**SCIENCE CHINA**
Information Sciences

• **Supplementary File** •

# A Formal Model for Multiagent $Q$-Learning on Graphs

Jinzhuo LIU[1,2], Guangchen JIANG[1], Chen CHU[2,3,4], Yong LI[5], Zhen WANG[2,6*] & Shuyue HU[7*]

[1]*Department of School of Software, Yunnan University, Kunming* 650091*, China;*
[2]*Department of School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi'an* 710072*, China;*
[3]*Department of Statistics, School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming* 650091*, China;*
[4]*Yunnan Key Laboratory of Service Computing, Yunnan University of Finance and Economics, Kunming* 650091*, China;*
[5]*Department of School of Information Science & Engineering, Yunnan University, Kunming* 650091*, China;*
[6]*Department of School of Cybersecurity, Northwestern Polytechnical University, Xi'an* 710072*, China;*
[7]*Shanghai Artificial Intelligence Laboratory, Shanghai* 200232*, China*

## Appendix A Proof of Lemma 1

*Proof.* First, we analyze the reward of an arbitrary agent $i$. For the agent's reward $r_t\left(a_j \mid \mathbf{a}^{-i}\right)$, we can expand

$$r_t^i\left(a_j \mid \mathbf{a}^{-i}\right) = \frac{1}{k^i} \sum_{\forall z \in \mathcal{N}_i, a_y \in \mathcal{A}} r_t^i\left(a_j \mid a_y^z\right),$$

and obtain:

$$\begin{aligned} r_t\left(a_j \mid \mathbf{a}^{-i}\right) = \frac{1}{k^i}\big[ & k_{a_1,t}^i r_t\left(a_j \mid a_1\right) + \ldots \\ & + k_{a_m,t}^i r_t\left(a_j \mid a_m\right)\big]. \end{aligned} \tag{A1}$$

According to

$$r\left(a_j \mid a_y\right) = \mathbf{a}_j^{\mathsf{T}} \mathbf{R} \mathbf{a}_y,$$

we can transform reward in Equation (A1) into a matrix multiplication form:

$$\begin{aligned} r_t\left(a_j \mid \mathbf{a}^{-i}\right) &= \frac{1}{k^i}\left[k_{a_1,t}^i \mathbf{a}_j^{\mathsf{T}} \mathbf{R} \mathbf{a}_1 \ldots + k_{a_m,t}^i \mathbf{a}_j^{\mathsf{T}} \mathbf{R} \mathbf{a}_m^{\mathsf{T}}\right] \\ &= \frac{1}{k^i} \mathbf{a}_j^{\mathsf{T}} \mathbf{R}\left(k_{a_1,t}^i \mathbf{a}_1 + \ldots + k_{a_m,t}^i \mathbf{a}_m\right) \\ &= \frac{1}{k^i} \mathbf{a}_j^{\mathsf{T}} \mathbf{R}\left[k_{a_1,t}^i, \ldots, k_{a_m,t}^i\right]^{\mathsf{T}} \\ &= \frac{1}{k^i} \mathbf{a}_j^{\mathsf{T}} \mathbf{R} \gamma_t^i \\ &= r_t\left(a_j \mid \gamma_t^i\right). \end{aligned} \tag{A2}$$

■

## Appendix B Proof of Lemma 2

*Proof.* For agent $i$, the action choices of their neighbors are generally different. With mean-field approximation [1, 2], we approximate the strategy of each neighbor by the same average strategy $\overline{\mathbf{x}}_t$ of the population. However, different from the previous approaches [1–3], we additionally characterize the effects of graph structure by explicitly considering the neighborhood size. Specifically, we observe that the decision-making process of each neighbor can be seen as an independent multinoulli trial:

- the $m$ available actions amounts to the $m$ events,
- each trial leads to success for exactly one event (a particular action choice),
- the success probability of an event is the probability of the corresponding action choice, which can be approximated by the average probability of choosing that action in the population.

