Learning Mean Field Control on Sparse Graphs

Christian Fabian [1 2] **Kai Cui** [1] **Heinz Koeppl** [1 2]

Abstract

Large agent networks are abundant in applications and nature and pose difficult challenges in the field of multi-agent reinforcement learning (MARL) due to their computational and theoretical complexity. While graphon mean field games and their extensions provide efficient learning algorithms for dense and moderately sparse agent networks, the case of realistic sparser graphs remains largely unsolved. Thus, we propose a novel mean field control model inspired by local weak convergence to include sparse graphs such as power law networks with coefficients above two. Besides a theoretical analysis, we design scalable learning algorithms which apply to the challenging class of graph sequences with finite first moment. We compare our model and algorithms for various examples on synthetic and real world networks with mean field algorithms based on Lp graphons and graphexes. As it turns out, our approach outperforms existing methods in many examples and on various networks due to the special design aiming at an important, but so far hard to solve class of MARL problems.

1. Introduction

Despite the rapid developments in the field of multi-agent reinforcement learning (MARL) over the last years, systems with many agents remain hard to solve in general (Canese et al., 2021; Gronauer & Diepold, 2022). Mean field games (MFGs) (Caines et al., 2006; Lasry & Lions, 2007) and mean field control (MFC) (Andersson & Djehiche, 2011; Bensoussan et al., 2013) are a promising way to model large agent problems in a computationally tractable manner and to provide a solid theoretical framework at the same time. While MFGs consider competitive agent populations, the focus of MFC are cooperative scenarios where agents optimize a common goal. The idea of MFC and MFGs is to abstract large, homogeneous crowds of small agents into a single probability distribution, the *mean field* (MF). While MFC and MFGs have been used in various areas ranging from pedestrian flows (Bagagiolo et al., 2019; Achdou & Laurière, 2020) to finance (Carmona & Delarue, 2018; Carmona & Laurière, 2023) and oil production (Bauso et al., 2016), the assumption of indistinguishable agents is not fulfilled in many applications.

A particularly important class of MARL problems are those with many connected agents. Initially, these agent networks were modeled by combining the graph theoretical concept of graphons (Lovász, 2012) with MFGs, resulting in graphon MFGs (GMFGs) (Caines & Huang, 2019; 2021; Cui & Koeppl, 2022; Zhang et al., 2024) and graphon MFC (Hu et al., 2023b). Since GMFGs only model often unrealistic dense graphs, subsequently mean field models based on Lp graphons (Borgs et al., 2018b; 2019) and graphexes (Veitch & Roy, 2015; Caron & Fox, 2017; Borgs et al., 2018a) were developed, called LPGMFGs and GXMFGs, respectively (Fabian et al., 2023; 2024). While these models facilitate learning algorithms in moderately sparse networks, they exclude sparser topologies. Formally, (LP)GMFGs and GXMFGs are designed exclusively for graphs with expected average degree going to infinity which, for example, excludes power laws with a coefficient above two.

The learning literature contains various approaches to finding optimal behavior in MFGs and MFC, see Laurière et al. (2022a) for an overview. For example, Subramanian et al. (2022) develop a decentralized learning algorithm for MFGs where agents independently learn policies, while Guo et al. (2019; 2023) focus on Q-learning methods for general MFGs. Various MFC learning approaches exist (Ruthotto et al., 2020; Carmona et al., 2023; Gu et al., 2023), but we are aware of only one work by Hu et al. (2023b) which learns policies for MFC on dense networks, but not on sparse ones.

Many empirical networks of high practical relevance are considerably sparser than the topologies covered by (LP)GMFGs and GXMFGs. Examples of sparse empirical networks which at least to some extent follow power laws with coefficients between two and three include the internet (Vázquez et al., 2002), coauthorship graphs (Goh et al.,

---
[1]Department of Electrical Engineering and Information Technology, Technische Universität Darmstadt, Germany [2]Hessian Center for Artificial Intelligence (hessian.AI). Correspondence to: Heinz Koeppl <heinz.koeppl@tu-darmstadt.de>.

2002) and biological networks (Dorogovtsev & Mendes, 2002). These topologies are particularly challenging to analyze (Dorogovtsev et al., 2002), especially in combination with particle dynamics (Lacker et al., 2023). Most important in our context, the high fraction of low degree nodes in these networks renders classical mean field approximations and their LPGMFG and GXMFG extensions highly inaccurate.

Both LPGMFGs and GXMFGs assume that the average expected degree diverges to infinity to ensure that the neighborhoods of crucial agents are accurately approximated by a mean field. Since this assumption is inaccurate for various empirical networks, we require a different modeling and learning approach. To learn policies for such very sparse networks, we employ a suitable graph theoretical convergence principle, *local weak convergence* (Van Der Hofstad, 2024). Many established graph theoretical models like the configuration model (Bollobás, 1980), the Barabási-Albert model (Barabási & Albert, 1999) and the Chung-Lu model (Chung & Lu, 2006) can model graph sequences converging in the local weak sense. Most importantly in our context, this includes graph sequences with finite average expected degree and power law networks with a coefficient above two, which are neither covered by Lp graphons nor graphexes.

Leveraging local weak convergence, we formulate our new local weak mean field control (LWMFC) model. LWMFC provides a theoretically motivated framework for learning agent behavior in challenging large empirical networks where the average expected degree is finite, but the degree variance may diverge to infinity. On the algorithmic side, we provide a *two systems approximation* for LWMFC and corresponding learning algorithms to approximately learn optimal behaviour in these complex agent networks. Finally, we evaluate our novel LWMFC learning approach for multiple problems on synthetic and real-world networks and compare it to different existing methods mentioned above. Overall, our contributions can be summarized as:

- We introduce LWMFC to model large cooperative agent populations on very sparse graphs with finite expected average degree;

- We give a rigorous theoretical analysis and motivation for LWMFC;

- We provide a two systems approximation and scalable learning algorithms for LWMFC;

- We show the capabilities of our LWMFC learning approach on synthetic and real world networks for different examplary problems.

2. Locally Weak Converging Graphs

In the following, let $(G_N)_{N \in \mathbb{N}} = (V_N, E_N)_{N \in \mathbb{N}}$ be a growing sequence of random graphs where V_N denotes the vertex set and E_N is the edge set of the corresponding graph G_N. In this paper, we focus on growing graph sequences where the expected average degree remains finite in the limit while the degree variance may diverge to infinity. To formalize the properties of the graph sequences we are focusing on, we first require a suitable graph convergence concept. We choose *local weak convergence in probability* which means that local node neighborhoods converge to neighborhoods in a limiting model. The definition below states local weak convergence, for details see e.g. Lacker et al. (2023).

Definition 2.1 (Local weak convergence in probability). A sequence of finite graphs $(G_N)_N$ converges in probability in the local weak sense to a random element G of $\mathcal{G}^*$ if for all continuous and bounded functions $f : \mathcal{G}^* \to \mathbb{R}$

$$\lim_{N \to \infty} \frac{1}{N} \sum_{i \in [N]} f(C_{v_i}(G_N)) = \mathbb{E}[f(G)] \quad \text{in probability}$$

where $C_{v_i}(G_N)$ denotes the connected component of $v_i \in G_N$ with root v_i and $\mathcal{G}^*$ is the set of isomorphism classes of connected rooted graphs.

To obtain meaningful theoretical results and hence practical approximations for large graphs in the next sections, we focus on graph sequences converging in the local weak sense, which we formalize with the next assumption.

Assumption 2.2. The sequence $(G_N)_{N \in \mathbb{N}}$ converges in probability in the local weak sense to some random element G of $\mathcal{G}^*$.

The class of random graph sequences fulfilling Assumption 2.2 covers many famous graph theoretical frameworks which are frequently used in the literature. We will briefly discuss three particularly important types of these models, namely configuration models, preferential attachment models and Chung-Lu graphs. We point to Van Der Hofstad (2024) for an extensive introduction and theoretical details and for more random graph models converging in the local weak sense.

Configuration models. The configuration model (Bender & Canfield, 1978; Bollobás, 1980; Molloy & Reed, 1995; 1998) (CM) is arguably one of the most established random graph models. The basic mechanism of the CM is to start with a fixed and arbitrary degree sequence. Then, a multigraph is randomly generated with the prescribed degree distribution which means that the graph can contain self-edges and double edges between pairs of nodes.

The CM is known to converge under suitable and moderate assumptions in the local weak sense in probability (Van Der Hofstad, 2024, Theorem 4.1). However, the CM generates multigraphs instead of simple graphs and the number of multiedges increases drastically as the vertex degrees

increase (Bollobás, 1998). Consequently, the CM is suboptimal for generating simple graphs with a significant fraction of high degree nodes such as power law networks.

Preferential attachment models. To model random graphs with power law features, Barabási & Albert (1999) introduced the famous Barabási-Albert (BA) model. The original BA model was subsequently extended in various ways for different applications, see Piva et al. (2021) for an overview. The BA model depicts graphs that grow over time by adding new nodes and edges to the topology. Since new nodes are more likely to be connected to highly connected nodes in the current graph, these models are often referred to as *preferential attachment* models.

Preferential attachment models under suitable conditions converge in the local weak sense in probability, see Van Der Hofstad (2024, Theorem 5.8) for details. The BA model, for example, generates power law networks with a coefficient of exactly three (Bollobás et al., 2001). In many applications, however, it is beneficial to also consider graphs with power law coefficient deviating from three to capture different empirical graph topologies.

Chung-Lu graphs. The Chung-Lu (CL) random graph model (Aiello et al., 2000; 2001; Chung & Lu, 2002; 2006) provides an efficient way to model large, sparse networks (Fasino et al., 2021). To generate a random CL graph with $N \in \mathbb{N}$ nodes, first specify a weight vector $\boldsymbol{w} \in \mathbb{R}_+^N$ with one weight $w_i \in \mathbb{R}_+$ for each node $i \in \{1, \ldots, N\}$. Then, two nodes i and j are connected with probability $w_i \cdot w_j / \bar{w}$, independently of all other node pairs and with normalization factor $\bar{w} := \sum_{1 \leq k \leq N} w_k$. Intuitively, a node with high weight is more likely to have many connections than a node with small weight.

Sequences of CL graphs fulfill Assumption 2.2 under mild technical assumptions, see Van Der Hofstad (2024, Theorem 3.18) for a formal statement. Most importantly, the average expected degree has to converge to a finite limit, which perfectly aligns with our goal to model very sparse networks. Throughout the paper, we use the running example of power law degree distributions with coefficient $\gamma > 2$ observed in many real world networks (Newman, 2003; Kaufmann & Zweig, 2009; Newman et al., 2011).

Example 1 (Power law). In our work, a power law is a zeta distribution with parameter $\gamma > 2$ such that $P(\deg(v) = k) = k^{-\gamma} / \zeta(\gamma)$, where $\zeta(\gamma)$ is the Riemann zeta function $\zeta(\gamma) := \sum_{j=1}^{\infty} j^{-\gamma}$. A power law degree distribution has a finite expectation $\mathbb{E}[\deg(v)] = \zeta(\gamma - 1) / \zeta(\gamma)$ for $\gamma > 2$.

Large, sparse power law networks of the above form can be efficiently generated by the CL framework (Fasino et al., 2021). Note that our methods apply to all distributions meeting Assumption 2.2 and perform well on many empirical

networks, as shown in the next sections.

Advantages over graphons and graphexes. Local weak converging graph sequences such as those generated by CM, BM or CL can model sparser, and thus often more realistic topologies than those captured by Lp graphons and graphexes. Figure 1 provides an illustration of how Lp graphons, graphexes, and CL graphs compare to a real world subsampled YouTube network (Mislove, 2009; Kunegis, 2013). Mathematically, both Lp graphons and graphexes are limited to graph sequences were the average degree diverges to infinity. Locally weak converging graph sequences, on the other hand, can capture sparse and often more realistic topologies. The usefulness of models like the CM, BA model and CL graphs is reflected in their frequent use in various research areas. However, formulating a mean field approach based on these graph theoretical models is challenging due to their high number of low degree nodes.

3. The Finite Model and Its Limit

Denote by $\mathcal{P}(\mathcal{X})$ the set of probability distributions over a finite set $\mathcal{X}$ and define $[N] := \{1, \ldots, N\}$ for any $N \in \mathbb{N}$.

Finite model. Assume some finite state space $\mathcal{X}$, finite action space $\mathcal{U}$ and finite and discrete time horizon $\mathcal{T} := \{0, \ldots, T-1\}$ with terminal time point T are given. Furthermore, there are $N \in \mathbb{N}$ agents connected by some graph $G_N = (V_N, E_N)$ with vertex set V_N and edge set E_N. Here, the random state of agent $i \in [N]$ at time $t \in \mathcal{T}$ is denoted by $X_{i,t}^N$. All agents $V_N^k \subseteq V_N$ with degree $k \in \mathbb{N}$ share a common policy π_t^k at all time points $t \in \mathcal{T}$. The empirical k-degree MF is defined as

$$\mu_t^{N,k} := \frac{1}{|V_N^k|} \sum_{i \in [N] : v_i \in V_N^k} \delta_{X_{i,t}^N} \in \mathcal{P}(\mathcal{X})$$

for all $t \in \mathcal{T}$ and $k \in \mathbb{N}$. Define the overall empirical MF sequence as $\mu_t^N := (\mu_t^{N,1}, \mu_t^{N,2}, \ldots) \in \mathcal{P}(\mathcal{X})^{\mathbb{N}}$. Each policy $\pi^k \in \mathcal{P}(\mathcal{U})^{\mathcal{T} \times \mathcal{X} \times \boldsymbol{\mathcal{G}}^k}$ in the policy ensemble $\pi = (\pi^1, \pi^2, \ldots) \in \mathcal{P}(\mathcal{U})^{\mathcal{T} \times \mathcal{X} \times \boldsymbol{\mathcal{G}}^k \times \mathbb{N}}$ takes into account the current state of the respective agent i with k neighbors and its neighborhood $\mathbb{G}_{i,t}^N \in \boldsymbol{\mathcal{G}}^k := \{G \in \mathcal{P}(\mathcal{X}) : k \cdot G \in \mathbb{N}_0^k\}$. Our learning algorithms also apply to other policy types, e.g., in our experiments we consider computationally efficient policies only depending on the current agent state. Then, the model dynamics are

$$U_{i,t}^N \sim \pi_t^k \left(\cdot | X_{i,t}^N, \mathbb{G}_{i,t}^N \right)$$

and

$$X_{i,t+1}^N \sim P \left(\cdot | X_{i,t}^N, U_{i,t}^N, \mathbb{G}_{i,t}^N \right)$$

for an agent i with degree k, $t \in \mathcal{T}$, i.i.d. initial distribution $\mu_0 \in \mathcal{P}(\mathcal{X})$, and transition kernel $P : \mathcal{X} \times \mathcal{U} \times \mathcal{P}(\mathcal{X}) \rightarrow$
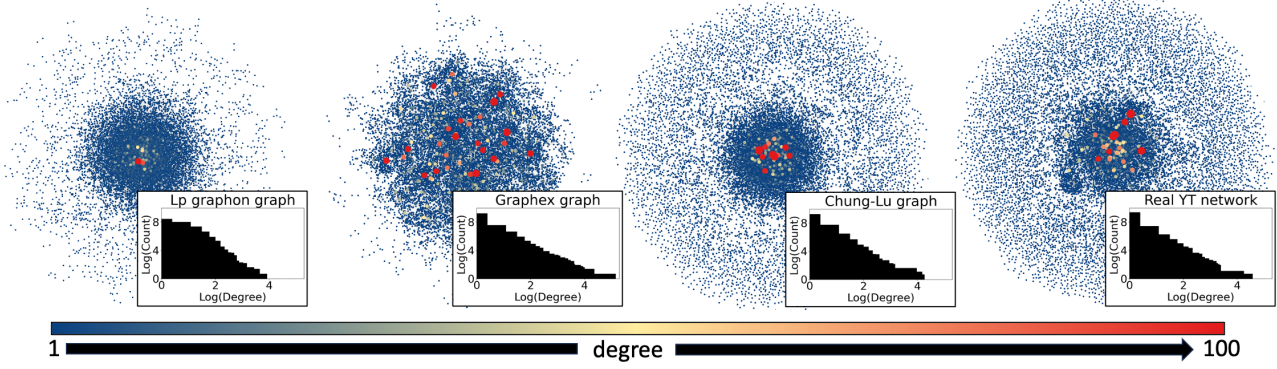
Figure 1. Four networks, first two generated by an Lp graphon and graphex, third is a CL graph and fourth is a real subsampled YouTube (YT) network (Mislove, 2009; Kunegis, 2013), highly connected nodes are depicted larger. Each network has around 14.5k nodes and 13k edges, except graphex has around 16.5k edges; all networks are plotted in the *prefuse force directed layout* (software: cytoscape). While the Lp graphon graph lacks sufficiently many high degree nodes, the tail of the graphex degree distribution is too heavy. In contrast, the CL graph is qualitatively close to the real YT network.

