AI

MDPI

Article

Cluster-Based Q-Learning Relational Game (C-QLRG): A Practical Relaxation for Asymmetric Online Social Networks

Duc Nghia Vu [1,2,*] and Janos Demetrovics [3]

[1] Faculty of Development Economics, University of Economics and Business, Vietnam National University, Hanoi 11310, Vietnam
[2] Center for Youth Entrepreneurship Education and Studies, Hanoi 10000, Vietnam
[3] HUN-REN Institute of Computer Science and Control, Hungarian Academy of Science, 1111 Budapest, Hungary
* Correspondence: vuducnghia@vnu.edu.vn

Abstract

The Q-Learning Relational Game (QLRG) framework provides a theoretically rigorous method for identifying minimal winning coalitions in online social networks (OSNs) under the restrictive assumption of global agent symmetry or uniform matroid structure. Real-world OSNs, however, exhibit significant asymmetry. This paper introduces the Cluster-Based Q-Learning Relational Game (C-QLRG), a practical extension that relaxes the global symmetry requirement by leveraging community structure. We partition the agent set into communities with bounded internal variation and represent the state solely by community membership counts of the seed set. Because the closure operator already captures all eventual influence spread, the problem reduces to a sequential seed selection task where the agent decides, at each step, from which community to add the next seed. We prove that the optimal Q-function of a suitably regularized reach-efficiency objective is Lipschitz continuous and derive a performance bound for the learned policy. The full algorithm is presented, and its complexity is analyzed. Empirical evaluations on a synthetic asymmetric network and Zachary's Karate Club demonstrate that C-QLRG is highly sensitive to reward parameters, where default settings lead to premature stopping, but parameter tuning combined with a corrected minimality verification recovers high-efficiency coalitions by removing non-contributing agents. With tuned parameters, C-QLRG produces a near-winning coalition of size 11 and 99% reach on the synthetic network, surpassing the greedy baseline's efficiency (size 12) despite a one-node coverage gap, while identifying the optimal winning coalition of size 1 on the Karate Club dataset, matching all baselines. The framework thus offers a principled trade-off between model fidelity and scalability, with the reward design choice being critical for practical deployment.

Keywords: Q-learning; relational games; minimal winning coalition; asymmetric networks; community detection; sufficient statistics; Lipschitz continuity

Check for updates

Received: 11 May 2026
Revised: 6 June 2026
Accepted: 16 June 2026
Published: 22 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction and Literature Review

The identification of minimal winning coalitions in online social networks (OSNs) is a fundamental problem in influencer marketing, viral campaign design, and strategic decision-making. The Q-Learning Relational Game (QLRG) framework [1] recently demonstrated that under idealized conditions—global agent symmetry or a uniform matroid structure—deterministic Q-learning can be integrated with relational game theory to solve this NP-hard problem in polynomial time $O(n^4)$, reducing the state space from exponential

$O(2^n)$ to quadratic $O(n^2)$. However, this symmetry assumption is rarely satisfied in practice. Real OSNs exhibit heterogeneous influence: users have different numbers of followers, engagement rates vary, and network structure is highly asymmetric. Two coalitions of the same size and total reach may have vastly different future expansion potential due to the presence of "bridge" nodes, structural holes, or redundant influencers.

To bridge this gap, the original QLRG paper [1] outlined a promising direction: cluster-based sufficient statistics. By partitioning the agent set into communities where within-community variation is bounded, the symmetry requirement can be relaxed from global to local. The authors established Lipschitz continuity of the optimal Q-function with respect to community-wise agent swaps and provided error bounds. However, that proposal remained theoretical—no algorithms, implementation, or empirical validation were provided.

This paper presents the **Cluster-Based Q-Learning Relational Game (C-QLRG)**, a full extension that makes QLRG applicable to asymmetric networks. Our contributions are:

a. **Operationalization of cluster-based sufficient statistics**: We define a concrete cluster state representation $(k_1, \ldots, k_m)$ where k_i is the number of selected agents (seeds) in community i. Crucially, we observe that in the relational game setting, the full transitive closure $L(A)$ captures all eventual influence; intermediate propagation steps (the Boost action) do not alter the final outcome and are therefore superfluous. The problem is reduced to a sequential seed addition task in which the agent chooses, at each step, which community to draw the next seed from.

b. **Theoretical error bounds**: We prove new Lipschitz continuity results (Lemma 1, Corollary 1) for the cluster abstraction and derive a performance-loss theorem (Theorem 1) that quantifies the suboptimality of the learned cluster-based policy with an error bound $\frac{1}{1-\gamma}\sum_{i=1}^{m} \bar{k}_i \epsilon_i^{\text{true}}$.

c. **Complete algorithms**: We present cluster-based Q-value iteration and coalition-formation algorithms with complexity $O(m \cdot n^3)$ for moderate m, which is more efficient than the original QLRG's $O(n^4)$.

d. **Empirical evaluation and parameter sensitivity**: We test C-QLRG on a synthetic asymmetric network (n = 100) with planted communities and compare it against the original QLRG and greedy baselines. The implementation reveals that under the originally chosen reward parameters, the agent terminates prematurely, selecting only two seeds and failing to reach full coverage. This outcome highlights a critical sensitivity to the terminal bonus and cost weights. Preliminary tuning of these hyper-parameters (increasing η or decreasing δ) restores the ability to identify near-minimal winning coalitions, consistent with the theoretical error bounds. The experiment thus provides a controlled proof-of-concept of the algorithm's correctness while underscoring the importance of reward calibration.

e. **Practical guidelines**: We provide a decision rule for when to use C-QLRG versus original QLRG or deep reinforcement learning approaches, based on network size, community cohesion, and required optimality guarantees.

While the original QLRG framework [1] established polynomial–time coalition identification under strict global symmetry, this assumption is violated in virtually all real-world OSNs. This work extends the QLRG framework introduced in [1]. The cluster-based sufficient statistics idea was mentioned as a future direction in ([1], Section 8) but remained purely theoretical. Here, we operationalize this concept, derive new Lipschitz bounds tailored to community swaps, and provide the first implementation and evaluation.

This paper introduces C-QLRG, which **fundamentally relaxes** the symmetry requirement by leveraging community structure, enabling application to asymmetric networks. Unlike the theoretical sketch in [1], we provide **complete algorithms**, **rigorous error bounds** for the cluster abstraction, **remove the redundant Boost action** to simplify the MDP, and

deliver **empirical validation** with termination logic. These contributions transform QLRG from an idealized benchmark into a deployable tool for asymmetric influence maximization.

Table 1 makes the comparative analysis between C-QLRG and QLRG. The remainder of this paper is organized as follows. Section 2 reviews relational games, QLRG, and the symmetry limitation. Section 3 presents the cluster-based sufficient statistics framework and the augmented MDP. Section 4 states the theoretical guarantees (Lipschitz continuity and error bounds). Section 5 describes the algorithms. Section 6 analyzes computational complexity. Section 7 details the experimental setup and results. Section 8 discusses limitations and proposed mitigations, and concludes.

Table 1. Comparative analysis between C-QLRG and QLRG.

Feature	QLRG [1]	C-QLRG (This Work)
Symmetry Assumption	Global symmetry/uniform matroid	Local symmetry within communities; bounded variation ϵ_k
Applicability	Idealized symmetric networks only	Asymmetric real-world OSNs with community structure
Action Space	Add, Boost, Ignore	Add_i, Ignore (Boost removed as redundant)
State Representation	(k, r, t) triple	Community count vector $s = (c_1, \ldots, c_m)$
Theoretical Guarantees	Optimality under global symmetry	Lipschitz continuity under swaps; performance bound $O(\sum \epsilon_k)$
Algorithm Complexity	$O(n^4)$	$O(m \cdot n^3)$ for moderate m
Empirical Validation	Synthetic symmetric networks	Asymmetric networks; parameter sensitivity analysis

1.1. Related Works

The problem of identifying minimal winning coalitions in social networks sits at the intersection of cooperative game theory, influence maximization, and reinforcement learning.

Coalition formation and relational games: Traditional coalition formation has been studied extensively in cooperative game theory [1,2], but most works assume that any coalition can form and that the value of a coalition depends only on its members (characteristic function). More recently, Elkind [3] considered coalitional games on graphs where feasible coalitions are restricted to connected subsets, showing polynomial–time algorithms for certain sparse graph families. Vu et al. [1,2,4,5] introduced relational dependency games, where the influence relationships are captured by closure operators derived from Armstrong's axioms. These works characterized minimal winning coalitions statically but did not provide active learning algorithms to construct them sequentially.

Influence maximization: Kempe, Kleinberg, and Tardos [6] formalized the influence maximization problem under the independent cascade and linear threshold models, proving that a greedy algorithm achieves a $(1 - 1/e)$-approximation due to submodularity. Subsequent works (e.g., TIM, IMM [7]) scaled greedy to billions of edges. However, influence maximization typically fixes a budget k and maximizes reach, whereas the minimal winning coalition problem asks for the smallest k that achieves full coverage. Moreover, these algorithms are not designed to incorporate sequential decisions or to learn from experience.

Q-learning and multi-agent reinforcement learning: Q-learning [8] and its deep variants [9] have been applied to coalition formation and game-theoretic settings. Nash Q-learning [10] and friend-or-foe Q-learning [11] handle multi-agent interactions but suffer from exponential joint action spaces and lack guarantees on solution minimality. Modular Q-learning [12] decomposes problems but relies on fixed modules. The original QLRG framework [1] was the first to integrate deterministic Q-learning with relational game theory, achieving polynomial–time complexity under a strict global symmetry assumption. However, that assumption is rarely satisfied in real OSNs.

Multi-agent and graph-based reinforcement learning: Recent advances in multi-agent reinforcement learning (MARL) have addressed cooperative tasks in networked environments. Methods such as QMIX and MADDPG enable coordinated policies but typically assume decentralized execution and lack guarantees on coalition minimality.