* Corresponding author (email: w-zhen@nwpu.edu.cn, hushuyue@pjlab.org.cn)

Thus, $\lambda\left(\gamma_t^i\right)$ can be understood as the probability distribution of a multinomial distribution with the number of independent trials being $k^i$ (the number of neighbors), the number of events being $m$ (the number of available actions), and the success probabilities of events being $\overline{\mathbf{x}}_t$ (the average strategy).

$$\lambda\left(\gamma_t^i\right)=\frac{k^i!}{\prod_{l=1}^m k_{a_l}^i!}\prod_{j=1}^m \overline{x}_t\left(a_j\right)^{k_{a_j}^i}$$

is the probability mass function of the above multinomial distribution. ■

## Appendix C Proof of Lemma 3

*Proof.* The expected change in the $Q$-value on the graph caused by the action $a_j$ of agent $i$ with degree $k^i$ is the product of the probability that these agents have neighbor configuration $\gamma_t^i$, the probability of adopting action, and the change in the $Q$-value vector $\mathbf{q}_t$ in the $j$-th dimension, $\Delta q_{t,j}^i\left(\gamma_t^i,\mathbf{q}_t^i\right)$:

$$\begin{aligned}
f_j\left(\mathbf{q}_t^i\right)&=x_t^i\left(a_j\right)\sum_{\gamma_t^i\in\Gamma^i}\lambda\left(\gamma_t^i\right)\Delta q_{t,j}^i\left(\gamma_t^i,\mathbf{q}_t^i\right)\\
&=x_t^i\left(a_j\right)\sum_{\gamma_t^i\in\Gamma^i}\lambda\left(\gamma_t^i\right)\left[q_{t+1}^i\left(a_j\right)-q_t^i\left(a_j\right)\right]\\
&=x_t^i\left(a_j\right)\sum_{\gamma_t^i\in\Gamma^i}\lambda\left(\gamma_t^i\right)\alpha\left[r_t^i\left(a_j\mid\gamma_t^i\right)-q_t^i\left(a_j\right)\right]\\
&=\alpha x_t^i\left(a_j\right)\left[\sum_{\gamma_t^i\in\Gamma^i}\left[\lambda\left(\gamma_t^i\right)r_t^i\left(a_j\mid\gamma_t^i\right)\right]-q_t^i\left(a_j\right)\right].
\end{aligned}$$

■

## Appendix D Proof of Lemma 4

*Proof.* The Lemma 3 corresponds to the case of an agent with a known degree $k^i$. However, in a population where agents can have different degrees, we must consider the proportion of these agents, i.e. $\rho\left(k\right)$. In this case, we can express the average reward of the entire population in round $t$ as follows:

$$\sum_{k=1}^{\tilde{k}}\sum_{\gamma\in\Gamma(k)}\left[\rho\left(k\right)\lambda\left(\gamma\right)r_t\left(a_j\mid\gamma_t\right)\right]$$

Therefore,

$$f_j^i\left(\mathbf{q}_t^i\right)=\alpha x_t^i\left(a_j\right)\left[\sum_{\gamma_t^i\in\Gamma^i}\left[\lambda\left(\gamma_t^i\right)r_t\left(a_j\mid\gamma_t^i\right)\right]-q_t^i\left(a_j\right)\right].$$

can be expanded to:

$$v_j\left(\mathbf{q}_t\right)=\alpha x_t\left(a_j\right)\left[\sum_{k=1}^{\tilde{k}}\sum_{\gamma\in\Gamma(k)}\left[\rho\left(k\right)\lambda\left(\gamma\right)r_t\left(a_j\mid\gamma_t\right)\right]-q_t\left(a_j\right)\right]$$

