Near-optimal decentralized algorithms for network dynamic optimization

Judy Gan YGAN23@GSB.COLUMBIA.EDU
Yash Kanoria YKANORIA@COLUMBIA.EDU
Decision, Risk, and Operations division, Columbia Business School

Xuan Zhang XZ2569@COLUMBIA.EDU
Department of Industrial Engineering and Operations Research, Columbia University

Abstract

We study dynamic decision-making problems in networks under stochastic uncertainty about future payoffs. The network has a bounded degree, and each node takes a discrete decision at each period, leading to a per-period payoff that is a sum of three parts: node rewards for individual node decisions, temporal interactions between individual node decisions from the current and previous periods, and spatial interactions between decisions from pairs of neighboring nodes. The objective is to maximize the expected total payoffs over a finite horizon. We propose a *decentralized* algorithm whose computational requirement is linear in the graph size and planning horizon, and characterize sufficient conditions under which our decentralized algorithm achieves near optimality compared to the centralized global optimal. The class of decentralized algorithms is parameterized by locality parameter L. An L-local algorithm makes its decision at each node v based on current and (simulated) future payoffs only up to L periods ahead, and only in an L-radius neighborhood around v. Given any permitted error $\epsilon > 0$, with $L = O(\log(1/\epsilon))$, we show that L-local algorithm has an average per-node-per-period optimality loss of up to ϵ when temporal and spatial interactions are relatively small compared to the randomness in the node rewards and the graph degree.

1. Introduction

Many real-world contexts call for dynamic decision making in networks with uncertainty about the future: At each period, a decision is made at each node in the network and a central planner aims to maximize the total payoff across the network. Examples of such settings include influence maximization in social networks [5, 34], multi-product pricing on product networks [8, 10], and logistics planning on transportation networks [13, 17]. In all these settings, the resulting payoffs arise from both individual decisions at each node and interactions among neighbors on the network. Moreover, when the decisions are made repeatedly over time, the current decision at each node influences the future payoff at the node.

Besides the goal of maximizing payoffs in such contexts, it is desirable to have decision rules that are "simple" in various ways such as computational efficiency, potential to be computed in a distributed manner, interpretability, and robustness to model misspecification. In a networked optimization setting, an attractive class of algorithms is *decentralized* algorithms which obtain the decisions of individual nodes based solely on information from the "nearby" part of the network [31]. Motivated by the ubiquitous setting of dynamic decision-making on networks and the practicality of decentralized algorithms, we aim to answer the following research question:

Can decentralized algorithms be near-optimal in terms of maximizing collective rewards on networks under stochastic uncertainty about the future?

In this paper, we propose a benchmark model of dynamic decision-making on bounded degree graphs with the global payoff consisting of: 1) per-period *individual* node rewards, which are random functions over individual decisions; 2) per-period *spatial* interactions between neighboring nodes, which are random functions of pairs of decisions; 3) *temporal* interactions for each individual node between consecutive periods, which are also random functions of pairs of decisions.

Contribution. We show that when temporal and spatial interactions are small with respect to the randomness in node rewards and the graph degree, we can construct a simple class of global near-optimal decentralized algorithms, which has computation requirement *linear* in the network size and the time horizon. For each focal node v, our algorithm simulates future reward realizations up to L periods ahead for a L-radius local neighborhood of the focal node, where L is the locality parameter determined based on desired precision. Our algorithm then solve the network optimization problem on the local graph and treats only the decision at the focal node as final. To the best of our knowledge, this is the first result on establishing sufficient conditions for approximating *global optimal solutions* via decentralized algorithms for network dynamic optimization problems under stochastic uncertainty.

Our work is inspired by the literature on *correlation decay* for static networks [21, 23], which studies how the effects of decisions at the boundary of a graph propagate towards the focal node. The correlation decay property has only been previously studied in *static* settings where reward functions consisting of node rewards and pairwise spatial interactions. However, when generalizing to dynamic setting, the value-to-go functions contain interactions between groups of nodes which are not neighbors of each other. Due to such *interaction-at-a-distance*, previous technical machinery used to establish correlation decay in static networks does not generalize to our dynamic setting. We develop a novel machinery to establish correlation decay in dynamic decision-making settings (outlined in Appendix C), which handles the interaction-at-a-distance phenomenon.

Related work. Our paper contributes to the following related research areas: (1) Correlation decay for decentralized algorithms; (2) Dynamic optimization in networks and (3) Multi-agent reinforcement learning. We include a detailed section on related literature in the Appendix B.

Notation and terminology. We denote our underlying graph as $G = (V, E)$ with node set V and edge set E. For two nodes $u, v \in G$, we let $\text{dist}_G(u, v)$ denote the length of a shortest path between u and v. If $uv \in E$, we say u is a *neighbor* of v. For any node $v \in V$, we denote by $\Gamma(v)$ its set of neighbors: $\Gamma(v) := \{u \in V : uv \in E\}$; and denote by $d_G(v) := |\Gamma(v)|$ its *degree* in G. We let d_G denote the *degree* of graph G, which is the maximum degree of nodes in graph G. For a subgraph M of G and a vector $y := \{y^v\}_{v\in V}$, we denote by y^M the subvector $\{y^v\}_{v\in V(M)}$. Let $B_G(v, R)$ denote the subgraph induced by all vertices whose distance to v is at most R. Often times, when the underlying graph is clear from context, we drop the subscript for the above notations.

Given a graph $G = (V, E)$ and time horizon $\mathcal{T}$, the spatial-temporal (ST) graph is constructed by making a clone of G for each time $t = 1, 2, \cdots, \mathcal{T}$, and connecting copies of the same node between consecutive times via edges. The ST graph distance is defined as follows: given two ST nodes (v_1, t_1) and (v_2, t_2), $\text{dist}^{\text{st}}((v_1, t_1), (v_2, t_2)) = \text{dist}(v_1, v_2) + |t_1 - t_2|$.

