transition probability matrix, if there are enough observations of resource reciprocation and the empirical frequency is deployed to identify the state transition probability functions,

the state transition probability functions can be well-identified, i.e., the more observations, the higher accuracy. However, it may require many peers' interactions to obtain accurate state transition probability functions. In contrast, the reciprocation model based approach can efficiently identify the state transition probability functions with fewer observations than the empirical frequency based approach. However, unlike the empirical frequency based approach, the improvement gained for more observations diminishes rapidly (before reaching the best performance of the empirical frequency based approach), as the estimation relies on predetermined reciprocation models. Therefore, it is important to decide an appropriate approach considering these tradeoffs. The detailed weight update processes for the reciprocation model based approach are shown in Fig. 9.

In order to show the weight update process, predetermined true state transition probability functions for the resource reciprocation models (i.e., neutral, pessimistic, optimistic, and general) are used in these simulations. As shown in Fig. 9, the proposed algorithm based on the reciprocation models effectively computes the weights with fewer observations, thereby providing a faster convergence. This convergence property becomes important when the state transition probabilities vary over time. Illustrative simulation results are shown in Fig. 10.

Fig. 10 shows the D_{J_l} obtained by two approaches. To study the effectiveness of the proposed algorithm in a dynamic environment, different state transition probability matrices of a peer are deployed every 10- or 20-resource reciprocation. As discussed, the proposed reciprocation model based approach provides a faster convergence, thereby enabling peers to efficiently capture the changes of the state transition probability. Therefore, the proposed approach can cope with a dynamic environment, thereby making it more suitable than the empirical frequency based approach for a peer.

B. Quantifying the Impact of the Number of Reciprocation Models

As discussed in Section V, various reciprocation models can be deployed, which enable peers to identify the state transition probabilities more accurately. To study the impact of the number of reciprocation models on the accuracy of the state transition probabilities, we assume that a true state transition probability of a peer is stationary and randomly generated. The achieved D_{J_l} when different number of reciprocation models are used are shown in Fig. 11(a).

In these simulation results, the number of reciprocation models is increased by symmetrically extending the pessimistic and optimistic reciprocation models. For the case where two reciprocation models are used, only the neutral and pessimistic resource models are used as an illustration. As expected, in general, the more resource reciprocation models are deployed, the higher accuracy for identifying the state transition probability matrices can be achieved. However, we can observe that the D_{J_l} improvement decreases as the number of deployed reciprocation models increases. Since the D_{J_l} decreases as more reciprocation models are deployed, the reward discrepancy due to the estimation error is reduced as shown in the Proposition 4.

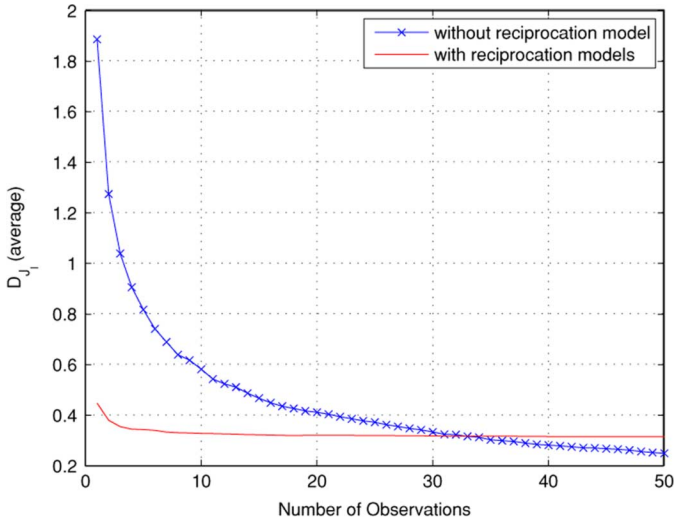

Fig. 8. (Averaged) D_{J_l} for estimated state transition probability matrices with/without reciprocation models.