■

## Appendix E Proof of Corollary 1

*Proof.* Intuitively, the probability of a neighbor configuration $\lambda\left(\gamma_t\right)$ occurring follows a multinomial distribution Multinom $\left(k,\overline{\mathbf{x}}_t\right)$ with

$$\overline{\mathbf{x}}_t\coloneqq\left[\overline{x}_t\left(a_1\right),\ldots,\overline{x}_t\left(a_m\right)\right]^{\mathsf{T}},\overline{\mathbf{x}}_t\in\mathbb{R}^m,\overline{x}_t\geqslant 0,\sum_{j=1}^m\overline{x}_t\left(a_j\right)=1,$$

where $\overline{x}_t\left(a_j\right)$ represents the average probability of all agents in the population adopting action $a_j$ at time step $t$.

Since the limit of a multinomial distribution with $m$ random variables can be approximated by a $(m-1)$-dimensional normal distribution, we can restrict $\gamma\in\mathbb{R}^m$ to its first $m-1$ rows, denote it as $\hat{\gamma}_t\in\mathbb{R}^{m-1}$. Then, the agents' average reward by adopting action $a_j$, denoted as $\bar{r}\left(a_j\right)$, can be expressed as follows in this population at time step $t$.

$$\sum_{\gamma_t\in\Gamma(k)}\int\cdots\int\varphi\left(\hat{\gamma}_t\right)r_t\left(a_j\mid\gamma_t\right)\mathrm{d}k_{a_1}\cdots\mathrm{d}k_{a_{m-1}},\tag{E1}$$

where $\varphi\left(\hat{\gamma}_t\right)$ is density function

$$\varphi\left(\hat{\gamma}_t\right)=\frac{\exp\left(-1/2\left(\hat{\gamma}_t-\hat{\mathbf{x}}_t\right)^{\mathsf{T}}\hat{\Sigma}_t^{-1}\left(\hat{\gamma}_t-\hat{\mathbf{x}}_t\right)\right)}{\sqrt{(2\pi)^{m-1}|\hat{\Sigma}_t|}}$$

and $\hat{\gamma}_t\sim\mathcal{N}_{m-1}\left(\hat{\mathbf{x}}_t,\hat{\Sigma}_t\right)$ with $(m-1)$-dimensional mean vector

$$\hat{\mathbf{x}}_t\coloneqq\left[\overline{x}_t\left(a_1\right),\ldots,\overline{x}_t\left(a_{m-1}\right)\right]^{\mathsf{T}},\ \hat{\mathbf{x}}_t\in\mathbb{R}^{m-1},$$

and $(m-1)\times(m-1)$ covariance matrix

$$\begin{aligned}\hat{\Sigma}_{t;i,j} &= \mathbb{E}\left[\left(\hat{\gamma}_{t;i}-\overline{x}_{t;i}\left(a_i\right)\right)\left(\hat{\gamma}_{t;j}-\overline{x}_{t;j}\left(a_j\right)\right)\right]\\ &= \mathrm{Cov}\left[\hat{\gamma}_{t;i},\hat{\gamma}_{t;j}\right]\end{aligned}$$

such that $1\leqslant i\leqslant(m-1)$ and $1\leqslant j\leqslant(m-1)$.

$v_j\left(\mathbf{q}_t\right)$ can be viewed as a function of the expected reward, i.e.

$$v_j\left(\mathbf{q}_t\right)=g\left(\mathbb{E}\left[r_t\left(a_j\right)\right]\right),$$

where

$$\mathbb{E}\left[r_t\left(a_j\right)\right]=\sum_{k=1}^{\tilde{k}}\sum_{\gamma\in\Gamma(k)}\left[\rho\left(k\right)\lambda\left(\gamma\right)r_t\left(a_j\mid\gamma\right)\right]. \tag{E2}$$

For a fixed $k$,

$$\mathbb{E}\left[r_t\left(a_j\right)\right]=\sum_{\gamma\in\Gamma(k)}\left[\lambda\left(\gamma\right)r_t\left(a_j\mid\gamma\right)\right].$$

For the case of the mean-field method [1,2], we can assume that the actions of agents' neighbors are the average strategy of the entire population. In other words, we consider the average payoff across the population for an agent's payoff, i.e.

$$g\left(\overline{r}_t\left(a_j\right)\right),$$

where

$$\overline{r}_t\left(a_j\right)=\mathbf{a}_j^{\mathsf{T}}\mathbf{R}\overline{\mathbf{x}}_t, \tag{E3}$$

which is the agents' average reward by adopting action $a_j$.