For a given integer $K \geq 1$, we use $[K]$ as a short-hand notation for set $\{1, 2, \cdots, K\}$. For a collection of random variables $Y_{[k]}$, $\sigma(Y_{[k]})$ denotes the smallest sigma algebra generated by $Y_{[k]}$. Given a random variable X, we write $\widetilde{X} \overset{d}{=} X$ to define $\widetilde{X}$ as an independent copy of X which follows the same distribution.

2. Model

We consider a dynamic decision network $(G = (V, E), \Phi, \mathcal{T}, \mathcal{A}, x_0)$ with future stochastic uncertainty. Here, G is the underlying undirected graph where individual decisions are made at each node. Φ denotes the joint stochastic reward functions over the graph G during the planning horizon. We consider a discrete-time model from time 0 to the planning horizon $\mathcal{T}$. We denote by $\mathcal{A}$ the discrete action set that the decision of each node must be chosen from. The initial decision vector taken on the network is given and is denoted by $x_0 \in \mathcal{A}^{|V|}$. The global objective is to maximize the collective payoff from the entire graph over the time horizon. At time t, the single-period reward is the sum of three types of (random) reward functions:

- **Node reward**: Each node $v \in V$ earns a random reward $\Phi_t^v(x_t^v) : \mathcal{A} \to \mathbb{R}$, which depends on its decision x_t^v at time t.
- **Temporal interaction reward**: Each node $v \in V$ at each time period t is associated with a random reward function $\Phi_{t-1,t}^v(x_{t-1}^v, x_t^v) : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$, which capture how consecutive decisions at node v interact with each other.
- **Spatial interaction reward**: Each edge $uv \in E$ at each time period t is associated with a random reward function $\Phi_t^{u,v}(x_t^u, x_t^v) : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$, which capture how neighboring nodes interact with each other at time t.

Collectively, we let $\Phi := \{\{\Phi_t^v\}_{t\in[\mathcal{T}],v\in V}, \{\Phi_{t-1,t}^v\}_{t\in[\mathcal{T}],v\in V}, \{\Phi_t^{u,v}\}_{uv\in E}\}$ denote joint random reward functions. Given any subgraph M, let $\Phi_t^M := \{\{\Phi_t^v\}_{v\in M}, \{\Phi_{t-1,t}^v\}_{v\in M}, \{\Phi_t^{u,v}\}_{uv\in M}\}$. At each time $t \in [\mathcal{T}]$, node $v \in V$ makes a decision $x_t^v \in \mathcal{A} := \{0, 1, \cdots, |\mathcal{A}|-1\}$. These reward functions are endowed with a probabilistic structure: their function values are assumed to follow known distributions. The random functions $\Phi_t^{\text{node}} := \{\Phi_t^v\}_{v\in V}$ and $\Phi_t^{\text{inter}} := \{\{\Phi_t^{u,v}\}_{uv\in E}, \{\Phi_{t-1,t}^v\}_{v\in V}\}$ are realized only at the beginning of time period t. We denote the realized reward functions as $\{\phi_t^v\}_{v\in V}$, $\{\phi_t^{u,v}\}_{uv\in E}$, and $\{\phi_{t-1,t}^v\}_{v\in V}$. Moreover, we denote the reward distribution and realization at time t collectively by Φ_t and ϕ_t, respectively.

We call $x_t := \{x_t^v\}_{v\in V}$ a *decision vector* at time period t. At each time period t, a decision vector x_t must be chosen after observing ϕ_t. We illustrate the dynamics under our model through an example in Figure 2. Given realized reward functions ϕ_t at time t, and decision vectors x_{t-1}, x_t, the *single-period reward* collected at period t is

$$f_t(x_t; x_{t-1}, \phi_t) := \sum_{v\in V} \phi_{t-1,t}^v(x_{t-1}^v, x_t^v) + \sum_{v\in V} \phi_t^v(x_t^v) + \sum_{uv\in E} \phi_t^{u,v}(x_t^u, x_t^v)\,. \tag{1}$$

The overall goal is to construct a dynamic decision-making policy x_t, which is adapted to the available information up to time t, i.e., $x_t \in \sigma(x_0, x_{[t-1]}, \Phi_{[t]})$, that maximizes the expected collected rewards over the time horizon: $\mathcal{R} := \mathbb{E}_\Phi \left[\sum_{t=1}^{\mathcal{T}} f_t(x_t; x_{t-1}, \Phi_t) \right]$.

Following the modeling convention on online stochastic optimization [6, 9], we assume that Φ_t is independent of past reward functions $\{\phi_{[t-1]}\}$. At period t, we observe the previous decision x_{t-1} and reward realization at time t, i.e., ϕ_t. By the principle of optimality, the optimal $x_t(x_{t-1}, \phi_t)$ maximizes the *realized value-to-go* function: $\mathrm{RV}_{t-1}(x_t; x_{t-1}, \phi_t) := f_t(x_t; x_{t-1}, \phi_t) + V_t(x_t; \phi_t)$, where the *expected value-to-go function* $V_t(x_t)$ is recursively defined as

$$V_t(x_t; \phi_t) := \mathop{\mathbb{E}}_{\Phi_{t+1}} \Big[\max_{x_{t+1}} \mathrm{RV}_t(x_{t+1}; x_t, \Phi_{t+1}) \Big], \tag{2}$$

with $V_{\mathcal{T}}(x_{\mathcal{T}}) = 0$ at the end of the horizon. We denote by $x^* := \{x_t^*\}_{1\le t\le \mathcal{T}}$ the optimal solution of (2) and denote the optimal expected global reward as $\mathcal{R}^* := \mathbb{E}_\Phi\left[\sum_{t=1}^{\mathcal{T}} f_t(x_t^*; x_{t-1}^*, \Phi_t)\right]$. For any *adaptive algorithm* which makes decisions $\text{Alg}_t \in \sigma(x_0, x_{[t-1]}, \Phi_{[t]})$ at time t, we define the expected total rewards under Alg as $\mathcal{R}(\text{Alg}) := \mathbb{E}_\Phi\left[\sum_{t=1}^{\mathcal{T}} f_t(\text{Alg}_t; \text{Alg}_{t-1}, \Phi_t)\right]$, with $\text{Alg}_0 = x_0$.

