The Price of Governance: A Middle Ground Solution to Coordination in Organizational Control

Chao Yu[1*] and **Guozhen Tan**[2]

[1]School of Data & Computer Science, Sun Yat-Sen University, Guangzhou, China
[2]School of Computer Science & Technology, Dalian University of Technology, Dalian, China
yuchao3@mail.sysu.edu.cn, gztan@dlut.edu.cn

Abstract

Achieving coordination is crucial in organizational control. This paper investigates a middle ground solution between decentralized interactions and centralized administrations for coordinating agents beyond inefficient behavior. We first propose the *price of governance* (PoG) to evaluate how such a middle ground solution performs in terms of effectiveness and cost. We then propose a hierarchical supervision framework to explicitly model the PoG, and define step by step how to realize the core principle of the framework and compute the optimal PoG for a control problem. Two illustrative case studies are carried out to exemplify the applications of the proposed framework and its methodology. Results show that the hierarchical supervision framework is capable of promoting coordination among agents while bounding administrative cost to a minimum in different kinds of organizational control problems.

1 Introduction

Imagine a fairy tale that you were the king of an ancient kingdom. Owning a large piece of land, you were now facing a tricky problem: how to divide the land into smaller administrative districts such that you can run your kingdom as efficiently as possible? You might choose to directly govern each citizen by yourself. While you could take full control of your kingdom in this way, a heavy burden would definitely be imposed. Or, you might prefer letting each district administrate itself. As the number (size) of districts is getting larger (smaller), however, your kingdom tends to be more fragmented and thus your authority is prone to be weakened. Facing this dilemma, you were puzzled: what is the best size for the partition such that your control is not impaired but at the same time the whole kingdom can function efficiently?

Although this fairy tale seems naive, it reveals a fundamental yet challenging issue in organizational control, where *centralization* and *decentralization* are two completely opposite solutions to guarantee system performance. Application domains include but are not limited to the management of supply chains [Giannoccaro, 2018], resource allocation in cognitive radio networks [Hasegawa *et al.*, 2014] or multiuser OFDMA networks [Yassin *et al.*, 2017], and multi-robot formation/consensus control [Oh *et al.*, 2015]. In these domains, it is crucial to design efficient *coordination mechanisms* that enable all the agents to reach an agreement in areas of common interest. While *centralized* mechanisms often rely on a dictatorial authority to formulate, specify and enforce how the agents should coordinate with each other, *decentralized* mechanisms enable agents to coordinate via local interactions and without relying on any centralized authority.

On one hand, it is generally believed that system performance could be improved upon given dictatorial control over agents' actions. Imposing such control, however, can be costly or even infeasible in large systems due to the expense of high administrative cost. Moreover, as the environments where agents reside in become even more dynamic and open, continuously monitoring and governing each agent's behavior will soon become intractable. On the other hand, pure decentralized mechanisms usually cannot guarantee satisfactory performance if no external interventions or explicit mechanisms are imposed. As in the *social dilemma*, for example, pure rational behavior based on best-response reinforcement learning will end up with non-cooperative defection, also known as *selfish equilibria*, which is suboptimal with respect to the social welfare [Bazzan *et al.*, 2011; Yu *et al.*, 2015]. In *coordination games*, high level of coordination among distributed learning agents are rarely achieved if no extra mechanisms are introduced, especially in situations with stochastic and partial information [Kapetanakis and Kudenko, 2002]. Similar results can also be observed in various *congestion games*, for example, resource allocation problems, in which inductive reasoning or regret-based learning would always lead to inefficient equilibria that are far worse than the optimum [Oh and Smith, 2008].

In this paper, we investigate the possibility of a middle ground solution between decentralized interactions and centralized administrations, by which self-interested agents can utilize a proper level of coordination to improve their performance beyond inefficient equilibria or uncoordinated behavior. Based on the two quantitative criteria of *price of anarchy* (PoA) and *price of monarchy* (PoM), we propose the *price of governance* (PoG) to evaluate how such a middle ground solution performs in terms of effectiveness and cost. The PoA

*Contact Author

measures the *inefficiency* of a decentralized solution with respect to the natural objective function, while PoM is defined as the practical *cost* of centralized administration. By combining these two criteria into an overall value of PoG using a combination function, an optimal middle ground solution can be properly discovered to make the best trade-off between *inefficiency/cheapness* of decentralization and *optimality/high cost* of centralization.