Please note that Equation (E2) and (E3) are different and should not be considered identical.

Then, we take the Taylor series expansion of $v_j\left(\mathbf{q}_t\right)=g\left(\mathbb{E}\left(r_t\left(a_j\right)\right)\right)$ at $\overline{r}_t\left(a_j\right)$, and obtain

$$\begin{aligned}g\left(\mathbb{E}\left[r_t\left(a_j\right)\right]\right) &= g\left(\overline{r}_t\left(a_j\right)\right)+\frac{g'\left(\overline{r}_t\left(a_j\right)\right)}{1!}\left(\mathbb{E}\left[r_t\left(a_j\right)\right]-\overline{r}_t\left(a_j\right)\right)\\ &+\frac{g''\left(\overline{r}_t\left(a_j\right)\right)}{2!}\left(\mathbb{E}\left[r_t\left(a_j\right)\right]-\overline{r}_t\left(a_j\right)\right)^2+\cdots\\ &= g\left(\mathbf{a}_j\mathbf{R}\overline{\mathbf{x}}_t\right)+\frac{g'\left(\mathbf{a}_j\mathbf{R}\overline{\mathbf{x}}_t\right)}{1!}\left(\sum_{\gamma\in\Gamma(k)}\left[\lambda\left(\gamma\right)\frac{1}{k}\mathbf{a}_j\mathbf{R}\gamma\right]-\mathbf{a}_j\mathbf{R}\overline{\mathbf{x}}_t\right)\\ &+\frac{g''\left(\mathbf{a}_j\mathbf{R}\overline{\mathbf{x}}_t\right)}{2!}\left(\sum_{\gamma\in\Gamma(k)}\left[\lambda\left(\gamma\right)\frac{1}{k}\mathbf{a}_j\mathbf{R}\gamma\right]-\mathbf{a}_j\mathbf{R}\overline{\mathbf{x}}_t\right)^2\\ &+\cdots\\ &= g\left(\mathbf{a}_j\mathbf{R}\overline{\mathbf{x}}_t\right)+\sum_{\gamma_t\in\Gamma(k)}\lambda\left(\gamma_t\right)\sum_{h=1}^{\infty}\frac{{\Delta_{\bar{\gamma}_t}}^h}{h!}g^{(h)}\left(\mathbf{a}_j\mathbf{R}\overline{\mathbf{x}}_t\right)\end{aligned} \tag{E4}$$

where $\Delta_{\bar{\gamma}_t}=\bar{\gamma}_t/k-\overline{\mathbf{x}}_t$, which is the difference between the actual neighborhood configuration and the average policy. Note that ${\Delta_{\bar{\gamma}_t}}^h$ is the $h$-th central moment of the normal distribution; therefore, ${\Delta_{\bar{\gamma}_t}}^h$ is 0 if and only if $h$ is odd. For this reason, we focus only on even $h$, meaning that Equation (E4) can be reduced to

$$\overbrace{g\left(\mathbf{a}_j\mathbf{R}\overline{\mathbf{x}}_t\right)}^{\mathsf{term1}}+\underbrace{\sum_{\gamma_t\in\Gamma(k)}\lambda\left(\gamma_t\right)\sum_{l=1}^{\infty}\frac{{\Delta_{\bar{\gamma}_t}}^{2l}}{2l!}g^{(2l)}\left(\mathbf{a}_j\mathbf{R}\overline{\mathbf{x}}_t\right)}_{\mathsf{term2}} \tag{E5}$$

It is obvious that the bias of the mean-field method [1,2] is mainly caused by the central moments ${\Delta_{\bar{\gamma}_t}}^{2l}$ of the distribution in term2 of Equation (E5).