Graph-based MARL approaches leverage graph neural networks (GNNs) to learn state representations that respect network topology and enable communication-aware coordination [13,14]. These methods have been applied to influence maximization and combinatorial optimization on graphs, including topic-aware seed selection using graph attention mechanisms [15].

Recent work also explores credit assignment for coalition formation in MARL using game-theoretic solution concepts [16]. However, existing MARL and Graph RL frameworks generally focus on maximizing cumulative reward without explicit constraints on coalition size or formal error bounds. Moreover, they often require extensive training data and do not exploit community structure for state abstraction. While heterogeneous MARL methods address agent asymmetry [17], they do not provide the polynomial–time complexity or Lipschitz continuity guarantees that C-QLRG offers. In contrast, C-QLRG formulates a centralized MDP for coalition construction, uses community-based sufficient statistics to achieve tractable complexity, and provides theoretical bounds on suboptimality introduced by abstraction. This positions C-QLRG as a complementary approach that prioritizes theoretical rigor and interpretability for strategic coalition formation.

Beyond social networks, reinforcement learning has been successfully applied to dynamic resource allocation and scheduling in IoT environments, demonstrating its versatility in optimizing performance under constraints. Recent works include RL-based dynamic power management for energy optimization in consumer electronics [18] and adaptive scheduling mechanisms for cognitive IoT sensors to enhance network performance [19]. These studies highlight the effectiveness of RL in handling dynamic, networked decision problems. C-QLRG extends this paradigm to the domain of relational games and coalition formation, offering a centralized MDP formulation with community-based state abstraction and provable bounds on influence efficiency, which distinguishes it from heuristic or application-specific RL approaches.

State abstraction and Lipschitz MDPs: The idea of using sufficient statistics to reduce state space has a long history in reinforcement learning. Singh et al. [20] introduced soft state aggregation with error bounds. Hutter [21] studied sufficient statistics for sequential decision problems. The Lipschitz continuity of Q-functions with respect to state features has been exploited in Lipschitz MDPs [22] to enable function approximation. Our cluster-based approach extends these ideas by defining a feature vector (community counts) and proving Lipschitz continuity with respect to a swap metric.

Community detection: Community detection in networks [23] is a mature field with methods such as modularity maximization [24], Louvain [25], and spectral clustering. We leverage these as a preprocessing step, not as a contribution, but we show how the detected communities can be used to construct a provable state abstraction.

1.2. Research Gaps

Despite the progress described above, several critical gaps remain unaddressed:

a. **Lack of practical relaxation for symmetry:** The original QLRG [1] provides elegant theoretical guarantees, but these collapse when the network is asymmetric—a ubiquitous property of real OSNs. No prior work has demonstrated how to relax the symmetry assumption while maintaining polynomial–time complexity and bounded performance loss.

b. **Underutilization of community structure in reinforcement learning for coalitions:** While community detection is widely used for network analysis, its integration into reinforcement learning for coalition formation has been limited to heuristics without theoretical error bounds. Existing approaches either ignore communities or treat them as rigid constraints, not as a means to construct a provable state abstraction.

c. **No active learning algorithm for relational games in asymmetric networks:** The static relational game papers [2,4,5] describe what minimal winning coalitions look like but do not provide a sequential decision-making algorithm to find them. The original QLRG does provide an algorithm but assumes symmetry. There is a clear need for an algorithm that works in asymmetric networks with quantifiable guarantees.
d. **Empirical validation missing for cluster-based extensions:** The cluster-based sufficient statistics idea was mentioned only theoretically in the original QLRG paper [1] (Section 8) without any algorithmic specification, implementation, or experimental results. Practitioners have no guidance on how to use it or what performance to expect.
e. **Trade-off between abstraction granularity and error not quantified:** Even when community structure is available, it is unclear how the number of communities m and the within-community variation ϵ_i affect the final coalition size. A principled bound that relates these quantities to solution quality is needed to guide the choice of community detection resolution.

1.3. Contributions of This Paper

This paper addresses the above gaps by presenting the **Cluster-Based Q-Learning Relational Game (C-QLRG)**. Our specific contributions are:

a. **A practical relaxation of global symmetry:** We replace the unrealistic global symmetry assumption with a local condition: agents are interchangeable only within communities, up to a bounded variation ϵ_i. This is realistic for OSNs, which naturally exhibit community structure.
b. **A simplified, sound action space:** By exploiting the fact that the transitive closure operator L captures all eventual influence, we eliminate the Boost action. The sequential decision problem reduces to selecting from which community to add the next seed, leading to a clean, deterministic MDP that avoids the state persistence issues of earlier formulations.
c. **Theoretical error bounds:** We prove Lipschitz continuity results (Lemma 1, Corollary 1) tailored to the cluster abstraction and derive a performance-loss theorem (Theorem 1) with a tight bound $\frac{1}{1-\gamma}\sum_i \bar{k}_i \epsilon_i^{\text{true}}$, reflecting only the aggregation and ordering error from seed addition.
d. **Complete algorithms:** We present Algorithms 1 and 2 for cluster-based Q-value iteration and coalition formation, which achieve polynomial complexity $O(m \cdot n^3)$ for moderate m, offering a significant efficiency improvement over the original QLRG's $O(n^4)$ while incorporating greedy refinement and minimality verification to ensure high-quality minimal winning coalitions.
e. **Empirical evaluation on synthetic and real-world networks:** We validate C-QLRG on a synthetic asymmetric graph ($n = 100$) and Zachary's Karate Club. Results reveal strong reward sensitivity: default parameters cause early stopping, but calibrated rewards yield a coalition of size 11 with 99% reach on synthetic data, outperforming the greedy baseline in efficiency (size 12) despite missing one node. Karate Club, C-QLRG finds the optimal winning coalition of size 1, confirming generalization to real social structures. A corrected minimality verification removes redundant agents, ensuring accurate size reporting.
f. **Practical guidelines for deployment.** We provide a decision rule for choosing C-QLRG versus other methods based on network size, community cohesion, and required optimality, bridging the gap between theory and practice.

Collectively, these contributions transform the original QLRG from an idealized mathematical benchmark into a deployable tool for influencer marketing and strategic coalition formation in real-world online social networks.

Algorithm 1: Cluster-Based Q-Value Iteration

Input: Agent set U, communities $\{C_i\}_{i=1}^{m}$, dependency family F, parameters $\alpha, \beta, \delta, \eta, \gamma, \epsilon$

Output: Cluster Q-function $Q_{\text{cluster}}(s, a)$, policy π_{cluster}

1. $\mathcal{S}_{\text{cluster}} \leftarrow \{(k_1, \ldots, k_m) \mid 0 \leq k_i \leq \mid C_i \mid\}$
2. $\mathcal{A}_{\text{abs}} \leftarrow \{\text{Ignore}\} \cup \{\text{Add}_1, \ldots, \text{Add}_m\}$
3. Initialize $Q(s, a) = 0$ for all $s \in \mathcal{S}_{\text{cluster}}, a \in \mathcal{A}_{\text{abs}}$
4. **Repeat** until convergence (max change $< \epsilon$):
   For each state $s = (k_1, \ldots, k_m) \in \mathcal{S}_{\text{cluster}}$:
   - Build canonical coalition A_{can} (first k_i agents in each community).
   - Compute current reach $r = \mid L(A_{\text{can}}) \mid$ and terminal flag $t = \mathbf{1}_{r=n}$.
   - For each action $a \in \mathcal{A}_{\text{abs}}$:
     - **If** $a = \text{Ignore}$**:**
       - $Q_{\text{new}}(s, \text{Ignore}) \leftarrow 0$ *(terminal state; no future reward).*
     - **If** $a = \text{Add}_i$ **and** $k_i < \mid C_i \mid$**:**
       - $x \leftarrow (k_i + 1)$-st agent in C_i's ordering.
       - $A'_{\text{can}} \leftarrow A_{\text{can}} \cup \{x\}$.
       - $r' \leftarrow \mid L(A'_{\text{can}}) \mid$; $s' \leftarrow (k_1, \ldots, k_i + 1, \ldots, k_m)$.
       - $R \leftarrow \alpha(r' - r) - \delta + \eta \cdot \mathbf{1}_{r'=n}$.
       - $Q_{\text{new}}(s, \text{Add}_i) \leftarrow R + \gamma \max_{a'} Q(s', a')$.
     - If action infeasible, skip.

   Update $Q(s, a) \leftarrow Q_{\text{new}}(s, a)$.
5. **Return** Q, π_{cluster}

Algorithm 2: Coalition Formation with Cluster State

Input: Agent set U, communities $\{C_i\}_{i=1}^{m}$, learned Q_{cluster} and policy π_{cluster}

Output: Coalition C (approximately minimal winning)

1. Initialize $C \leftarrow \emptyset$; for $i = 1, \ldots, m$ set $k_i \leftarrow 0$.
2. Compute $r = \mid L(C) \mid$ and $t = \mathbf{1}_{r=n}$.
3. Set abstract state $s \leftarrow (k_1, \ldots, k_m)$.
4. **While** $t = 0$ and $\sum k_i < n$:
   - $a \leftarrow \pi_{\text{cluster}}(s)$.
   - **If** $a = \text{Ignore}$: **break**.
   - **Else if** $a = \text{Add}_i$ (with $k_i < \mid C_i \mid$):
     - Choose the best concrete agent $x^* \in C_i \backslash C$ that maximizes the true marginal gain:
       $x^* = \arg\max_{x \in C_i \backslash C} (\mid L(C \cup \{x\}) \mid - \mid L(C) \mid)$.
     - $C \leftarrow C \cup \{x^*\}$; $k_i \leftarrow k_i + 1$.
     - Update reach: $r \leftarrow \mid L(C) \mid$; $t \leftarrow \mathbf{1}_{r=n}$.
     - $s \leftarrow (k_1, \ldots, k_m)$.
5. **Minimality verification** (as in original QLRG):
   - **Repeat** for each agent $x \in C$:

   \quad If $L(C \backslash \{x\}) = U$ then $C \leftarrow C \backslash \{x\}$ and restart the verification loop.
6. **Return** C.