Motivated to explicitly model PoG and seek its optimal value, we then introduce a hierarchical supervision framework, which nicely reflects the features of organizational structure and hierarchical governance in human societies. We define step by step how to realize the core principle of hierarchical supervision in the framework and compute the optimal PoG for a control problem. We then carry out a preliminary set of simulations in two case studies: the *norm learning* (N-L) problem [Delgado, 2002] and *multi-agent resource selection* (MARS) problem [Oh and Smith, 2008], to evaluate the efficacy of our framework. Results show that the hierarchical supervision framework can facilitate coordination among agents (i.e., reducing the PoA) compared to a pure decentralized solution. At the same time, an optimal PoG can be achieved to bring out the maximum coordination promotion while bounding the PoM significantly lower than that of a centrally administrated system.

2 Seeking the Optimal PoG

2.1 The Price of Governance

In its original form, the *price of anarchy* (PoA) was defined as the worst-case ratio between the value of social cost in a *Nash equilibrium* and that of some *social optimum*. To expand the realm of its applications, PoA has also been generalized to many other contexts, such as for defining the inefficiency of a *multiagent learning* algorithm in resource allocation problems [Oh and Smith, 2008], and more broadly the inefficiency of decentralization [Youn *et al.*, 2008; Cole *et al.*, 2015]. Following this, we define PoA as the ratio of performance loss at an equilibrium (convergence) to the optimal performance that could possibly be achieved by a centralized optimization approach. More formally,

$$PoA = \frac{\psi_{opt} - \psi_{dis}}{\psi_{opt}}, \quad PoA \in [0, 1] \tag{1}$$

where ψ_{opt} is the optimal performance using a centralized solution and ψ_{dis} is the performance of a decentralized solution. The performance can be any criterion that evaluates a solution, e.g., coordination level or convergence speed.

Analogous to the price of anarchy, we can define the price of monarchy as the practical cost of maintaining centralization in a system. To simplify illustration, we mainly discuss managerial cost in terms of communication cost. Thus, the lower bound of the price of monarchy is found in a fully decentralized non-communicating system, and the upper bound of the price of monarchy is found in a fully centralized system. Let φ_{dis} and φ_{opt} denote a communication cost function of a decentralized solution and a centralized solution, respectively. *Price of Monarchy* (PoM) is given by:

$$PoM = \frac{\varphi_{dis}}{\varphi_{opt}}, \quad PoM \in [0, 1] \tag{2}$$

The *Price of Governance* (PoG) then can be computed using a combination function Γ of PoA and PoM:

$$PoG = \Gamma(PoA, PoM) \tag{3}$$

By defining different function Γ, one can capture various patterns of behavior towards totalitarianism or liberalism, depending on the specific purpose of solving a target problem.

2.2 The Hierarchical Supervision Framework

Inspired to model PoG and seek its optimal value, we then introduce the hierarchical supervision framework, which is composed of the following five steps.

Step 1: Segmentation of social groups. To model hierarchical supervision and organizational governance in human societies, a social group is first divided into a set of subgroups according to some predefined methods. Interactions of subgroup members are purely local and decentralized, and may be constrained by certain external factors such as network topologies or social relationships. In each sub-group, a superior governor monitors and administrates the behavior of its subordinates. The governor can be any one of the subordinate agents in the sub-group or another dedicated agent. A governor can also interact with another governor or other governors to exchange their information or learn from each other using social learning strategies.

Step 2: Aggregation of public opinions. In each subgroup, agents make decisions in a fully decentralized manner. Agents may

- learn from the outcome of interaction with another randomly chosen member using reinforcement learning;
- copy another member's behavior using some imitation rules; or simply
- make decisions independently.

Each agent then reports its decision to its governor, who then aggregates all the information from its subordinates to form a public opinion using various democratic mechanisms. This public opinion summarizes the overall attitude towards the members' behavior in the governor's group.

Step 3: Generation of supervision policies. After obtaining the public opinion, each governor then generates a supervision policy by exchanging its information with another governor and learning from situations in other subgroups. The generated supervision policy is deemed as the most successful behavior in the neighbourhood, and can be in different forms such as being the majority action adopted by the members, the action that performs the best (i.e., with the highest reward) or their combinations.

Step 4: Adaption of local behavior. The supervision policy is then passed down to the group members by the governor in order to entrench its influence in the group. According to the targeted problem, this integration process can be conducted in distinct manners. The supervision policy can be used to dictate the policy for group members directly, or as a suggestive guide to adapt members' behavior through modifying their behaviorial parameters (e.g., learning speed or exploration mode), transforming the environmental components (e.g., states or rewards), or changing the way how members interact with each other (e.g., to whom to interact).