In the scenario we are considering, where the actions are independent of each other, we can obtain

$${\Delta_{\bar{\gamma}_t}}^{2l}=\frac{\sum_{j=1}^{m}\left[\overline{x}_t\left(a_j\right)\left(1-\overline{x}_t\left(a_j\right)\right)\right]^l}{k^l}\left(2l-1\right)!!$$

[4] and

$$\overbrace{g\left(\mathbf{a}_j\mathbf{R}\overline{\mathbf{x}}_t\right)}^{\mathsf{term1}}+\underbrace{\sum_{\gamma_t\in\Gamma(k)}\lambda\left(\gamma_t\right)\sum_{l=1}^{\infty}\left[\frac{(2l-1)!!}{(2l)!}g^{(2l)}\left(\mathbf{a}_j\mathbf{R}\overline{\mathbf{x}}_t\right)\frac{\sum_{j=1}^{m}\left[\overline{x}_t\left(a_j\right)\left(1-\overline{x}_t\left(a_j\right)\right)\right]^l}{k^l}\right]}_{\mathsf{term2}}. \tag{E6}$$

It is evident that the difference $\mathbb{E}\left[r_t\left(a_j\right)\right]$ and $\overline{r}_t\left(a_j\right)$ is only related to $\sum_{j=1}^{m}[\overline{x}_t(a_j)(1-\overline{x}_t(a_j))]^l/k^l$ in term2 of Equation (E6). For Equation (E6), $\mathsf{term2}\to 0$ with $k\to 0$. Hence, we can obtain

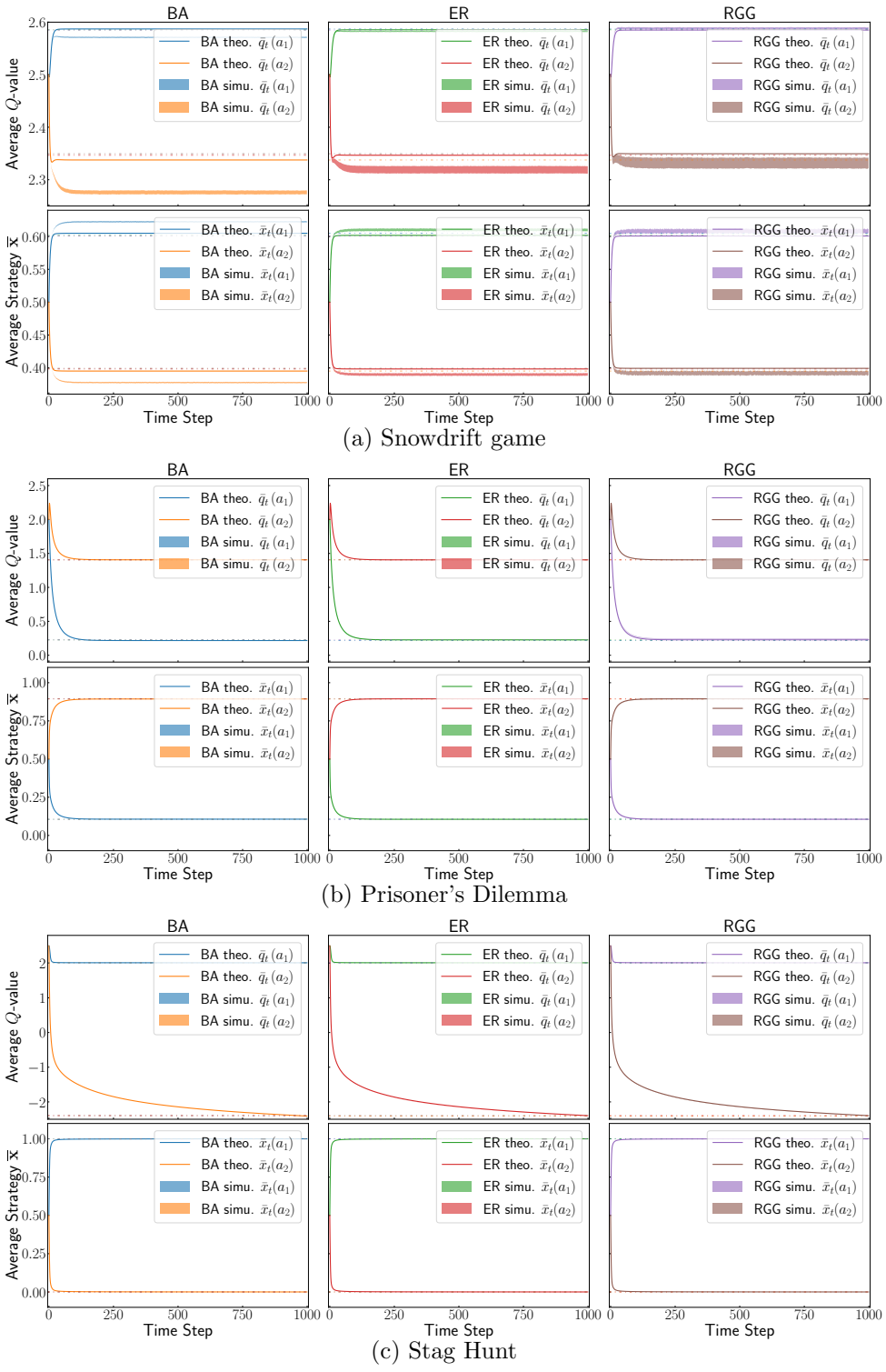
$$\lim_{k\to\infty}\frac{\partial p\left(\mathbf{q}_t,t,k\right)}{\partial t}=\frac{\partial\hat{p}\left(\mathbf{q}_t,t\right)}{\partial t},$$