$\mathcal{P}(\mathcal{X})$. Note that the theory and subsequent learning algorithms extend to degree dependent transition kernels P^k. The policies are chosen to maximize the common objective

$$J^N(\pi) := \sum_{t=1}^{T} r(\mu_t^N)$$

with reward function $r : \mathcal{P}(\mathcal{X})^{\mathbb{N}} \mapsto \mathbb{R}$. Our model also covers reward functions with actions as inputs by using an extended state space $\mathcal{X} \cup (\mathcal{X} \times \mathcal{U})$ and splitting each time step $t \in \mathcal{T}$ into two.

Limiting LWMFC system. In the limiting LWMFC system, the MF for each degree $k \in \mathbb{N}$ evolves according to

$$
\begin{aligned}
\mu_{t+1}^k &:= \mu_t^k P_{t,\boldsymbol{\mu}',W}^{\pi,k} \\
&:= \sum_{x \in \mathcal{X}} \mu_t^k(x) \sum_{G \in \mathcal{G}^k} P_{\boldsymbol{\pi}}\left(\mathbb{G}_t^k(\boldsymbol{\mu}_t') = G \mid x_t = x\right) \\
&\qquad \cdot \sum_{u \in \mathcal{U}} \pi_t^k(u \mid x, G) P(\cdot \mid x, u, G)
\end{aligned}
$$

with i.i.d. initial distribution $\mu_0^k \in \mathcal{P}(\mathcal{X})$ and where $\mathcal{G}^k$ is the set of k-neighborhood distributions as before. As in the finite system, define the limiting MF ensemble $\mu_t := (\mu_t^1, \mu_t^2, \ldots) \in \mathcal{P}(\mathcal{X})^{\mathbb{N}}$ and the corresponding reward in the limiting system is $J(\pi) := \sum_{t=1}^{T} r(\mu_t)$.

Theoretical results. Next, we show the strong theoretical connection between the finite and limiting LWMFC system. The following theoretical results built on the assumption that the underlying graph sequence converges in the local weak sense, formalized by Assumption 2.2. The proofs are in Appendix A. We first state empirical MF convergence to the limiting MFs.

Theorem 3.1 (MF convergence). *Under Assumption 2.2, for any fixed policy ensemble π, the empirical MFs converge to the limiting MFs such that for all $k \in \mathbb{N}$ and all $t \in \mathcal{T}$*

$$\mu_t^{N,k} \to \mu_t^k \quad \text{in probability for} \quad N \to \infty.$$

The MF convergence from Theorem 3.1 enables us to derive a corresponding convergence result for the objective function under a standard continuity assumption on the reward.

Assumption 3.2. The reward function $r : \mathcal{P}(\mathcal{X})^{\mathbb{N}} \mapsto \mathbb{R}$ is continuous.

With the above assumption in place, we establish the convergence of the objective function in the finite system to the one in the limiting LWMFC model.

Proposition 3.3 (Objective convergence). *Under Assumptions 2.2 and 3.2 and for any fixed policy ensemble π, the common objective in the finite system converges to the limiting objective, i.e.*

$$J^N(\pi) \to J(\pi) \quad \text{in probability for} \quad N \to \infty.$$

We leverage these findings to show that for a finite set of policy ensembles, the optimal policy for the limiting system in the set is also optimal in all sufficiently large finite systems. Therefore, if one wants to know the optimal ensemble policy for an arbitrary, large agent system, it suffices to find the optimal ensemble policy in the limiting system once which is formalized by Corollary 3.4.

Corollary 3.4 (Optimal policy). *Assume some set $\{\pi_1, \ldots, \pi_M\}$ of $M < \infty$ policy ensembles is given and that w.l.o.g. $J(\pi_1) > J(\pi_i)$ for all $i \in [M]$ with $i \neq 1$. Under Assumptions 2.2 and 3.2 and for some $N^* \in \mathbb{N}$, π_1 is optimal in all finite systems of size $N > N^*$ such that*

$$J^N(\pi_1) > \max_{i \in [M], i \neq 1} J^N(\pi_i).$$

4. The Two Systems Approximation

In limiting systems on sparse graphs, the state evolution and optimal policy of an agent potentially depend on the entire network (Lacker & Soret, 2022). Calculating $P_{\boldsymbol{\pi}}\left(\mathbb{G}_t^k\left(\boldsymbol{\mu}_t'\right)=G \mid x_t=x\right)$ at time $t \in \mathcal{T}$ in the limiting system requires all possible t-hop neighborhood degree-state distributions where t-hop neighborhoods include all agents with a distance of at most t edges to the initial agent. Unfortunately, by Lemma 4.1 the number of t-hop neighborhoods grows at least exponentially with the degree k in important classes of locally weak converging graphs sequences, such as CL graphs with power laws above two.

Lemma 4.1. *In the limiting system, the number of possible t-hop degree-state neighborhood distributions of agents with degree $k \in \mathbb{N}$ at time $t \in \mathcal{T}$ in the worst case, e.g. CL power law, is $\Omega\left(2^{\mathrm{poly}(k)}\right)$.*

Just neglecting high degree nodes in the model might appear as a reasonable approximation to reduce computational complexity. However, the heavy tail of a degree distribution with finite expectation and infinite variance makes this approach highly inaccurate, as Example 2 illustrates.

Example 2. In a power law graph with $\gamma=2.5$, around 96% of nodes have a degree of at most five. However, these 96% of low degree nodes only account for roughly two thirds of the expected degree, formally $\sum_{h=1}^5 h^{1-\gamma}/\zeta(\gamma-1) < 0.68$. Nodes with a degree of at most ten still only account for around 76% of the expected degree.

Two systems approximation. For the subsequent two systems approximation, we first require a heuristic on the neighbor degree distribution for a given node.

Heuristic 1. For an arbitrary node $v' \in V$ the degree distribution of its neighbor $v \in V$ is approximately

$$P(\deg(v)=k \mid \deg(v')=k', (v',v) \in E)$$
$$\approx \frac{k \cdot P(\deg(v)=k)}{\sum_{k'' \in \mathbb{N}} k'' \cdot P(\deg(v)=k'')}.$$

Heuristic 1 is a good approximation for some sequences of locally weak converging graphs, such as CL graphs (Jackson et al., 2008, Chapter 4), and thus reasonable in our setup. The idea of Heuristic 1 is the following: if one fixes any node $v' \in V$ and considers its neighbors, high degree nodes are more likely to be connected to v' than lowly connected ones. Instead of the overall degree distribution, we thus weight each probability by its degree and normalize accordingly. The result is an approximate neighbor degree distribution accounting for the increased probability of highly connected neighbors.

To address the complexity of the limiting system, we provide an approximate limiting system based on Heuristic 1 and

the underlying sparse graph structure. Our two systems approximation consists of a system for small degree agents with at most k^* neighbors and another one for agents with more than k^* connections, where $k^* \in \mathbb{N}$ is some arbitrary, but fixed finite threshold. Define an approximate MF $\hat{\mu}^k$ for each $k \in [k^*]$ and furthermore summarize all agents with more than k^* connections into the infinite approximate MF $\hat{\mu}^\infty$ and define $\hat{\boldsymbol{\mu}} := (\hat{\mu}^1, \ldots, \hat{\mu}^{k^*}, \hat{\mu}^\infty)$. Based on Heuristic 1, we assume that all agents with more than k^* neighbors observe the same neighborhood state distribution

$$\hat{\mathbb{G}}_t^\infty(\hat{\boldsymbol{\mu}}) := \frac{1}{\mathbb{E}[\deg(v)]}\left(\sum_{k=k^*+1}^\infty k P(\deg(v)=k)\right)\hat{\mu}_t^\infty$$
$$+ \frac{1}{\mathbb{E}[\deg(v)]}\sum_{h=1}^{k^*} h P(\deg(v)=h)\hat{\mu}_t^h.$$

The unified approximate neighborhood state distribution $\hat{\mathbb{G}}_t^\infty$ allows us to state an approximate, simplified version of the MF forward dynamics for high degree agents given by

$$\hat{\mu}_{t+1}^\infty := \hat{\mu}_t^\infty \hat{P}_{t,\boldsymbol{\mu}',W}^{\pi,\infty} :=$$
$$\sum_{x,u} \hat{\mu}_t^\infty(x)\pi_t^\infty\left(u \mid x, \hat{\mathbb{G}}_t^\infty(\boldsymbol{\mu}')\right)P\left(\cdot \mid x, u, \hat{\mathbb{G}}_t^\infty(\boldsymbol{\mu}')\right)$$

where all agents with more than k^* connections follow the same policy $\pi_t^\infty \in \mathcal{P}(\mathcal{U})^{\mathcal{T} \times \mathcal{X} \times \mathcal{P}(\mathcal{X})}$ and where the sum is over all $(x,u) \in \mathcal{X} \times \mathcal{U}$. The approximate neighborhood of an agent with degree $k \in [k^*]$ at each time $t \in \mathcal{T}$ is sampled from $\hat{\mathbb{G}}_t^k(\hat{\boldsymbol{\mu}}) \sim \mathrm{Mult}(k, \hat{\mathbb{G}}_t^\infty(\hat{\boldsymbol{\mu}}))$, i.e. $\hat{\mathbb{G}}_t^k(\hat{\boldsymbol{\mu}})$ is multinomial with k trials and probabilities $\hat{\mathbb{G}}_t^\infty(\hat{\boldsymbol{\mu}})(x)$ for each $x \in \mathcal{X}$. Using Heuristic 1, the approximation yields for each $k \in [k^*]$ the MF forward dynamics

$$\hat{\mu}_{t+1}^k := \hat{\mu}_t^k \hat{P}_{t,\boldsymbol{\mu}',W}^{\pi,k} :=$$
$$\sum_{x,u,G} \hat{\mu}_t^k(x) P_{\mathrm{Mult}}\left(\hat{\mathbb{G}}_t^k = G\right)\pi_t^k(u \mid x, G)P(\cdot \mid x, u, G)$$

where the sum is over all $(x,u,G) \in \mathcal{X} \times \mathcal{U} \times \boldsymbol{\mathcal{G}}^k$.

Extensive approximation. In Appendix B we derive a second, extensive approximation

$$P_{\boldsymbol{\pi},\boldsymbol{\mu}}\left(\mathbb{G}_{t+1}^k(\boldsymbol{\mu}_t)=G, x_{t+1}=x\right)$$
$$\approx \sum_{G' \in \boldsymbol{\mathcal{G}}^k}\sum_{x' \in \mathcal{X}}\sum_{c \in \boldsymbol{\mathcal{C}}^k} P_{\boldsymbol{\pi},\boldsymbol{\mu}}\left(\mathbb{G}_t^k(\boldsymbol{\mu}_t)=G', x_t=x'\right)$$
$$\cdot \left[\sum_{u \in \mathcal{U}}\pi^k(u \mid x')P(x \mid x', u, G')\right]$$
$$\cdot \left[\sum_{\boldsymbol{a}_2 \in \boldsymbol{\mathcal{A}}_2^k(G',c)}\prod_j \mathrm{Mult}_{\boldsymbol{p}_{2,j}}(\boldsymbol{a}_{2,j})\right]$$
$$\cdot \frac{\sum_{\boldsymbol{a}_3 \in \boldsymbol{\mathcal{A}}_3^k(G,G',c)}\prod_{j,m}\mathrm{Mult}_{\boldsymbol{p}_{3,jm}}(\boldsymbol{a}_{3,jm})}{\sum_{\boldsymbol{a}_2 \in \boldsymbol{\mathcal{A}}_2^k(G',c)}\prod_{j,m}p_{j,m}(\boldsymbol{a}_2)}.$$

of the finite agent neighborhoods in the LWMFC model where we use the abbreviated notation $p_{j,m}(\boldsymbol{a}_2) := \left(P\left(\deg(v) = m \mid (v', v) \in E\right) \mu_t^m(s_j)\right)^{a_{jm}}$. Here, the idea is to go beyond the previous multinomial assumption $\hat{\mathbb{G}}_t^k(\hat{\boldsymbol{\mu}}) \sim \mathrm{Mult}(k, \hat{\mathbb{G}}_t^\infty(\hat{\boldsymbol{\mu}}))$ and to use state-degree neighborhood distributions $\boldsymbol{a}_2 \in \mathcal{A}_2^k(G', c)$ and state-state-degree neighborhood distributions $\boldsymbol{a}_3 \in \mathcal{A}_3^k(G, G', c)$ to capture agents changing from $x \in \mathcal{X}$ to $x' \in \mathcal{X}$ at a time step. We provide the extensive approximation derivation and corresponding definitions of sets like $\mathcal{A}_2^k(G', c)$ and $\mathcal{A}_3^k(G, G', c)$ in Appendix B. As we will see in the following, the extensive approximation often shows a moderately higher accuracy than our first approximation. However, the accuracy boost entails a significantly higher computational complexity due to multiple sums over sets like $\mathcal{A}_2^k(G', c)$ and $\mathcal{A}_3^k(G, G', c)$. Thus, our first approximation is more practical since it combines reasonable accuracy with low computational complexity while the extensive approximation is computationally too expensive for our purposes.

5. Learning Algorithms

To solve the MARL problem of finding optimal policies for each class of k-degree nodes, we propose two methods based on reducing the otherwise intractable many-agent graphical system to a single-agent MFC MDP. The first approach in Algorithm 1 is based on solving the resulting limiting MFC MDP under the parameters of the real graph, using the previously established two systems approximation. The second approach in Algorithm 2 instead directly learns according to single-agent RL that solves the MFC MDP by interacting with the real graph.

RL in MFC MDP. The two system approximation reduces the complexity of otherwise intractable large interacting systems on networks to the MFs of each degree. The system state at any time is then given by low-degree MFs $\mu_t^1, \mu_t^2, \ldots, \mu_t^{k^*}$ and high-degree MF μ_t^∞, briefly $\boldsymbol{\mu}_t := (\mu_t^1, \mu_t^2, \ldots, \mu_t^{k^*}, \mu_t^\infty)$. Given a state $\boldsymbol{\mu}_t$, the possible state evolutions depend only on the analogous set of low-degree and high-degree policies at that time, $\boldsymbol{\pi}_t := (\pi_t^1, \pi_t^2, \ldots, \pi_t^{k^*}, \pi_t^\infty)$. Therefore, choosing a $\boldsymbol{\pi}_t$ fully defines the state transition of the overall system, and is thus considered as the *high-level action* in the MFC MDP. Introducing a high-level policy $\hat{\pi}$ to output $\boldsymbol{\pi}_t \sim \hat{\pi}_t(\boldsymbol{\pi}_t \mid \boldsymbol{\mu}_t)$ allows us to solve for an optimal set of policies by solving the MFC MDP for optimal $\hat{\pi}$, since the limiting MF dynamics are deterministic. Finally, the MFC MDP is solved by applying single-agent policy gradient RL, resulting in Algorithm 1. In practice, we use proximal policy optimization (Schulman et al., 2017). To lower the complexity of the resulting MDP, we parametrize policies as distributions over actions given the node state, $\pi_t^k \in \mathcal{P}(\mathcal{U})^{\mathcal{X}}$.