Step 5: Calculation of the optimal PoG. PoA can be computed as the ratio of the performance value at convergence to the optimal performance value using a centralized solution that a single governor supervises the whole group. PoM can be computed as the practical communication cost of maintaining centralization in the group. The communication cost can be in different forms such as the number of message exchange or the geometrical distance between group members and the governor. Then, PoG can be calculated using a predefined combination function Γ that depends on the specific purpose of solving a target problem. As the size of subgroups indicates different levels of centralization and thus different PoG, the whole problem is then reduced to seeking the optimal size of subgroups in which case the minimal PoG can be achieved. Let $p \in \mathcal{P}$ be a partition of the group. The problem is now transformed into the following optimization problem:

$$\begin{aligned} &\min_{p\in\mathcal{P}} \ PoG = \Gamma(PoA, PoM) \\ &s.t. \ \ \urcorner_p(PoA, PoM) = 0, \quad \forall p \in \mathcal{P} \end{aligned} \tag{4}$$

where $\urcorner$ is the PoA and PoM relationship function that is determined by the chosen coordination solution. The constraints in Eq. (4) means that for each partition of the group, its PoA and PoM value should satisfy their relationship function. For a problem that PoA, PoM and their relationship function can be computed in a closed form, general optimization methods can be applied to compute the solution of this optimization problem. The other more straightforward way is to sample in the partition space in different levels of granularity, and then apply approximation methods to estimate the relationship between PoA and PoM. The optimal PoG then can be easily derived by solving the combinatorial equations of relationship function $\urcorner$ and PoG function Γ.

3 Two Case Studies

3.1 The Norm Learning Problem

Social norm is an important concept in multiagent systems to facilitate coordination among agents by posing constraints on agents' behavior [Shoham and Tennenholtz, 1997]. The *norm learning* (NL) problem deals with how a social norm can be established in a bottom-up manner via agents' local learning interactions. This problem has attracted a great interest in recent years and extensive investigations have been conducted under various assumptions about agent interaction protocols, societal topologies, and observation capabilities [Yu *et al.*, 2014; Vouros, 2017; Hasan *et al.*, 2015].

Problem Description

Considering a typical setting of network topology, a group of agents are organized in a social network and each agent can only interact with its neighbors, using either reinforcement learning approaches or some predefined imitation rules. The interactions between two agents can be modeled as a pure *Coordination Game* (CG) [Sen and Airiau, 2007], in which the agents are rewarded positively when their actions are consistent and penalized otherwise. The goal is to enable all the agents to reach an agreement (*social norm*) in the whole system. Although this problem seems simple, successfully solving it is a challenging task due to the widely recognized existence of *sub-norms*, which prevents completely consistent social norm in the whole group [Mihaylov *et al.*, 2014].

Application of the Methodology

We now provide an illustration of how to apply the proposed methodology in solving NL problems in a structured system where agents interact with each other using basic reinforcement learning algorithms. Some results in this section appeared in an earlier version of our work in [Yu *et al.*, 2018].

Step 1: Group segmentation. We use an $R * R$ grid network by default, and separate it into $n * n$ $(1 \leq n \leq R)$ subgroups (In case of n being not divisible by R, the remaining agents on the border are included in a single subgroup.), each of which is denoted as C_x. We imagine a governor located in the geometrical center of each subgroup.

Step 2: Aggregation of public opinion. At time step t, in each subgroup C_x, agent i chooses an action a_i with the highest Q-value or randomly chooses an action with an exploration probability ϵ_i^t. Agent i then interacts with a random neighbor j and receives a payoff r_i. The learning experience in terms of action-reward pair (a_i, r_i) is reported to agent i's governor x, and the governor aggregates all the information from its subordinates into two values F_x and R_x. Value $F_x(a)$ indicates the overall acceptance (i.e., frequency) of action a in subgroup C_x and value R_x(a) indicates the overall reward of action a in C_x. $F_x(a)$ can be calculated as $F_x(a) = \sum_{i \in C_x} \delta(a, a_i)$, where $\delta(a, a_i)$ is the Kronecker delta function, which equals to 1 if $a = a_i$, and 0 otherwise. $R_x(a)$ can be calculated by $R_x(a) = \frac{1}{F_x(a)} \sum_{i \in C_x, a_i = a} r_i$. Especially, $R_x(a)$ is set to 0 if $F_x(a) = 0$. Each governor x then combines the actions of each subgroup into a public opinion o_x using democratic voting mechanism ($o_x = arg\ max_a\ F_x(a)$ is the action most accepted by the subgroup).

Step 3: Generation of supervision policies. After generating the public opinion, each governor then generates a supervision policy a_x, which indicates the *social norm*, i.e., the most successful behavior, in the neighborhood. To this end, the governor resorts to *social learning* with another governor by changing their information and comparing the performance of their public opinions. Many *social learning* strategies can be applied for this purpose, and we here employ the widely used imitation rules from the *evolutionary game theory* (EGT) [Szabó and Fath, 2007], given by $p_{x \to y} = \frac{1}{1+e^{-\beta(u_y - u_x)}}$, where $p_{x \to y}$ is a probability for governor x to imitate the action of neighboring governor y, $u_x = R_x(o_x)$ is the fitness of the public opinion of governor x, $u_y = R_y(o_y)$ is the fitness of neighboring governor y, and $\beta > 0$ is a parameter to control selection bias.