3. Main Results and Algorithms

Assumption 1 *For some constants $C_{\text{node}}, g, c_{\text{time}}, c_{\text{edge}} \in (0,\infty)$, the distributions of reward functions $\{\Phi_t\}_{t\in\mathcal{T}}$ satisfy the following:*

- *For every $v \in V$ and $t \in [\mathcal{T}]$, $\sup_{a\in\mathcal{A}} |\Phi_t^v(a)| \le C_{\text{node}}$.*
- *For every $v \in V$ and $t \in [\mathcal{T}]$, Φ_t are independent of past decisions $x_{[t-1]}$ and past reward realizations $\phi_{[t-1]}$.*
- *There exists a constant $g > 0$ such that for any $v \in V$, $t \in [\mathcal{T}]$, decisions $a \neq a' \in \mathcal{A}$, given Φ_t^{inter} and $\{\Phi_t^u\}_{u\neq v}$,*

$$\mathbb{P}(\Phi_t^v(a) - \Phi_t^v(a') \in [b_1, b_2) \mid \Phi_t^{\text{inter}}, \{\Phi_t^u\}_{u\neq v}) \le g(b_2 - b_1), \text{for any } b_1 < b_2.$$

- *With probability 1, for any $v \in V$, $(u,v) \in E$, $t \in [\mathcal{T}]$ and decisions $a \neq a' \in \mathcal{A}$, the temporal interaction $\Phi_{t-1,t}^v(a,a') \in [-c_{\text{time}}, c_{\text{time}}]$, and the edge interaction $\Phi_t^{u,v}(a,a') \in [-c_{\text{edge}}, c_{\text{edge}}]$. Moreover, we require the constants to satisfy*

$$\rho := 4g(dc_{\text{edge}} + 2c_{\text{time}}) \le \frac{1}{2(d+2)}.$$

3.1. Main theorem

Definition 1 *An algorithm for the dynamic decision network $(G, \Phi, \mathcal{T}, \mathcal{A}, x_0)$ is said to be an L-*local algorithm *if the decision of node v at time t only relies on the local information up to its L-radius neighborhood in the ST graph, i.e., $x_t^v \in \sigma(x_{t-1}^{B(v,L)}, \Phi_t^{B(v,L)}, \widetilde{\Phi}_{t+1}^{B(v,L)} \cdots, \widetilde{\Phi}_{t+L}^{B(v,L)})$, where each $\widetilde{\Phi}_{t'}^{B(v,L)} \overset{d}{=} \Phi_{t'}^{B(v,L)}$ for $t+1 \le t' \le t+L$ is an independent copy of $\Phi_{t'}^{B(v,L)}$.*

Note that L is a parameter in both spatial and temporal dimension. In an L-local algorithm, although the future realizations of reward functions are not revealed, L-step simulations are used to approximate the future value-to-go functions in the corresponding local neighborhood.

Definition 2 *Consider a dynamic decision network $(G, \Phi, \mathcal{T}, \mathcal{A}, x_0)$. An algorithm* Alg *is an ϵ(-*additive)-approximation *algorithm if $\mathcal{R}^* - \mathcal{R}(\text{Alg}) \le |V|\mathcal{T}\epsilon$, where $\mathcal{R}^*$ is the optimal payoff, and $\mathcal{R}(\text{Alg})$ is the payoff collected by* Alg.

Note that there is a $|V|\mathcal{T}$ factor in the loss permitted because the total reward scales up linearly with the number of nodes $|V|$ times the time horizon $\mathcal{T}$; in other words, we permit an average per-node-per-period loss of up to ϵ.

We introduce a model parameter C which is the largest possible change in total rewards when one node changes its decision at one time. For any $a, a' \in \mathcal{A}$, changing from $x_t^v = a$ to $x_t^v = a'$ can cause at most $2 \cdot C_{\text{node}}$ difference in the node reward, $d \cdot 2c_{\text{edge}}$ difference in the edge rewards, and $2 \cdot 2c_{\text{time}}$ difference in the temporal rewards. Hence, we define the constant

$$C := 2C_{\text{node}} + 2dc_{\text{edge}} + 4c_{\text{time}}. \tag{3}$$

Theorem 3 *Consider a dynamic decision network $(G, \Phi, \mathcal{T}, \mathcal{A}, x_0)$ where underlying graph G has degree $d \geq 2$. Suppose the reward functions Φ satisfy Assumption 1. Then, given any $\epsilon > 0$ and $L \triangleq \lfloor \log_2 \frac{4C}{\epsilon} \rfloor$, we can construct an L-local algorithm for the dynamic decision network problem that is an ϵ-approximation algorithm.*

The main contribution of Theorem 3 is to establish the global near-optimality property of a *decentralized* algorithm under Assumption 1. Our local algorithm (presented in Algorithm 1) has the advantage of being computationally efficient: the computational requirement of Algorithm 1 is $O(|V|\mathcal{T}e^{\text{poly}(\frac{1}{\epsilon})})$, where the dependence on model parameters d, g, C and $|\mathcal{A}|$ is suppressed in the $O(\cdot)$ notation. The proof of Theorem 3 and the details on the computational requirement are presented in Appendix C and Appendix G, respectively.