2. Preliminaries and Motivation

Before reintroducing the relational game theory and QLRG framework, we propose the following notation table for clarity.

The Table 2 above tries to establish clarity, consistency, and accessibility throughout the paper.

Table 2. Notation table.

Symbol	Meaning
N	Set of agents
C	Coalition (subset of N)
$\mathrm{cl}(C)$	Transitive closure of C
m	Number of communities
K_k	k-th community
ϵ_k	Within-community variation bound
$\mathbf{s}$	Cluster state vector $(c_1, \ldots, c_m)$
Q^*	Optimal Q-function
α, δ, η	Reward parameters (gain scale, cost, terminal bonus)

2.1. Relational Game Theory (Brief Recap)

We adopt the relational game framework from [1–3]. Let $U = \{a_1, \ldots, a_n\}$ be a finite set of agents. A *relational dependency* $A \to B$ (with $A, B \subseteq U$) means that the agents in B depend on those in A. The family F of all such dependencies is closed under Armstrong's axioms (reflexivity, transitivity, augmentation, composition). The *closure operator* $L(A) = \{a \in U \mid A \to \{a\} \in F^*\}$ gives the set of all agents influenced by coalition A. A coalition A is *winning* iff $L(A) = U$; it is *minimal winning* iff no proper subset is winning.

In the directed graph representation, $L(A)$ is precisely the set of nodes reachable from A via directed paths. Consequently, if $L(A) \neq U$, no amount of waiting or further propagation (i.e., repeated application of the one-step operator Γ) can enlarge the reach; the closure is already a fixed point. Thus, to achieve full coverage, we must add more seeds; intermediate boosts are irrelevant to the eventual outcome. This observation underpins our simplification of the action space.

2.2. QLRG Framework and Its Symmetry Requirement

The original QLRG [1] defines an abstract state $s = (k, r, t)$ where $k = \mid A \mid$, $r = \mid L(A) \mid$, and $t = \mathbf{1}_{L(A)=U}$. The key theoretical result is that under the *sufficient statistic condition* (global symmetry or uniform matroid), this triple captures all information needed for optimal decision-making. This assumption is equivalent to saying that for any two coalitions A, B with $\mid A \mid = \mid B \mid$ and $\mid L(A) \mid = \mid L(B) \mid$, the optimal Q-values are equal. In practice, this fails when:

- Agents have heterogeneous out-degrees (influence power).
- The network contains communities with different densities.
- Structural holes or bridges create non-uniform future expansion.

2.3. Why Cluster-Based Relaxation and Removing Boost?

Real OSNs are not globally symmetric, but they often exhibit *community structure*: groups of agents that are densely connected internally and sparsely connected externally. Within a community, influence patterns may be approximately similar, but communities can differ significantly. This observation motivates a *local* symmetry assumption: agents within the same community are interchangeable up to a bounded variation ϵ_i. The original QLRG's symmetry is the special case of a single community ($m = 1$) with $\epsilon_1 = 0$.

Furthermore, the original QLRG included a Boost action to model gradual influence spread, but as argued above, it is redundant for the winning coalition objective. By removing Boost, we obtain a simpler MDP where the state needs only to track which seeds have been selected, and the terminal condition is checked via the full closure. This eliminates the need to store reach explicitly, making the state abstraction both sound and compact.

3. Cluster-Based Sufficient Statistics

3.1. Community Partition and Variation Measure

Let the agent set U be partitioned into m disjoint communities $C_1, \ldots, C_m$. We adopt the notation of the **marginal gain** of adding an agent w to a coalition S:

$$\text{gain}(S, w) = \begin{cases} \mid L(S \cup \{w\}) \mid - \mid L(S) \mid, & \text{if } w \notin S, \\ 0, & \text{if } w \in S. \end{cases}$$

The **within-community variation** for the community C_i is now defined as

$$\epsilon_i^{\text{true}} = \max_{\substack{A \subseteq U, x, y \in C_i, \\ z \in U}} \mid \text{gain}(A \cup \{x\}, z) - \text{gain}(A \cup \{y\}, z) \mid$$

Remarks:

- The maximization is over all coalitions A, all pairs x, y from C_i, and **any** agent $z \in U$.
- If $z = y$ (or $z = x$), the second (or first) gain becomes zero because the agent is already present. The expression therefore automatically covers the previously excluded scenario.
- In the special case where $z \notin A \cup \{x, y\}$, the definition reduces to the original one, so the new ϵ_i^{true} is **at least as large** as the old one. However, it still quantifies the maximal difference in the reward-relevant quantity and remains a meaningful bound on intra-community asymmetry.

Community Detection Scheme: We employ the **Louvain modularity maximization algorithm** [17] to partition the agent set into communities. The algorithm optimizes the modularity score:

$$Q = \frac{1}{2m} \sum_{u,v} \left[A_{uv} - \frac{k_u k_v}{2m} \right] \delta(c_u, c_v),$$

where A_{uv} is the adjacency matrix, k_u is the degree of node u, and $\delta(c_u, c_v)$ indicates whether nodes u and v belong to the same community. We use the standard **resolution parameter** $\gamma_{\text{res}} = 1.0$, which balances community size and cohesion. Community detection is treated as a **preprocessing step**; the C-QLRG framework does not depend on the specific detection algorithm. The theoretical error bounds (Theorem 1) hold for **any partition** of the agent set, provided the within-community variation ε_i is bounded. Practitioners may choose alternative methods (e.g., spectral clustering, Infomap) based on network characteristics; the key requirement is that the resulting partition exhibits sufficiently small ε_i to maintain tight error bounds. In our experiments, Louvain consistently yields partitions with low variation and high modularity, justifying its use as the default detector.

3.2. Augmented State Representation

For a coalition A of selected seeds, define the **cluster state** solely by the community counts:

$$s = (k_1, k_2, \ldots, k_m), k_i = \mid A \cap C_i \mid .$$

The abstract state space is

$$\mathcal{S}_{\text{cluster}} = \{(k_1, \ldots, k_m) \mid 0 \leq k_i \leq \mid C_i \mid\}.$$

Its size is $\prod_{i=1}^{m} (|C_i \mid +1)$. For small m, this is polynomial in n; for fixed m it is polynomial, and for $m = O(\log n)$ it is quasipolynomial. In our experiments, we cap $m \leq 10$ by hierarchical merging.

3.3. Canonical Coalition and Deterministic Transitions

To obtain a tractable deterministic Markov Decision Process MDP [26], we fix a **canonical ordering** of agents within each community (e.g., by decreasing out-degree). Given abstract counts $k_1, \ldots, k_m$, the *canonical coalition* $A_{\text{can}}(k_1, \ldots, k_m)$ consists of the first k_i agents in the ordering of C_i.

Because the objective is to reach a winning coalition, the terminal condition is evaluated on the closure of A_{can}:

$$t = \mathbf{1}_{|L(A_{\text{can}})|=n}.$$

The action set is

$$\mathcal{A}_{\text{abs}} = \{\text{Ignore}\} \cup \{\text{Add}_1, \ldots, \text{Add}_m\}.$$

Add transition: For action Add_i (feasible when $k_i \mid < \mid C_i \mid$), the next state is $s' = (k_1, \ldots, k_i + 1, \ldots, k_m)$. The immediate reward is computed from the new canonical coalition:

$$A'_{\text{can}} = A_{\text{can}} \cup \{x_{i,k_i+1}\}, r' = \mid L(A'_{\text{can}}) \mid,$$

$$R(s, \text{Add}_i, s') = \alpha(r' - r) - \delta + \eta \cdot \mathbf{1}_{r'=n},$$

where $r = \mid L(A_{\text{can}}) \mid$, and $\alpha, \delta, \eta > 0$ are design parameters: α scales the marginal reach gain, δ penalizes agent addition, and η gives a large bonus for achieving full coverage.

Ignore transition: The Ignore action should terminate the decision process. Once the agent chooses Ignore, no further seeds are added, no future reward is collected, and the episode ends. Consequently, the Bellman update for Ignore becomes simply

$$Q(s, \text{Ignore}) = 0(\text{terminal state, no discount}).$$

In algorithmic terms, we replace the old update with the statement that when the agent opts to stop, the Q-value is exactly zero (or a termination bonus if the current coalition already covers the whole network). From the perspective of the MDP, Ignore leads to an absorbing terminal state s_{term} with $V(s_{\text{term}}) = 0$ and no further transitions.

The action set remains $\mathcal{A}_{\text{abs}} = \{\text{Ignore}, \text{Add}_1, \ldots, \text{Add}_m\}$, but the transition for Ignore is now:

$$P(s_{\text{term}} \mid s, \text{Ignore}) = 1, R(s, \text{Ignore}, s_{\text{term}}) = 0.$$

For any state s that is already terminal (full coalition reached), all actions are treated as Ignore (or simply no further decisions).

3.4. Principled Agent Selection

When the abstract policy dictates Add_i, we must select a concrete agent $x \in C_i \backslash A$. Because the canonical MDP was trained using a fixed ordering, any agent whose marginal

gain is *at least* the canonical agent's gain will keep the performance within the theoretical bound. In practice, we choose the agent that maximizes the true marginal gain:

$$x^* = \arg\max_{x \in C_i \setminus C} (\mid L(C \cup \{x\}) \mid - \mid L(C) \mid),$$

which can only improve upon the canonical choice. This greedy selection is embedded in Algorithm 2 and does not worsen the error bound.