Step 4: Adaption of local behavior. Each agent i in a subgroup adjusts its learning behavior in order to comply with the generated supervision policy from its governor. By comparing its action a_i^t with the supervision policy a_x, agent i can

evaluate whether it is performing well or not so that its learning behavior can be dynamically adapted to fit the supervision policy. Learning rate and exploration rate are two fundamental tuning parameters in RL. Heuristic adaption of these two parameters thus models the adaptive learning behavior of agents. More specifically, when agent i has chosen the same action with the supervision policy (i.e., $a_i^t = a_x$), it decreases its *learning rate* to maintain its current state ($\alpha_i^{t+1} = (1-\lambda)\alpha_i^t$), otherwise, it increases its learning rate to learn faster from its interaction experience ($\alpha_i^{t+1} = (1-\alpha_i^t)\lambda + \lambda$), where $\lambda \in [0, 1]$ is a parameter to control the adaption rate. The *exploration rate* can be updated likewise. Finally, agent i updates its Q-value using the new learning rate α_i^{t+1} and/or exploration rate ϵ_i^{t+1}. The proposed mechanisms are based on the concept of "winning/losing" in the well-known multi-agent learning algorithm WoLF (Win-or-Learn-Fast) [Bowling and Veloso, 2002]. While the original meaning of "winning/losing" in WoLF and its variants is to indicate whether an agent is doing better or worse than its Nash-Equilibrium policy, this heuristic is gracefully borrowed here to evaluate the agent's performance against the supervision policy.

Step 5: Calculation of optimal PoG. We sample the size of subgroup from $n = 1$ to $n = R$ and derive PoA and PoM for each size of subgroup through simulations. The PoA indicates the ratio of performance loss at an equilibrium (convergence) to the optimal performance that could possibly be achieved by a centralized optimization approach. In the NL problem, the performance loss can be reflected by the consensus level of the whole group, i.e., the proportion of agents in the whole system that have not achieved a consensus. Thus, PoA can be computed as the proportion of agents with subnorms in the system. As for the PoM, the geometric distance between a group agent and its governor is used to represent the communication cost. For each case of subgroup size, the PoA and PoM pair can be obtained accordingly. Function approximation methods then can be applied to fit all the pairs to derive the relationship function $\urcorner$ between PoA and PoM. Finally, given a predefined function Γ, the optimal PoG value and its associated subgroup size can be found accordingly.

Experiments and Results

First, we would like to test whether the proposed hierarchical supervision framework is capable of facilitating coordination among agents (i.e., reducing the PoA), compared on a pure decentralized learning approach. We conduct the investigation on a $10 * 10$ grid network, which is separated into several $4 * 4$ clusters (the remaining 2 agents on the border are included in a single group). We consider stateless version of Q-learning, and each agent can choose from 4 actions as default. Parameters α and ϵ are initially set to 0.1 and 0.01, respectively. Moreover, parameter β and γ are both set to 0.1. The final results are averaged over 1000 independent runs. We compare our *hierarchical learning* approaches (denoted as HL) to the fully decentralized *individual learning* (IL) approach that agents learn randomly with another agent and update their strategies independently.

The left part in Figure 1 [Yu *et al.*, 2018] shows that the coordination ratio of the whole group using different approaches increases as learning proceeds, but the hierarchical learning

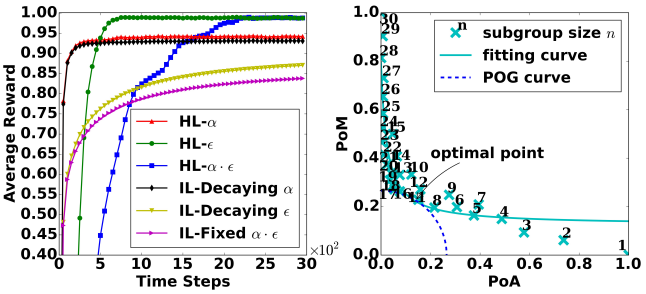

Figure 1: The left part plots the dynamics of PoA in terms of average reward using different learning approaches. HL-α, HL-ϵ, and HL-$\alpha \cdot \epsilon$ denote, respectively, the three approaches under the proposed hierarchical supervision framework when agents adapt their learning rate α, exploration rate ϵ, and both rates at the same time. The other three approaches represent the IL approaches with a decaying α, a decaying ϵ or a fixed α and ϵ. The right part shows the relationship between PoA and PoM, and the calculation of the optimal PoG in the NL problem.