3.2. Local Algorithm

In this section, we present our local algorithm. Given $t \in [\mathcal{T}]$, the global decision problem is

$$\max_{x_t} \; f_t(x_t; x_{t-1}, \phi_t) + V_t(x_t). \tag{4}$$

The algorithm, outlined in Algorithm 1, determines the decision of each node by solving a *decentralized* version of (4). For each node v, the local algorithm utilizes all available reward information from its local neighborhood $B(v, L)$ and fixes the decision of each boundary node $u \in B(v, L) \setminus B(v, L-1)$ as the default decision 0. We define $f_t^L(x_t; x_{t-1}, \phi_t) := \sum_{u \in B(v,L)} (\phi_t^u(x_t^u) + \phi_{t-1,t}^u(x_{t-1}^u, x_t^u)) + \sum_{uu' \in B(v,L)} \phi_t^{u,u'}(x_t^u, x_t^{u'})$ as the single-period payoff on $B(v, L)$ and $V_t^L(x_t) := \mathbb{E}_{\Phi_{t+1}}[\max_{x_{t+1}} f_{t+1}^L(x_{t+1}; x_t, \phi_{t+1}) + V_{t+1}^L(x_{t+1})]$ (with terminal condition $V_{\min\{t+L,\mathcal{T}\}}^L = 0$) as the expected value-to-go function up to L-step look-ahead on $B(v, L)$. We omit the dependency on v as the focal node for the decentralized algorithm is usually clear from context. In addition, we denote by $\widehat{V}_t^{L,n}(x_t)$ the sample average estimate of $V_t^L(x_t)$ by simulating independent samples of $\Phi_{t+1}, \Phi_{t+2}, \cdots, \Phi_{\min(\mathcal{T},t+L)}$.

4. Numerical Experiments

To test the performance of our local algorithm and the presence (or absence) of correlation decay while varying interaction strength, we conducted a simulation experiment. We summarize the simulation environment and main findings here and defer the details to Appendix I.

In our experiment, we first generate multiple dynamic decision network, parameterized by interaction strength c in both the temporal and spatial dimensions. These decision networks share all other components (e.g., a random 3-regular graph as the underlying graph, binary action set $\mathcal{A} = \{0, 1\}$, uniform distribution on $[-1, 1]$ as the node reward distribution when taking action 1) so that the differences in performance can be solely attributed to the interaction strength. For each decision network, multiple instances are generated by sampling the first period node rewards (allowing us to compute confidence intervals for the performance). Then, for each instance, we compute

Algorithm 1 Obtain a near-optimal solution to the decision problem (4) at time t.

```
Input: decision network $(G, \Phi, \mathcal{T}, \mathcal{A}, \mathrm{Alg}_{t-1})$, realized reward function $\phi_t$, precision level $\epsilon$.
Output: a near-optimal solution $\mathrm{Alg}_t$ for the problem in (4).
set the locality parameter $L = \lfloor \log_2 \frac{4C}{\epsilon} \rfloor$ and sample size $n = O((\frac{1}{\epsilon})^{2\log_2 d})$
for all $v \in V$ do
  restrict to subgraph $B(v, L)$
  let $y_t \in \mathcal{A}^{B(v,L)}$ be an optimal solution to the following problem
      $$\begin{aligned} \widehat{\mathrm{RV}}_{t-1}^{L,n}(x_{t-1};\phi_t) \coloneqq \max_{x_t} \quad & f_t^L(x_t; x_{t-1}, \phi_t) + \widehat{V}_t^{L,n}(x_t) \\ \text{s.t.} \quad & x_t^u = 0, \quad \text{if } \mathrm{dist}(v,u) = L \end{aligned} \tag{5}$$
  where $\widehat{V}_t^{L,n}(\cdot)$ is an estimate of $V_t(\cdot)$ defined recursively in function $\widehat{V}_\tau^{L,n}(\cdot)$ for $t \le \tau \le t + L$
  set $\mathrm{Alg}_t^v = y_t^v$
end for
```

```
function $\widehat{V}_\tau^{L,n}(x_\tau)$                                ▷ Input: $L$, $n$, $v$, $\tau$.
  if $\tau = \min\{t + L, \mathcal{T}\}$ then
    set $\widehat{V}_\tau^{L,n}(x_\tau) = 0$ for any $x_\tau \in \mathcal{A}^{B(v,L)}$
  else
    sample $\{\phi_{\tau+1}^{(s)}\}_{s\in[n]}$ independently from $\Phi_{\tau+1}$
    for any $x_\tau \in \mathcal{A}^{B(v,L)}$, compute $\widehat{V}_\tau^{L,n}(x_\tau) \coloneqq \frac{1}{n}\sum_{s=1}^n \widehat{\mathrm{RV}}_\tau^{L,n}(x_\tau; \phi_{\tau+1}^{(s)})$ where the summand $\widehat{\mathrm{RV}}_\tau^{L,n}(x_\tau; \phi_{\tau+1}^{(s)})$ is defined as in Equation (5) with $t = \tau + 1$.
  end if
end function
```

several solutions: one being the solution to the global optimization problem, and the others obtained by our local algorithms with different choices of the locality parameter. We formulate each global or local network optimization problem as a Mixed Integer Program (MIP) and solve it through Gurobi [22]. Lastly, for each solution, we compute its *relative payoff*, which is the ratio between the total payoff under the local solution to that under the global optimal solution.

We summarize the results in Figure 1, with one plot showing simply the relative payoffs (the higher, the better) and the other showing the relative payoff gaps (1 − relative payoff) in log scale (the lower, the better). Our experimental results corroborate our theoretical finding in Theorem 3; when the interaction strength is small (or even medium-sized), the error in payoffs is seen to decay exponentially in the locality parameter. This is especially prominent on the second plot in Figure 1. In addition, we observe that when the interaction strength is large $c \geq 0.4$, the optimality gap ceases to improve (and remains non-trivial) for locality parameter values larger than 4.