Algorithm 1 LWMFC Policy Gradient

1: **for** iterations $n = 1, 2, \ldots$ **do**
2:    **for** time steps $t = 0, \ldots, B_{\mathrm{len}} - 1$ **do**
3:       Sample LWMFC MDP action $\boldsymbol{\pi}_t \sim \hat{\pi}^\theta(\boldsymbol{\pi}_t \mid \boldsymbol{\mu}_t)$.
4:       Compute reward $r(\boldsymbol{\mu}_t)$, next MF $\boldsymbol{\mu}_{t+1}$, termination flag $d_{t+1} \in \{0, 1\}$.
5:    **end for**
6:    Update policy $\hat{\pi}^\theta$ on minibatches $b \subseteq \{(\boldsymbol{\mu}_t, \boldsymbol{\pi}_t, r_t, d_{t+1}, \boldsymbol{\mu}_{t+1})\}_{t \geq 0}$ of length b_{len}.
7: **end for**

MARL on real networks. In addition to assuming knowledge of the model and computing the limiting MFC MDP equations, we may also directly learn on real network data without such model knowledge in a MARL manner. To do so, we still apply policy gradient RL to solve an assumed MFC MDP, but substitute samples from the real network into $\boldsymbol{\mu}_t$. At the same time, we let each node perform its actions according to the sampled $\boldsymbol{\pi}_t \sim \hat{\pi}_t(\boldsymbol{\pi}_t \mid \boldsymbol{\mu}_t)$. This approach is well justified by the previous theory and approximation, as for sufficiently large networks the limiting system and therefore also its limiting policy gradients are well approximated by this procedure.

Algorithm 2 LWMFMARL Policy Gradient

1: **for** iterations $n = 1, 2, \ldots$ **do**
2:    **for** time steps $t = 0, \ldots, B_{\mathrm{len}} - 1$ **do**
3:       Sample LWMFC MDP action $\boldsymbol{\pi}_t \sim \hat{\pi}^\theta(\boldsymbol{\pi}_t \mid \boldsymbol{\mu}_t)$.
4:       **for** node $i = 1, \ldots, N$ **do**
5:          Sample per-node action $U_{i,t} \sim \pi_t^{k_i}(U_{i,t} \mid X_{i,t})$ with degree $k_i = \infty$ if $k_i > k^*$.
6:       **end for**
7:       Perform actions, observe reward r_t, next MF $\boldsymbol{\mu}_{t+1}$, termination flag $d_{t+1} \in \{0, 1\}$.
8:    **end for**
9:    Update policy $\hat{\pi}^\theta$ on minibatches $b \subseteq \{(\boldsymbol{\mu}_t, \boldsymbol{\pi}_t, r_t, d_{t+1}, \boldsymbol{\mu}_{t+1})\}_{t \geq 0}$ of length b_{len}.
10: **end for**

The approach results in Algorithm 2 and has advantages. Firstly, the algorithm does not assume model knowledge and is therefore a true MARL algorithm, in contrast to solving the limiting MFC MDP. Secondly, the algorithm avoids potential inaccuracies of the two systems approximation, as we will see in Section 7, since it directly interacts with a real network of interest. Lastly, in contrast to standard independent and joint learning MARL methods, the method is rigorously justified by single-agent RL theory and avoids exponential complexity in the number of agents respectively.

Table 1. Average expected total variation $\Delta\mu = \frac{1}{2T}\mathbb{E}\left[\sum_t \|\hat{\mu}_t - \mu_t\|_1\right] \in [0,1]$ of MF μ_t and empirical MF $\hat{\mu}_t = \sum_i \delta_{X_t^i}$ ($\pm$ standard deviation, 50 trials), for the four models for four problems on eight real-world networks. Extensive LWMFC* not displayed for last two problems since calculations exceed maximum runtime. Best result for each network-problem combination in bold.

	Model	Average expected total variation $\Delta\mu$ in %, standard deviation in brackets							
		CAIDA	Cities	Digg Friends	Enron	Flixster	Slashdot	Yahoo	YouTube
SIS	LPGMFG	24.02 (1.25)	28.16 (0.41)	21.98 (0.26)	24.77 (0.32)	22.48 (0.07)	23.70 (0.43)	10.11 (2.10)	22.94 (0.25)
	GXMFG	9.07 (1.25)	10.90 (0.41)	4.72 (0.26)	4.73 (0.32)	3.78 (0.07)	5.48 (0.43)	9.31 (2.10)	6.43 (0.25)
	LWMFC	2.59 (1.14)	5.00 (0.40)	3.57 (0.26)	3.39 (0.31)	1.60 (0.07)	2.41 (0.43)	**3.59 (1.59)**	3.53 (0.25)
	LWMFC*	**1.75 (0.90)**	**4.20 (0.40)**	**3.02 (0.26)**	**2.67 (0.31)**	**0.90 (0.07)**	**1.70 (0.42)**	3.81 (1.70)	**2.93 (0.25)**
SIR	LPGMFG	9.11 (1.40)	10.01 (0.34)	8.68 (0.31)	9.51 (0.32)	8.99 (0.09)	9.37 (0.38)	4.88 (1.82)	8.90 (0.23)
	GXMFG	2.81 (1.10)	2.63 (0.31)	1.27 (0.29)	0.99 (0.30)	0.99 (0.09)	1.58 (0.36)	4.60 (1.71)	1.79 (0.23)
	LWMFC	1.31 (0.87)	1.36 (0.27)	1.08 (0.28)	0.91 (0.30)	0.58 (0.08)	0.99 (0.33)	**2.62 (1.30)**	1.07 (0.23)
	LWMFC*	**1.18 (0.82)**	**1.10 (0.26)**	**0.80 (0.27)**	**0.59 (0.28)**	**0.26 (0.08)**	**0.71 (0.29)**	2.63 (1.30)	**0.78 (0.23)**
Color	LPGMFG	38.73 (0.17)	38.59 (0.09)	38.70 (0.04)	39.83 (0.06)	39.55 (0.02)	39.07 (0.06)	34.18 (0.26)	38.52 (0.04)
	GXMFG	11.33 (0.13)	7.90 (0.06)	7.85 (0.02)	4.91 (0.03)	6.38 (0.01)	6.81 (0.03)	32.62 (0.24)	8.76 (0.02)
	LWMFC	**0.70 (0.12)**	**0.48 (0.05)**	**0.19 (0.02)**	**0.36 (0.04)**	**0.39 (0.02)**	**0.33 (0.04)**	**1.05 (0.19)**	**0.19 (0.03)**
Rumor	LPGMFG	20.03 (2.15)	22.56 (0.50)	18.39 (0.55)	20.27 (0.61)	18.94 (0.16)	19.70 (0.82)	9.68 (3.76)	19.23 (0.47)
	GXMFG	6.98 (2.06)	7.49 (0.49)	3.33 (0.54)	2.86 (0.58)	2.65 (0.16)	3.82 (0.79)	9.01 (3.69)	4.79 (0.47)
	LWMFC	**3.06 (1.59)**	**4.31 (0.48)**	**3.00 (0.53)**	**2.62 (0.57)**	**1.73 (0.15)**	**2.41 (0.75)**	**5.01 (2.21)**	**3.27 (0.46)**

6. Examples

For a general overview of many applications, see Laurière et al. (2022a). We consider four problems briefly described here. Problem details can be found in Appendix C.

Susceptible-Infected-Susceptible/Recovered (SIS/SIR). The classical SIS model (Kermack & McKendrick, 1927; Brauer, 2005) is a benchmark in the MF learning literature (Laurière et al., 2022b; Zhou et al., 2024). Agents are infected or susceptible, resulting in the state space $\mathcal{X} := \{S, I\}$, and decide to protect themselves or not. The infection probability increases without protection, and with the number of infected neighbors. Furthermore, there is a constant probability for recovering from an infection. The SIR model (Hethcote, 2000; Doncel et al., 2022) is an extension of SIS where agents can also be in a recovered state R where they are immune to a reinfection. Consequently, the state space for the SIR model is $\mathcal{X} := \{S, I, R\}$.

Graph coloring (Color). Inspired by graph coloring (Jensen & Toft, 2011; Barenboim & Elkin, 2022), the states
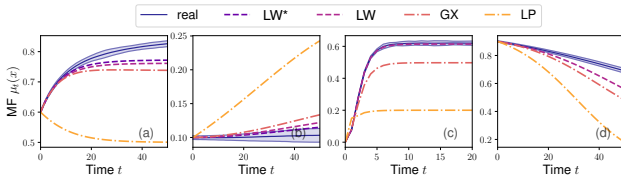

Figure 2. Overall MF evolution on real networks (50 trials, with two std. devs.), for our approx. (LW), our extensive approx. (LW*), graphex (GX), and Lp graphon (LP) models: (a) SIS on Enron, (b) SIR on Slashdot, (c) Color on CAIDA (without LW*), (d) Rumor on Cities (without LW*).

are finitely many colors on a circle and a target color distribution is given. Agents stay at their color or costly move to a neighboring color. The objective decreases for deviations from the target color distribution and if neighbors of an agent have neighboring colors to the agent's color on the respective color ring.

Rumor. In the rumor model (Maki & Thompson, 1973; Gani, 2000; Cui et al., 2022), agents are either aware of a rumor and are consequently in the aware state A or they have not heard the rumor and are in the ignorant state I. Aware agents decide whether they spread the rumor to their neighbors or keep it to themselves. They are awarded for spreading the rumor to unaware agents but loose reputation for telling the rumor to already aware agents.

7. Simulation & Results

In this section, we numerically verify the two system approximation as well as the proposed learning algorithms by comparing them with baselines from the literature. The two systems approximation is compared with previous graph approximations such as graphex or Lp graphon MF equations, and the learning algorithms are verified against standard scalable independent learning methods such as IPPO (Tan, 1993; Papoudakis et al., 2021), due to the large scale of networks considered here. To generate artificial networks of different sizes we employ a CL-based graph sampling algorithm (Chung & Lu, 2002; Miller & Hagberg, 2011) from the Python NetworkX package.

We compare the accuracy of our model on different empirical datasets with Lp graphon and graphex based models and with our extensive approximation LWMFC*, where computationally feasible, to see how much information is lost

Table 2. (LW)MFC policy gradient, (LW)MFMARL policy gradient, and IPPO for four problems on synthetic CL graphs of size N. Best objective after 24 hours of training on 96 CPUs. Best result for each problem-graph tuple in bold.

Problem	$N = 167$			$N = 406$			$N = 860$			$N = 1598$		
	IPPO	MFC	MFMARL	IPPO	MFC	MFMARL	IPPO	MFC	MFMARL	IPPO	MFC	MFMARL
SIS	-20.80	-14.56	**-12.50**	-21.40	-14.18	**-11.64**	-19.70	-12.42	**-9.11**	-22.42	-13.51	**-11.13**
SIR	-7.45	-7.84	**-6.99**	-7.18	-7.42	**-6.55**	-10.64	-6.86	**-5.15**	-7.73	-7.42	**-6.32**
Color	-8.20	-6.84	**-6.74**	-8.05	-7.04	**-6.98**	-8.48	-7.08	**-5.85**	-8.15	-6.97	**-6.94**
Rumor	0.24	**1.19**	0.27	0.16	**1.33**	0.19	0.25	**1.47**	1.35	0.12	**1.33**	0.17

in the LWMFC approximation. We use eight datasets from the KONECT database (Kunegis, 2013), where we substitute directed or weighted edges by simple undirected edges: CAIDA (Leskovec et al., 2007)($N \approx 26k$), Cities (Kunegis, 2013) ($N \approx 14k$), Digg Friends (Hogg & Lerman, 2012) ($N \approx 280k$), Enron (Klimt & Yang, 2004) ($N \approx 87k$), Flixster (Zafarani & Liu, 2009) ($N \approx 2.5mm$), Slashdot (Gómez et al., 2008) ($N \approx 50k$), Yahoo (Kunegis, 2013) ($N \approx 653k$), and YouTube (Mislove, 2009) ($N \approx 3.2mm$). See the references for details on the respective empirical networks.

Results. First, we establish the usefulness of LWMFC and LWMFC* by comparing their dynamics to those of LPGMFGs and GXMFGs (Fabian et al., 2023; 2024) on eight real-world networks, see Figure 2 for examplary dynamics over time. As Table 1 shows, our LWMFC approach clearly outperforms the current LPGMFGs and GXMFGs methods for all empirical networks and problems. The extensive approximation LWMFC* moderately outperforms LWMFC across datasets, except Yahoo. Since the extensive approximation is more detailed, it is often more accurate than the LWMFC approximation. However, Table 1 lacks an evaluation for LWMFC* on the Color and Rumor problem because the extensive approximation is computationally too expensive for these problems. Consequently, LWMFC dynamics are the more practical choice since they are computationally tractable and yield a good performance across problems and datasets.
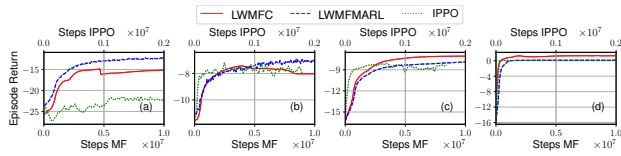
Figure 3. Training curves of LWMFC policy gradient, LWMF-MARL, and IPPO on a random CL graph with 406 nodes for: (a) SIS, (b) SIR, (c) Color, (d) Rumor.

The second part of our results focuses on our two learning algorithms LWMFC and LWMFMARL and compares them to the well-known IPPO algorithm. In Table 2, our algorithms

outperform IPPO for all problems on the two larger graphs with 860 and 1598 nodes, respectively. On the two smaller graphs, LWMFC and LWMFMARL still yield an at least competitive performance compared to IPPO, where IPPO is only marginally better than LWMFC on two problem instances, namely SIR on $N = 167$ and $N = 406$. We point out that LWMFC, in contrast to IPPO and LWMFMARL, is not evaluated on the empirical system, but by design on the limiting LWMFC model, which may differ from the true system behavior. These findings are complemented by the corresponding training curves in Figure 3. Finally, Figure 4 depicts how the training curves of our LWMFC and LWMF-MARL algorithms converge on different empirical networks for different problems.
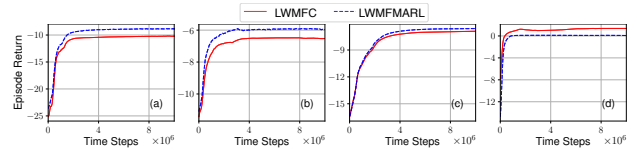
Figure 4. Training curves of LWMFC policy gradient and LWMF-MARL for four different examples: (a) SIS on Enron, (b) SIR on Slashdot, (c) Color on CAIDA, (d) Rumor on Cities.

8. Conclusion

We have introduced the novel LWMFC framework which can depict agent networks with finite expected degree and diverging variance. After a theoretical analysis, we provided a practical two systems approximation which was then leveraged to design scalable learning algorithms. Finally, we evaluated the performance of our model and learning algorithms for different problems on synthetic and real-world datasets and compared them to existing methods. For future work, one could extend the LWMFC model to various types of specific mean field models, e.g. to partial observability or agents under bounded rationality. We hope that LWMFC and the corresponding learning approaches prove to be a versatile and useful tool for researchers across various applied research areas.

Acknowledgements

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Achdou, Y. and Laurière, M. *Mean Field Games and Applications: Numerical Aspects*, pp. 249–307. Springer International Publishing, 2020.