approaches (especially HL-ϵ and HL-$\alpha \cdot \epsilon$) can reach a higher convergence level than the individual learning approaches. This result indicates that by introducing a certain level of centralized control, the PoA can be greatly reduced. The right part in Figure 1 plots the relationship between PoM and PoA when the subgroup size takes different values of n in a 30×30 grid network. As we can see, a larger size can result in a higher consensus level (i.e., lower PoA). This is easy to understand because each governor can have a more powerful control force over the group comparatively when the subgroup size is larger. The communication cost, however, also increases as the subgroup size becomes larger, causing a higher PoM. It is obvious that the PoA and PoM are two contradictory criteria that evaluate the coordination performance. PoA indicates the consensus level while PoM indicates the cost for achieving this performance. Higher PoM indicates a more centralized system (i.e., higher cost) and thus a better coordination performance (i.e., lower PoA) can be achieved. We can observe that the PoA and PoM exhibit a monotonous relationship with a long tail phenomenon. This indicates that the PoA can be reduced significantly by only introducing a bit of centralized control, e.g., when subgroup size is less than 10. This improvement, however, is only at the expense of very low communication cost, as reflected by the low value of PoM. We then apply simple function approximation methods to fit all the PoM-PoA pairs to derive the relationship function $\urcorner$, and derive the optimal PoG value and its associated subgroup size, on the tangent point between the curve of PoG function and the fitting curve of relationship function $\urcorner$.

3.2 The MARS Problem

The *multi-agent resource selection* problem (MARS) is a class of *congestion games* characterized by a large number of self-interested agents competing for common resources [Oh and Smith, 2008]. This so called congestion effect is apparent in many real-world situations, ranging from traffic routing in transportation systems, bandwidth allocation in communication networks, to other versions of *tragedy of the commons*

that are characterized by negative externalities.

Problem Description

Formally, an MARS problem can be defined as a quadruple of (N, Θ, A, R), in which $N = 1, 2, ..., n$ is a set of agents, $\Theta = r_1, ..., r_m$ denotes a set of resources available for agents in N, $A_t = a_1 \times ... \times a_n$ denotes the resource choices of the agents at time t where $a_i \in \Theta, \forall i \in N$, and $R_t : \Theta \times A_t \rightarrow \Re$ is a reward function. In MARS, a reward associated with using a resource is defined as a function of the number of concurrent users of the resource, and all users using the same resource share the same reward. So, agents' valuations of congested resources are not exogenously-determined, but rather are endogenous functions of one each other's actions.

The *El Farol bar problem* (EFBP) is a simple example of MARS, which was introduced in [Arthur, 1994]. In EFBP, a set of n agents repeatedly make decisions of whether to attend a bar or not on certain nights. The only observations available to the agents are the past history of attendance at the bar. The bar is a congested resource so that the payoff of attending a bar is high only if the number of attendees at the bar on the night is less than a certain threshold ζ. The agents receive the worst payoff if the bar is over crowded. Thus, an agent needs to reason about the attendance so as to decide whether to attend it or not. However, a rational agent always fails to learn the best decision based on its expected reward, since all agents are simultaneously learning the same information and reason in the same manner. This leads to the *rationality paradox* in general MARS problems.

Application of the Methodology

In the bar problem, we apply the methodology of our hierarchical supervision framework in the following five steps.

Step 1: Group segmentation. We imagine 900 people living in a district with 30*30 blocks, which can be separated into several subgroups with $n*n$ blocks. The different values of n thus indicate different levels of centralized governance.

Step 2: Aggregation of public opinion. Each member i has a probability of p_i to attend the bar. A community governor in each subgroup collects the information of how many subordinates went to the bar last night (i.e., attendance radio in the subgroup) and the average reward of subordinates in the subgroup. The attendance radio in the subgroup is the public opinion and the average reward is its performance

Step 3: Generation of supervision policies. Based on the public opinion from the subgroup in terms of attendance radio and its reward, the governor applies reinforcement learning to update her knowledge about whether to attend the bar. To apply the tabular form of Q-learning, we transform the attendance radio between $[0, 1]$ into a set of discrete actions. Moreover, the governors need to learn from other governors by comparing their performance. Simple imitation rule (e.g., the Fermi rule) can be employed for this purpose. If the governor accepts the action of another governor in the social learning process, she informs the new action to her subordinates as the subversion policy, otherwise, the governor chooses an action based on the Q values using an exploration strategy, and informs her subordinates the chosen action accordingly