5. Concluding Remarks

We introduced a benchmark model of a dynamic optimization problem in networks where the global payoff includes spatial interactions and temporal interactions as well as node rewards. At each time

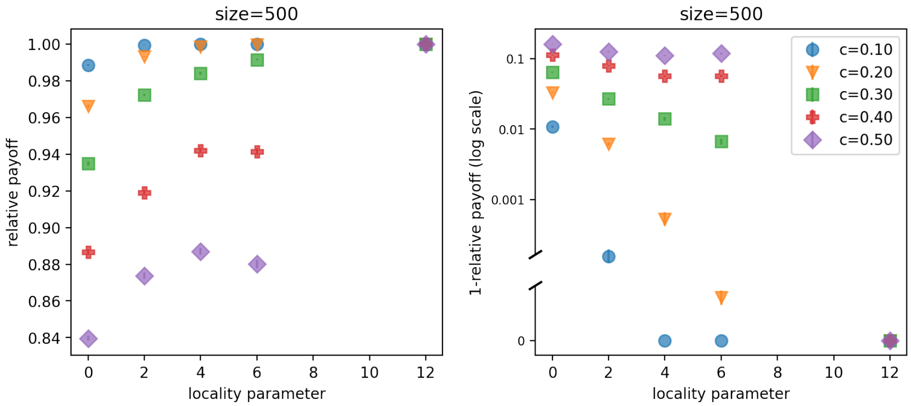

Figure 1: Compare payoffs from solutions under local algorithms and the global optimal solution. The experiment controls the sampling loss and thus, the loss in total reward is solely due to the locality loss. The vertical line on each data point represents its 95% confidence interval. The global optimal solutions correspond to the local solutions when the locality parameter equals 12, which is the *diameter* of the underlying graph.

step, a decision vector has to be chosen before observing the realizations of future rewards. We propose a class of (computationally efficient) decentralized algorithms – which make decisions only using information about the nearby part of the network. We showed that if spatial and temporal interactions are relatively weak, then the decision maker can employ decentralized algorithms, which essentially optimize on small subgraphs of the network, to first estimate the value-to-go and then to obtain near-optimal decisions.

References

[1] Mohammad Akbarpour, Suraj Malladi, and Amin Saberi. Diffusion, seeding, and the value of network information. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 641–641, 2018.

[2] Ross Anderson, Itai Ashlagi, David Gamarnik, and Yash Kanoria. Efficient dynamic barter exchange. *Operations Research*, 65(6):1446–1459, 2017.

[3] Ali Aouad and Ömer Saritaç. Dynamic stochastic matching under limited time. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 789–790, 2020.

[4] Antar Bandyopadhyay and David Gamarnik. Counting without sampling: Asymptotics of the log-partition function for certain statistical physics models. *Random Structures & Algorithms*, 33(4):452–479, 2008.

[5] Suman Banerjee, Mamata Jenamani, and Dilip Kumar Pratihar. A survey on influence maximization in a social network. *Knowledge and Information Systems*, 2020. ISSN 02193116. doi: 10.1007/s10115-020-01461-4.

[6] Russell Bent and Pascal Van Hentenryck. Online stochastic and robust optimization. pages 286–300, 01 2004. ISBN 978-3-540-24087-7. doi: 10.1007/978-3-540-30502-6_21.

[7] Omar Besbes, Yonatan Gur, and Assaf Zeevi. Optimization in online content recommendation services: Beyond click-through rates. *Manufacturing and Service Operations Management*, 2016. ISSN 15265498. doi: 10.1287/msom.2015.0548.

[8] Ozan Candogan, Kostas Bimpikis, and Asuman Ozdaglar. Optimal pricing in networks with externalities. *Operations Research*, 60(4):883–905, 2012.

[9] Xuanyu Cao, Junshan Zhang, and H. Vincent Poor. Online stochastic optimization with time-varying distributions. *IEEE Transactions on Automatic Control*, 66(4):1840–1847, 2021. doi: 10.1109/TAC.2020.2996178.

[10] Felipe Caro and Jérémie Gallien. Clearance pricing optimization for a fast-fashion retailer. *Operations Research*, 60(6):1404–1422, 2012.

[11] Zongchen Chen, Kuikui Liu, and Eric Vigoda. Rapid Mixing of Glauber Dynamics up to Uniqueness via Contraction. 2020. URL http://arxiv.org/abs/2004.09083.

[12] Natalie Collina, Nicole Immorlica, Kevin Leyton-Brown, Brendan Lucier, and Neil Newman. Dynamic weighted matching with heterogeneous arrival and departure rates. In *International Conference on Web and Internet Economics*, pages 17–30. Springer, 2020.

[13] Aashwinikumar Devari, Alexander G. Nikolaev, and Qing He. Crowdsourcing the last mile delivery of online orders by exploiting the social networks of retail store customers. *Transportation Research Part E: Logistics and Transportation Review*, 2017. ISSN 13665545. doi: 10.1016/j.tre.2017.06.011.

[14] Jian Ding, Allan Sly, and Nike Sun. Proof of the satisfiability conjecture for large k. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 59–68, 2015.

[15] Roland L Dobrushin. Prescribing a system of random variables by conditional distributions. *Theory of Probability & Its Applications*, 15(3):458–486, 1970.

[16] Pablo Fajgelbaum, Amit Khandelwal, Wookun Kim, Cristiano Mantovani, and Edouard Schaal. Optimal lockdown in a commuting network. Technical report, National Bureau of Economic Research, 2020.

[17] Soraya Fatehi and Michael R. Wagner. Crowdsourcing Last-Mile Deliveries. *Manufacturing & Service Operations Management*, 2021. ISSN 1523-4614. doi: 10.1287/msom.2021.0973.

[18] Joel Friedman. *A proof of Alon's second eigenvalue conjecture and related problems*. American Mathematical Soc., 2008.

[19] David Gamarnik and David A. Goldberg. Randomized greedy algorithms for independent sets and matchings in regular graphs: Exact results and finite girth corrections. *Combinatorics Probability and Computing*, 19(1):61–85, 2010. ISSN 09635483. doi: 10.1017/S0963548309990186.

[20] David Gamarnik and Dmitriy Katz. Sequential cavity method for computing free energy and surface pressure. *Journal of Statistical Physics*, 137(2):205, 2009.

[21] David Gamarnik, David A. Goldberg, and Theophane Weber. Correlation decay in random decision networks. *Mathematics of Operations Research*, 39(2):229–261, 2014. ISSN 15265471. doi: 10.1287/moor.2013.0609.

[22] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2022. URL https://www.gurobi.com.

[23] Azer Kerimov. A disagreement-percolation type uniqueness condition for Gibbs states in models with long-range interactions. *Journal of Statistical Mechanics: Theory and Experiment*, 2014. ISSN 17425468. doi: 10.1088/1742-5468/2014/10/P10014.