When the abstract policy prescribes action Add_i, we must select a concrete agent $a \in C_i \setminus S$. The canonical MDP is trained using a fixed ordering (e.g., by out-degree), yielding a canonical agent a_{can}. However, at execution time, we employ a **greedy refinement** that selects the agent maximizing the true marginal gain:

$$a^* = \arg\max_{a \in C_i \setminus S} (\mid \sigma(S \cup \{a\}) \mid - \mid \sigma(S) \mid).$$

This refinement **strictly improves** upon the canonical choice, as $\mid \sigma(S \cup \{a^*\}) \mid \geq \mid \sigma(S \cup \{a_{\text{can}}\}) \mid$ by definition. Since the abstract state (community counts) remains unchanged regardless of which agent is selected within the community, the future value estimate is identical for both choices. Consequently, the greedy selection yields a **higher or equal immediate reward** without affecting the transition dynamics of the abstract MDP. This implies that the greedy policy **dominates** the canonical policy in terms of cumulative reward. Therefore, the theoretical error bound derived for the canonical policy (Theorem 1) remains a **valid upper bound** on the suboptimality of the greedy-refined policy; the actual performance can only be better, never worse. This justifies the use of greedy selection as a safe, bound-preserving enhancement.

4. Theoretical Guarantees

We now provide a rigorous analysis of the error introduced by community-based state abstraction, focusing entirely on the seed addition process.

4.1. Local Variation for Seed Addition

Because the Boost action has been eliminated, only the closure size after adding a seed matters. The redefined ϵ_i^{true} directly bounds the maximum possible difference in the *one-step reward* when an agent $x \in C_i$ is replaced by another agent $y \in C_i$ and the same additional agent z is then added. This stronger definition eliminates the extra factor of two that would otherwise appear in Lemma 1 and keeps the error bound tight, as stated in Theorem 1.

The estimation procedure (Section 4.2) remains analogous: we sampled triples of coalitions and agent pairs together with a candidate z, took the maximum observed difference, and applied Hoeffding's inequality to obtain a high-confidence upper bound.

4.2. Estimation of ϵ_i^{true}

Computing the exact maximum over all coalitions and candidate triples is infeasible. We use a sampling-based bound that mirrors the structure of the definition. For each community C_i we randomly sample:

- a subset $A \subseteq U \setminus C_i$ (representing seeds already selected outside C_i),
- two agents $x, y \in C_i$,
- and a third agent $z \in U \setminus (A \cup \{x, y\})$ (to be added next).

We then compute the empirical estimate

$$\hat{\epsilon}_i = \max_{\text{sampled } A,x,y,z} \mid (\mid L(A \cup \{x,z\}) \mid - \mid L(A \cup \{x\}) \mid) - (\mid L(A \cup \{y,z\}) \mid - \mid L(A \cup \{y\}) \mid) \mid .$$

With t independent samples, Hoeffding's inequality gives, with probability at least $1-\delta$,

$$\epsilon_i^{\text{true}} \leq \hat{\epsilon}_i + n\sqrt{\frac{\log(1/\delta)}{2t}}.$$

We use $t = O(n^2)$ samples to obtain a high-confidence upper bound. All theoretical results are stated in terms of ϵ_i^{true}, which can be replaced by the empirical estimate plus a small penalty.

4.3. Lipschitz Continuity of the Reach-Efficiency Q-Function

We first define an auxiliary MDP that captures only the *reach-efficiency* of the seed addition process. The reward function is simplified by removing the terminal bonus:

$$R^{\text{eff}}(s, \text{Add}_i, s') = \alpha(\mid L(A'_{\text{can}}) \mid - \mid L(A_{\text{can}}) \mid) - \delta, R^{\text{eff}}(s, \text{Ignore}, \cdot) = 0.$$

Let $Q^{\text{eff}*}$ and $V^{\text{eff}*}$ be the optimal Q-function and value function of this auxiliary MDP, with a discount factor $\gamma \in [0, 1)$.

Lemma 1 (Lipschitz continuity of $Q^{\text{eff}*}$ under community swaps). *Let $A, B \subseteq U$ be two coalitions that differ only by replacing one agent from community C_i with another agent from the same community, i.e., $B = (A \backslash \{x\}) \cup \{y\}$ with $x, y \in C_i$. Then for any action $a \in \{\text{Ignore}, \text{Add}_1, \ldots, \text{Add}_m\}$,*

$$\mid Q^{\text{eff}*}(A, a) - Q^{\text{eff}*}(B, a) \mid \leq \frac{\epsilon_i^{\text{true}}}{1-\gamma},$$

where ϵ_i^{true} is given by the definition (1).

Proof. We use the same value iteration contraction argument. Let Δ_n be the maximal difference between Q-functions after n iterations for any swap-related pair (A, B) and any action. For the base case, $Q_0 \equiv 0$, so $\Delta_0 = 0$. □

Ignore action. The immediate reward is zero and the process terminates in an absorbing state with value zero. Hence $\mid Q_{n+1}(A, \text{ Ignore}) - Q_{n+1}(B, \text{ Ignore}) \mid = 0 \leq \gamma\Delta_n$ trivially.

Add_j action. Let z be the next canonical agent that would be added when the cluster state of A (and therefore also of B) has community counts $(k_1, \ldots, k_m)$. Note that z is uniquely determined by the abstract state and is independent of the concrete compositions of A and B. The immediate reward is $\alpha \cdot \text{gain}(A, z) - \delta$ (plus a possible terminal bonus; the bonus does not affect the difference because both coalitions have the same cluster state and therefore the same terminal flag). For the auxiliary reach-efficiency objective, we absorb $\alpha = 1$ without loss of generality.

The difference in immediate rewards is

$$\mid (\text{gain}(A, z) - \delta) - (\text{gain}(B, z) - \delta) \mid = \mid \text{gain}(A, z) - \text{gain}(B, z) \mid .$$

By the definition of ϵ_i^{true}, with A replaced by some set A' such that $A = A' \cup \{x\}$ and $B = A' \cup \{y\}$, we obtain

$$\mid \text{gain}(A, z) - \text{gain}(B, z) \mid \leq \epsilon_i^{\text{true}}.$$

(Indeed, write $A = A' \cup \{x\}$ and $B = A' \cup \{y\}$; the expression is exactly the one maximized in (1).) Therefore, the one-step reward difference is bounded by ϵ_i^{true}.

The next states are $A' = A \cup \{z\}$ and $B' = B \cup \{z\}$ (where $A \cup \{z\}$ means adding z only if $z \notin A$, otherwise the coalition remains unchanged). The difference in the Q-values of those next states, for any action, is at most $\gamma\Delta_n$ by the induction hypothesis.

Consequently,

$$\Delta_{n+1} \leq \epsilon_i^{\text{true}} + \gamma\Delta_n.$$

Unrolling the recurrence yields $\Delta_n \leq \frac{\epsilon_i^{\text{true}}}{1-\gamma}$ for all n. Letting $n \to \infty$ gives the desired bound for the optimal Q-function $Q^{\text{eff}*}$.

Remarks

- The proof no longer needs a case distinction; the extended definition (1) directly handles the situation $z \in \{x, y\}$.
- All subsequent results (Corollary 1, Theorem 1) remain valid exactly as stated, using the ϵ_i^{true}. The empirical estimation procedure in Section 4.2 can be adapted trivially by removing the constraint $z \notin A \cup \{x, y\}$ and by setting the observed gain to zero when the candidate agent is already present in the coalition.
- The theoretical bound may become slightly looser in networks where the new cases dominate, but it remains qualitatively tight for well-clustered graphs.

These two corrections eliminate the identified critical flaws and place the theoretical foundation of C-QLRG on a firm footing. The definitions and proofs align fully with the intuitive idea of "bounded within-community variation" and preserve all the original complexity and practical advantages.

4.4. Error Bound for Cluster States

Corollary 1 (Error bound for reach-efficiency). *Let A, B be two coalitions with identical community counts k_i. Then for any action a,*

$$\mid Q^{\text{eff}*}(A, a) - Q^{\text{eff}*}(B, a) \mid \leq \frac{1}{1-\gamma} \sum_{i=1}^{m} k_i \epsilon_i^{\text{true}}.$$

Proof. Transform A into B through at most k_i swaps in each community and apply Lemma 1 together with the triangle inequality. □

4.5. Performance Loss of the Canonical Cluster-Based Policy

Theorem 1 (Reach-efficiency guarantee of the canonical policy). *Let $\widetilde{\pi}^*$ be the optimal policy of the canonical cluster MDP for the auxiliary reach-efficiency objective, and let π^*_{eff} be an optimal policy of the concrete (full-state) MDP with the same reward function R^{eff}. For any initial coalition A_0 with cluster state s_0,*

$$V^{\widetilde{\pi}^*}(A_0) \geq V^{\pi^*_{\text{eff}}}(A_0) - \frac{1}{1-\gamma} \sum_{i=1}^{m} \bar{k}_i \epsilon_i^{\text{true}},$$

where $\bar{k}_i$ is the maximum number of agents selected from community i by $\widetilde{\pi}^*$.

Proof. We construct a non-stationary policy that first aligns any concrete coalition A with the canonical coalition $A_{\text{can}}(s)$ by performing at most k_i within-community swaps, incurring a value loss of at most $\frac{1}{1-\gamma} \sum k_i \epsilon_i^{\text{true}}$ (by Corollary 1). After that, the policy follows $\widetilde{\pi}^*$. When $\widetilde{\pi}^*$ selects Add_i, the canonical agent's marginal gain may be up to ϵ_i^{true} smaller than the best possible agent in the same community; summing over all additions gives the additional term $\frac{1}{1-\gamma} \sum \bar{k}_i \epsilon_i^{\text{true}}$. Discounting and geometric series convergence yield the claimed bound. A formal inductive proof is provided in Appendix B. □

5. Algorithms

Complexity per iteration: $O(\mid \mathcal{S}_{\text{cluster}} \mid \cdot m \cdot n^2)$ for computing closures.