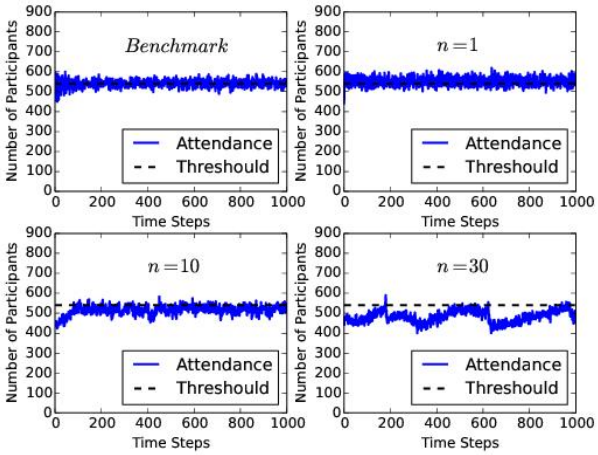

Figure 2: The dynamics of attendance in the bar problem. In the benchmark case, agents make their decisions without any hierarchial control. In the hierarchical supervision case, n indicates the size of subgroups in the 30*30 group.

Step 4: Adaption of local behavior. When receiving the supervision policy in terms of attendance probability a_i, a subordinate agent i updates its policy of attendance probability p_i directly by $p_i = (1 - \mu) * p_i + \mu * a_i$, where μ is a learning rate. In the next round, agent i decides whether to attend the bar based on the updated attendance probability.

Step 5: Calculation of optimal PoG. We normalize the PoA by $PoA_n = 1 - \frac{r_n - r_{min}}{r_{max} - r_{min}}$, where r_n is the average reward of group size n, r_{min} and r_{max} are the minimum and maximum average reward for different sizes of subgroups, respectively. The calculation of PoM is: $PoM_n = \frac{c_n - c_{min}}{c_{max} - c_{min}}$, where c_n is the communication cost of group size n, c_{min} and c_{max} are the minimum and maximum communication cost for different sizes of subgroups, respectively. The geometric distance is still used to represent the communication cost. The relationship function between PoA and PoM, as well as the optimal size and PoG value then can be derived in the same way as in the NL problem.

3.3 Experiment Results

We set the threshold of attendance ζ to 540, which means that the bar can only accommodate 60% of total 900 agents at most. If more than 540 agents attend the bar, they will receive a penalized reward to indicate a crowded situation. The learning parameter μ is set to 0.1. In Q-learning, the range of attendance probability in between $[0, 1]$ is divided into 50 discrete actions, with the interval between two adjacent action being probability of 0.02. Exploration rate ε is set to 0.01. In each trial, the experiment runs for 1000 nights (i.e., time steps), and we take the average of rewards in the last 100 time steps for evaluation. As a benchmark, agents maintain the probability of attendance, and update the policy by learning with an arbitrary agent in the system using the imitation rule. The results are averaged over 1000 trials.

From Figure 2, it is clear that the number of participants in

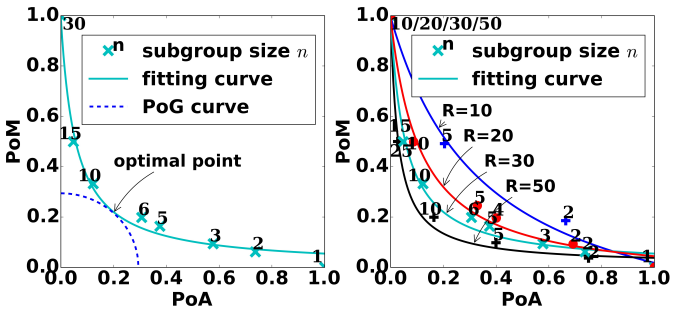

Figure 3: The relationship of PoA and PoM in the bar problem in the case of 30*30 group (the left part), and the relationship of PoA and PoM in different sizes of agent groups (the right part)

the benchmark model is fluctuating along the threshold, indicating an inefficient equilibrium among agents. When an agent predicts the attendance at the bar is lower than ζ, then the agent decides to attend the bar. Since the other agents also reason in the same manner, the entire population decides to attend the bar, ending up with the worst payoffs. Therefore, agents face contradicting outcomes by making decisions based on their rationality. This *rationality paradox* can be greatly alleviated using the proposed framework by imposing a certain level of centralized control on the agents. As can be seen, the number of overcrowded nights is significantly reduced when $n = 10$. The performance is further promoted when the governor has a wider control range when $n = 30$ in a fully centralized manner.

The left part in Figure 3 presents the relationship curve when the 30*30 group is divided to different sizes of subgroups. It shows the same pattern of result as the NL problem, and the optimal value of PoG can be computed in the same way as described before. The right part in Figure 3 presents the impact of different sizes of agent population. It is apparent that a higher population size R generates a lower optimal PoG. This is an interesting phenomenon that is a bit of counterintuitive. It demonstrates that the proposed methodology is more suitable for larger systems, where decentralized control methods cannot perform well because of the narrow vision of agents and lack of centralized control. This also provides an explanation on why in real-life situations, large organizations and systems such as countries and companies usually embody the feature of hierarchical supervision structures to make an elegant balance between centralized governance and decentralized administration.