Aiello, W., Chung, F., and Lu, L. A random graph model for massive graphs. In *Proceedings of the Annual ACM Symposium on Theory of Computing*, pp. 171–180, 2000.

Aiello, W., Chung, F., and Lu, L. A random graph model for power law graphs. *Experimental Mathematics*, 10(1): 53–66, 2001.

Andersson, D. and Djehiche, B. A maximum principle for SDEs of mean-field type. *Applied Mathematics & Optimization*, 63:341–356, 2011.

Bagagiolo, F., Faggian, S., Maggistro, R., and Pesenti, R. Optimal control of the mean field equilibrium for a pedestrian tourists' flow model. *Networks and Spatial Economics*, pp. 1–24, 2019.

Barabási, A.-L. and Albert, R. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

Barenboim, L. and Elkin, M. *Distributed graph coloring: Fundamentals and recent developments*. Springer Nature, 2022.

Bauso, D., Tembine, H., and Başar, T. Robust mean field games. *Dynamic Games and Applications*, 6(3):277–303, 2016.

Beck, M. and Robins, S. *Computing the continuous discretely: Integer-point enumeration in polyhedra*, volume 2. Springer, 2007.

Bender, E. A. and Canfield, E. R. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, 24(3):296–307, 1978.

Bensoussan, A., Frehse, J., Yam, P., et al. *Mean field games and mean field type control theory*, volume 101. Springer, 2013.

Bollobás, B. A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *European Journal of Combinatorics*, 1(4):311–316, 1980.

Bollobás, B. *Random graphs*. Springer, 1998.

Bollobás, B., Riordan, O., Spencer, J., and Tusnády, G. The degree sequence of a scale-free random graph process. *Random Structures & Algorithms*, 18(3):279–290, 2001.

Borgs, C., Chayes, J., Cohn, H., and Holden, N. Sparse exchangeable graphs and their limits via graphon processes. *Journal of Machine Learning Research*, 18(210):1–71, 2018a.

Borgs, C., Chayes, J., Cohn, H., and Zhao, Y. An lp theory of sparse graph convergence ii: Ld convergence, quotients and right convergence. *The Annals of Probability*, 46(1): 337–396, 2018b.

Borgs, C., Chayes, J., Cohn, H., and Zhao, Y. An lp theory of sparse graph convergence i: Limits, sparse random graph models, and power law distributions. *Transactions of the American Mathematical Society*, 372(5):3019–3062, 2019.

Brauer, F. The Kermack–McKendrick epidemic model revisited. *Mathematical Biosciences*, 198(2):119–131, 2005.

Caines, P. E. and Huang, M. Graphon mean field games and the GMFG equations: ε-Nash equilibria. In *IEEE 58th Conference on Decision and Control (CDC)*, pp. 286–292. IEEE, 2019.

Caines, P. E. and Huang, M. Graphon mean field games and their equations. *SIAM Journal on Control and Optimization*, 59(6):4373–4399, 2021.

Caines, P. E., Huang, M., and Malhamé, R. P. Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle. *Communications in Information and Systems*, 6(3): 221–252, 2006.

Canese, L., Cardarilli, G. C., Di Nunzio, L., Fazzolari, R., Giardino, D., Re, M., and Spanò, S. Multi-agent reinforcement learning: A review of challenges and applications. *Applied Sciences*, 11(11):4948, 2021.

Carmona, R. and Delarue, F. Probabilistic theory of mean field games with applications. i, volume 83 of probability theory and stochastic modelling, 2018.

Carmona, R. and Laurière, M. Deep learning for mean field games and mean field control with applications to finance. *Machine Learning and Data Sciences for Financial Markets: A Guide to Contemporary Practices*, pp. 369, 2023.

Carmona, R., Laurière, M., and Tan, Z. Model-free mean-field reinforcement learning: mean-field MDP and mean-field Q-learning. *The Annals of Applied Probability*, 33 (6B):5334–5381, 2023.

Caron, F. and Fox, E. B. Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(5):1295–1366, 2017.

Chung, F. and Lu, L. Connected components in random graphs with given expected degree sequences. *Annals of Combinatorics*, 6(2):125–145, 2002.

Chung, F. and Lu, L. *Complex graphs and networks*. American Mathematical Soc., 2006.

Cui, K. and Koeppl, H. Learning graphon mean field games and approximate Nash equilibria. In *International Conference on Learning Representations (ICLR)*, 2022.

Cui, K., KhudaBukhsh, W. R., and Koeppl, H. Hypergraphon mean field games. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 32(11), 2022.

Doncel, J., Gast, N., and Gaujal, B. A mean field game analysis of SIR dynamics with vaccination. *Probability in the Engineering and Informational Sciences*, 36(2): 482–499, 2022.

Dorogovtsev, S. N. and Mendes, J. F. Evolution of networks. *Advances in Physics*, 51(4):1079–1187, 2002.

Dorogovtsev, S. N., Goltsev, A. V., and Mendes, J. F. F. Pseudofractal scale-free web. *Physical Review E*, 65(6): 066122, 2002.

Fabian, C., Cui, K., and Koeppl, H. Learning sparse graphon mean field games. In *International Conference on Artificial Intelligence and Statistics*, pp. 4486–4514. PMLR, 2023.

Fabian, C., Cui, K., and Koeppl, H. Learning mean field games on sparse graphs: A hybrid graphex approach. In *International Conference on Learning Representations (ICLR)*, 2024.

Fasino, D., Tonetto, A., and Tudisco, F. Generating large scale-free networks with the Chung–Lu random graph model. *Networks*, 78(2):174–187, 2021.

Gani, J. The Maki–Thompson rumour model: a detailed analysis. *Environmental Modelling & Software*, 15(8): 721–725, 2000.

Goh, K.-I., Oh, E., Jeong, H., Kahng, B., and Kim, D. Classification of scale-free networks. *Proceedings of the National Academy of Sciences*, 99(20):12583–12588, 2002.

Gómez, V., Kaltenbrunner, A., and López, V. Statistical analysis of the social network and discussion threads in slashdot. In *Proceedings of the International Conference on World Wide Web*, pp. 645–654, 2008.

Gronauer, S. and Diepold, K. Multi-agent deep reinforcement learning: a survey. *Artificial Intelligence Review*, 55(2):895–943, 2022.

Gu, H., Guo, X., Wei, X., and Xu, R. Dynamic programming principles for mean-field controls with learning. *Operations Research*, 71(4):1040–1054, 2023.

Guo, X., Hu, A., Xu, R., and Zhang, J. Learning mean-field games. *Advances in Neural Information Processing Systems*, 32, 2019.

Guo, X., Hu, A., Xu, R., and Zhang, J. A general framework for learning mean-field games. *Mathematics of Operations Research*, 48(2):656–686, 2023.

Hethcote, H. W. The mathematics of infectious diseases. *SIAM Review*, 42(4):599–653, 2000.

Hogg, T. and Lerman, K. Social dynamics of digg. *EPJ Data Science*, 1(1):1–26, 2012.

Hu, S., Zhong, Y., Gao, M., Wang, W., Dong, H., Liang, X., Li, Z., Chang, X., and Yang, Y. MARLlib: A scalable and efficient library for multi-agent reinforcement learning. *JMLR*, 24:1–23, 2023a.

Hu, Y., Wei, X., Yan, J., and Zhang, H. Graphon mean-field control for cooperative multi-agent reinforcement learning. *Journal of the Franklin Institute*, 360(18):14783–14805, 2023b.

Jackson, M. O. et al. *Social and economic networks*, volume 3. Princeton University Press, 2008.

Jensen, T. R. and Toft, B. *Graph coloring problems*. John Wiley & Sons, 2011.

Kaufmann, M. and Zweig, K. Modeling and designing real–world networks. In *Algorithmics of Large and Complex Networks: Design, Analysis, and Simulation*, pp. 359–379. Springer, 2009.

Kermack, W. O. and McKendrick, A. G. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

Klimt, B. and Yang, Y. The enron corpus: A new dataset for email classification research. In *European Conference on Machine Learning*, pp. 217–226. Springer, 2004.

Kunegis, J. Konect: the Koblenz network collection. In *Proceedings of the International Conference on World Wide Web*, pp. 1343–1350, 2013.

Lacker, D. and Soret, A. A case study on stochastic games on large graphs in mean field and sparse regimes. *Mathematics of Operations Research*, 47(2):1530–1565, 2022.

Lacker, D., Ramanan, K., and Wu, R. Local weak convergence for sparse networks of interacting processes. *The Annals of Applied Probability*, 33(2):843–888, 2023.

Lasry, J.-M. and Lions, P.-L. Mean field games. *Japanese Journal of Mathematics*, 2(1):229–260, 2007.

Laurière, M., Perrin, S., Geist, M., and Pietquin, O. Learning mean field games: A survey. *arXiv preprint arXiv:2205.12944*, 2022a.

Laurière, M., Perrin, S., Girgin, S., Muller, P., Jain, A., Cabannes, T., Piliouras, G., Pérolat, J., Elie, R., Pietquin, O., et al. Scalable deep reinforcement learning algorithms for mean field games. In *International Conference on Machine Learning*, pp. 12078–12095. PMLR, 2022b.

Leskovec, J., Kleinberg, J., and Faloutsos, C. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2–es, 2007.

Liang, E., Liaw, R., Nishihara, R., Moritz, P., Fox, R., Goldberg, K., Gonzalez, J., Jordan, M., and Stoica, I. RLlib: Abstractions for distributed reinforcement learning. In *Proc. ICML*, pp. 3053–3062, 2018.

Lovász, L. *Large networks and graph limits*, volume 60. American Mathematical Soc., 2012.

Maki, D. P. and Thompson, M. *Mathematical Models and Applications: With Emphasis on the Social, Life, and Management Sciences*. Prentice Hall, 1973.

Mann, H. B. and Wald, A. On stochastic limit and order relationships. *The Annals of Mathematical Statistics*, 14(3):217–226, 1943.

Marsaglia, G. and Marsaglia, J. C. A new derivation of Stirling's approximation to n! *The American Mathematical Monthly*, 97(9):826–829, 1990.

Miller, J. C. and Hagberg, A. Efficient generation of networks with given expected degrees. In *International Workshop on Algorithms and Models for the Web-Graph*, pp. 115–126. Springer, 2011.

Mislove, A. E. *Online social networks: measurement, analysis, and applications to distributed information systems*. Rice University, 2009.

Molloy, M. and Reed, B. A critical point for random graphs with a given degree sequence. *Random Structures & Algorithms*, 6(2-3):161–180, 1995.

Molloy, M. and Reed, B. The size of the giant component of a random graph with a given degree sequence. *Combinatorics, Probability and Computing*, 7(3):295–305, 1998.

Newman, M. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

Newman, M., Barabási, A.-L., and Watts, D. J. *The structure and dynamics of networks*. Princeton University Press, 2011.

Papoudakis, G., Christianos, F., Schäfer, L., and Albrecht, S. V. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Piva, G. G., Ribeiro, F. L., and Mata, A. S. Networks with growth and preferential attachment: modelling and applications. *Journal of Complex Networks*, 9(1):cnab008, 2021.

Ruthotto, L., Osher, S. J., Li, W., Nurbekyan, L., and Fung, S. W. A machine learning framework for solving high-dimensional mean field game and mean field control problems. *Proceedings of the National Academy of Sciences*, 117(17):9183–9193, 2020.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Subramanian, S. G., Taylor, M. E., Crowley, M., and Poupart, P. Decentralized mean field games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 9439–9447, 2022.

Tan, M. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 330–337, 1993.

Van Der Hofstad, R. *Random graphs and complex networks*, volume 54. Cambridge University Press, 2024.

Van der Vaart, A. W. *Asymptotic statistics*, volume 3. Cambridge University Press, 2000.

Vázquez, A., Pastor-Satorras, R., and Vespignani, A. Large-scale topological and dynamical properties of the internet. *Physical Review E*, 65(6):066130, 2002.

Veitch, V. and Roy, D. M. The class of random graphs arising from exchangeable random measures. *arXiv preprint arXiv:1512.03099*, 2015.

Zafarani, R. and Liu, H. Social computing data repository at ASU, 2009. URL http://socialcomputing.asu.edu.

Zhang, F., Tan, V., Wang, Z., and Yang, Z. Learning regularized monotone graphon mean-field games. *Advances in Neural Information Processing Systems*, 36, 2024.

Zhou, F., Zhang, C., Chen, X., and Di, X. Graphon mean field games with a representative player: Analysis and learning algorithm. In *International Conference on Machine Learning (ICML)*, 2024.

A. Appendix: Proofs for the Theoretical Results

A.1. Proof of Theorem 3.1

Proof. We aim to eventually apply Lacker et al. (2023, Theorem 3.6) and therefore have to check that the respective conditions hold in our model. Keeping in mind Assumption 2.2 and the i.i.d. initial distribution μ_0, we leverage Lacker et al. (2023, Corollary 2.16) to obtain convergence in probability in the local weak sense of the marked graphs (G_N, X^{G_N}) to the limiting marked graph (G, X^G).

Since the theory in Lacker et al. (2023) is only formulated in terms of particle systems without including actions in the form of policies, we provide a suitable reformulation of our LWMFC model. Thus, define an auxiliary extended state space $\mathcal{X}_e := \mathcal{X} \cup (\mathcal{X} \times \mathcal{U})$ which serves as the state space for the extended particle system for some fixed policy ensemble π. The idea behind the extended state space $\mathcal{X}_e$ is to define an extended particle system where the state transition in $\mathcal{X}$ and the choice of the next action $u_{t+1} \in \mathcal{U}$ are separated into two different time steps.

Using the notations from Lacker et al. (2023), denote by $\mathcal{S}^{\sqcup}(\mathcal{X})$ the set of finite unordered sequences of arbitrary length with values in $\mathcal{X}$ and by $\Xi := \mathcal{X}^{\mathcal{X} \times \mathcal{S}^{\sqcup}(\mathcal{X})} \times \mathcal{U}^{\mathcal{X} \times \mathcal{U} \times \mathcal{S}^{\sqcup}(\mathcal{X})}$ the set of possible noise values. Next, specify a transition function $F^\tau : \mathcal{X}_e \times \mathcal{S}^{\sqcup}(\mathcal{X}_e) \times \Xi \to \mathcal{X}_e$ for each $\tau \in \mathcal{T}_e := \{0\} \cup [2T-1]$ by

$$X_{e,i,\tau+1}^N = F^\tau(X_{e,i,\tau}^N, \mathbb{G}_{e,i,\tau}^N, \xi_{i,\tau+1}) := \begin{cases} (X_{e,i,\tau}^N, \xi_{i,\tau+1}^0(X_{e,i,\tau}^N, \mathbb{G}_{e,i,\tau}^N)) & \text{if} \quad \tau/2 \in \{0\} \cup \mathbb{N} \\ \xi_{i,\tau+1}^1(X_{e,i,\tau}^N, \mathbb{G}_{e,i,\tau}^N) & \text{otherwise,} \end{cases}$$

where the neighborhood in the extended particle system $\mathbb{G}_{e,i,\tau}^N$ corresponds to $\mathbb{G}_{i,\lfloor \tau/2 \rfloor}^N$ in the original system. Here, the noise terms $\xi_{i,\tau+1} = (\xi_{i,\tau+1}^0, \xi_{i,\tau+1}^1)$ depict the used noise depending on whether τ is an even or odd number. If τ is an even number, i.e. $\tau/2 \in \{0\} \cup \mathbb{N}$, we use $\xi_{i,\tau+1}^0(X_{e,i,\tau}^N, \mathbb{G}_{e,i,\tau}^N)$ which is a $\mathcal{U}$-valued random variable with distribution

$$\xi_{i,\tau+1}^0(x, G) \sim \pi_{\tau/2}^k(\cdot \mid x, G)$$

for each neighborhood G and state $x \in \mathcal{X}$, where k is the degree of agent i. If τ is odd, i.e. we have $X_{e,i,\tau}^N \in \mathcal{X} \times \mathcal{U}$, we choose the first, $\mathcal{X}$-valued entry of $X_{e,i,\tau}^N$ as the x in the above probability distribution.