Moreover, if the information about the relationship between the number of deployed reciprocation models and the resulting D_{J_l} is available for peers, they can select the number of reciprocation models and can expect the number of required observations for the resource reciprocation. As discussed in Proposition 4, the reward deviation is bounded by the D_{J_l}, which can also be bounded by a peer's desired maximum reward deviation δ, i.e.,

$$\left\| (\mathbf{p}_l - \mathbf{p}'_l) + \frac{\gamma}{1-\gamma}(\boldsymbol{\nu} - \boldsymbol{\nu}') \right\|_L \|\mathbf{r}\|_{L'} \leq \delta \cdot \|\mathbf{r}\|_{L'}. \quad (34)$$

For instance, in Fig. 11(a), if a desired maximum reward deviation is 0.15, i.e., $\delta = 0.15$, a peer can select 3, 5, 7, or 9 number of reciprocation models, expecting 14, 7, 4, or 2 observations of resource reciprocation, respectively, to achieve δ, as shown in Table III. Hence, peers can select the appropriate number of reciprocation models, by explicitly considering their tolerable durations for resource reciprocation and their desired maximum reward deviation.

If a priori information about the associated peers' resource reciprocation behavior is available (e.g., using reputation [25]), deploying the minimum number of reciprocation models that closely approximate their behaviors will result in the best performance to identify the state transition probabilities. As an illustration, we consider four cases where a different number of reciprocation models are used: i) two models for the pessimistic and optimistic reciprocation; and these models are extended to ii) four, iii) six, and iv) eight reciprocation models by considering different degrees of pessimism/optimism. We assume that the true state transition probability of an associated peer is well matched by a set of reciprocation models included in the cases of ii), i.e., a linear combination of the deployed reciprocation models in ii) can lead to the true state transition probability. Note that the cases of iii) and iv) can also include a set of reciprocation models that approximates the associated peer's resource reciprocation as they are extended from the case of ii). The simulation results are shown in Fig. 11(b).