4 Related Work

PoA has been extensively studied in the area of computational economics and computer science to study the inefficiency of selfish behavior [Koutsoupias and Papadimitriou, 1999; Andelman *et al.*, 2009]. The mainstream research in this direction focuses on the computational analysis of upper or lower bound on PoA under various conditions of congestion games [Wang *et al.*, 2016; Feldman *et al.*, 2016] and real-world applications [Youn *et al.*, 2008; Chen and Zhang, 2012]. However, most of these studies do not consider an explicit cost associated with the decision making process, which is unrealistic in real-life problems where manageable cost or communication cost is inevitable in sustaining the global system order. While the work in [Oh and Smith, 2008] has extended the definition of PoA as a measure of inefficiency of a multiagent learning algorithm in MARS problems, and considered adminstration cost in such contexts, our work highlights a general hierarchical supervision framework to explicitly model and trade off PoA and its associated cost.

There is also tremendous amount of work that aims to solve coordination issues in the two case problems. For the NL problem, numerous mechanisms have been proposed for efficient emergence of social norms while agents interact with each other using learning (particularly reinforcement learning) methods. These mechanisms include the social learning strategy [Sen and Airiau, 2007], the collective interaction protocol [Yu *et al.*, 2014], the utilization of topological knowledge [Hasan *et al.*, 2018] and agents' observation capabilities [Villatoro *et al.*, 2011]. Several solutions have also been proposed to solve the *selfish equilibria* problem in MARS, or more broadly, social dilemmas, when agents use rational learning strategies for interaction. For example, Bazzan *et al.* resorted to social instruments of hierarchy and coalition to promote cooperation in Iterated Prisoner's Dilemmas [Bazzan *et al.*, 2011]. Oh and Smith applied social learning to promote the social welfare in MARS problems [Oh and Smith, 2008]. Our work supplements the literature by providing new effective solutions to these two challenging problems.

5 Conclusions

In this paper, we argue for the benefits of considering both centralized control and decentralized interactions in solving a coordination problem. By trading off between these two aspects, an optimal middle ground solution, represented by the lowest PoG, can be discovered. We implement the framework using sampling-based simulations to fit the PoA and PoM relationship curve before computing the lowest PoG. This proof-of-concept validation is reasonable for the two case problems where theoretical analysis and proof over the solution and its performance are still open issues in this area. However, building on the rich literature in game theoretical analysis on PoA, it is possible to derive the optimal PoG in a closed form by introducing a cost function into the problem formulation and solving the optimization problem given by Eq. (4) directly. Moreover, richer phenomenon may be revealed under various problem settings or specifications, e.g., combination functions, or cost measurement etc. Also, we are expecting implementations of this framework and its methodology in solving other real-world coordination problems, in which efficiency and cost are the two main optimized objectives (e.g., resources allocation in cognitive radio networks).

References

[Andelman *et al.*, 2009] Nir Andelman, Michal Feldman, and Yishay Mansour. Strong price of anarchy. *Games and Economic Behavior*, 65(2):289–317, 2009.

[Arthur, 1994] William Brian Arthur. Inductive reasoning and bounded rationality. *The American economic review*, 84(2):406–411, 1994.

[Bazzan *et al.*, 2011] Ana L.C. Bazzan, Ana Peleteiro, and Juan C. Burguillo. Learning to cooperate in the iterated prisoner's dilemma by means of social attachments. *Journal of the Brazilian Computer Society*, 17(3):163–174, 2011.

[Bowling and Veloso, 2002] Michael Bowling and Manuela Veloso. Multiagent learning using a variable learning rate. *Artificial Intelligence*, 136(2):215–250, 2002.

[Chen and Zhang, 2012] Ying-Ju Chen and Jiawei Zhang. Design of price mechanisms for network resource allocation via price of anarchy. *Mathematical programming*, 131(1-2):333–364, 2012.

[Cole *et al.*, 2015] Richard Cole, José R Correa, Vasilis Gkatzelis, Vahab Mirrokni, and Neil Olver. Decentralized utilitarian mechanisms for scheduling games. *Games and Economic Behavior*, 92:306–326, 2015.

[Delgado, 2002] Jordi Delgado. Emergence of social conventions in complex networks. *Artificial intelligence*, 141(1):171–185, 2002.

[Feldman *et al.*, 2016] Michal Feldman, Nicole Immorlica, Brendan Lucier, Tim Roughgarden, and Vasilis Syrgkanis. The price of anarchy in large games. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 963–976. ACM, 2016.