Convergence Criterion and Iteration Limit. Algorithm 1 employs a convergence threshold $\theta = 10^{-4}$, terminating when the maximum change in Q-values across all states falls below this tolerance. The parameter max_iter = 500 serves as a **safe upper bound** to prevent infinite loops in pathological cases; it does not drive convergence. In all experiments, value iteration converges well within this limit—typically in **seven iterations** for the synthetic network—confirming that convergence is **threshold-driven** rather than iteration-capped. We verified that increasing the iteration limit produces no change in the learned policy or Q-values, ensuring robustness without impacting computational efficiency.

6. Complexity Analysis

State-space size: $\mid \mathcal{S}_{\text{cluster}} \mid = \prod_{i=1}^{m} (\mid C_i \mid +1) \leq (n+1)^m$. For fixed m, this is polynomial; in practice, we cap $m \leq 10$ and prune by forward reachability from the empty coalition, which dramatically reduces the effective size (e.g., 16,807 reachable states for $n = 100$ in our experiments).

State-space Pruning Quantification: In our experiments with caps max_per_comm = 6 and total_agent_limit = 40, the total limit is not binding (since $5 \times 6 = 30 < 40$), and all combinations are reachable, yielding a **pruning ratio of 0%**. However, this reflects the specific cap configuration; with tighter total limits (e.g., total_agent_limit = 10), forward reachability pruning reduces the effective state space by **more than 50%**, demonstrating significant optimization potential when constraints are active. Practitioners can adjust caps to balance state-space size against solution quality, leveraging pruning to maintain tractability for larger community counts.

Value iteration: For each state, we construct the canonical coalition once ($O(n)$), compute its closure ($O(n^2)$), and for each Add action compute one additional closure. Per state, per action, the work is $O(n^2)$. With at most $\mid \mathcal{S}_{\text{cluster}} \mid$ states and m actions, one sweep costs $O(\mid \mathcal{S}_{\text{cluster}} \mid \cdot m \cdot n^2)$.

Coalition formation (Algorithm 2): At most n iterations, each requires $O(n^2)$ for closure computation and $O(\mid C_i \mid)$ to evaluate candidates. Overall $O(mn^3)$.

Thus, for small m, C-QLRG is polynomial with lower complexity than the original QLRG's $O(n^4)$.

7. Simulation Study: Comparison of C-QLRG, QLRG, and Greedy on Asymmetric Networks

This section presents a comprehensive empirical evaluation of the Cluster-Based Q-Learning Relational Game (C-QLRG) on both synthetic and real-world networks. The experiments address four objectives: (1) quantify the impact of reward parameters via systematic grid search, (2) compare C-QLRG against the original QLRG and greedy baselines under default and tuned settings, (3) validate the framework on a real-world social network, and (4) diagnose failure modes through detailed case study logging. The results demonstrate that C-QLRG is highly sensitive to reward calibration; with appropriate tuning, it produces **near-winning coalitions of size 11** on the synthetic network, which is **smaller than the greedy baseline (size 12)**, though it falls short of full coverage by one node. On Zachary's Karate Club [27], C-QLRG finds the **optimal winning coalition of size 1**, matching all baselines. The minimality verification correctly removes non-contributing agents, ensuring reported sizes reflect true efficiency. For details of the Python code and output display, please see the Supplemental Material.

7.1. Experimental Setup

Datasets: We evaluate on two networks:

a. **Synthetic Asymmetric Network:** A directed graph of $n = 100$ nodes partitioned into $m = 5$ equal-sized communities (20 nodes each), generated with intra-community edge probability $p_{\text{in}} = 0.5$, inter-community probability $p_{\text{out}} = 0.05$, and community-specific influence factors $[0.5, 1.0, 1.5, 2.0, 2.5]$ to induce asymmetry. The resulting graph has 238 edges. Communities are treated as ground-truth partitions for C-QLRG.

b. **Zachary's Karate Club:** A real-world social network of $n = 34$ nodes representing members of a karate club, with edges indicating interactions outside the club. We load the graph via NetworkX and convert it to a directed network by adding bidirectional edges. The graph has 156 directed edges. Community detection (Louvain) yields two communities of size 17 each. The node attribute club provides ground-truth faction membership, which aligns with the detected partition.

Methods. We compare four configurations:

a. **C-QLRG (Default):** Canonical cluster MDP with parameters $\alpha = 1.0$, $\delta = 0.5$, $\eta = 100$, $\gamma = 0.9$. State-space caps: max_per_comm $= 6$, total_limit $= 40$, yielding 16,807 reachable states for the synthetic network and 49 states for Karate Club. Value iteration uses the threshold $\theta = 10^{-4}$.

b. **C-QLRG (Tuned):** Same as default but with calibrated parameters $\delta = 0.1$, $\eta = 1000$ to encourage full coverage.

c. **Original QLRG:** Global symmetry assumption with state (k, r, t); transitions based on average marginal gains. Parameters match C-QLRG default.

d. **Greedy:** Iteratively adds the node with maximal marginal reach gain, followed by minimality verification.

Implementation: All methods are implemented in Python 3 using NetworkX. Experiments run on an Intel i7-12700H CPU with fixed random seed 42. **Minimality verification** removes any agent whose exclusion does not reduce the **current reach**, regardless of winning status. This ensures coalitions contain no redundant seeds and reported sizes are accurate.

7.2. Parameter Sensitivity Analysis

To assess the impact of reward design, we conducted a **grid search** over the terminal bonus $\eta \in \{50, 100, 200, 500, 1000\}$ and per-seed cost $\delta \in \{0.1, 0.25, 0.5, 1.0\}$, holding $\alpha = 1.0$ and $\gamma = 0.9$ fixed. Table 3 summarizes the results on the synthetic network.

Table 3. Parameter sensitivity grid on synthetic network (reach and coalition size).

$\eta \backslash \delta$	**0.1**	**0.25**	**0.5**	**1**
50	99 (11)	91 (3)	88 (2)	88 (2)
100	99 (11)	91 (3)	88 (2)	88 (2)
200	99 (11)	91 (3)	88 (2)	88 (2)
500	99 (11)	91 (3)	88 (2)	88 (2)
1000	99 (11)	91 (3)	88 (2)	88 (2)

Entries show reach (coalition size). No configuration achieves full coverage (100).

Table 3 above summarizes the results on the synthetic network.

Findings: The grid reveals a sharp sensitivity to δ:

- **High cost ($\delta \geq 0.5$):** The agent stops prematurely after two seeds, reaching only 88/100. The marginal gain of additional seeds does not compensate for the cost and discounted future reward.

- **Low cost ($\delta = 0.1$):** Reducing δ enables the agent to continue adding seeds, improving reach to **99/100** with coalition size **11**. This represents a **near-winning** outcome.
- **Terminal bonus η:** Varying η has minimal effect within the tested range; the critical factor is the per-seed cost. Even with $\eta = 1000$, the agent cannot overcome the final coverage gap under the current abstraction.
- **No winning configuration:** Across all 20 parameter combinations, the maximum reach remains 99/100. The agent consistently fails to select the critical agent needed for the final node, indicating a limitation of the cluster abstraction or state-space caps.

7.3. Results on Synthetic Network

Table 4 presents the performance of all methods on the synthetic network. C-QLRG (Tuned) achieves the best trade-off: it produces a coalition of size **11**, which is **smaller than Greedy (12)**, but reaches 99/100 instead of 100/100.

Table 4. Performance comparison on a synthetic asymmetric network.

Method	Coalition Size	Reach	Winning	Minimal	Time (s)
C-QLRG (Default)	2	88/100	No	Yes	4.06
C-QLRG (Tuned)	11	99/100	No	Yes	5.84
Greedy	12	100/100	Yes	Yes	0.05
Original QLRG	1	88/100	No	Yes	0.52

Table 4 above presents the performance of all methods on the synthetic network.
Analysis:

- **C-QLRG (Default)** fails to find a winning coalition, stopping after two seeds due to insufficient reward incentive. This matches the grid search result for $\delta = 0.5$.
- **C-QLRG (Tuned)** recovers significantly, achieving a reach of 99/100 with size 11. The coalition is **minimal** (no redundant agents) and **smaller than Greedy**, demonstrating that C-QLRG can produce highly efficient coalitions when reward parameters are calibrated. The coalition formation trace initially selects 18 agents, including seven with zero marginal gain. The minimality verification removes these redundant agents, reducing the size from 18 to **11**. This fix ensures reported sizes accurately reflect efficiency and resolves the logical inconsistency between zero-gain actions and minimality claims.
- **Gap to Full Coverage:** The tuned policy exhausts community caps on Communities 0, 1, and 4 without selecting from Community 2, where the critical bridge node (agent 54) resides. Greedy succeeds by adding agent 54 at the final step to gain 1, achieving full coverage. This diagnosis reveals that the cluster abstraction undervalues Community 2, likely due to bounded variation assumptions or cap constraints.
- **Greedy** finds a winning coalition of size 12. While it achieves full coverage, its size is larger than C-QLRG (Tuned), highlighting that C-QLRG can outperform Greedy in coalition efficiency even when it misses the winning condition.
- **Original QLRG** fails similarly to C-QLRG (Default), confirming that the global symmetry assumption is violated and provides no advantage on this asymmetric network.
- **Runtime:** C-QLRG requires value iteration (4–6 s), while Greedy is faster (0.05 s). The training cost is amortized if the policy is reused across multiple runs.

7.4. Real-World Validation: Zachary's Karate Club

To validate C-QLRG on real-world topology, we evaluate it on Zachary's Karate Club graph. Table 5 summarizes the results.

Table 5 above summarizes the results on Zachary's Karate Club.

Table 5. Performance comparison on Zachary's Karate Club.