The $\mathcal{X}$-valued noise component $\xi_{i,\tau+1}^1$ is distributed as follows: if τ is an odd number, the noise term is sampled from

$$\xi_{i,\tau+1}^1(x, u, G) \sim P(\cdot | x, u, G)$$

where $X_{e,i,\tau}^N = (x, u)$. If τ is even and thus $X_{e,i,\tau}^N \in \mathcal{X}$, we just choose some arbitrary, but fixed action $u' \in \mathcal{U}$ instead of u in the above sampling process.

Now, it remains to check that Lacker et al. (2023, Assumption A) is satisfied by the extended particle system defined above. First, the noise terms $\xi_{i,\tau}$ are i.i.d. distributed for all agents $i \in [N]$ and with respect to all time points $\tau \in \mathcal{T}_e$ by construction. Finally, keeping in mind that the respective spaces are discrete, the map F^τ is continuous for each $\tau \in \mathcal{T}_e$. Therefore, Lacker et al. (2023, Theorem 3.6) yields the desired result. $\square$

A.2. Proof of Proposition 3.3

Proof. We want to show

$$J^N(\pi) \to J(\pi) \quad \text{in probability for} \quad N \to \infty$$

which is equivalent to

$$\sum_{t=1}^T r(\mu_t^N) \to \sum_{t=1}^T r(\mu_t) \quad \text{in probability for} \quad N \to \infty.$$

The reward r is a continuous function by Assumption 3.2. Furthermore, by Theorem 3.1 we know that the empirical mean fields converge in probability to the limiting mean fields. Hence, we can apply the continuous mapping theorem (Mann & Wald, 1943; Van der Vaart, 2000) to obtain the desired result. $\square$

A.3. Proof of Corollary 3.4

Proof. Quantify the gap Δ between the optimal and the second best solution as

$$\Delta := J(\pi_1) - \max_{i \in [M], i \neq 1} J(\pi_i) > 0.$$

Keeping in mind Proposition 3.3, we know that the objectives of the finite systems eventually converge to the limiting mean field objectives as N approaches infinity. Thus, there exists some N^* such that

$$\max_{i \in [M]} |J^N(\pi_i) - J(\pi_i)| < \frac{\Delta}{2}$$

holds for all $N > N^*$. Finally, the above considerations allow us to bound the difference of interest

$$
\begin{aligned}
J^N&(\pi_1) - \max_{i \in [M], i \neq 1} J^N(\pi_i) \\
&= J^N(\pi_1) - J(\pi_1) + J(\pi_1) - \max_{i \in [M], i \neq 1} J^N(\pi_i) \\
&= \underbrace{J^N(\pi_1) - J(\pi_1)}_{> -\Delta/2} + \underbrace{J(\pi_1) - \left(\max_{i \in [M], i \neq 1} J(\pi_i) \right)}_{= \Delta} + \left(\max_{i \in [M], i \neq 1} J(\pi_i) \right) - \max_{i \in [M], i \neq 1} J^N(\pi_i) \\
&> \frac{\Delta}{2} + \min_{i \in [M], i \neq 1} J(\pi_i) - J^N(\pi_i) > \frac{\Delta}{2} - \frac{\Delta}{2} = 0,
\end{aligned}
$$

for all $N > N^*$ which implies the desired statement

$$J^N(\pi_1) > \max_{i \in [M], i \neq 1} J^N(\pi_i)$$

and thereby concludes the proof. $\square$

A.4. Proof of Lemma 4.1

Proof. Since we want to lower bound the number of possible t-hop neighborhoods $N_{G,t}$, we assume for simplicity that t-hop neighbors of the initial agent have at most degree k themselves. Furthermore, we keep in mind the fact (Beck & Robins, 2007, Theorem 2.2) that in a d-dimensional simplex with edge length $\ell \in \mathbb{N}$, the number of integer points contained in the simplex is

$$\binom{d + \ell}{d} = \frac{(d + \ell)!}{d! \ell!}. \tag{1}$$

Since Lemma 4.1 considers the worst case, it suffices to prove the lower bound $\Omega\left(2^{\mathrm{poly}(k)}\right)$ for one class of CL graphs. We choose our running example of CL power law graphs with coefficient above two. It is well known that these large power law CL graphs are locally tree-like (Van Der Hofstad, 2024, Theorem 3.18) which we tacitly exploit in the following induction proof.

The proof is via induction over t. We start with $t = 1$ and the corresponding 1-hop neighborhood. The neighborhood consists of k agents where each one has a degree in $[k]$ and a state in $\mathcal{X}$. Since the agents themselves are indistinguishable in our model, we focus on the degree-state neighborhood distributions. Then, the set of possible state-degree neighborhood distributions can be seen as the integer points in a $(k + |\mathcal{X}| - 1)$-dimensional simplex with edge length k. Keeping in mind

Equation (1) and the well-known Stirling approximation, see e.g. Marsaglia & Marsaglia (1990), we obtain

$$
N_{G,t} \geq \binom{k + |\mathcal{X}| - 1 + k}{k + |\mathcal{X}| - 1}
$$

$$
= \frac{(2k + |\mathcal{X}| - 1)!}{(k + |\mathcal{X}| - 1)!k!}
$$

$$
\overset{\text{Stirling}}{\sim} \sqrt{\frac{2\pi(2k + |\mathcal{X}| - 1)}{2\pi(k + |\mathcal{X}| - 1)2\pi k}} \frac{(2k + |\mathcal{X}| - 1)^{2k+|\mathcal{X}|-1}}{(k + |\mathcal{X}| - 1)^{k+|\mathcal{X}|-1}k^k}
$$

$$
\geq \frac{1}{\sqrt{2\pi k}} \frac{(2k + |\mathcal{X}| - 1)^k}{k^k}
$$

$$
\geq \frac{2^k}{\sqrt{2\pi k}}
$$

$$
= 2^{k-1/2 \cdot \log_2(2\pi k)} \in \Theta\left(2^{\text{poly}(k)}\right).
$$

Now, it remains to establish the induction step from t to $t + 1$ where we assume that $N_{G,t} \in \Omega\left(2^{\text{poly}(k)}\right)$ holds. Then, instead of looking at the $(t + 1)$-hop neighborhoods of the initial agent, we can equivalently look at his or her 1-hop neighborhoods where each neighbor's 'extended state' now consists of the neighbor's t-hop neighborhood, where we ignore the edge between the neighbor and initial agent. Thus, the simplex edge length decreases by one from k to $k - 1$ which is negligible for large k. Leveraging the induction assumption, we obtain

$$
N_{G,t+1} = \Omega\left(\binom{N_{G,t} + k}{k}\right)
$$

$$
= \Omega\left(\frac{(N_{G,t} + k)!}{k!N_{G,t}!}\right)
$$

$$
\overset{\text{Stirling}}{=} \Omega\left(\sqrt{\frac{N_{G,t} + k}{kN_{G,t}}} \frac{(N_{G,t} + k)^{N_{G,t}+k}}{N_{G,t}^{N_{G,t}}k^k}\right)
$$

$$
= \Omega\left(\frac{1}{\sqrt{k}} \frac{(N_{G,t} + k)^k}{k^k}\right)
$$

$$
\overset{\text{(IA)}}{=} \Omega\left(\frac{1}{\sqrt{k}} \frac{\left(2^{\text{poly}(k)}\right)^k}{k^k}\right)
$$

$$
= \Omega\left(2^{k \cdot \text{poly}(k) - (k+1/2) \log_2(k)}\right)
$$

$$
= \Omega\left(2^{\text{poly}(k)}\right)
$$

which concludes the proof. $\qquad\square$

B. Extensive Approximation Derivation

The goal of the following section is to establish a detailed approximation of the probability

$$
P_{\boldsymbol{\pi},\boldsymbol{\mu}}\left(\mathbb{G}_{t+1}^k\left(\boldsymbol{\mu}_t\right) = G, x_{t+1} = x\right).
$$

First, we condition on the previous neighborhood distribution and the previous state of the agent at time t

$$
P_{\boldsymbol{\pi},\boldsymbol{\mu}}\left(\mathbb{G}_{t+1}^k\left(\boldsymbol{\mu}_t\right) = G, x_{t+1} = x\right)
$$

$$
= \sum_{x' \in \mathcal{X}} P_{\boldsymbol{\pi},\boldsymbol{\mu}}\left(\mathbb{G}_{t+1}^k\left(\boldsymbol{\mu}_t\right) = G, x_{t+1} = x, x_t = x'\right)
$$

$$
= \sum_{G' \in \boldsymbol{\mathcal{G}}^k} \sum_{x' \in \mathcal{X}} P_{\boldsymbol{\pi},\boldsymbol{\mu}}\left(\mathbb{G}_{t+1}^k\left(\boldsymbol{\mu}_t\right) = G, \mathbb{G}_t^k\left(\boldsymbol{\mu}_t\right) = G', x_{t+1} = x, x_t = x'\right).
$$

Now, we can decompose the above expression into three separate terms

$$P_{\boldsymbol{\pi},\boldsymbol{\mu}} \left(\mathbb{G}_{t+1}^k \left(\boldsymbol{\mu}_t \right) = G, \mathbb{G}_t^k \left(\boldsymbol{\mu}_t \right) = G', x_{t+1} = x, x_t = x' \right)$$
$$= \underbrace{P_{\boldsymbol{\pi},\boldsymbol{\mu}} \left(\mathbb{G}_t^k \left(\boldsymbol{\mu}_t \right) = G', x_t = x' \right)}_{\text{(I)}} \cdot \underbrace{P_{\boldsymbol{\pi},\boldsymbol{\mu}} \left(x_{t+1} = x \mid \mathbb{G}_t^k \left(\boldsymbol{\mu}_t \right) = G', x_t = x' \right)}_{\text{(II)}}$$
$$\cdot \underbrace{P_{\boldsymbol{\pi},\boldsymbol{\mu}} \left(\mathbb{G}_{t+1}^k \left(\boldsymbol{\mu}_t \right) = G \mid \mathbb{G}_t^k \left(\boldsymbol{\mu}_t \right) = G', x_{t+1} = x, x_t = x' \right)}_{\text{(III)}}$$

which allows us to handle each term individually. Since we require a recursive computation of the probability $P_{\boldsymbol{\pi},\boldsymbol{\mu}} \left(\mathbb{G}_{t+1}^k \left(\boldsymbol{\mu}_t \right) = G, x_{t+1} = x \right)$, the first term (I) will not be reformulated any further. The computation of the second term (II) is straight-forward, i.e.

$$P_{\boldsymbol{\pi},\boldsymbol{\mu}} \left(x_{t+1} = x \mid \mathbb{G}_t^k \left(\boldsymbol{\mu}_t \right) = G', x_t = x' \right) = \sum_{u \in \mathcal{U}} \pi^k \left(u \mid x' \right) P \left(x \mid x', u, G' \right).$$

Thus, it remains to approximate the third term (III)

$$P_{\boldsymbol{\pi},\boldsymbol{\mu}} \left(\mathbb{G}_{t+1}^k \left(\boldsymbol{\mu}_t \right) = G \mid \mathbb{G}_t^k \left(\boldsymbol{\mu}_t \right) = G', x_{t+1} = x, x_t = x' \right)$$
$$= P_{\boldsymbol{\pi},\boldsymbol{\mu}} \left(\mathbb{G}_{t+1}^k \left(\boldsymbol{\mu}_t \right) = G \mid \mathbb{G}_t^k \left(\boldsymbol{\mu}_t \right) = G', x_t = x' \right).$$

To ensure a reasonable approximation complexity, we make the simplifying assumption that the neighborhood distribution does not (crucially) depend on the current state of the agent of interest, i.e.

$$P_{\boldsymbol{\pi},\boldsymbol{\mu}} \left(\mathbb{G}_{t+1}^k \left(\boldsymbol{\mu}_t \right) = G \mid \mathbb{G}_t^k \left(\boldsymbol{\mu}_t \right) = G', x_t = x' \right) \approx P_{\boldsymbol{\pi},\boldsymbol{\mu}} \left(\mathbb{G}_{t+1}^k \left(\boldsymbol{\mu}_t \right) = G \mid \mathbb{G}_t^k \left(\boldsymbol{\mu}_t \right) = G' \right).$$

Thus, we focus on

$$P_{\boldsymbol{\pi},\boldsymbol{\mu}} \left(\mathbb{G}_{t+1}^k \left(\boldsymbol{\mu}_t \right) = G \mid \mathbb{G}_t^k \left(\boldsymbol{\mu}_t \right) = G' \right)$$

which requires an involved combinatorial argument to be calculated. The main difficulty in the calculation stems from the fact that the k neighbors of the initial agent in general have different degrees, different states at time t as well as different states at time $t + 1$. For notational convenience, we denote by $x_{1,t}, \ldots, x_{k,t}$ the states of the k neighbors of the initial agent at time t and by $\deg_1, \ldots, \deg_{k^*}, \deg_\infty \in \{1, \ldots, k\}$ the number of neighbors with the respective degree. Also, define $\mathcal{C}^k := \{c = (c_1, \ldots, c_{k^*}, c_\infty) \in \mathbb{N}_0^{k^*+1} : c_1 + \ldots + c_{k^*} + c_\infty = k\}$ for notational convenience. Then, the above probability can be expressed as

$$P_{\boldsymbol{\pi},\boldsymbol{\mu}} \left(\mathbb{G}_{t+1}^k \left(\boldsymbol{\mu}_t \right) = G \mid \mathbb{G}_t^k \left(\boldsymbol{\mu}_t \right) = G' \right)$$
$$= \sum_{c \in \mathcal{C}^k} P_{\boldsymbol{\pi},\boldsymbol{\mu}} \left(\mathbb{G}_{t+1}^k \left(\boldsymbol{\mu}_t \right) = G, \deg_1 = c_1, \ldots, \deg_{k^*} = c_{k^*}, \deg_\infty = c_\infty \mid \mathbb{G}_t^k \left(\boldsymbol{\mu}_t \right) = G' \right)$$
$$= \sum_{c \in \mathcal{C}^k} P_{\boldsymbol{\pi},\boldsymbol{\mu}} \left(\deg_1 = c_1, \ldots, \deg_{k^*} = c_{k^*}, \deg_\infty = c_\infty \mid \mathbb{G}_t^k \left(\boldsymbol{\mu}_t \right) = G' \right)$$
$$\cdot P_{\boldsymbol{\pi},\boldsymbol{\mu}} \left(\mathbb{G}_{t+1}^k \left(\boldsymbol{\mu}_t \right) = G \mid \deg_1 = c_1, \ldots, \deg_{k^*} = c_{k^*}, \deg_\infty = c_\infty, \mathbb{G}_t^k \left(\boldsymbol{\mu}_t \right) = G' \right).$$

In the remainder of the derivation, we will frequently use for all $s \in \mathcal{X}, m \in \mathbb{N}$, and $t \in \mathcal{T}$ the approximation

$$P \left(x_t^1 = s \mid \deg(v_1) = m, (v_0, v_1) \in E \right) \approx P \left(x_t^1 = s \mid \deg(v_1) = m \right) = \mu_t^m(s). \tag{2}$$