and

$$p\left(\mathbf{q}_t,t,k\right)\xrightarrow[\mathcal{D}]{k\to\infty}\hat{p}\left(\mathbf{q}_t,t\right)$$

■

## Appendix F Extended Validation of Figure 2. (a): Results from 100 to 1000 Steps




(a) Snowdrift game

(b) Prisoner's Dilemma

(c) Stag Hunt

**Figure F1** Extended experimental results from the Snowdrift (SD), Prisoner's Dilemma (PD), and Stag Hunt (SH) games for populations in three graph structures — Barabási–Albert (BA), Erdős–Rényi (ER), and Random Geometric Graph (RGG) - extend from 100-steps to 1,000-steps. The results clearly demonstrate that the discrepancies between theoretical predictions and experimental outcomes stabilize and do not increase beyond a certain point. Specifically, in SD, which shows the largest difference among the games tested, the discrepancies remain relatively small even in the BA network, with errors $\approx 0.016$ for $\bar{q}(a_1)$ and $\approx 0.061$ for $\bar{q}(a_2)$. Similarly, in the ER network and RGG, these errors are about $\approx 0.0033$ and $\approx 0.02$, respectively.

**References**


1 Hu S, Leung C, Leung H. Modelling the dynamics of multiagent q-learning in repeated symmetric games: a mean field theoretic approach. In: Advances in Neural Information Processing Systems, 2019, 32

2 Hu S, Leung C W, Leung H, et al. The Dynamics of Q-learning in Population Games: a Physics-Inspired Continuity Equation Model. In: 21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022, Auckland, New Zealand, 2022: 615-623

3 Leung C W, Hu S, Leung H F. Formal Modeling of Reinforcement Learning with Many Agents through Repeated Local Interactions. In: 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI), IEEE, 2021: 714-718

4 Shiryaev A N. Probability-1. Springer, 2016