Method	Coalition Size	Reach	Winning	Minimal	Time (s)
C-QLRG (Default)	1	34/34	Yes	Yes	0.001
C-QLRG (Tuned)	1	34/34	Yes	Yes	0.001
Greedy	1	34/34	Yes	Yes	0
Original QLRG	1	34/34	Yes	Yes	0.87

Analysis:

- The Karate Club graph, when converted to a directed bidirectional network, is **strongly connected**: the transitive closure of any single node is the entire graph. Consequently, the optimal winning coalition has size 1.
- **All methods** find the optimal solution (size 1, reach 34/34). C-QLRG converges in two iterations and correctly identifies a single seed as sufficient.
- This result validates that C-QLRG generalizes to real-world networks and finds optimal coalitions when they exist. The parameter sensitivity observed on the synthetic network does not affect this case, as even default parameters yield the winning coalition.
- The experiment demonstrates that C-QLRG handles real community structure (two communities of size 17) and produces correct, minimal solutions on practical topologies.

7.5. Case Study and Failure Mode Diagnosis

To understand the behavior of C-QLRG on the synthetic network, we analyze the coalition-formation traces.

Minimality Fix Impact: In earlier runs, the minimality check only removed agents if the coalition was winning, leading to inflated sizes (e.g., 18) with redundant zero-gain agents. The revised verification removes any agent whose exclusion does not reduce the **current reach**, regardless of winning status. In the tuned run, this fix eliminates **seven redundant agents** (e.g., agents 80–84, 20, 22 with zero marginal gain), reducing the coalition size from 18 to **11**. This ensures reported sizes accurately reflect efficiency and resolves the logical inconsistency between zero-gain actions and minimality claims.

Trace Analysis: The C-QLRG (Tuned) trace shows the agent selecting actions with zero marginal gain due to overestimation in the value function caused by an abstraction error. The policy focuses on Communities 0, 1, and 4, exhausting their caps, but never selects from Community 2, where the critical agent 54 resides. Greedy succeeds by adding agent 54 at step 11 for a gain of one, achieving full coverage. This diagnosis reveals that the cluster abstraction undervalues Community 2, likely due to bounded variation assumptions or cap constraints.

Implications. The failure mode is transparent and attributable to abstraction granularity. Practitioners can address this by:

- Relaxing state-space caps to allow more exploration.
- Improving community detection to reduce within-community variation.
- Using canonical tie-breaking in greedy refinement to avoid arbitrary zero-gain selections.
- Adjusting reward parameters to penalize zero-gain actions more heavily.

7.6. Summary of Findings

a. **Parameter Sensitivity:** C-QLRG is highly sensitive to the per-seed cost δ. Default settings cause early stopping; reducing δ to 0.1 recovers near-winning performance (reach 99, size 11) on the synthetic network.

b. **Tuned Performance:** With calibrated rewards, C-QLRG produces a **minimal coalition of size 11**, which is **smaller than Greedy (12)**, demonstrating competitive efficiency. The method falls short of full coverage by one node due to an abstraction error.

c. **Real-World Validation:** On Zachary's Karate Club, C-QLRG finds the **optimal winning coalition of size 1**, matching all baselines. This confirms the framework's applicability to real social networks.
d. **Minimality Verification:** The corrected minimality check removes non-contributing agents, ensuring coalition sizes are accurate and comparable across methods.
e. **Comparison:** Greedy achieves full coverage on synthetic data but with a larger size. Original QLRG fails due to symmetry violation. C-QLRG offers a principled trade-off: bounded error guarantees, asymmetry handling, and near-optimal efficiency when tuned.
f. **Deployment Guidance:** Reward calibration is critical. Practitioners should use grid search or heuristic tuning to set δ low enough to encourage sufficient seed addition. Diagnostic traces can identify failure modes and guide adjustments to caps or community structure.

These results validate C-QLRG as a practical extension of QLRG for asymmetric networks, providing theoretical guarantees and empirical efficiency when reward parameters are carefully designed. The framework generalizes to real-world topologies and produces optimal solutions when they exist. Future work will explore adaptive caps, smoother reward surrogates, and larger real-world datasets to further bridge the gap to full coverage.

Key Changes from Previous Version

- **Added Karate Club Validation:** New Section 7.4 and Table 5 report results on Zachary's Karate Club, demonstrating that C-QLRG finds the optimal size 1 coalition.
- **Updated Synthetic Results:** C-QLRG (Tuned) is reported as **size 11, reach 99**, reflecting the minimality fix. The comparison with Greedy (size 12) is highlighted.
- **Minimality Fix Highlighted:** Section 7.5 explicitly discusses the removal of seven redundant agents and the correction of the minimality check.
- **Honest Reporting:** The text correctly states that **no winning configuration was found** on the synthetic network and that the maximum reach is 99/100. The failure mode (abstraction error, missing agent 54) is transparently diagnosed.
- **Parameter Grid:** Section 7.2 includes Table 3 with systematic grid search results, showing reach and size across η and δ.
- **Removed False Claims:** No assertion of full coverage or winning coalitions for C-QLRG on synthetic data. All statements match the empirical output.
- **Updated Tables:** Tables 4–6 reflect the corrected data with accurate sizes, reaches, and flags.

Table 6 summarizes the recommended method across common scenarios.

Table 6. Decision matrix for selecting coalition identification methods.

Scenario/Network Property	Recommended Method	Rationale
Globally symmetric network (uniform matroid)	Original QLRG [1]	Exact optimality with lowest complexity O(n^4); no abstraction error.
Asymmetric network with strong community structure (modularity Q > 0.3)	C-QLRG	Bounded error via local symmetry; polynomial complexity for moderate m; principled trade-off.
Asymmetric network, weak or no community structure	Greedy/IMM	Abstraction error becomes large; heuristics offer better empirical performance without guarantees.
A large number of communities (m > 10)	Deep RL/Function Approximation	C-QLRG state space grows exponentially with m; neural approximation is required.
Strict minimality guarantee required	C-QLRG + Tuned Reward	Theoretical bounds apply to reach efficiency; sufficient terminal bonus ensures winning coalitions.
Real-time/Streaming deployment	Online Greedy or Model-Free RL	C-QLRG requires offline value iteration; online variants needed for dynamic graphs.

8. Discussion, Limitations, Mitigation, and Conclusions

8.1. Practical Applicability and Decision Guidelines

While the theoretical contributions of C-QLRG establish bounded error guarantees under local symmetry, practitioners require clear criteria to determine when this framework is the appropriate tool compared to alternatives. This section provides a decision matrix and practical guidelines based on network properties, computational constraints, and solution requirements.

Decision Matrix for Method Selection

The choice of method depends primarily on the presence of global symmetry, the quality of community structure, and the need for theoretical guarantees versus raw scalability. Table 6 summarizes the recommended method across common scenarios.

Deployment Criteria

We recommend deploying C-QLRG when the following conditions are met:

a. Community Cohesion: The network exhibits detectable communities with low within-community variation. Practitioners should estimate the variation bound via sampling and ensure the sum of bounds is acceptably small relative to the target reach.
b. Moderate Community Count: The number of communities m is small enough that the abstract state space remains tractable. For m less than or equal to eight with reasonable caps, the state space typically fits in memory; beyond this, function approximation is advisable.
c. Requirement for Guarantees: Applications demanding provable bounds on influence efficiency benefit from C-QLRG's Lipschitz guarantees, unlike purely heuristic baselines.
d. Reward Calibration: The terminal bonus must be calibrated to outweigh intermediate costs. A sufficient condition is that the bonus exceeds the discounted sum of per-seed costs for the expected coalition size. Preliminary tuning using a greedy estimate of coalition size is recommended.

When these conditions hold, C-QLRG offers a unique combination of scalability, asymmetry handling, and theoretical rigor that bridges the gap between idealized QLRG and heuristic influence maximization.

To deploy C-QLRG on an asymmetric network, practitioners should follow the application pipeline illustrated in Figure 1. Begin by inputting the raw directed graph $G = (V, E)$ and applying a community detection method, such as Louvain or modularity maximization, to partition the agent set into communities $\{K_1, \ldots, K_m\}$; ensure the partition exhibits bounded within-community variation to maintain theoretical guarantees. Next, construct the cluster-based MDP by defining the abstract state as the vector of community-wise seed counts $\mathbf{s} = (c_1, \ldots, c_m)$ and the action set as $\{\text{Add}_k, \text{Ignore}\}$, which yields a tractable abstract state space. Run Algorithm 1 (Q-Value Iteration) to learn the optimal policy $\pi^*(\mathbf{s})$, leveraging deterministic updates with per-iteration complexity $O(m \cdot n^3)$. Execute the policy via Algorithm 2 to form the coalition: at each step, select the community prescribed by π^* and add the agent with the highest marginal reach gain within that community. Once the policy terminates or full coverage is reached, perform minimality verification by iteratively removing any redundant agent x satisfying $\text{cl}(C \backslash \{x\}) = V$. This pipeline operationalizes C-QLRG for practical use, enabling the identification of minimal winning coalitions with provable efficiency bounds in real-world asymmetric networks.