Next, we make an auxiliary calculation to calculate the degree distribution of a (uniformly at random picked) node v_1

conditional on its state x_t^1 and that it is a neighbor of the initial node v_0 of interest

$$P\left(\deg(v_1) = m \mid x_t^1 = s, (v_0, v_1) \in E\right)$$

$$= \frac{P\left(\deg(v_1) = m \cap x_t^1 = s \mid (v_0, v_1) \in E\right)}{P\left(x_t^1 = s \mid (v_0, v_1) \in E\right)}$$

$$= \frac{P\left(\deg(v_1) = m \mid (v_0, v_1) \in E\right) P\left(x_t^1 = s \mid \deg(v_1) = m, (v_0, v_1) \in E\right)}{P\left(\deg(v_1) > k^* \cap x_t^1 = s \mid (v_0, v_1) \in E\right) + \sum_{k=1}^{k^*} P\left(\deg(v_1) = k \cap x_t^1 = s \mid (v_0, v_1) \in E\right)}$$

$$\overset{(2)}{\approx} \frac{P\left(\deg(v_1) = m \mid (v_0, v_1) \in E\right) \mu_t^m(s)}{P\left(\deg(v_1) > k^* \cap x_t^1 = s \mid (v_0, v_1) \in E\right) + \sum_{k=1}^{k^*} P\left(\deg(v_1) = k \cap x_t^1 = s \mid (v_0, v_1) \in E\right)}$$

$$= \frac{P\left(\deg(v_1) = m \mid (v_0, v_1) \in E\right) \mu_t^m(s)}{P\left(\deg(v_1) > k^* \mid (v_0, v_1) \in E\right) \mu_t^\infty(s) + \sum_{k=1}^{k^*} P\left(\deg(v_1) = k \mid (v_0, v_1) \in E\right) \mu_t^k(s)}$$

where we exploit that

$$P\left(\deg(v_1) > k^* \cap x_t^1 = s \mid (v_0, v_1) \in E\right) + \sum_{k=1}^{k^*} P\left(\deg(v_1) = k \cap x_t^1 = s \mid (v_0, v_1) \in E\right)$$

$$= P\left(\deg(v_1) > k^* \mid (v_0, v_1) \in E\right) P\left(x_t^1 = s \mid \deg(v_1) > k^*, (v_0, v_1) \in E\right)$$

$$+ \sum_{k=1}^{k^*} P\left(\deg(v_1) = k \mid (v_0, v_1) \in E\right) P\left(x_t^1 = s \mid \deg(v_1) = k, (v_0, v_1) \in E\right)$$

$$\overset{(2)}{\approx} P\left(\deg(v_1) > k^* \mid (v_0, v_1) \in E\right) \mu_t^\infty(s) + \sum_{k=1}^{k^*} P\left(\deg(v_1) = k \mid (v_0, v_1) \in E\right) \mu_t^k(s).$$

For the running example of power law degree distributions with exponent $\gamma \in (2, 3)$, the conditional degree distribution is approximately

$$P\left(\deg(v_1) = m \mid x_t^1 = s_j, (v_0, v_1) \in E\right)$$

$$\approx \frac{\frac{m^{1-\gamma}}{\zeta(\gamma-1)} \mu_t^m(s_j)}{\frac{1}{\zeta(\gamma-1)} \left[\sum_{\ell=k^*+1}^\infty \ell^{1-\gamma}\right] \mu_t^\infty(s_j) + \frac{1}{\zeta(\gamma-1)} \sum_{h=1}^{k^*} h^{1-\gamma} \mu_t^h(s_j)}$$

$$= \frac{m^{1-\gamma} \mu_t^m(s_j)}{\left[\sum_{\ell=k^*+1}^\infty \ell^{1-\gamma}\right] \mu_t^\infty(s_j) + \sum_{h=1}^{k^*} h^{1-\gamma} \mu_t^h(s_j)}.$$

Based on the above probability and by the symmetry of the model, we obtain

$$P_{\boldsymbol{\pi},\boldsymbol{\mu}}\left(\deg_1 = c_1, \ldots, \deg_{k^*} = c_{k^*}, \deg_\infty = c_\infty \mid \mathbb{G}_t^k(\boldsymbol{\mu}_t) = G'\right)$$

$$= \sum_{\boldsymbol{a}_2 \in \mathcal{A}_2^k(G', c)} P_{\boldsymbol{\pi},\boldsymbol{\mu}}\left(A_2 = \boldsymbol{a}_2, \deg_1 = c_1, \ldots, \deg_{k^*} = c_{k^*}, \deg_\infty = c_\infty \mid \mathbb{G}_t^k(\boldsymbol{\mu}_t) = G'\right)$$

$$= \sum_{\boldsymbol{a}_2 \in \mathcal{A}_2^k(G', c)} P_{\boldsymbol{\pi},\boldsymbol{\mu}}\left(A_2 = \boldsymbol{a}_2 \mid \mathbb{G}_t^k(\boldsymbol{\mu}_t) = G'\right)$$

$$\approx \sum_{\boldsymbol{a}_2 \in \mathcal{A}_2^k(G', c)} \prod_{j=1}^d \binom{g_j'}{a_{j1}, \ldots, a_{j\infty}} \prod_{m \in [k^*] \cup \{\infty\}} \left(P\left(\deg(v_1) = m \mid x_t^1 = s_j, (v_0, v_1) \in E\right)\right)^{a_{jm}}$$

where we neglect dependencies between the nodes in the last line and define the matrix set $\mathcal{A}_2^k(G', c)$ for given $G' \in \mathcal{G}^k$ and $c \in \mathcal{C}^k$ as

$$\mathcal{A}_2^k(G', c) := \left\{\boldsymbol{a}_2 = (a_{jm})_{j \in [d], m \in [k^*] \cup \{\infty\}} \in \mathbb{N}_0^{d \times (k^*+1)} : \right.$$

$$\left. \sum_{m' \in [k^*] \cup \{\infty\}} a_{jm'} = g_j', \forall j \in [d] \quad \text{and} \quad \sum_{\ell=1}^d a_{\ell m} = c_m, \forall m \in [k^*] \cup \{\infty\}\right\}.$$

Therefore, it remains to calculate the conditional probability

$$P_{\boldsymbol{\pi},\boldsymbol{\mu}} \left(\mathbb{G}_{t+1}^k \left(\boldsymbol{\mu}_t \right) = G \mid \deg_1 = c_1, \ldots, \deg_{k^*} = c_{k^*}, \deg_\infty = c_\infty, \mathbb{G}_t^k \left(\boldsymbol{\mu}_t \right) = G' \right) .$$

As a first step, we define the set of matrices $\boldsymbol{\mathcal{A}}_3^k(G, G', c)$ for a given triple of vectors $G, G' \in \boldsymbol{\mathcal{G}}^k$ and $c \in \boldsymbol{\mathcal{C}}^k$ as

$$\boldsymbol{\mathcal{A}}_3^k(G, G', c) := \left\{ \boldsymbol{a}_3 = (a_{ijm})_{i,j \in [d], m \in [k^*] \cup \{\infty\}} \in \mathbb{N}_0^{d \times d \times (k^*+1)} : \right.$$

$$\sum_{m' \in [k^*] \cup \{\infty\}} \sum_{\ell=1}^d a_{i\ell m'} = g_i \text{ and } \sum_{m' \in [k^*] \cup \{\infty\}} \sum_{\ell=1}^d a_{\ell j m'} = g_j', \quad \forall i, j \in [d]$$

$$\left. \text{and } \sum_{\ell, \ell'=1}^d a_{\ell \ell' m} = c_m, \forall m \in [k^*] \cup \{\infty\} \right\} .$$

where $d := |\mathcal{X}|$ is the finite number of states. Intuitively, the matrix set $\boldsymbol{\mathcal{A}}_3^k(G, G', c)$ for an agent with degree k contains all possible numbers $(a_{ijm})_{i,j \in [d], m \in [k^*] \cup \{\infty\}}$ of neighbors whose degree is m and current state is x_i and who transition to state x_j in the next time step. For notational convenience, let A denote the random variable taking values in $\boldsymbol{\mathcal{A}}_3^k(G, G', c)$ and analogously let A_2 be the random variable with values in $\boldsymbol{\mathcal{A}}_2^k(G', c)$. We continue with the reformulation

$$P_{\boldsymbol{\pi},\boldsymbol{\mu}} \left(\mathbb{G}_{t+1}^k \left(\boldsymbol{\mu}_t \right) = G \mid \deg_1 = c_1, \ldots, \deg_{k^*} = c_{k^*}, \deg_\infty = c_\infty, \mathbb{G}_t^k \left(\boldsymbol{\mu}_t \right) = G' \right)$$

$$= \sum_{\boldsymbol{a}_2 \in \boldsymbol{\mathcal{A}}_2^k(G', c)} P_{\boldsymbol{\pi},\boldsymbol{\mu}} \left(A_2 = \boldsymbol{a}_2 \mid \deg_1 = c_1, \ldots, \deg_{k^*} = c_{k^*}, \deg_\infty = c_\infty, \mathbb{G}_t^k \left(\boldsymbol{\mu}_t \right) = G' \right)$$

$$\cdot P_{\boldsymbol{\pi},\boldsymbol{\mu}} \left(\mathbb{G}_{t+1}^k \left(\boldsymbol{\mu}_t \right) = G \mid A_2 = \boldsymbol{a}_2, \deg_1 = c_1, \ldots, \deg_{k^*} = c_{k^*}, \deg_\infty = c_\infty, \mathbb{G}_t^k \left(\boldsymbol{\mu}_t \right) = G' \right)$$

$$= \sum_{\boldsymbol{a}_2 \in \boldsymbol{\mathcal{A}}_2^k(G', c)} P_{\boldsymbol{\pi},\boldsymbol{\mu}} \left(A_2 = \boldsymbol{a}_2 \mid \deg_1 = c_1, \ldots, \deg_{k^*} = c_{k^*}, \deg_\infty = c_\infty, \mathbb{G}_t^k \left(\boldsymbol{\mu}_t \right) = G' \right)$$

$$\cdot P_{\boldsymbol{\pi},\boldsymbol{\mu}} \left(\mathbb{G}_{t+1}^k \left(\boldsymbol{\mu}_t \right) = G \mid A_2 = \boldsymbol{a}_2 \right) .$$

Next, we consider the two conditional probabilities separately. We start with

$$P_{\boldsymbol{\pi},\boldsymbol{\mu}} \left(A_2 = \boldsymbol{a}_2 \mid \deg_1 = c_1, \ldots, \deg_{k^*} = c_{k^*}, \deg_\infty = c_\infty, \mathbb{G}_t^k \left(\boldsymbol{\mu}_t \right) = G' \right)$$

$$= \frac{P_{\boldsymbol{\pi},\boldsymbol{\mu}} \left(A_2 = \boldsymbol{a}_2 \cap \deg_1 = c_1, \ldots, \deg_{k^*} = c_{k^*}, \deg_\infty = c_\infty, \mathbb{G}_t^k \left(\boldsymbol{\mu}_t \right) = G' \right)}{P_{\boldsymbol{\pi},\boldsymbol{\mu}} \left(\deg_1 = c_1, \ldots, \deg_{k^*} = c_{k^*}, \deg_\infty = c_\infty, \mathbb{G}_t^k \left(\boldsymbol{\mu}_t \right) = G' \right)}$$

$$= \frac{P_{\boldsymbol{\pi},\boldsymbol{\mu}} \left(A_2 = \boldsymbol{a}_2 \right)}{P_{\boldsymbol{\pi},\boldsymbol{\mu}} \left(\deg_1 = c_1, \ldots, \deg_{k^*} = c_{k^*}, \deg_\infty = c_\infty, \mathbb{G}_t^k \left(\boldsymbol{\mu}_t \right) = G' \right)} .$$

Keeping in mind both

$$P_{\boldsymbol{\pi},\boldsymbol{\mu}} \left(A_2 = \boldsymbol{a}_2 = (a_{jm})_{j,m} \right) \approx \prod_{j=1}^d \prod_{m \in [k^*] \cup \{\infty\}} \left(P \left(\deg(v_1) = m \mid (v_0, v_1) \in E \right) \mu_t^m(s_j) \right)^{a_{jm}}$$

by neglecting dependencies between the nodes and

$$P_{\boldsymbol{\pi},\boldsymbol{\mu}} \left(\deg_1 = c_1, \ldots, \deg_{k^*} = c_{k^*}, \deg_\infty = c_\infty, \mathbb{G}_t^k \left(\boldsymbol{\mu}_t \right) = G' \right)$$

$$= \sum_{\boldsymbol{a}_2 \in \boldsymbol{\mathcal{A}}_2^k(G', c)} P_{\boldsymbol{\pi},\boldsymbol{\mu}} \left(\deg_1 = c_1, \ldots, \deg_{k^*} = c_{k^*}, \deg_\infty = c_\infty, \mathbb{G}_t^k \left(\boldsymbol{\mu}_t \right) = G', A_2 = \boldsymbol{a}_2 \right)$$

$$= \sum_{\boldsymbol{a}_2 \in \boldsymbol{\mathcal{A}}_2^k(G', c)} P_{\boldsymbol{\pi},\boldsymbol{\mu}} \left(A_2 = \boldsymbol{a}_2 \right)$$

$$= \sum_{\boldsymbol{a}_2 \in \boldsymbol{\mathcal{A}}_2^k(G', c)} \prod_{j=1}^d \prod_{m \in [k^*] \cup \{\infty\}} \left(P \left(\deg(v_1) = m \mid (v_0, v_1) \in E \right) \mu_t^m(s_j) \right)^{a_{jm}}$$

we obtain

$$P_{\boldsymbol{\pi},\boldsymbol{\mu}}\left(A_2 = \boldsymbol{a}_2 \mid \deg_1 = c_1, \ldots, \deg_{k^*} = c_{k^*}, \deg_\infty = c_\infty, \mathbb{G}_t^k\left(\boldsymbol{\mu}_t\right) = G'\right)$$

$$= \frac{P_{\boldsymbol{\pi},\boldsymbol{\mu}}\left(A_2 = \boldsymbol{a}_2\right)}{P_{\boldsymbol{\pi},\boldsymbol{\mu}}\left(\deg_1 = c_1, \ldots, \deg_{k^*} = c_{k^*}, \deg_\infty = c_\infty, \mathbb{G}_t^k\left(\boldsymbol{\mu}_t\right) = G'\right)}$$

$$\approx \frac{\prod_{j=1}^d \prod_{m\in[k^*]\cup\{\infty\}}\left(P\left(\deg(v_1) = m \mid (v_0,v_1)\in E\right)\mu_t^m(s_j)\right)^{a_{jm}}}{\sum_{\boldsymbol{a}_2'\in\boldsymbol{\mathcal{A}}_2^k(G',c)}\prod_{j=1}^d \prod_{m\in[k^*]\cup\{\infty\}}\left(P\left(\deg(v_1) = m \mid (v_0,v_1)\in E\right)\mu_t^m(s_j)\right)^{a_{jm}'}}$$

and especially, for the case of a power law degree distribution with $\gamma \in (2,3)$, we have

$$P_{\boldsymbol{\pi},\boldsymbol{\mu}}\left(A_2 = \boldsymbol{a}_2 \mid \deg_1 = c_1, \ldots, \deg_{k^*} = c_{k^*}, \deg_\infty = c_\infty, \mathbb{G}_t^k\left(\boldsymbol{\mu}_t\right) = G'\right)$$

$$= \frac{\prod_{j=1}^d\left(\mu_t^\infty(s_j)\left(1 - \sum_{m'=1}^{k^*}\frac{(m')^{1-\gamma}}{\zeta(\gamma-1)}\right)\right)^{a_{j\infty}}\prod_{m=1}^{k^*}\left(\frac{m^{1-\gamma}\mu_t^m(s_j)}{\zeta(\gamma-1)}\right)^{a_{jm}}}{\sum_{\boldsymbol{a}_2'\in\boldsymbol{\mathcal{A}}_2^k(G',c)}\prod_{j=1}^d\left(\mu_t^\infty(s_j)\left(1 - \sum_{m'=1}^{k^*}\frac{(m')^{1-\gamma}}{\zeta(\gamma-1)}\right)\right)^{a_{j\infty}'}\prod_{m=1}^{k^*}\left(\frac{m^{1-\gamma}\mu_t^m(s_j)}{\zeta(\gamma-1)}\right)^{a_{jm}'}}$$

$$\approx \frac{\prod_{j=1}^d\left(\mu_t^\infty(s_j)\left(\zeta(\gamma-1) - \sum_{m'=1}^{k^*}(m')^{1-\gamma}\right)\right)^{a_{j\infty}}\prod_{m=1}^{k^*}\left(m^{1-\gamma}\mu_t^m(s_j)\right)^{a_{jm}}}{\sum_{\boldsymbol{a}_2'\in\boldsymbol{\mathcal{A}}_2^k(G',c)}\prod_{j=1}^d\left(\mu_t^\infty(s_j)\left(\zeta(\gamma-1) - \sum_{m'=1}^{k^*}(m')^{1-\gamma}\right)\right)^{a_{j\infty}'}\prod_{m=1}^{k^*}\left(m^{1-\gamma}\mu_t^m(s_j)\right)^{a_{jm}'}}.$$