[24] Matt V Leduc, Matthew O Jackson, and Ramesh Johari. Pricing and referrals in diffusion on networks. *Games and Economic Behavior*, 104:568–594, 2017.

[25] Yiheng Lin, Guannan Qu, Longbo Huang, and Adam Wierman. Distributed reinforcement learning in multi-agent networked systems. *arXiv preprint arXiv:2006.06555*, 2020.

[26] Yiheng Lin, Judy Gan, Guannan Qu, Yash Kanoria, and Adam Wierman. Decentralized online convex optimization in networked systems, 2022. URL https://arxiv.org/abs/2207.05950.

[27] Vahideh Manshadi, Sidhant Misra, and Scott Rodilitz. Diffusion in random networks: Impact of degree distribution. *Operations Research*, 68(6):1722–1741, 2020.

[28] Andrea Montanari. Optimization of the sherrington-kirkpatrick hamiltonian. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1417–1433. IEEE, 2019.

[29] Guannan Qu, Yiheng Lin, Adam Wierman, and Na Li. Scalable Multi-Agent Reinforcement Learning for Networked Systems with Average Reward. 2020. URL http://arxiv.org/abs/2006.06626.

[30] Balasubramanian Sivan. *Prior robust optimization*. PhD thesis, The University of Wisconsin-Madison, 2013.

[31] Jukka Suomela. Survey of local algorithms. *ACM Computing Surveys (CSUR)*, 45(2):1–40, 2013.

[32] Kalyan T Talluri and Garrett J Van Ryzin. *The theory and practice of revenue management*, volume 68. Springer Science & Business Media, 2006.

[33] Sekhar C. Tatikonda and Michael I. Jordan. Loopy belief propagation and gibbs measures. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, UAI'02, page 493–500, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. ISBN 1558608974.

[34] Guangmo Tong, Weili Wu, Shaojie Tang, and Ding Zhu Du. Adaptive Influence Maximization in Dynamic Social Networks. *IEEE/ACM Transactions on Networking*, 2017. ISSN 10636692. doi: 10.1109/TNET.2016.2563397.

[35] Salil P Vadhan et al. *Pseudorandomness*, volume 7. Now Delft, 2012.

[36] Dror Weitz. Counting independent sets up to the tree threshold. In *Proceedings of the Annual ACM Symposium on Theory of Computing*, 2006. ISBN 1595931341. doi: 10.1145/1132516.1132538.

Appendix A. Figures

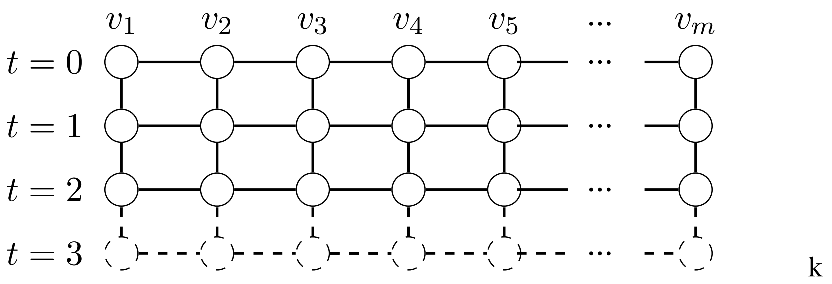

Figure 2: Decision dynamics. An example of a dynamic decision network $(G = (V, E), \Phi, \mathcal{T}, \mathcal{A}, x_0)$ with G being a line graph, $V = \{v_1, v_2, \cdots, v_m\}$, $E = \{v_{i-1}v_i : i \in \{2, 3, \cdots, m\}\}$, and $\mathcal{T} = 3$. At time $t = 2$, with the previous decision vector x_1, realized rewards ϕ_1, ϕ_2 (represented by solid lines and circles), and unrealized reward Φ_3 (represented by dotted lines and circles), decision vector x_2 needs to be chosen.

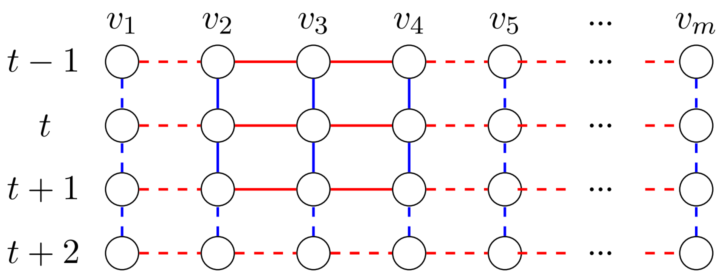

Figure 3: L-local algorithm. Illustration of Algorithm 1 for focal node v_3 at time t with $L = 1$, for the same underlying line graph G.

Appendix B. Related Literature

Correlation decay for decentralized algorithms. Correlation decay is the cornerstone for the success of numerous decentralized algorithms for static network optimization in the literature. It was first studied in the domain of statistical physics. The seminal work of Dobrushin [15] studied graphical models (e.g., a Markov chain is a graphical model on a line graph) on infinite graphs via correlation decay methods, investigating whether the joint distribution –the Gibbs measure– is uniquely determined by the distribution of each random variable conditional on its neighbors. Since then, the concept of correlation decay has expanded to applications beyond statistical physics [11, 14, 28, 36], including wireless communication [4, 20, 36], combinatorial optimization [19, 21], marginal inference on graphical models [33], etc. The typical regime for static decision problems under which the correlation decay property holds is when the underlying graph has a bounded degree (i.e., each node interacts with a constant number of other nodes) [19, 21, 36]. However, in our multi-period network model involving uncertainty about the future, there exists an implicit interaction between *any* pair of nodes due to the fact that their decisions for future time periods could be correlated, resulting in new challenges.