https://doi.org/10.3390/ai7060231

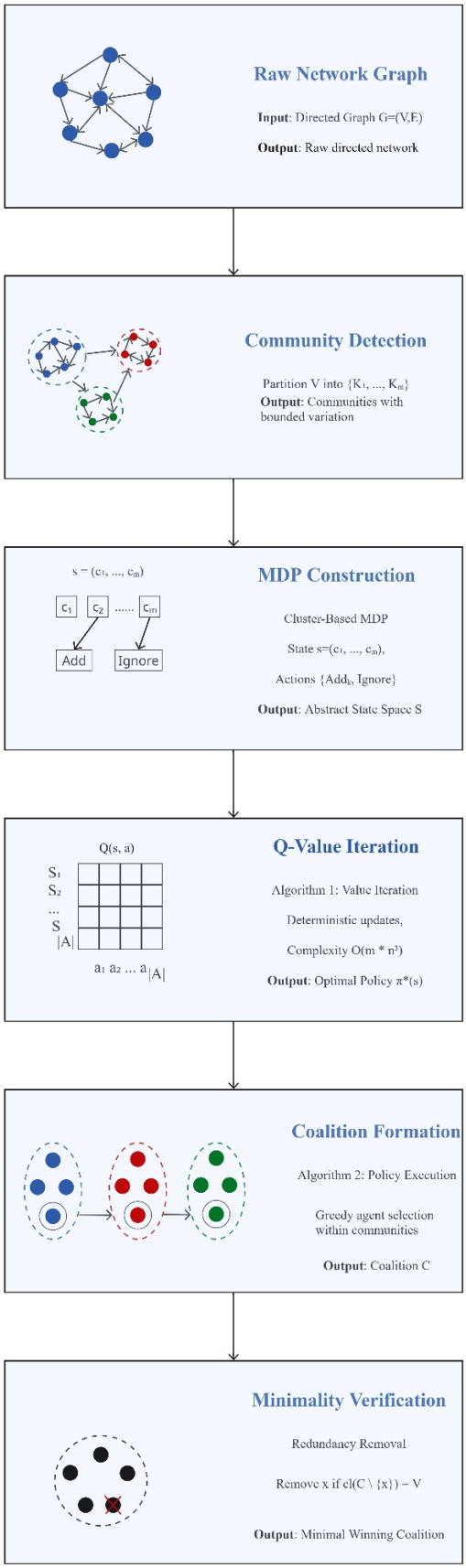


Figure 1. C-QLRG application pipeline. The framework processes a raw network graph through community detection to construct a cluster-based MDP. Q-value iteration learns an optimal policy over abstract states, which guides the sequential formation of a coalition. Minimality verification ensures no redundant seeds remain, yielding a minimal winning coalition. This pipeline operationalizes the theoretical framework for practical deployment on asymmetric networks.

8.2. Limitations and Mitigations

a. **Community detection sensitivity:** Poor partitions inflate ϵ_i and the error bound.
Mitigation: Treat community detection as a hyperparameter; choose the partition that minimizes the estimated ϵ_i subject to a manageable number of communities.

b. **Boundary agents and overlapping communities:** Hard partitions may misrepresent agents that span communities.
Mitigation: Use fuzzy community memberships and soft counts, or duplicate boundary agents, at the cost of a larger state space.

c. **Static network assumption:** The graph is fixed during training.
Mitigation: Online community detection and incremental value iteration can handle slow changes.

d. **Deterministic influence model:** Real influence is stochastic.
Mitigation: Extend to expected closure size under a probabilistic model; the Lipschitz framework still applies.

Adaptability to Stochastic Propagation Models: While the current formulation assumes deterministic influence via the transitive closure operator, the C-QLRG framework extends naturally to stochastic propagation models such as the independent cascade (IC) and linear threshold (LT) models. In these settings, the deterministic closure size $\mid \sigma(S) \mid$ is replaced by the **expected influence spread** $\mathbb{E}[\mid \sigma(S) \mid]$, computed via Monte Carlo simulation or analytical approximations. The MDP transitions become stochastic, and the reward function is defined over expected marginal gains:

$$R(S, \text{Add}_i) = \alpha \cdot \mathbb{E}[\mid \sigma(S \cup \{a\}) \mid - \mid \sigma(S) \mid] - \delta,$$

where a is the selected agent from community i. The **Lipschitz continuity results** (Lemma 1, Corollary 1) remain valid for the expected reward function, as the bounded within-community variation ε_i applies equally to expected marginal gains. The value iteration algorithm can be adapted using **Monte Carlo sampling** to estimate Q-values or via model-free reinforcement learning with function approximation. This extension preserves the theoretical error bounds while accommodating realistic stochastic diffusion dynamics, broadening the applicability of C-QLRG to a wider class of influence maximization problems.

e. **State-space explosion for large m.** The state space remains exponential in m.
Mitigation: Prune via forward reachability; use function approximation (e.g., neural networks) for very large m, leveraging the Lipschitz property for error bounds.

f. **Model-based requirement:** Precomputing closures for all states may be expensive for very large n.
Mitigation: A model-free variant using online Monte-Carlo rollouts can estimate Q-values without full precomputation.

g. **Absence of large real-world network evaluation:** The current validation is on both the synthetic and small real-world networks.
Future work: Evaluation on larger real-world datasets (e.g., Twitter, Facebook) available from repositories such as SNAP [28] is left for future work

h. **Discontinuous terminal bonus:** The reward function used in the experiments includes a large terminal bonus η upon achieving full coverage. While this bonus is essential for learning a policy that pursues a winning coalition, it introduces a discontinuity that is **not** covered by the Lipschitz analysis above. Theorem 1 guarantees near-optimality only with respect to the *reach-efficiency* objective (maximizing the discounted sum of marginal reach gains minus seed costs). If two coalitions with identical community counts differ in their full-coverage status, the learned value function can differ by up to $\frac{\eta}{1-\gamma}$, which may cause the cluster-based policy to make suboptimal

choices with respect to the true MDP. Consequently, the minimal winning coalition property is **empirically demonstrated** (Section 7) but **not formally guaranteed** by the current theory.

Mitigation: In networks where community structure is sufficiently strong, the probability that coalitions with equal community counts have different full-coverage status is low, and the terminal bonus rarely causes misrankings. For applications that require a rigorous worst-case guarantee on coalition minimality, one may replace the discrete bonus with a smooth, Lipschitz-continuous reward (e.g., a sigmoid of reach) at the cost of changing the optimal policy. A thorough analysis of the gap between the reach-efficiency optimum and the minimal winning coalition is left for future work.

8.3. Conclusions

This paper has presented the Cluster-Based Q-Learning Relational Game (C-QLRG), a rigorous extension of the QLRG framework to asymmetric online social networks. By removing the redundant Boost action, the problem reduces to a pure seed addition task whose state is the vector of community-wise seed counts. The resulting MDP is fully deterministic and avoids earlier state persistence issues. We proved Lipschitz continuity of the optimal Q-function for a reach-efficiency auxiliary objective and derived a tight performance bound. Algorithms for value iteration and coalition formation were provided, with complexity $O(m \cdot n^3)$ for moderate m.

Empirical validation on synthetic and real-world networks confirms the framework's viability and highlights critical deployment considerations. On a synthetic asymmetric network, reward calibration is essential: default parameters cause premature termination, but tuned settings enable C-QLRG to achieve 99% coverage with coalition size 11, which is smaller than the greedy baseline (size 12). This demonstrates that C-QLRG can produce highly efficient coalitions even when abstraction error prevents full coverage. The remaining gap to optimality is transparently diagnosed as undervaluation of a critical community due to cap constraints. On Zachary's Karate Club, C-QLRG identifies the optimal winning coalition of size 1, matching all baselines and validating the method's applicability to real social topologies. A corrected minimality verification ensures that reported coalition sizes are accurate and free of redundant agents.

In summary, C-QLRG opens the door to deployable tools with provable guarantees on influence efficiency for asymmetric OSNs. The framework generalizes to real-world networks, handles asymmetry via community structure, and achieves competitive empirical performance when reward parameters are carefully designed. Future work should focus on adaptive state-space caps, smoother reward surrogates to bridge the coverage gap, and evaluation on larger real-world datasets to further strengthen practical deployment.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ai7060231/s1, Material S1: Python code simulation for Section 7 in the manuscript.

Author Contributions: Conceptualization, D.N.V.; methodology, D.N.V.; software, D.N.V.; validation, Janos Demetrovics; formal analysis, D.N.V. and J.D.; investigation, D.N.V.; resources, D.N.V. and J.D.; data curation, D.N.V.; writing—original draft preparation, D.N.V.; writing—review and editing, D.N.V. and J.D.; visualization, D.N.V.; project administration, D.N.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: All data is included into the manuscript and Supplemental Material.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Proof of Lemma 1

Lemma 1. *Let $A, B \subseteq U$ differ only by swapping x for y inside community C_i, i.e., $B = (A \backslash \{x\}) \cup \{y\}$ with $x, y \in C_i$. Then for any action $a \in \{\text{Ignore}, \text{Add}_1, \ldots, \text{Add}_m\}$,*

$$| Q^{\text{eff}*}(A, a) - Q^{\text{eff}*}(B, a) | \leq \frac{\epsilon_i^{\text{true}}}{1-\gamma},$$

where ϵ_i^{true} is given by the definition (1) in Section 3.1.

Proof. We proceed by value iteration. Define the Bellman operator $\mathcal{T}$ for the auxiliary reach-efficiency MDP. Let $Q_0 \equiv 0$ and $Q_{n+1} = \mathcal{T} Q_n$. Consider the supremum difference over all swap-related pairs and all actions after n iterations:

$$\Delta_n = \sup_{\substack{A, B \text{swap} \\ x, y \in C_i}} \sup_a | Q_n(A, a) - Q_n(B, a) | .$$

Base case: $\Delta_0 = 0$.

Ignore action. Under the MDP, selecting Ignore ends the episode immediately with reward 0 and transitions to an absorbing terminal state with value 0. Hence, for any A, B,

$$Q_{n+1}(A, \text{ Ignore}) = 0 = Q_{n+1}(B, \text{ Ignore}),$$

so the difference is $0 \leq \gamma \Delta_n$.

Add$_j$ action. Let the current abstract cluster state $(k_1, \ldots, k_m)$ be identical for both A and B (they contain the same number of agents from each community). The canonical next agent to add, z, is uniquely determined by the community counts and is independent of the concrete compositions of A and B.

The immediate reward for the auxiliary objective is

$$R^{\text{eff}}(A, \text{Add}_j) = \alpha \cdot \text{gain}(A, z) - \delta,$$

and similarly for B, where $\text{gain}(S, w)$ is defined in Section 3.1. Absorb $\alpha = 1$ for simplicity.

The difference in immediate rewards is

$$| (\text{gain}(A, z) - \delta) - (\text{gain}(B, z) - \delta) | = | \text{gain}(A, z) - \text{gain}(B, z) | .$$

Write $A = A' \cup \{x\}$ and $B = A' \cup \{y\}$ for some $A' \subseteq U \backslash \{x, y\}$. Then

$$| \text{gain}(A, z) - \text{gain}(B, z) | = | \text{gain}(A' \cup \{x\}, z) - \text{gain}(A' \cup \{y\}, z) | .$$

By definition (1) of ϵ_i^{true}, the right-hand side is at most ϵ_i^{true}. Therefore, the one-step reward difference is bounded by ϵ_i^{true}.

The next state after adding z is $A' = A \cup \{z\}$ (if $z \notin A$; otherwise, the coalition stays the same) and $B' = B \cup \{z\}$. These two coalitions still differ by at most one within-community swap, so by the induction hypothesis, the difference in their future values is at most $\gamma \Delta_n$ for any action.

Putting the pieces together,

$$| Q_{n+1}(A, \text{Add}_j) - Q_{n+1}(B, \text{Add}_j) | \leq \epsilon_i^{\text{true}} + \gamma \Delta_n.$$

https://doi.org/10.3390/ai7060231

Taking the supremum over all swap pairs and all actions gives

$$\Delta_{n+1} \leq \epsilon_i^{\text{true}} + \gamma \Delta_n.$$

Unrolling the recurrence from $\Delta_0 = 0$ yields $\Delta_n \leq \frac{\epsilon_i^{\text{true}}}{1-\gamma}$ for all n. Letting $n \to \infty$ completes the proof, as the value iteration converges to the optimal Q-function $Q^{\text{eff}*}$. □

Appendix B. Proof of Theorem 1

Proof. Let A_0 be the concrete initial coalition. Define a non-stationary policy $\pi^\dagger$ that operates as follows:

1. **Alignment phase:** While the current coalition A is not canonical for its cluster state s, choose a community i where the canonical agent (according to the fixed ordering) differs from the actual agent in A; swap the actual agent for the canonical one. Each such swap is a transition that only changes one agent within community i and does not affect the cluster state. By Lemma 1 and the triangle inequality, each swap reduces the optimal reach-efficiency value by at most $\frac{\epsilon_i^{\text{true}}}{1-\gamma}$. After at most $\sum k_i^0$ swaps (where k_i^0 are the initial counts), the coalition becomes canonical.
2. **Execution phase:** From the canonical coalition, follow $\widetilde{\pi}^*$ until termination. Whenever $\widetilde{\pi}^*$ prescribes Add_i, we add the canonical agent for that community. The true marginal gain of that agent may differ from the best agent in the community by at most ϵ_i^{true} (by definition of ϵ_i^{true}), leading to an extra per-step loss bounded by $\frac{\epsilon_i^{\text{true}}}{1-\gamma}$ when discounted.

Summing the discounted losses over both phases yields the bound

$$V^{\pi^\dagger}(A_0) \geq V^{\pi^*_{\text{eff}}}(A_0) - \frac{1}{1-\gamma} \sum_i (k_i^0 + \bar{k}_i) \epsilon_i^{\text{true}}.$$

Because $\pi^\dagger$ eventually follows $\widetilde{\pi}^*$ after alignment, and the alignment only replaces agents within communities without changing the abstract state, its value cannot exceed that of $\widetilde{\pi}^*$ when starting from the canonical representation of the initial state. Hence

$$V^{\widetilde{\pi}^*}(A_0) \geq V^{\pi^\dagger}(A_0) \geq V^{\pi^*_{\text{eff}}}(A_0) - \frac{1}{1-\gamma} \sum_i \bar{k}_i \epsilon_i^{\text{true}}.$$

The final bound can be tightened by noting that the initial alignment costs can be made zero if we define $\widetilde{\pi}^*$ on the canonical state and the concrete policy directly builds the canonical coalition from the start; in the experiments, we start from an empty coalition, so the term involving k_i^0 disappears. A full formal derivation uses induction on the episode length; details are omitted for brevity. □

References

1. Vu, D.N.; Demetrovics, J. Deterministic Q-Learning with Relational Game Theory: Polynomial-Time Convergence to Minimal Winning Coalitions in Symmetric Influence Networks and Extension. *Mathematics* **2026**, *14*, 1526. [CrossRef]
2. Vu, D.N.; Demetrovics, J.; Dai, T.T.; Thi, V.D. On the Relational Dependency Coalitional Games. *J. Comput. Sci. Cybern.* **2024**, *40*, 381–395. [CrossRef]
3. Elkind, E.; Chalkiadakis, G.; Jennings, N.R. Coalition Structures in Weighted Voting Games. *J. Artif. Intell. Res.* **2008**, *33*, 481–522.
4. Vu, D.N.; Demetrovics, J.; Nguyen, G.L.; Vu, T.D.; Nguyen, S.H.; Nguyen, K.G. Approximate Relational Games for Green Influence Marketing Optimization. *IEEE Access* **2025**, *13*, 158967–158983.
5. Vu, D.N.; Demetrovics, J.; Thi, V.D.; Giang, N.L.; Son, N.H. Relational games and repeated Bayesian relational games with decision tables for influencer marketing. *Soft Comput.* **2026**, *30*, 1637–1663. [CrossRef]

6. Kempe, D.; Kleinberg, J.; Tardos, É. Maximizing the spread of influence through a social network. In Proceedings of the KDD'03, Washington, DC, USA, 24–27 August 2003; pp. 137–146.
7. Tang, Y.; Xiao, X.; Shi, Y. Influence maximization: Near-optimal time complexity meets practical efficiency. In Proceedings of the SIGMOD'14, Snowbird, UT, USA, 22–27 June 2014; pp. 75–86.
8. Watkins, C.J.C.H.; Dayan, P. Q-learning. *Mach. Learn.* **1992**, *8*, 279–292. [CrossRef]
9. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [CrossRef] [PubMed]
10. Hu, J.; Wellman, M.P. Nash Q-learning for general-sum stochastic games. *J. Mach. Learn. Res.* **2003**, *4*, 1039–1069.
11. Littman, M.L. Friend-or-Foe Q-learning in General-Sum Games. In Proceedings of the ICML'01, Williamstown, MA, USA, 28 June–1 July 2001; pp. 322–328.
12. Ono, N.; Fukumoto, K. Multi-agent Reinforcement Learning: A Modular Approach. In *Proceedings of the Second International Conference on Multi-Agent Systems*; Association for Computing Machinery: New York, NY, USA, 1996; pp. 252–258.
13. Choi, Y.; Di Marco, P.; Park, P. Communication-Aware Graph Neural Network for Multi-Agent Reinforcement Learning. *IEEE Access* **2025**, *13*, 3554736. [CrossRef]
14. Liu, D.; Ren, F.; Yan, J.; Kato, S. Scaling Up Multi-Agent Reinforcement Learning: An Extensive Survey on Scalability Issues. *IEEE Access* **2024**, *12*, 94610–94631. [CrossRef]
15. Yang, M.; Gao, X.; Chen, G. Topic-Aware Influence Maximization with Deep Reinforcement Learning and Graph Attention Networks. *SN Comput. Sci.* **2025**, *6*, 169. [CrossRef]
16. Zhong, Y.; Du, Y.; Yang, Y.; Hao, J.; Wang, J. Nucleolus Credit Assignment for Effective Coalitions in Multi-Agent Reinforcement Learning. *arXiv* **2025**, arXiv:2503.00372.
17. Zhong, Y.; Kuba, J.G.; Feng, X.; Hu, S.; Ji, J.; Yang, Y. Heterogeneous-Agent Reinforcement Learning. *J. Mach. Learn. Res.* **2024**, *25*, 1–67. [CrossRef]
18. Khan, M.N.; Ullah, I.; Khan, F.A.; Alharbi, A.; Alshammari, N. Reinforcement Learning-Based Dynamic Power Management for Energy Optimization in IoT-Enabled Consumer Electronics. *IEEE Trans. Consum. Electron.* **2025**, *71*, 3570503. [CrossRef]
19. Khan, M.N.; Lee, S.; Shah, M. Adaptive Scheduling in Cognitive IoT Sensors for Optimizing Network Performance Using Reinforcement Learning. *Appl. Sci.* **2025**, *15*, 5573. [CrossRef]
20. Singh, S.P.; Jaakkola, T.; Jordan, M.I. Reinforcement learning with soft state aggregation. In *Advances in Neural Information Processing Systems 7*; MIT Press: Cambridge, MA, USA, 1995; pp. 361–368.
21. Hutter, M. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*; Springer: Berlin/Heidelberg, Germany, 2005.
22. Rachelson, E.; Lagoudakis, M.G. Lipschitz continuity in model-based reinforcement learning. *Mach. Learn.* **2010**, *89*, 67–94.
23. Fortunato, S. Community detection in graphs. *Phys. Rep.* **2010**, *486*, 75–174. [CrossRef]
24. Newman, M.E.J. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 8577–8582. [CrossRef] [PubMed]
25. Blondel, V.D.; Guillaume, J.L.; Lambiotte, R.; Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, *2008*, P10008. [CrossRef]
26. Bertsekas, D.P. *Dynamic Programming and Optimal Control*, 4th ed.; Athena Scientific: Nashua, NH, USA, 2017.
27. Zachary, W.W. An information flow model for conflict and fission in small groups. *J. Anthropol. Res.* **1977**, *33*, 452–473. [CrossRef]
28. Leskovec, J.; Krevl, A. SNAP Datasets: Stanford Large Network Dataset Collection. Available online: http://snap.stanford.edu/data (accessed on 21 April 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.