Now, it remains to calculate the second probability term, namely

$$P_{\boldsymbol{\pi},\boldsymbol{\mu}}\left(\mathbb{G}_{t+1}^k\left(\boldsymbol{\mu}_t\right) = G \mid A_2 = \boldsymbol{a}_2\right).$$

Exploiting the symmetry of the problem, we obtain

$$P_{\boldsymbol{\pi},\boldsymbol{\mu}}\left(\mathbb{G}_{t+1}^k\left(\boldsymbol{\mu}_t\right) = G \mid A_2 = \boldsymbol{a}_2\right)$$

$$\approx \sum_{\boldsymbol{a}_3\in\boldsymbol{\mathcal{A}}^k(G,G',c)}\prod_{j=1}^d\prod_{m\in[k^*]\cup\{\infty\}}\binom{\sum_i a_{ijm}}{a_{1jm},\ldots,a_{djm}}\mathbf{1}_{\{\sum_i a_{ijm}=a_{jm}\}}$$

$$\cdot\prod_{i=1}^d\left(P_{\boldsymbol{\pi},\boldsymbol{\mu}}\left(x_{t+1}^1 = x_i \mid x_t^1 = x_j, \deg(v_1) = m\right)\right)^{a_{ijm}}$$

$$\approx \sum_{\boldsymbol{a}_3\in\boldsymbol{\mathcal{A}}^k(G,G',c)}\prod_{j=1}^d\prod_{m\in[k^*]\cup\{\infty\}}\binom{\sum_i a_{ijm}}{a_{1jm},\ldots,a_{djm}}\mathbf{1}_{\{\sum_i a_{ijm}=a_{jm}\}}$$

$$\cdot\prod_{i=1}^d\left(\sum_{G''\in\boldsymbol{\mathcal{G}}^m}P_{\boldsymbol{\pi}}\left(\mathbb{G}_t^m\left(\boldsymbol{\mu}_t\right) = G'' \mid x_t'' = s_j\right)\sum_{u\in\mathcal{U}}\pi_t^m\left(u \mid s_j\right)\cdot P\left(s_i \mid s_j, u, G''\right)\right)^{a_{ijm}}$$

where $\mathbf{1}_{\{\ldots\}}$ denotes the indicator function and where we neglect the potential dependencies between the neighbors of the

initial node in the second line. Finally, we arrive at

$$
\begin{aligned}
&P_{\boldsymbol{\pi},\boldsymbol{\mu}}\left(\mathbb{G}_{t+1}^k\left(\boldsymbol{\mu}_t\right)=G \mid \deg_1=c_1,\ldots,\deg_{k^*}=c_{k^*},\deg_\infty=c_\infty,\mathbb{G}_t^k\left(\boldsymbol{\mu}_t\right)=G'\right) \\
&=\sum_{\boldsymbol{a}_2\in\mathcal{A}_2^k(G',c)} P_{\boldsymbol{\pi},\boldsymbol{\mu}}\left(A_2=\boldsymbol{a}_2 \mid \deg_1=c_1,\ldots,\deg_{k^*}=c_{k^*},\deg_\infty=c_\infty,\mathbb{G}_t^k\left(\boldsymbol{\mu}_t\right)=G'\right) \\
&\hspace{6cm}\cdot P_{\boldsymbol{\pi},\boldsymbol{\mu}}\left(\mathbb{G}_{t+1}^k\left(\boldsymbol{\mu}_t\right)=G \mid A_2=\boldsymbol{a}_2\right) \\
&\approx \sum_{\boldsymbol{a}_2\in\mathcal{A}_2^k(G',c)} \frac{\prod_{j=1}^d\prod_{m\in[k^*]\cup\{\infty\}}\left(P\left(\deg(v_1)=m \mid (v_0,v_1)\in E\right)\mu_t^m(s_j)\right)^{a_{jm}}}{\sum_{\boldsymbol{a}_2'\in\mathcal{A}_2^k(G',c)}\prod_{j=1}^d\prod_{m\in[k^*]\cup\{\infty\}}\left(P\left(\deg(v_1)=m \mid (v_0,v_1)\in E\right)\mu_t^m(s_j)\right)^{a_{jm}'}} \\
&\hspace{6cm}\cdot P_{\boldsymbol{\pi},\boldsymbol{\mu}}\left(\mathbb{G}_{t+1}^k\left(\boldsymbol{\mu}_t\right)=G \mid A_2=\boldsymbol{a}_2\right) \\
&\approx \sum_{\boldsymbol{a}_2\in\mathcal{A}_2^k(G',c)} \frac{\prod_{j=1}^d\prod_{m\in[k^*]\cup\{\infty\}}\left(P\left(\deg(v_1)=m \mid (v_0,v_1)\in E\right)\mu_t^m(s_j)\right)^{a_{jm}}}{\sum_{\boldsymbol{a}_2'\in\mathcal{A}_2^k(G',c)}\prod_{j=1}^d\prod_{m\in[k^*]\cup\{\infty\}}\left(P\left(\deg(v_1)=m \mid (v_0,v_1)\in E\right)\mu_t^m(s_j)\right)^{a_{jm}'}} \\
&\quad \sum_{\boldsymbol{a}_3\in\mathcal{A}_3^k(G,G',c)}\prod_{j=1}^d\prod_{m\in[k^*]\cup\{\infty\}}\binom{\sum_i a_{ijm}}{a_{1jm},\ldots,a_{djm}}\mathbf{1}_{\{\sum_i a_{ijm}=a_{jm}\}} \\
&\quad\cdot\prod_{i=1}^d\left(\sum_{G''\in\mathcal{G}^m} P_{\boldsymbol{\pi}}\left(\mathbb{G}_t^m\left(\boldsymbol{\mu}_t\right)=G'' \mid x_t''=s_j\right)\sum_{u\in\mathcal{U}}\pi_t^m\left(u \mid s_j\right)\cdot P\left(s_i \mid s_j,u,G''\right)\right)^{a_{ijm}}
\end{aligned}
$$

and for the running example of power law graphs we especially obtain

$$
\begin{aligned}
&P_{\boldsymbol{\pi},\boldsymbol{\mu}}\left(\mathbb{G}_{t+1}^k\left(\boldsymbol{\mu}_t\right)=G \mid \deg_1=c_1,\ldots,\deg_{k^*}=c_{k^*},\deg_\infty=c_\infty,\mathbb{G}_t^k\left(\boldsymbol{\mu}_t\right)=G'\right) \\
&\approx \sum_{\boldsymbol{a}_2\in\mathcal{A}_2^k(G',c)} \frac{\prod_{j=1}^d\left(\mu_t^\infty(s_j)\left(1-\sum_{m'=1}^{k^*}\frac{(m')^{1-\gamma}}{\zeta(\gamma-1)}\right)\right)^{a_{j\infty}}\prod_{m=1}^{k^*}\left(\frac{m^{1-\gamma}\mu_t^m(s_j)}{\zeta(\gamma-1)}\right)^{a_{jm}}}{\sum_{\boldsymbol{a}_2'\in\mathcal{A}_2^k(G',c)}\prod_j\left(\mu_t^\infty(s_j)\left(1-\sum_{m'=1}^{k^*}\frac{(m')^{1-\gamma}}{\zeta(\gamma-1)}\right)\right)^{a_{j\infty}'}\prod_{m=1}^{k^*}\left(\frac{m^{1-\gamma}\mu_t^m(s_j)}{\zeta(\gamma-1)}\right)^{a_{jm}'}} \\
&\quad \sum_{\boldsymbol{a}_3\in\mathcal{A}_3^k(G,G',c)}\prod_{j=1}^d\prod_{m\in[k^*]\cup\{\infty\}}\binom{\sum_i a_{ijm}}{a_{1jm},\ldots,a_{djm}}\mathbf{1}_{\{\sum_i a_{ijm}=a_{jm}\}} \\
&\quad\cdot\prod_{i=1}^d\left(\sum_{G''\in\mathcal{G}^m} P_{\boldsymbol{\pi}}\left(\mathbb{G}_t^m\left(\boldsymbol{\mu}_t\right)=G'' \mid x_t''=s_j\right)\sum_{u\in\mathcal{U}}\pi_t^m\left(u \mid s_j\right)\cdot P\left(s_i \mid s_j,u,G''\right)\right)^{a_{ijm}} \\
&=\sum_{\boldsymbol{a}_3\in\mathcal{A}_3^k(G,G',c)} \frac{\prod_{j=1}^d\left(\mu_t^\infty(s_j)\left(1-\sum_{m'=1}^{k^*}\frac{(m')^{1-\gamma}}{\zeta(\gamma-1)}\right)\right)^{a_{j\infty}}\prod_{m=1}^{k^*}\left(\frac{m^{1-\gamma}\mu_t^m(s_j)}{\zeta(\gamma-1)}\right)^{a_{jm}}}{\sum_{\boldsymbol{a}_2'\in\mathcal{A}_2^k(G',c)}\prod_j\left(\mu_t^\infty(s_j)\left(1-\sum_{m'=1}^{k^*}\frac{(m')^{1-\gamma}}{\zeta(\gamma-1)}\right)\right)^{a_{j\infty}'}\prod_{m=1}^{k^*}\left(\frac{m^{1-\gamma}\mu_t^m(s_j)}{\zeta(\gamma-1)}\right)^{a_{jm}'}} \\
&\quad \prod_{j=1}^d\prod_{m\in[k^*]\cup\{\infty\}}\binom{\sum_i a_{ijm}}{a_{1jm},\ldots,a_{djm}} \\
&\quad\cdot\prod_{i=1}^d\left(\sum_{G''\in\mathcal{G}^m} P_{\boldsymbol{\pi}}\left(\mathbb{G}_t^m\left(\boldsymbol{\mu}_t\right)=G'' \mid x_t''=s_j\right)\sum_{u\in\mathcal{U}}\pi_t^m\left(u \mid s_j\right)\cdot P\left(s_i \mid s_j,u,G''\right)\right)^{a_{ijm}}.
\end{aligned}
$$

Resulting Approximation   Eventually, we obtain the approximation

$$P_{\boldsymbol{\pi},\boldsymbol{\mu}}\left(\mathbb{G}_{t+1}^k\left(\boldsymbol{\mu}_t\right)=G, x_{t+1}=x\right)$$

$$\approx \sum_{G'\in\mathcal{G}^k}\sum_{x'\in\mathcal{X}} P_{\boldsymbol{\pi},\boldsymbol{\mu}}\left(\mathbb{G}_t^k\left(\boldsymbol{\mu}_t\right)=G', x_t=x'\right)\left[\sum_{u\in\mathcal{U}}\pi^k\left(u\mid x'\right)P\left(x\mid x',u,G'\right)\right]$$

$$\cdot \sum_{c\in\mathcal{C}^k}\left[\sum_{\boldsymbol{a}_2\in\mathcal{A}^k(G',c)}\prod_{j=1}^d\binom{g_j'}{a_{j1},\ldots,a_{j\infty}}\prod_{m\in[k^*]\cup\{\infty\}}\right.$$

$$\left.\cdot\left(\frac{P\left(\deg(v_1)=m\mid(v_0,v_1)\in E\right)\mu_t^m(s_j)}{P\left(\deg(v_1)>k^*\mid(v_0,v_1)\in E\right)\mu_t^\infty(s_j)+\sum_{k=1}^{k^*}P\left(\deg(v_1)=k\mid(v_0,v_1)\in E\right)\mu_t^k(s_j)}\right)^{a_{jm}}\right]$$

$$\cdot\frac{1}{\sum_{\boldsymbol{a}_2'\in\mathcal{A}_2^k(G',c)}\prod_{j,m}\left(P\left(\deg(v_1)=m\mid(v_0,v_1)\in E\right)\mu_t^m(s_j)\right)^{a_{jm}'}}$$

$$\sum_{\boldsymbol{a}_3\in\mathcal{A}_3^k(G,G',c)}\prod_{j=1}^d\prod_{m\in[k^*]\cup\{\infty\}}\binom{\sum_i a_{ijm}}{a_{1jm},\ldots,a_{djm}}\prod_{i=1}^d\left(P\left(\deg(v_1)=m\mid(v_0,v_1)\in E\right)\mu_t^m(s_j)\right)^{a_{ijm}}$$

$$\cdot\left(\sum_{G''\in\mathcal{G}^m}P_{\boldsymbol{\pi}}\left(\mathbb{G}_t^m\left(\boldsymbol{\mu}_t\right)=G''\mid x_t''=s_j\right)\sum_{u\in\mathcal{U}}\pi_t^m\left(u\mid s_j\right)\cdot P\left(s_i\mid s_j,u,G''\right)\right)^{a_{ijm}}$$

which, for the power law running example, can be reformulated as

$$P_{\boldsymbol{\pi},\boldsymbol{\mu}}\left(\mathbb{G}_{t+1}^k\left(\boldsymbol{\mu}_t\right)=G, x_{t+1}=x\right)$$

$$\approx \sum_{G'\in\mathcal{G}^k}\sum_{x'\in\mathcal{X}} P_{\boldsymbol{\pi},\boldsymbol{\mu}}\left(\mathbb{G}_t^k\left(\boldsymbol{\mu}_t\right)=G', x_t=x'\right)\left[\sum_{u\in\mathcal{U}}\pi^k\left(u\mid x'\right)P\left(x\mid x',u,G'\right)\right]$$

$$\cdot \sum_{c\in\mathcal{C}^k}\left[\sum_{\boldsymbol{a}_2\in\mathcal{A}^k(G',c)}\prod_{j=1}^d\binom{g_j'}{a_{j1},\ldots,a_{j\infty}}\right.$$

$$\left.\cdot\prod_{m\in[k^*]\cup\{\infty\}}\left(\frac{m^{1-\gamma}\mu_t^m(s_j)}{\left[\sum_{\ell=k^*+1}^\infty \ell^{1-\gamma}\right]\mu_t^\infty(s_j)+\sum_{h=1}^{k^*}h^{1-\gamma}\mu_t^h(s_j)}\right)^{a_{jm}}\right]$$

$$\cdot\sum_{\boldsymbol{a}_3\in\mathcal{A}_3^k(G,G',c)}\frac{\prod_{j=1}^d\left(\mu_t^\infty(s_j)\left(1-\sum_{m'=1}^{k^*}\frac{(m')^{1-\gamma}}{\zeta(\gamma-1)}\right)\right)^{a_{j\infty}}\prod_{m=1}^{k^*}\left(\frac{m^{1-\gamma}\mu_t^m(s_j)}{\zeta(\gamma-1)}\right)^{a_{jm}}}{\sum_{\boldsymbol{a}_2'\in\mathcal{A}_2^k(G',c)}\prod_j\left(\mu_t^\infty(s_j)\left(1-\sum_{m'=1}^{k^*}\frac{(m')^{1-\gamma}}{\zeta(\gamma-1)}\right)\right)^{a_{j\infty}'}\prod_{m=1}^{k^*}\left(\frac{m^{1-\gamma}\mu_t^m(s_j)}{\zeta(\gamma-1)}\right)^{a_{jm}'}}$$

$$\prod_{j=1}^d\prod_{m\in[k^*]\cup\{\infty\}}\binom{\sum_i a_{ijm}}{a_{1jm},\ldots,a_{djm}}$$

$$\cdot\prod_{i=1}^d\left(\sum_{G''\in\mathcal{G}^m}P_{\boldsymbol{\pi}}\left(\mathbb{G}_t^m\left(\boldsymbol{\mu}_t\right)=G''\mid x_t''=s_j\right)\sum_{u\in\mathcal{U}}\pi_t^m\left(u\mid s_j\right)\cdot P\left(s_i\mid s_j,u,G''\right)\right)^{a_{ijm}}$$