[Giannoccaro, 2018] Ilaria Giannoccaro. Centralized vs. decentralized supply chains: The importance of decision maker's cognitive ability and resistance to change. *Industrial Marketing Management*, 2018.

[Hasan *et al.*, 2015] Mohammad Rashedul Hasan, Anita Raja, and Ana Bazzan. Fast convention formation in dynamic networks using topological knowledge. In *AAAI2015*, pages 2067–2073, 2015.

[Hasan *et al.*, 2018] Mohammad Rashedul Hasan, Anita Raja, and Ana Bazzan. A context-aware convention formation framework for large-scale networks. *Autonomous Agents and Multi-Agent Systems*, pages 1–34, 2018.

[Hasegawa *et al.*, 2014] Mikio Hasegawa, Hiroshi Hirai, Kiyohito Nagano, Hiroshi Harada, and Kazuyuki Aihara. Optimization for centralized and decentralized cognitive radio networks. *Proceedings of the IEEE*, 102(4):574–584, 2014.

[Kapetanakis and Kudenko, 2002] Spiros Kapetanakis and Daniel Kudenko. Reinforcement learning of coordination in cooperative multi-agent systems. *AAAI/IAAI*, 2002:326–331, 2002.

[Koutsoupias and Papadimitriou, 1999] Elias Koutsoupias and Christos Papadimitriou. Worst-case equilibria. In *Annual Symposium on Theoretical Aspects of Computer Science*, pages 404–413. Springer, 1999.

[Mihaylov *et al.*, 2014] Mihail Mihaylov, Karl Tuyls, and Ann Nowé. A decentralized approach for convention emergence in multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 28(5):749–778, 2014.

[Oh and Smith, 2008] Jean Oh and Stephen F Smith. A few good agents: multi-agent social learning. In *AAMAS2008*, pages 339–346, 2008.

[Oh *et al.*, 2015] Kwang-Kyo Oh, Myoung-Chul Park, and Hyo-Sung Ahn. A survey of multi-agent formation control. *Automatica*, 53:424–440, 2015.

[Sen and Airiau, 2007] Sandip Sen and Stephane Airiau. Emergence of norms through social learning. In *Proc. of 20th IJCAI*, pages 1507–1512, 2007.

[Shoham and Tennenholtz, 1997] Yoav Shoham and Moshe Tennenholtz. On the emergence of social conventions: modeling, analysis, and simulations. *Artificial Intelligence*, 94(1-2):139–166, 1997.

[Szabó and Fath, 2007] György Szabó and Gabor Fath. Evolutionary games on graphs. *Physics reports*, 446(4-6):97–216, 2007.

[Villatoro *et al.*, 2011] Daniel Villatoro, Jordi Sabater-Mir, and Sandip Sen. Social instruments for robust convention emergence. In *IJCAI*, volume 11, pages 420–425, 2011.

[Vouros, 2017] George A Vouros. Learning conventions via a social reinforcement learning in complex and open settings. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 455–463. International Foundation for Autonomous Agents and Multiagent Systems, 2017.

[Wang *et al.*, 2016] Xuehe Wang, Nan Xiao, Lihua Xie, Emilio Frazzoli, and Daniela Rus. Analysis of price of total anarchy in congestion games via smoothness arguments. *IEEE Transactions on Control of Network Systems*, pages 876–885, 2016.

[Yassin *et al.*, 2017] Mohamad Yassin, Samer Lahoud, Kinda Khawam, Marc Ibrahim, Dany Mezher, and Bernard Cousin. Centralized versus decentralized multi-cell resource and power allocation for multiuser ofdma networks. *Computer Communications*, 107:112–124, 2017.

[Youn *et al.*, 2008] Hyejin Youn, Michael T Gastner, and Hawoong Jeong. Price of anarchy in transportation networks: efficiency and optimality control. *Physical review letters*, 101(12):128701, 2008.

[Yu *et al.*, 2014] Chao Yu, Minjie Zhang, and Fenghui Ren. Collective learning for the emergence of social norms in networked multiagent systems. *IEEE Transactions on Cybernetics*, 44(12):2342–2355, 2014.

[Yu *et al.*, 2015] Chao Yu, Minjie Zhang, Fenghui Ren, and Guozhen Tan. Emotional multiagent reinforcement learning in spatial social dilemmas. *IEEE transactions on neural networks and learning systems*, 26(12):3083–3096, 2015.

[Yu *et al.*, 2018] Chao Yu, Hongtao Lv, Hongwei Ge, Liang Sun, Jun Meng, and Bingcai Chen. Centralization or decentralization? a compromising solution toward coordination in multiagent systems. *Interactions in Multiagent Systems*, pages 19–42, 2018.