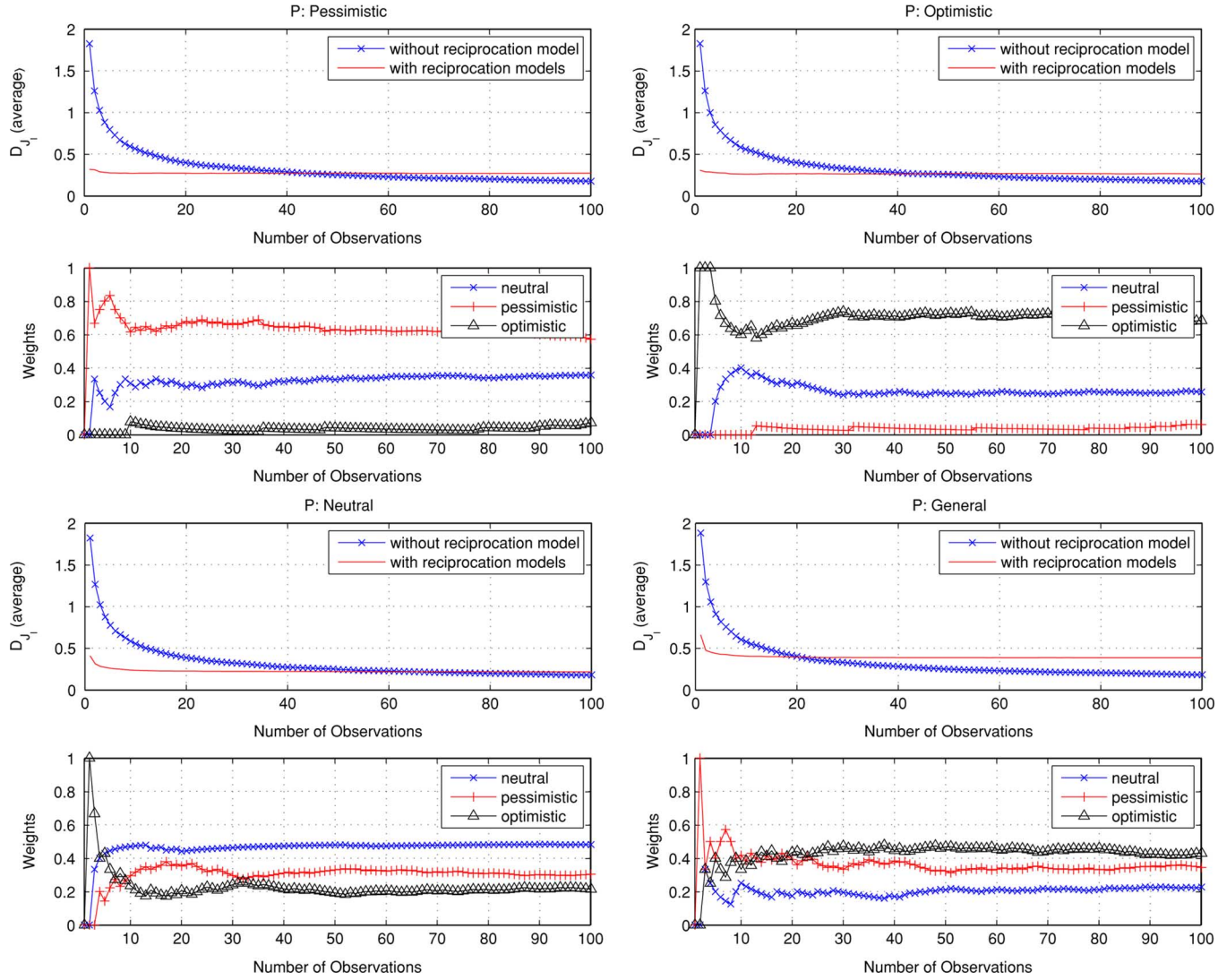

Fig. 9. Different types of true state transition probability functions and the weights for resource reciprocation models.

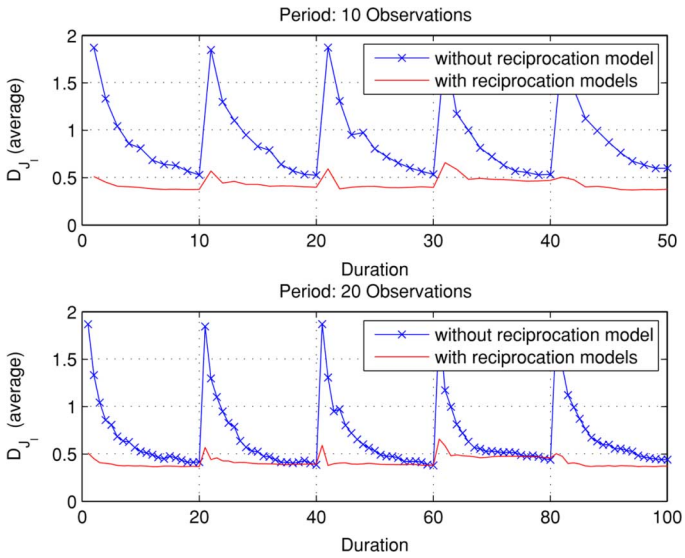

Fig. 10. D_{J_l} for time varying true state the transition probability.

Since the true state transition probability function is well identified by case ii), the lowest D_{J_l} can be achieved with four reciprocation models. However, D_{J_l} becomes the largest for the case of i), as the reciprocation models cannot efficiently estimate the true state transition probability function. Interestingly, we can observe that D_{J_l} with more reciprocation models such as cases iii) and iv) are larger than case ii), although case iii) and iv) also include the reciprocation models that are included in case ii). This is because the extended reciprocation models from case ii) become redundant, and do not improve the accuracy. Rather, it prevents the peer from identifying the state transition probability functions accurately. Hence, we can conclude that if *a priori* information about the associated peers' resource reciprocation behaviors is available, the minimum number of reciprocation models that can capture the peers resource reciprocation behaviors provides the best result for identifying the state transition probability functions.

C. Impact of Myopic and Foresighted Policies on Utilities

In this section, we quantitatively compare the impact of the myopic and foresighted policies on the achieved cumulative utilities. In these simulations, the reciprocation model based approach is used to identify the state transition probability

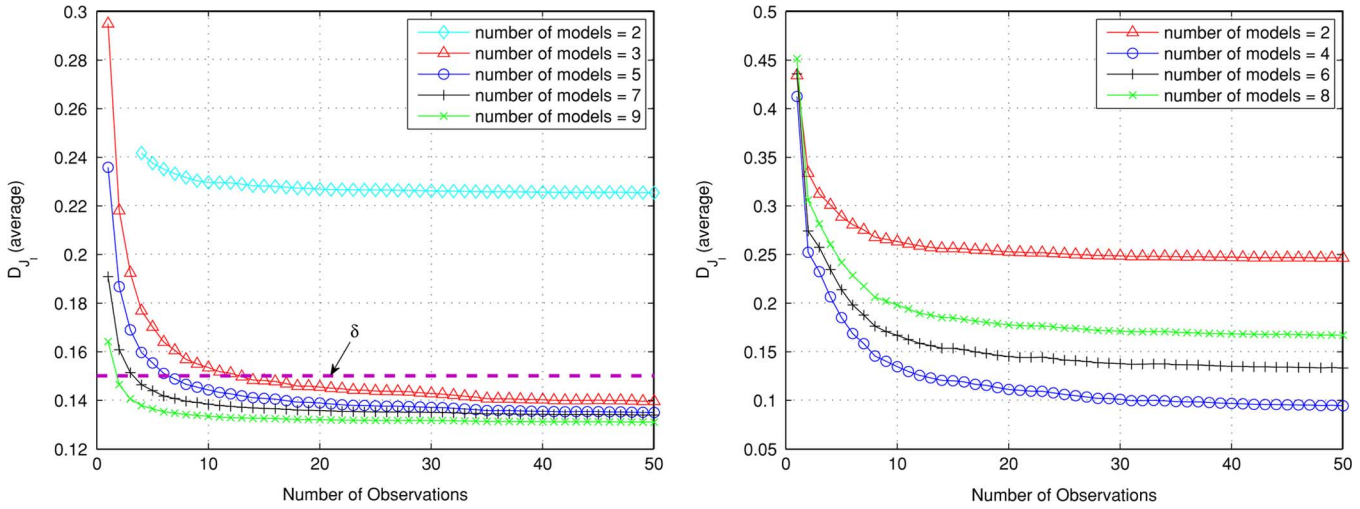

Fig. 11. D_{J_l} for different number of reciprocation models: (a) randomly generated true state transition probability function. $\delta = 0.15$. (b) Predetermined true state transition probability function.

TABLE III
NUMBER OF RECIPROCATION MODELS AND REQUIRED RESOURCE RECIPROCATION FOR $\delta = 0.15$

No. of Reciprocation Models	No. of Required Resource Reciprocation
2	N/A
3	14 (100%)
5	7 (50%)
7	4 (28.6%)
9	2 (14.3%)

functions. The myopic policies are made based on (11) and the foresighted policies are made based on maximizing R_i^{fore} in (12) with the discounted factor $\gamma = 0.8$ as an illustration. The solution to the MDP is implemented based on a well-known policy iteration method [18], which performs policy improvement and policy evaluation iteratively. In addition, as an illustration, we compare the TFT strategy implemented in BitTorrent-like system supporting two leechers simultaneously. The simulation results are shown in Fig. 12.

Fig. 12 shows the immediate expected rewards and cumulative discounted expected rewards of a peer i obtained based on the actions determined by the myopic (including TFT) and foresighted policies. A state $s_i = (s_{i1}, s_{i2}, s_{i3})$ of peer i is represented by four state descriptions of three associated peers, i.e., $s_{ik} \in \{s_{ik}^1, s_{ik}^2, s_{ik}^3, s_{ik}^4\}$, where $E(r(s_{ik}^m)) < E(r(s_{ik}^n))$ if $m < n$, for $k = 1, 2, 3$. The state indexes indicate the initial state of peer i, where it determines its optimal policy (and the corresponding actions). The size of the state space is 64 in this case. Each state is enumerated, and then, represented by state index from 1 to 64. The y-axis represents the normalized (immediate or cumulative discounted expected) rewards. We assume that each state description is represented by a number, i.e., $s_{ik}^l = l$ for $l = 1, \ldots, 4$, and the expected reward in each state is proportional to the numbers that corresponds to the states, i.e., $E(r(s_{ik}^l)) = W_k \cdot l$ for a constant W_k for peer k. Hence, the expected rewards in a state $s_i = (s_{i1}^{l_1}, s_{i2}^{l_2}, s_{i3}^{l_3})$ is represented by $E(r(s_i)) = \sum_{k=1}^{3} E(r(s_{ik}^{l_k})) = \sum_{k=1}^{3} l_k \cdot W_k$. This can be easily extended to represent actual rewards by assigning actual received resources to each state.

We can observe that the obtained utilities based on the TFT policy are worse than those based on the myopic or foresighted policies, since the actions determined by the TFT policy do not consider the expected utilities. Moreover, the constraints of fixed concurrent allowable uploads to the leechers can prevent the decision process from selecting better actions. The proposed approaches can enhance the resource reciprocation decisions of TFT strategy, which is currently implemented in current BitTorrent systems. By deploying the proposed approaches, peers can efficiently model their associated peers' behavior (in e.g., every rechoke period [6], [7]), and thus, the peers can allocate their resources to their associated peers based on their *levels of cooperation*. Hence, peers in the current BitTorrent systems can have multiple actions, rather than two simple actions (i.e., allowing or rejecting downloads), thereby efficiently adjusting their resource reciprocation and improving their performance.

As discussed previously, the myopic decisions are made based on (11), which maximize the immediate expected rewards. Hence, we can verify that the immediate expected rewards obtained by the actions of myopic policy are always higher (or equal) than the other policies in Fig. 12(a). However, as shown in Fig. 12(b), the foresighted decisions are made based on (12), which maximize the cumulative expected discounted rewards (i.e., R_i^{fore}). Therefore, the foresighted policy enables the peers to determine their decisions that lead to the highest cumulative discounted expected rewards.

A higher cumulative discounted expected reward can lead to a shorter downloading time or a higher multimedia quality. Fig. 13 shows illustrative examples of downloading time and achieved multimedia quality based on the proposed foresighted policy and the TFT strategy. In Fig. 13(a) and (b), we assume that a peer is downloading a general file with size of 5 Mbytes, and *Foreman* sequence (CIF, 30 frames/s) from its associated peers, respectively. We assume that the associated peers have 250Kbps maximum available upload bandwidth and have enough chunks to transmit. Moreover, the associated peers use five state descriptions. For Fig. 13(b), we assume that a peer downloads the packets that have higher quality impact

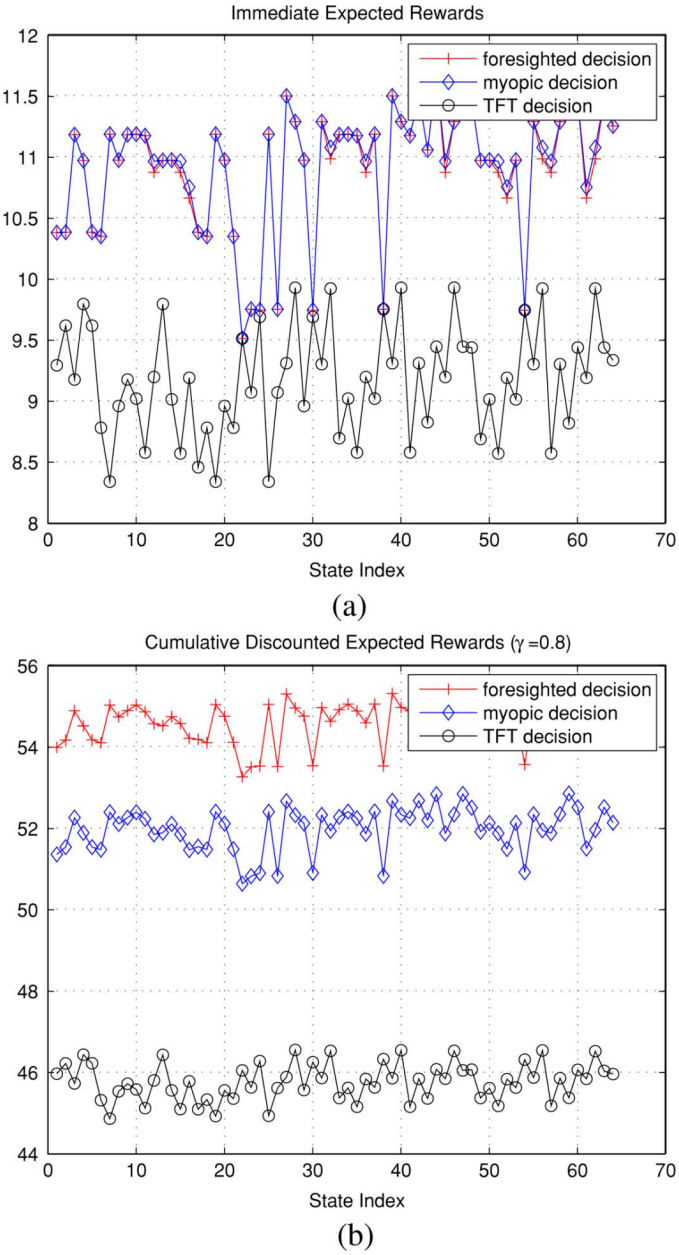

Fig. 12. Immediate and cumulative discounted expected rewards achieved by different policies. (a) Immediate expected rewards. (b) Cumulative discounted expected rewards.

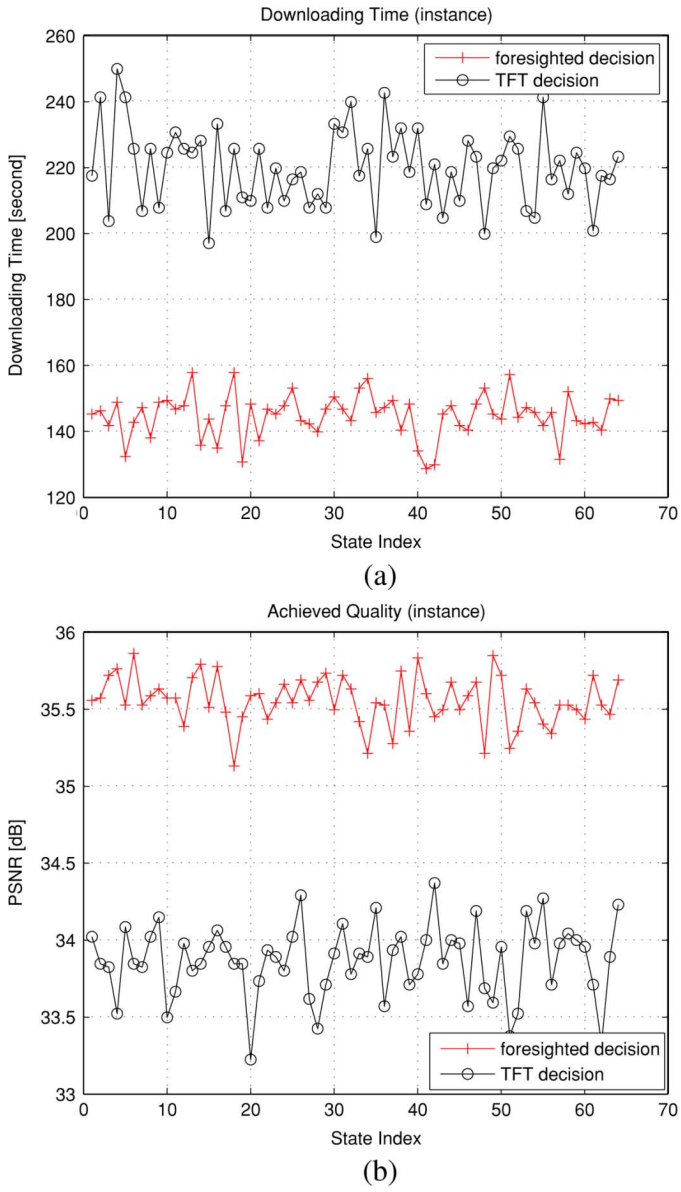

Fig. 13. An illustrative examples for immediate and cumulative discounted expected rewards achieved by different policies. (a) Downloading Time. (b) Achieved Multimedia Quality.

first. Simulation results in Fig. 13 clearly show that a higher cumulative discounted expected reward leads to a shorter downloading time for file sharing applications or a higher quality for multimedia content sharing applications.

VII. Conclusion

In this paper, we consider the resource reciprocation among heterogeneous and self-interested peers that negotiate the amount of resources that they will provide to each other. The resource reciprocation among the peers is modeled as a reciprocation game, and the game is played by foresighted peers using an MDP framework. To successfully implement the MDP framework in the dynamic P2P environment, we propose to model the resource reciprocation of peers. We study the tradeoffs between efficiency and accuracy when different numbers of reciprocation models are deployed. We show that if *a priori* information about the associated peers' behaviors is available, deploying the minimum number of reciprocation models that closely approximate their behaviors results in the minimum reward deviation. We also analytically show the impact of the estimation error between the true and modeled state transition probability function on each peer's reciprocation policy and its resulting rewards. Moreover, we analytically study how bounded rationality can impact the interactions among the peers and the resulting resource reciprocation. In the simulations, we show that the proposed reciprocation based approach is suitable for a dynamic environment. Finally, we show that the proposed foresighted decisions lead to the best performance in terms of the cumulative expected utilities as opposed to the

currently deployed strategy (i.e., TFT) in BitTorrent system or the myopic decisions.

REFERENCES

[1] *Napster*, [Online]. Available: http://www.napster.com, [Online]. Available

[2] *Gnutella*, [Online]. Available: http://www.gnutella.com, [Online]. Available

[3] *KaZaA*, [Online]. Available: http://www.kazaa.com, [Online]. Available

[4] S. Androutsellis-Theotokis and D. Spinellis, "A survey of peer-to-peer content distribution technologies," *ACM Comp. Surv.*, vol. 36, no. 4, pp. 335–371, Dec. 2004.

[5] J. Liu, S. G. Rao, B. Li, and H. Zhang, "Opportunities and challenges of peer-to-peer internet video broadcast," in *Proc. IEEE Special Issue on Recent Advances in Distributed Multimedia Communications*, 2007.

[6] B. Cohen, "Incentives build robustnessin bittorrent," in *Proc. P2P Economics Workshop*, Berkerly, CA, 2003.

[7] A. Legout, N. Liogkas, E. Kohler, and L. Zhang, "Clustering and sharing incentives in bittorrent systems," *SIGMETRICS Perform. Eval. Rev.*, vol. 35, no. 1, pp. 301–312, 2007.

[8] X. Zhang, J. Liu, B. Li, and T. S. P. Yum, "CoolStreaming/DONet: A data-driven overlay network for efficient live media streaming," in *Proc. INFOCOM'05*, 2005.

[9] V. Pai, K. Kumar, K. Tamilmani, V. Sambamurthy, and A. E. Mohr, "Chainsaw: Eliminating trees from overlay multicast," in *Proc. 4th Int. Workshop on Peer-to-Peer Systems (IPTPS)*, Feb. 2005.

[10] Z. Xiang, Q. Zhang, W. Zhu, Z. Zhang, and Y.-Q. Zhang, "Peer-to-peer based multimedia distribution service," *IEEE Trans. Multimedia*, vol. 6, no. 2, pp. 343–355, Apr. 2004.

[11] X. Jiang, Y. Dong, D. Xu, and B. Bhargava, "GnuStream: A P2P media streaming system prototype," in *Proc. of 4th Int. Conf. Multimedia and Expo.*, Jul. 2003.

[12] Y. Cui, B. Li, and K. Nahrstedt, "oStream: Asynchronous streaming multicast in application-layer overlay networks," *IEEE J. Sel. Areas Commun.*, vol. 22, no. 1, pp. 91–106, Jan. 2004.

[13] B. Yu and M. Singh, "Incentive mechanisms for agent-based peer-to-peer systems," in *Proc. 2nd Int. Joint Conf. Autonomous Agents and Multiagent Systems*, 2003.

[14] J. Shneidman and D. C. Parkes, "Rationality and self-interest in peer to peer networks," *Lecture Notes in Computer Science*, vol. 2735, pp. 139–148, 2003.

[15] C. Buragohain, D. Agrawal, and S. Suri, "A game theoretic framework for incentives in P2P systems," in *Proc. 3rd Int. Conf. on Peer-to-Peer Computing (P2P'03)*, Sept. 2003, pp. 48–56.

[16] K. Lai, M. Feldman, I. Stoica, and J. Chuang, "Incentives for cooperation in peer-to-peer networks," in *Proc. Workshop on Economics of Peer-to-Peer Systems*, 2003.

[17] D. Fudenberg and J. Tirole*, Game Theory*. Cambridge, MA: MIT Press, 1991.

[18] D. P. Bertsekas*, Dynamic Programming and Stochastic Control*. New York: Academic , 1976.

[19] E. Haruvy, D. O. Stahl, and P. W. Wilson, "Evidence for optimistic and pessimistic behavior in normal-form games," *Econ. Lett.*, vol. 63, pp. 255–259, 1999.

[20] H. A. Simon, "A behavioral model of rational choice," *Quart. J. Econ.*, vol. 69, pp. 99–118, 1955.

[21] K. Jain, L. Lovász, and P. A. Chou, "Building scalable and robust peer-to-peer overlay networks for broadcasting using network coding," *Distrib. Comput.*, vol. 19, no. 4, pp. 301–311, 2007.

[22] *Multimedia Over IP and Wireless Networks*, M. van der Schaar and P. A. Chou, Eds. New York: Academic, 2007.

[23] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, no. 3-4, pp. 279–292, May 1992.

[24] G. de Veciana and X. Yang, "Fairness, incentives and performance in peer-to-peer networks," in *41th Annu. Allerton Conf. Communication, Control and Computing*, 2003.

[25] M. Gupta, P. Judge, and M. Ammar, "A reputation system for peer-to-peer networks," in *Proc. 13th Int. Workshop on Netw. and Operating Systems Support for Digital Audio and Video (NOSSDAV'03)*, 2003, pp. 144–152, ACM Press.

[26] R. G. Gallager*, Discrete Stochastic Processes*. Norwell, MA: Kluwer, 1995.

Hyunggon Park received the B.S. degree (magna cum laude) in electronics and electrical engineering from the Pohang University of Science and Technology (POSTECH), Korea, in 2004, and the M.S. and Ph.D. degrees in electrical engineering from the University of California, Los Angeles (UCLA), in 2006 and 2008, respectively.

His research interests are game theoretic approaches for distributed resource management (resource reciprocation and resource allocation) strategies for multi-user systems and multi-user transmission over wireless/wired/peer-to-peer (P2P) networks. In 2008, he was an intern at IBM T. J. Watson Research Center, Hawthorne, NY.

Dr. Park was a recipient of the Graduate Study Abroad Scholarship from the Korea Science and Engineering Foundation during 2004-2006, and a recipient of the Electrical Engineering Department Fellowship at UCLA in 2008.

Mihaela van der Schaar is currently an Associate Professor in the Electrical Engineering Department at University of California, Los Angeles. Her research interests are in multimedia communications, networking, processing and systems. Dr. van der Schaar received the NSF Career Award in 2004, the Best Paper Award from IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY in 2006, the Okawa Foundation Award in 2006, the IBM Faculty Award in 2005, 2007 and 2008, and the Most Cited Paper Award from *EURASIP: Image Communications* journal in 2005, 2007 and 2008.