For notational convenience, define for each $j\in[d]$ and $m\in[k^*]\cup\{\infty\}$

$$p_{jm}\coloneqq\frac{P\left(\deg(v_1)=m\mid(v_0,v_1)\in E\right)\mu_t^m(s_j)}{P\left(\deg(v_1)>k^*\mid(v_0,v_1)\in E\right)\mu_t^\infty(s_j)+\sum_{k=1}^{k^*}P\left(\deg(v_1)=k\mid(v_0,v_1)\in E\right)\mu_t^k(s_j)}$$

and for each $i,j\in[d]$ and $m\in[k^*]\cup\{\infty\}$

$$p_{ijm}\coloneqq P\left(\deg(v_1)=m\mid(v_0,v_1)\in E\right)\mu_t^m(s_j)$$

$$\cdot\sum_{G''\in\mathcal{G}^m}P_{\boldsymbol{\pi}}\left(\mathbb{G}_t^m\left(\boldsymbol{\mu}_t\right)=G''\mid x_t''=s_j\right)\sum_{u\in\mathcal{U}}\pi_t^m\left(u\mid s_j\right)\cdot P\left(s_i\mid s_j,u,G''\right).$$

Then, the extensive approximation can be rewritten more compactly as

$$
P_{\boldsymbol{\pi},\boldsymbol{\mu}}\left(\mathbb{G}_{t+1}^k\left(\boldsymbol{\mu}_t\right)=G, x_{t+1}=x\right)
$$

$$
\approx \sum_{G'\in\boldsymbol{\mathcal{G}}^k}\sum_{x'\in\mathcal{X}} P_{\boldsymbol{\pi},\boldsymbol{\mu}}\left(\mathbb{G}_t^k\left(\boldsymbol{\mu}_t\right)=G', x_t=x'\right)\left[\sum_{u\in\mathcal{U}}\pi^k\left(u\mid x'\right)P\left(x\mid x',u,G'\right)\right]
$$

$$
\cdot \sum_{c\in\boldsymbol{\mathcal{C}}^k}\left[\sum_{\boldsymbol{a}_2\in\boldsymbol{\mathcal{A}}_2^k(G',c)}\prod_{j=1}^d\binom{g_j'}{a_{j1},\ldots,a_{j\infty}}\prod_{m\in[k^*]\cup\{\infty\}}p_{jm}^{a_{jm}}\right]
$$

$$
\cdot \frac{\sum_{\boldsymbol{a}_3\in\boldsymbol{\mathcal{A}}_3^k(G,G',c)}\prod_{j=1}^d\prod_{m\in[k^*]\cup\{\infty\}}\binom{\sum_i a_{ijm}}{a_{1jm},\ldots,a_{djm}}\prod_{i=1}^d p_{ijm}^{a_{ijm}}}{\sum_{\boldsymbol{a}_2\in\boldsymbol{\mathcal{A}}_2^k(G',c)}\prod_{j,m}\left(P\left(\deg(v_1)=m\mid(v_0,v_1)\in E\right)\mu_t^m(s_j)\right)^{a_{jm}}}.
$$

Furthermore, we introduce

$$
\boldsymbol{p}_{2,j}:=(p_{j1},\ldots,p_{jk^*},p_{j\infty})\quad\text{and}\quad\boldsymbol{a}_{2,j}:=(a_{j1},\ldots,a_{jk^*},a_{j\infty})
$$

for every $j\in[d]$ and similarly we define

$$
\boldsymbol{p}_{3,jm}:=(p_{1jm},\ldots,p_{djm})\quad\text{and}\quad\boldsymbol{a}_{3,jm}:=(a_{1jm},\ldots,a_{djm})
$$

for every tuple $(j,m)\in[d]\times([k^*]\cup\{\infty\})$. Then, the extensive approximation can be formulated as

$$
P_{\boldsymbol{\pi},\boldsymbol{\mu}}\left(\mathbb{G}_{t+1}^k\left(\boldsymbol{\mu}_t\right)=G, x_{t+1}=x\right)
$$

$$
\approx \sum_{G'\in\boldsymbol{\mathcal{G}}^k}\sum_{x'\in\mathcal{X}}\sum_{c\in\boldsymbol{\mathcal{C}}^k} P_{\boldsymbol{\pi},\boldsymbol{\mu}}\left(\mathbb{G}_t^k\left(\boldsymbol{\mu}_t\right)=G', x_t=x'\right)\left[\sum_{u\in\mathcal{U}}\pi^k\left(u\mid x'\right)P\left(x\mid x',u,G'\right)\right]
$$

$$
\cdot \frac{\left[\sum_{\boldsymbol{a}_2\in\boldsymbol{\mathcal{A}}_2^k(G',c)}\prod_j\mathrm{Mult}_{\boldsymbol{p}_{2,j}}(\boldsymbol{a}_{2,j})\right]\sum_{\boldsymbol{a}_3\in\boldsymbol{\mathcal{A}}_3^k(G,G',c)}\prod_{j,m}\mathrm{Mult}_{\boldsymbol{p}_{3,jm}}(\boldsymbol{a}_{3,jm})}{\sum_{\boldsymbol{a}_2\in\boldsymbol{\mathcal{A}}_2^k(G',c)}\prod_{j,m}\left(P\left(\deg(v_1)=m\mid(v_0,v_1)\in E\right)\mu_t^m(s_j)\right)^{a_{jm}}}.
$$

C. Simulation Details

We use MARLlib 1.0.0 (Hu et al., 2023a) building on RLlib 1.8.0 (Apache-2.0 license) (Liang et al., 2018) and its PPO implementation (Schulman et al., 2017) for IPPO and our algorithms. For our experiments, we used around $80\,000$ core hours on CPUs, and each training run usually took a single day of training on up to 96 parallel CPU cores. For the policies we used two hidden layers of 256 nodes with $\tanh$ activations. We used a discount factor of $\gamma=0.99$ with GAE $\lambda=1.0$, and training and minibatch sizes of 4000 and 1000, performing 5 updates per training batch. The KL coefficient and clip parameter were set to 0.2, with a KL target of 0.03. The learning rate was set to 0.00005. The problem details are found in the following.

Susceptible-Infected-Susceptible (SIS). In the SIS model with state space $\mathcal{X}:=\{S,I\}$, agents are either infected (I) or susceptible to a virus (S). At each time step $t\in\mathcal{T}$, agents either protect themselves (P) or not ($\bar{P}$) which is formalized by the action space $\mathcal{U}:=\{P,\bar{P}\}$. As usual, the game terminates at finite terminal time $T\in\mathbb{N}$ which can be interpreted as the time when a cure for the virus is found. Therefore, it remains to specify the transition dynamics. Susceptible agents who protect themselves at time t also remain susceptible at time $t+1$, i.e.

$$
P^k(S\mid S,P,G)=1\quad\text{and}\quad P^k(I\mid S,P,G)=0,
$$

irrespective of their degree k and neighborhood G. On the other hand, if a susceptible agent chooses action $\bar{P}$, the transition dynamics are

$$
P^k(I\mid S,\bar{P},G)=\rho_I\cdot G(I)\cdot\left(\frac{2}{1+\exp(-k/2)}-1\right)
$$

and $P^k(S \mid S, \bar{P}, G) = 1 - P^k(I \mid S, \bar{P}, G)$, correspondingly, and where $\rho_I > 0$ is a fixed infection rate. Apart from that, infected agents recover with some fixed recovery rate $1 \geq \rho_R \geq 0$, independent of their action and degree, which means that

$$P^k(S \mid I, \bar{P}, G) = P^k(S \mid I, P, G) = \rho_R.$$

To complete the model, the reward per agent taking action $u \in \mathcal{U}$ in state $x \in \mathcal{X}$ at each time t is

$$r(x, u) = -c_P \cdot \mathbf{1}_P(u) - c_I \cdot \mathbf{1}_I(x),$$

where the cooperative objective J is obtained by talking the average reward over all agents and summing up over all time points. Here, c_P and c_I denote the constant costs of protecting oneself and being infected, respectively. In our experiments from the main text, the chosen parameter values are $\mu_0(I) = 0.4, \mu_0(S) = 0.6, T = 50, \rho_I = 0.4, \rho_R = 0.1, c_P = 0.5$, and $c_I = 1$.

Susceptible-Infected-Recovered (SIR). In the SIR model, we extend the state space from the SIS by the recovered state R and obtain $\mathcal{X} := \{S, I, R\}$. As only infected agents can recover, the transition dynamics of the SIS model are modified by

$$P^k(R \mid I, \bar{P}, G) = P^k(R \mid I, P, G) = \rho_R$$

and

$$P^k(R \mid R, \bar{P}, G) = P^k(R \mid R, P, G) = 1,$$

to formalize that recovered agents cannot become susceptible or infected again. The rewards and hence objective remain the same as in the SIS model. In the experiments, we set the parameter values $\mu_0(I) = 0.1, \mu_0(S) = 0.9, T = 50, \rho_I = 0.1, \rho_R = 0.02, c_P = 0.25$, and $c_I = 1$.

Graph coloring (Color). In this problem, the state space consists of five colors $\mathcal{X} := \{x_1, x_2, x_3, x_4, x_5\}$ allocated on a circle. Agents can move from the current color to the next color on the left (ℓ), to the next one on the right (r), or stay at their current color (s) such that the action space is $\mathcal{U} := \{\ell, r, s\}$. The group of agents is also supposed to come close to a target distribution $\nu \in \mathcal{P}(\mathcal{X})$. To keep notations manageable, we make the auxiliary definition

$$\tilde{G}_k := \min(1, G^2 \cdot \rho_d \cdot \exp(-2/k)),$$

where $\rho_d > 0$ is a constant noise factor. The following three matrices specify the transition dynamics, where the row is the current agent color and the column is the next agent color:

$$P^k(\cdot \mid \cdot, \ell, G) = \begin{pmatrix} \tilde{G}_k(x_1)/2 & 0 & 0 & \tilde{G}_k(x_1)/2 & 1 - \tilde{G}_k(x_1) \\ 1 - \tilde{G}_k(x_2) & \tilde{G}_k(x_2)/2 & 0 & 0 & \tilde{G}_k(x_2)/2 \\ \tilde{G}_k(x_3)/2 & 1 - \tilde{G}_k(x_3) & \tilde{G}_k(x_3)/2 & 0 & 0 \\ 0 & \tilde{G}_k(x_4)/2 & 1 - \tilde{G}_k(x_4) & \tilde{G}_k(x_4)/2 & 0 \\ 0 & 0 & \tilde{G}_k(x_5)/2 & 1 - \tilde{G}_k(x_5) & \tilde{G}_k(x_5)/2 \end{pmatrix}$$

and

$$P^k(\cdot \mid \cdot, s, G) = \begin{pmatrix} 1 - \tilde{G}_k(x_1) & \tilde{G}_k(x_1)/2 & 0 & 0 & \tilde{G}_k(x_1)/2 \\ \tilde{G}_k(x_2)/2 & 1 - \tilde{G}_k(x_2) & \tilde{G}_k(x_2)/2 & 0 & 0 \\ 0 & \tilde{G}_k(x_3)/2 & 1 - \tilde{G}_k(x_3) & \tilde{G}_k(x_3)/2 & 0 \\ 0 & 0 & \tilde{G}_k(x_4)/2 & 1 - \tilde{G}_k(x_4) & \tilde{G}_k(x_4)/2 \\ \tilde{G}_k(x_5)/2 & 0 & 0 & \tilde{G}_k(x_5)/2 & 1 - \tilde{G}_k(x_5) \end{pmatrix}$$

and

$$P^k(\cdot \mid \cdot, r, G) = \begin{pmatrix} \tilde{G}_k(x_1)/2 & 1 - \tilde{G}_k(x_1) & \tilde{G}_k(x_1)/2 & 0 & 0 \\ 0 & \tilde{G}_k(x_2)/2 & 1 - \tilde{G}_k(x_2) & \tilde{G}_k(x_2)/2 & 0 \\ 0 & 0 & \tilde{G}_k(x_3)/2 & 1 - \tilde{G}_k(x_3) & \tilde{G}_k(x_3)/2 \\ \tilde{G}_k(x_4)/2 & 0 & 0 & \tilde{G}_k(x_4)/2 & 1 - \tilde{G}_k(x_4) \\ 1 - \tilde{G}_k(x_5) & \tilde{G}_k(x_5)/2 & 0 & 0 & \tilde{G}_k(x_5)/2 \end{pmatrix}.$$

The reward in our graph coloring model is defined as

$$r(x_j, u, G) := -(\mathbf{1}_\ell(u) + \mathbf{1}_r(u)) \cdot c_m - (G(x_{j-1}) + G(x_{j+1})) \cdot c_d - \sum_{i=1}^{5} |\mu(x_i) - \nu(x_i)| \cdot c_\nu,$$

where $c_m, c_d, c_\nu > 0$ are the costs of moving, having neighbors with neighboring colors, and deviating from the target distribution ν, respectively. In our experiments, we choose the parameters $\mu_0 = (1, 0, 0, 0, 0), \nu = (0.1, 0.2, 0.4, 0.2, 0.1), T = 20, \rho_d = 0.9, c_m = 0.1, c_d = 0.5$, and $c_\nu = 1$.

Rumor. The state space $\mathcal{X} := \{I, A\}$ in the rumor model consists of the state A where an agent is aware of a rumor and state I where the agent does not know the rumor and is therefore ignorant of the rumor. Agents either spread the rumor S or decide not to do so $\bar{S}$ which results in the action space $\mathcal{U} := \{S, \bar{S}\}$. Since the rumor spreading probability increases with the number of aware numbers who decide to spread the rumor, we work with the extended state space $\mathcal{X}' := \mathcal{X} \cup (\mathcal{X} \times \mathcal{U})$. Then, the transition dynamics are

$$P^k((A, u) \mid A, u, G) = P^k(A \mid (A, u), u, G) = 1, \quad \forall u \in \mathcal{U}, G \in \mathcal{G}^k, k \in \mathbb{N}$$

meaning that aware agents remain aware, and furthermore

$$P^k((I, u) \mid I, u, G) = 1, \quad \forall u \in \mathcal{U}, G \in \mathcal{G}^k, k \in \mathbb{N}$$

and

$$P^k(A \mid (I, u), u, G) = \min\left(1, \rho_A \cdot G((A, S)) \cdot \left(\frac{2}{1 + \exp(-k/2)} - 1\right)\right)$$
$$P^k(I \mid (I, u), u, G) = 1 - P^k(A \mid (I, u), u, G).$$

To complete the rumor model, the reward is given by

$$r(x, u, G) = \mathbf{1}_{(A,S)}(x) \cdot \left(r_S \cdot G((I, S)) + r_S \cdot G((I, \bar{S})) - c_S \cdot G((A, S)) - c_S \cdot G((A, \bar{S}))\right)$$

for each agent, where we obtain the overall objective by averaging over the individual rewards. In our experiments, the parameters are chosen as $\mu_0(A) = 0.1, \mu_0(I) = 0.9, T = 50, \rho_A = 0.3, c_S = 16$, and $r_S = 4$.