Dynamic optimization in networks. Dynamic decision making in networks under uncertainty about the future has been studied in a variety of contexts including network revenue management [32], network diffusion models [1, 24, 27], online matching under stochastic arrivals [2, 3, 12, 30], and choosing lockdown policies in a commuting network [16], to name a few. In most of these works the ideas used (e.g., shadow prices) are very different from ours. Some of these previous work [3, 7] make use of local decompositions that rely on their specific settings, or requires convexity in the reward functions [26]. We adopt a general framework, towards developing a foundational understanding regarding the sufficiency of decentralized algorithms for obtaining near optimality in dynamic stochastic optimization problems. Our paper contributes to this literature by providing an important theoretical foundation for decision-making problems on large networks: that is, even though the network interactions may be complicated and evolve over time, considering only the local neighborhood around the focal node already gives near-optimal performance when the strength of interactions are relatively weak.

Multi-agent reinforcement learning. In the setting of multi-agent reinforcement learning (MARL), computational issues are central due to large global state and decision spaces (exponential in the number of agents). A promising approach is to exploit local dependency structures (i.e., agents only interact with neighboring agents in the network). Lin et al. [25], Qu et al. [29] consider a class of MARL problems where the evolution of the state has only local dependencies, and propose a localized policy that converges to an *approximately stationary point*. In general, multi-agent reinforcement learning is a hard problem and most results focus on convergence to a stationary point if interactions are not too strong. We contribute to this literature by showing, in a special case where the state transitions are deterministic (i.e., the current state is the previous decision), *global near optimality* of a local policy under weak interactions. Our work may serve as a starting point for developing an understanding of sufficient conditions for achieving global near optimality in network MDP settings.

Appendix C. Proof Outline of Theorem 3

We first provide some explanation and justification for Assumption 1. 1) The first assumption gives us a uniform bound for the change in the global reward when any single node switches its single-period decision. This assumption ensures that there is no single node whose decision at a certain period has a dominant impact on the global reward. 2) The second assumption demands independence of the rewards across time. E.g., considering the following the reward functions where $\mathcal{A} = \{0,1\}$, $\Phi_t^v(0) = 0$, $\Phi_t^v(1) = 1 + \epsilon_t^v$ where $\epsilon_t^v \sim N(0,1)$, $\Phi_{t-1,t}^v(0,0) = \Phi_{t-1,t}^v(1,1) = c$, $\Phi_{t-1,t}^v(0,1) = \Phi_{t-1,t}^v(1,0) = 0$, $\Phi_t^{u,v}(0,0) = \Phi_t^{u,v}(1,1) = c$, and $\Phi_t^{u,v}(0,1) = \Phi_t^{u,v}(1,0) = 0$, this assumption requires the joint distributions of $\{\epsilon_t^v : v \in V\}_t$ are independent across time periods. 3) The third assumption guarantees sufficient randomness in the single node reward function at each period. E.g., considering the same reward functions, this assumption does not restrict $(\epsilon_t^v)_{v\in V}$ to be independent across nodes as long as there exists a $g > 0$ such that for any node $v \in V$, $t \in [\mathcal{T}]$ and $b_1 < b_2$, $\mathbb{P}(\epsilon_t^v \in [b_1, b_2) \mid \epsilon_t^u : u \neq v) \leq g(b_2 - b_1)$. 4) The last assumption requires that interactions are small compared to the graph degree, which is crucial for the *correlation decay* property to emerge. In Appendix H, we explicitly construct (static) decision networks with $c_{\text{edge}} = \Theta(1/d)$

which exhibit long-range correlations and show that local algorithms can perform poorly on such networks.

To establish the ϵ-approximation results in Theorem 3, we show that with high probability, Algorithm 1 takes optimal decisions in two steps. The main technical contribution is the first step, where we construct a sequence of local dynamic optimization problems with increasing local radius. We use the term *locality loss* to refer to the probability of making a suboptimal decision due to fixing the boundary nodes of the local neighborhood to the default decision 0. The second step is to bound the probability, termed the *sampling loss*, of making a suboptimal decision at the focal node as a result of using approximate (local) value-to-go functions estimated from sample averages. The second step relies on standard techniques such as Hoeffding's inequality, and we defer the details to Appendix E.1.

C.1. Bounding the Locality Loss

In this subsection, we bound the loss that is unavoidable from local decision making, even if one is able to perfectly estimate the local value-to-go functions. We define a sequence of decentralized policies $\{\pi(H)\}_{H\geq L}$, indexed by the locality parameter H. Note that a policy defines a mapping from available information so far to decisions. We use $\pi_t^v(H)$ to denote the decision of node v at time t under policy $\pi(H)$, and collectively, we use $\pi_t(H)$ to denote the decision vector at time t.

When solving for the decision at a focal node $v \in V$ and time $t \in [\mathcal{T}]$, the policy $\pi(H)$ focuses on the subgraph $B(v, H)$. It makes nodes outside of $B(v, H)$ taking the default decision 0 at any time. Along the temporal dimension, the policy $\pi(H)$ computes $V_t^H(\cdot)$, an estimate of the value-to-go function, via an H-*step look-ahead* with the terminal expected value-to-go $V^H_{\min(\mathcal{T},t+H)}(x) = 0$ for all decision vectors x. Formally, for a given focal node v at time t, $\pi(H)$ solves the following:

$$\begin{aligned} \mathrm{RV}^H_{t-1}(\pi_{t-1}(H);\phi_t) \coloneqq \max_{x_t\in\mathcal{A}^{B(v,H)}} \quad & f_t^H(x_t;\pi_{t-1}(H),\phi_t) + V_t^H(x_t) \\ \text{s.t.} \quad & x_t^u = 0 \quad \text{ if } \operatorname{dist}(v,u) = H\,. \end{aligned} \tag{6}$$

where the H-step look-ahead value-to-go $V_t^H(x_t;\phi_t)$ in the objective is defined recursively via

$$V_\tau^H(x_\tau;\phi_\tau) \coloneqq \mathop{\mathbb{E}}_{\Phi_{\tau+1}}[\mathrm{RV}_\tau^H(x_\tau;\Phi_{\tau+1})], \tag{7}$$

for $t \leq \tau \leq \min\{t+H, \mathcal{T}\}$ with terminal condition $V^H_{\min(t+H,\mathcal{T})}(x) = 0$ for any x.

Recall that $C \coloneqq 2C_{\text{node}} + 2dc_{\text{edge}} + 4c_{\text{time}}$ and x^* is the optimal decision. In this subsection, probability is over all reward distributions $\Phi = (\Phi_1, \cdots, \Phi_{\mathcal{T}})$. We write $\mathbb{P}$ as a short hand for $\mathbb{P}_\Phi$.

Proposition 4 *Given any $\epsilon > 0$, with $L = \lfloor \log_2 \frac{4C}{\epsilon} \rfloor$, we have for any $v \in V$ and $t \in [\mathcal{T}]$,*

$$\mathbb{P}(\pi_t^v(L) \neq (x_t^v)^*) \leq \epsilon/(2C).$$

Proposition 4 establishes that the probability of $\pi(L)$ making a suboptimal decision at the focal node v is exponentially small in the locality parameter L. In the remaining subsection, we outline two important lemmas that prove Proposition 4. For the following, we consider a fixed $t \in [\mathcal{T}]$ and focal node $v \in V$. We also consider a fixed value of H and compare the node decisions we obtain under policies $\pi_t(H)$ and $\pi_t(H+1)$. That is, we compare the solutions of the optimization problems in (6) when setting the locality parameter as H and $H+1$. For $t \leq \tau \leq t+H$, we let

w_τ denote the optimal solution of (6) at time τ when the locality parameter is H; and similarly, we let z_τ denote the optimal solution of (6) at time τ when the locality parameter is $H+1$. Moreover, we use w_{t-1} and z_{t-1} to denote the decision vectors $\pi_{t-1}(H)$ and $\pi_{t-1}(H+1)$, respectively. Note that $\{w_\tau\}_{t+1\le\tau\le t+H}$ and $\{z_\tau\}_{t+1\le\tau\le t+H}$ are the optimal "tentative" decisions from time $t+1$ to time $t+H$ under $\pi_t(H)$ and $\pi_t(H+1)$. That is, at time t, $\pi_t(H)$ (resp. $\pi_t(H+1)$) only executes w_t (resp. z_t) and discards the other decision vectors $\{w_{t'} : t' > t\}$ (resp. $\{z_{t'} : t' > t\}$). Since we restrict to adaptive policies, z_τ and w_τ are random variables which are measurable with respect to $\sigma(x_0, x_{[\tau-1]}, \Phi_{[\tau]})$. For convenience, we extend the definition of z_τ^u (resp. w_τ^u) to the entire network by setting $z_\tau^u = 0$ for $u \in V \setminus B(v, H+1)$ (resp. $w_\tau^u = 0$ for $u \in V \setminus B(v, H)$), and this does not change our original optimization problem in Equation (6). Recall that $\rho \coloneqq 4g(dc_{\text{edge}} + 2c_{\text{time}})$ and $\Gamma(v)$ denotes the neighbors of v.

Lemma 5 *For time $t \le \tau \le t+H$, and $u \in V$,*

$$\mathbb{P}(w_\tau^u \neq z_\tau^u, w_\tau^{\Gamma(u)} = z_\tau^{\Gamma(u)}) \le (\mathbb{P}(w_{\tau-1}^u \neq z_{\tau-1}^u) + \mathbb{P}(w_{\tau+1}^u \neq z_{\tau+1}^u))\rho.$$

We first look at a special case to get some intuitive understanding for the above lemma: if both $\mathbb{P}(w_{\tau-1}^u \neq z_{\tau-1}^u)$ and $\mathbb{P}(w_{\tau+1}^u \neq z_{\tau+1}^u)$ are equal to zero, then Lemma 5 implies $\mathbb{P}(w_\tau^u \neq z_\tau^u, w_\tau^{\Gamma(u)} = z_\tau^{\Gamma(u)}) = 0$. This reflects the fact given a ST node (τ, u) in the ST graph, if all immediate neighbors (i.e., spatial neighbors $\Gamma(u)$, temporal neighbors $(\tau-1, u)$ and $(\tau+1, u)$) take the same decisions under $\pi(H)$ and $\pi(H+1)$, then by principle of optimality, ST node (τ, u) take the same optimal decision under the above two policies. The lemma constitutes the key component of our analysis where we circumvent the challenge of analyzing dynamics with uncertainty about the future. Instead of bounding the probability of the focal node taking a suboptimal decision when k-hop neighbors ($2 \le k \le H$) in the (static) spatial graph take suboptimal decisions, we bound this probability in the ST graph since the interactions among nodes in the ST graph are easier to track. In the ST graph, a node makes a suboptimal decision only if a spatial or temporal neighbor is fixed suboptimally. The rigorous proof of Lemma 5 is quite involved. It argues that the event $(w_\tau^u \neq z_\tau^u, w_\tau^{\Gamma(u)} = z_\tau^{\Gamma(u)})$ happens only if the difference of node reward functions, i.e., $\Phi_\tau^u(w_\tau^u) - \Phi_\tau^u(z_\tau^u)$ falls in a small interval whose length is proportional to $c_{\text{time}}(\mathbb{I}\{w_{\tau-1}^u \neq z_{\tau-1}^u\} + \mathbb{E}_{\Phi_{\tau+1}}[\mathbb{I}\{w_{\tau+1}^u \neq z_{\tau+1}^u\}])$. By the third condition in Assumption 1, the probability of the event $(w_\tau^u \neq z_\tau^u, w_\tau^{\Gamma(u)} = z_\tau^{\Gamma(u)})$ is proportional to $gc_{\text{time}}(\mathbb{I}\{w_{\tau-1}^u \neq z_{\tau-1}^u\} + \mathbb{E}_{\Phi_{\tau+1}}[\mathbb{I}\{w_{\tau+1}^u \neq z_{\tau+1}^u\}])$, which further leads to the inequality in Lemma 5. We present the details in Appendix D.1.

After obtaining Lemma 5, we use induction on the ST graph distance to node (v, t) to upper bound the probability of making different node decisions under $\pi_t(H)$ and $\pi_t(H+1)$. We illustrate our proof ideas in Figure 4 and defer the proof of Lemma 6 to Appendix D.2. