Game Theoretic Markov Decision Processes for Optimal Decision Making in Social Systems

Yan Chen, Yang Gao, Chunxiao Jiang, and K. J. Ray Liu
Department of Electrical and Computer Engineering,
University of Maryland, College Park, MD 20742, USA.
E-mail:{yan, yanggao, jcx, and kjrliu}@umd.edu

***Abstract*—One key problem in social systems is to understand how users learn and make decision. Since the values of social systems are created by user participation while the user-generated data is the outcome of users' decisions, actions and their social-economic interactions, it is very important to take into account users' local behaviors and interests when analyzing a social system. In this paper, we propose a game-theoretic Markov decision process (GTMDP) framework to study how users make optimal decisions in a social system. By explicitly considering users' local interactions and interests, we show that the proposed GTMDP can correctly derive the optimal decision and thus achieve much better expected long-term utility compared with the traditional MDP. We also discuss how to design mechanism to steer users' behavior under the proposed GTMDP framework.**

***Index Terms*—Game theory, Markov decision process, Symmetric Nash equilibrium**

I. Introduction

The rapid development of social media has greatly reduced the barrier for people to participate in online activities and create online content, which consequently leads to a proliferation of social systems. Typical examples include online social networks like Facebook [1] or Twitter [2], crowdsourcing sites like Amazon Mechanical Turk [3], and online question and answering (Q&A) sites like Quora [4] or Stack Overflow [5]. These social systems have enabled and provided easy access to large-scale user generated content. Such abundant and still growing real life data, known as "big data", open a tremendous research opportunity in many fields, and in this paper we focus on the problem of utilizing the large-scale user generated content for better decision making.

Regarding analyzing and learning from big data, machine learning has been an important tool and various machine learning algorithms have been developed [6], [7]. However, since the user-generated big data is the outcome of users' decisions, actions and their social-economic interactions, which are highly dynamic, without considering users' local behaviors and interests, existing learning approaches tend to focus on optimizing a global objective function at the macroeconomic level, while totally ignore users' local decisions at the microeconomic level [8]–[10]. As such there is a need in bridging learning with strategic decision making to be more effective in mining, reasoning and extracting knowledge and information from the user-generated big data.

In this paper, we conduct a case study of the Markov decision process (MDP) modeling [11]. Specifically, we find that the traditional MDP cannot be directly used to model user behavior in social system since it implicitly assumes that the decisions made by the decision maker in an MDP have no influence on the observed information and thus the underlying model, i.e., the stationary transition probability table and reward function is independent with the actions of the decision maker. By explicitly involving users' decisions on the model, we propose a game-theoretic MDP to analyze users' optimal decisions in social systems. We show that compared with the traditional MDP, the proposed game-theoretic MDP can better predict the optimal decision and thus greatly improve users' expected long-term utility.

The rest of the paper is organized as follows. In section II, we briefly review the traditional MDP. Then, the proposed game-theoretic MDP is introduced in details in section III. Finally, simulation results are discussed in section IV and conclusions are drawn in section V.

II. Markov Decision Processes

A Markov decision process (MDP) models how a decision maker makes a sequence of decisions in a stochastic environment to maximize its long-term reward, as shown in Fig. 1. More precisely, an MDP is a discrete time stochastic control process. At each time step, the process is in some state and the decision maker may choose any action available at that state. Then, the process responds at the next time step by moving into a new state according a stationary transition probability and giving the decision maker a corresponding reward. Given current state and the action, the stationary transition probability is conditionally independent of all previous states and actions as well as the time step.

Formally, an MDP is a 4-tuple $\langle \mathbf{S}, \mathbf{A}, \mathbf{P}, \mathbf{R} \rangle$ [11], where $\mathbf{S}$ is the state space, $\mathbf{A}$ is the action space, $\mathbf{P}$ is the table of transition probabilities with $P(s'|s,a)$ being the probability that action $a \in \mathbf{A}$ in state $s \in \mathbf{S}$ will lead to state $s' \in \mathbf{S}$, and $\mathbf{R}$ is the reward function with $R(s,a)$ being the immediate expected reward received by taking action a at state s.

The goal of the decision maker is to choose a sequence of actions, i.e., the optimal policy, to influence the system to perform optimally to maximize its long-term reward. There are several different ways to define the long-term reward such as total rewards over finite horizon and total discounted rewards over infinite horizon. In this paper, we consider the latter case, i.e., the decision maker should maximize the long-term reward

978-1-4799-7088-9/14/$31.00 ©2014 IEEE

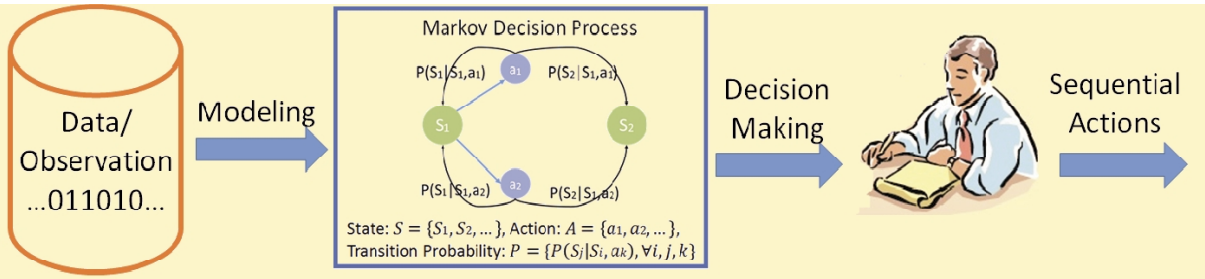

Fig. 1. System model of Markov decision process.

$\sum_{t=0}^{\infty}\gamma^t R(s^t,a^t)$, where γ is the discount factor and satisfies $0 \leq \gamma < 1$, and $R(s^t,a^t)$ is the expected reward received at time step t $R(s^t,a^t)=\sum_{s^{t+1}} P(s^{t+1}|s^t,a^t)R(s^{t+1}|s^t,a^t)$ with $R(s^{t+1}|s^t,a^t)$ being the immediate reward the decision maker can receive when transiting from state s^t to state s^{t+1} by taking action a^t.

Note that the decision maker can also take randomized actions. However, according to Theorem 1 shown as follows, in case of finite state and action space, there is an optimal deterministic stationary policy, i.e., the reward that the optimal deterministic policy will achieve is at least as high as that of the optimal randomized policy and such optimal deterministic policy is stationary and thus independent of time step.

Theorem 1: When both the state space **S** and action space **A** are finite, there exists an optimal deterministic stationary policy [11, Theorem 6.2.10 on p. 154].

With Theorem 1, it suffices to focus only on the deterministic stationary policy. Let $V(s)$ be the optimal long-term reward the decision maker can obtain when the current state is s, then we have the following Bellman equation

$$V(s) = R(s,a^*(s)) + \gamma \sum_{s'} P(s'|s,a^*(s))V(s'), \qquad (1)$$

where $a^*(s)$ is the optimal action that maximizes $V(s)$.

By solving (1), we can obtain optimal long-term reward $V(s)$ and optimal stationary deterministic policy $a^*(s)$. Value iteration is the most widely used and best understood algorithm for solving the Bellman equation. It starts with any initialization of $V^0(s)$ and iteratively updates $V^{t+1}(s)$ using $V^t(s)$ with the corresponding optimal action. The iteration terminates when the sum of the difference between $V^{t+1}(s)$ and $V^t(s)$ is smaller than a pre-defined tolerance. For the discounted MDP problem, the value iteration algorithm is guaranteed to converge and will lead to the optimal deterministic stationary policy with a sufficiently small tolerance [11].

III. Game Theoretic Markov Decision Processes

In an MDP, the whole process is controlled by a single decision maker as shown in Fig. 1. Specifically, the decision maker builds an MDP by determining the state and action space and training the transition probability table as well as the reward function based on the observed information. Then, with the value iteration algorithm, the optimal policy can be derived, according to which the decision maker can make a sequence of decisions at different time steps to maximize its long-term reward. Note that the sequential decisions made by the decision maker in an MDP have no influence on the observed information and thus the underlying model, i.e., the stationary transition probability table and reward function is independent with the actions of the decision maker. Therefore, the MDP formulation is only suitable to the scenario where there is a centralized controller that can control the whole process. To tackle the distributed decision-making scenario, which is generally the case in social systems with user-generated content, in this section, we extend the MDP formulation and propose a game theoretic Markov decision process (GTMDP).

A. Problem Formulation

As shown in Fig. 2, we consider a system with a homogeneous population, where the population can be finite, infinite or even dynamic. Every player in the population can observe the same information and build an MDP. Since players interact with each other, e.g., competing with each other for a certain resource or cooperating with each other to achieve a certain objective, the reward function and transition probability in the MDP are jointly determined by all players' decisions. Due to the homogeneous population assumption, here we focus on the symmetric scenarios. Therefore, the long-term reward of a player is evaluated under the assumption that other players choose a unified action which may be different from the action of the player under consideration.

Let $\hat{\mathbf{a}} = \{\hat{a}(s), \forall s \in \mathbf{S}\}$ be an action rule of the player under consideration, and $-\mathbf{a}$ denotes that all other players in the population choose action rule $\mathbf{a}$. Let $R(s,\hat{a}(s),-\mathbf{a})$ be the immediate expected reward received by the player at state s with action $\hat{a}(s)$ while other players use the action rule $\mathbf{a}$. Let $P(s'|s,\hat{a}(s),-\mathbf{a})$ be the stationary transition probability that the system will transit from state s to state s' when the player takes an action $\hat{a}(s)$ while other players use the action rule $\mathbf{a}$. Let $V(s)$ be the optimal long-term reward the player under consideration can obtain when the current state is s and all other players use the optimal action rule $\mathbf{a}^*$, then we have the following Bellman equation

$$V(s) = R(s,a^*(s),-\mathbf{a}^*) + \gamma \sum_{s'} P(s'|s,a^*(s),-\mathbf{a}^*)V(s'), \quad (2)$$

where the optimal policy $a^*(s)$ can be found as follows

$$a^*(s)=\arg\max_a \left\{ R(s,a,-\mathbf{a}^*)+\gamma \sum_{s'} P(s'|s,a,-\mathbf{a}^*)V(s') \right\}. \quad (3)$$

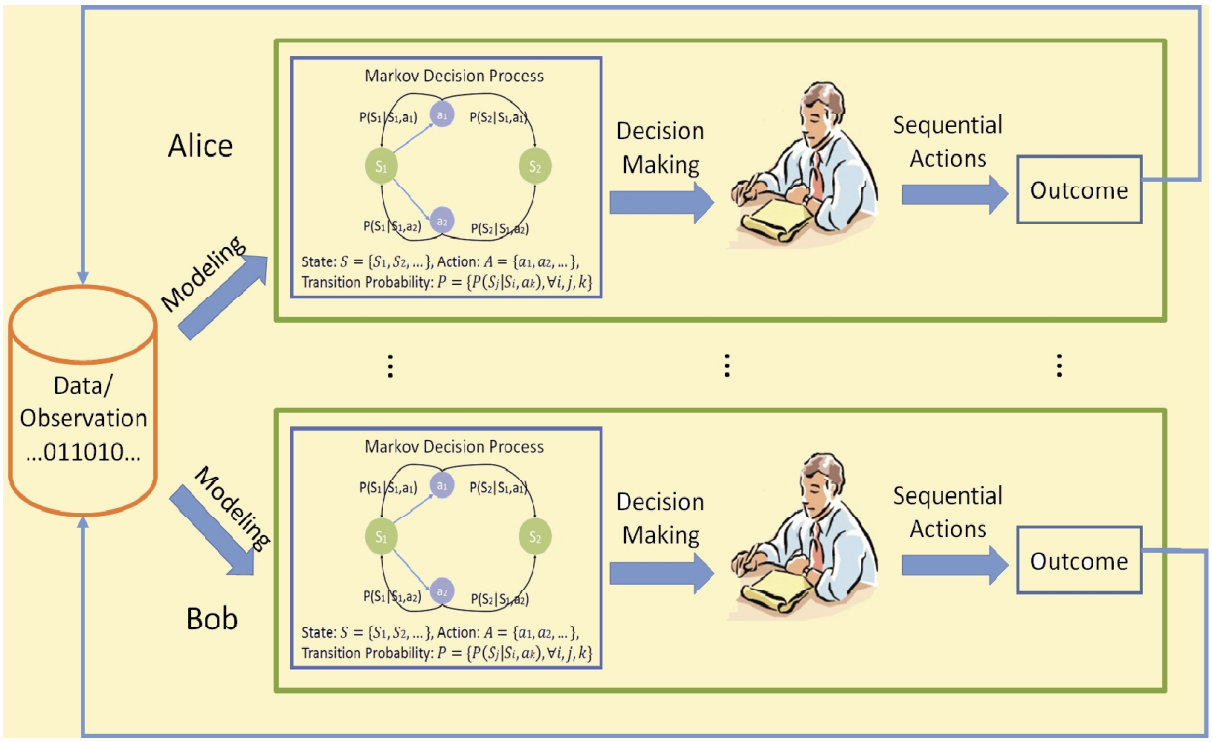

Fig. 2. System model of game theoretic Markov decision process.

Definition 1 (Symmetric Nash Equilibrium): an action rule $\mathbf{a}^*$ is a symmetric Nash equilibrium in GTMDP if $\mathbf{a}^*$ is the best response of a player when all other players are using action rule $\mathbf{a}^*$.

Theorem 2: An action rule $\mathbf{a}^*$ is a symmetric Nash equilibrium if and only if $\mathbf{a}^*$ is a solution to (2) and (3).

Proof: To prove that $\mathbf{a}^*$ is a symmetric Nash equilibrium, we first assume that all other players adopt action rule $\mathbf{a}^*$ except one player under consideration. Then, according to ***Definition 1***, we need to show that $\mathbf{a}^*$ is the best response of the player under consideration. Given that all other players adopt action rule $\mathbf{a}^*$, the problem of finding the optimal action for a certain player can be modeled as an MDP with the immediate reward function $R(s, a, -\mathbf{a}^*)$ and stationary transition probability $P(s'|s, a, -\mathbf{a}^*)$. According to the one shot deviation principle for MDP [12], the sufficient and necessary condition for $\mathbf{a}^*$ to be the best response of the player under consideration is

$$R(s, a^*(s), -\mathbf{a}^*) + \gamma \sum_{s'} P(s'|s, a^*(s), -\mathbf{a}^*)V(s') \geq R(s, a, -\mathbf{a}^*) + \gamma \sum_{s'} P(s'|s, a, -\mathbf{a}^*)V(s'), \ \forall a, \quad (4)$$

where $V(s)$ is the long-term reward the player under consideration can obtain at state s when all players in the population use the action rule $\mathbf{a}^*$, i.e., $V(s)$ can be obtained with (2).

Comparing (3) with (4), we can see that the sufficient and necessary condition for an action rule $\mathbf{a}^*$ to be a symmetric Nash equilibrium is that $\mathbf{a}^*$ satisfies (2) and (3). This completes the proof. ■

From ***Theorem 2***, we can see that the solution to (2) and (3) is the sufficient and necessary condition for an action rule to be a symmetric Nash equilibrium. Therefore, to find the symmetric Nash equilibrium, we need to solve (2) and (3). The major difference between GTMDP and MDP is that the immediate reward function $R(s, a, -\mathbf{a}^*)$ and stationary transition probability $P(s'|s, a, -\mathbf{a}^*)$ in GTMDP are dependent with the optimal action rule $\mathbf{a}^*$. In such a case, the value iteration algorithm cannot be directly applied here. To solve (2) and (3), we propose a modified value iteration algorithm. The algorithm is motivated by the fact that the problem of finding the optimal action for a certain player can be modeled as an MDP when the common action rule adopted by all other players is given. It starts with any initialization of the common action rule $\mathbf{a}^0$ for all other players and iteratively finds the optimal action rule $\mathbf{a}^{t+1}$ using value iteration algorithm. The iteration terminates when the optimal action rule $\mathbf{a}^{t+1}$ is equal to the common action rule adopted by all other players $\mathbf{a}^t$. Moreover, to avoid the algorithm being trapped into local oscillation, we record the history of common action rule and do not allow any common action rule being re-used. Since the action space is finite, the modified value iteration algorithm is guaranteed to converge to a symmetric Nash equilibrium if symmetric Nash equilibrium exists. If there are multiple symmetric Nash equilibria, the converged symmetric Nash equilibrium depends on the initial common action rule.

B. Mechanism Design

As we discussed above, if there exists multiple symmetric Nash equilibria, the proposed modified value iteration

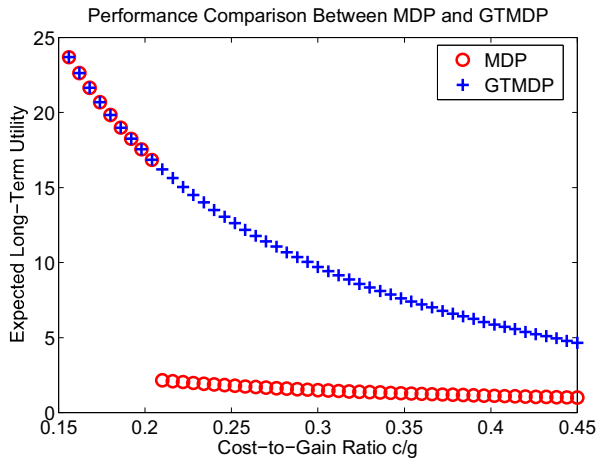

Fig. 3. Performance comparison between MDP and GTDMP.

algorithm will converge to a symmetric Nash equilibrium. Nevertheless, the existence of the symmetric Nash equilibria has not been discussed, which is typically very difficult [13]. Moreover, such action rules are generally not deterministic but randomized and their performance cannot be guaranteed. Therefore, in this paper, instead of focusing on the existence of the symmetric Nash equilibrium, we adopt the philosophy of mechanism design by first choosing a desired action rule and then finding the sufficient and necessary condition for the action rule to be a symmetric Nash equilibrium.

Specifically, let $\mathbf{a}^d$ be the desired action rule. Then, according to ***Theorem 2***, the sufficient and necessary condition for $\mathbf{a}^d$ to be a symmetric Nash equilibrium is

$$R(s,a^d(s),-\mathbf{a}^d)+\gamma\sum_{s'}P(s'|s,a^d(s),-\mathbf{a}^d)V(s')$$
$$\geq R(s,a(s),-\mathbf{a}^d)+\gamma\sum_{s'}P(s'|s,a(s),-\mathbf{a}^d)V(s'),\forall a(s)\&s. \quad (5)$$

By solving (5), we can find the constraints on the system parameters for the desired action rule $\mathbf{a}^d$ to be a symmetric Nash equilibrium.

IV. Simulations

To evaluate the performance of the proposed GTMDP algorithm, we consider a scenario with a sufficient large population where players interact with each other and produce user-generated data. Each player in the population is labeled with "+" or "-". At each time instant, a fraction of players is chosen from the population to form pairs to play a packet forwarding game. Within each pair, one player acts as a transmitter and the other player acts as a receiver. The transmitter can choose to either forward a packet or deny forwarding, i.e., the action of the transmitter is $\{0,1\}$ where "1" stands for forwarding a packet and "0" stands for not forwarding. With such a setting, we can use the labels of a pair of players to represent the system state. Since there is only two possible label for each player, "+" or "-", the state space is $\mathbf{S}=\{s_{11}=++,s_{12}=+-,s_{21}=-+,s_{22}=--\}$. Note that the transmitter may take different action at different system state. Therefore, an action rule can be defined as $\mathbf{a}=\{a(s_{11}),a(s_{12}),a(s_{21}),a(s_{22})\}$ where $a(s_{ij})\in\{0,1\}$ is the action taken at state s_{ij}.

Assuming that the receiver can obtain a gain g at a cost c to the transmitter. In such a case, the immediate payoff that a player can receive when taking action $a(s_{ij})$ while all other players use action rule $\tilde{\mathbf{a}}$ is $R(s_{ij},a(s_{ij}),-\tilde{\mathbf{a}})=-\frac{1}{2}a(s_{ij})c+\frac{1}{2}\tilde{a}(s_{ji})g$ where the factor $\frac{1}{2}$ comes from the equal probability of a player acting as a transmitter or a receiver. We further assume that there is a social norm, $\mathbf{Q}$, for updating players' label, which specifies what new label players will have according to their actions and current system state

$$\mathbf{Q}=\begin{array}{c}\begin{matrix}s_{11} & s_{12} & s_{21} & s_{22}\end{matrix}\\ \begin{bmatrix}\lambda & 1 & 0 & 1-\lambda\\ 1 & \lambda & 1-\lambda & 0\end{bmatrix}\end{array}\begin{matrix}a(s)=0\\ a(s)=1\end{matrix} \quad (6)$$

where each element $Q(a(s_{ij}),s_{ij})$ stands for the probability of the transmitter being labelled as "+" after taking action $a(s_{ij})$, and the parameter $\lambda\in[0,1]$ is used to control the weight of current label in determining the new label.

We compare the proposed GTMDP with the traditional MDP algorithm. The desired action rule of the proposed GTMDP is set to be $\mathbf{a}^d=\{a^d(s_{11})=1,a^d(s_{12})=0,a^d(s_{21})=1,a^d(s_{22})=0\}$. The MDP algorithm directly learns the model from the observation where we assume that 80% of players are with label "+", and half population uses the desired action rule while the rest adopts other action rules randomly and uniformly. Other parameters are set as follows: $c=1$, $\lambda=0.5$, $\gamma=0.9$, and $\epsilon=0.01$.

The expected long-term utility versus the cost-to-gain ratio under different schemes is shown in Fig. 3. We can see that the MDP achieves the same performance with the GTMDP at small cost-to-gain ratios. This is because when the cost-to-gain ratio is small, users tend to play cooperatively and the optimal action derived by MDP is the desired action rule $\mathbf{a}^d$. However, when cost-to-gain ratio increases, i.e., the cost of cooperation increases, users tend to take advantage of others and the action derived by MDP is to play non-cooperatively while the optimal action should be the desired action $\mathbf{a}^d$. By explicitly involving other users' decision, the proposed GTMDP can correctly derive the true optimal decision and thus achieves much better expected long-term utility.

V. Conclusions

In this paper, we proposed a general game-theoretic Markov decision process (GTMDP) framework to analyze and study users' decision making in social systems. Moreover, we developed a modified value iteration algorithm to compute optimal actions for users. Such an algorithm is guaranteed to converge to the optimal action when the optimal actions exist. We then used mechanism design to derive the sufficient and necessary conditions for a desired action rule to be optimal. Simulations show that compared with the traditional MDP, the proposed GTMDP can achieve much better expected long-term utility for users.

REFERENCES

[1] Facebook, https://www.facebook.com/.

[2] Twitter, https://twitter.com/.

[3] Amazon Mechanical Turk, https://www.mturk.com/mturk/welcome.

[4] Quora, http://www.quora.com/.

[5] Stack Overflow, http://stackoverflow.com/.

[6] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[7] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. The MIT Press, 2012.

[8] H. V. Zhao, W. Lin, and K. J. R. Liu, *Behavior Dynamics in Media-Sharing Social Networks*. Cambridge University Press, 2011.

[9] Y. Chen and K. J. R. Liu, "Understanding microeconomic behaviors in social networking: An engineering view," *IEEE Signal Processing Magazine*, vol. 29, no. 2, pp. 53–64, 2012.

[10] C. Jiang, Y. Chen, and K. J. R. Liu, "Graphical evolutionary game for information diffusion over social networks," *to appear in special issue on Signal Processing for Social Networks,IEEE Journal of Selected Topics in Signal Processing*.

[11] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., 2005.

[12] D. Blackwell, "Discounted dynamic programming," *Annals of Mathematical Statistics*, vol. 36, pp. 226–235, 1965.

[13] S. Adlakha and R. Johari, "Mean field equilibrium in dynamic games with strategic complementarities," *Operations Research*, vol. 61, no. 4, pp. 971–989, 2013.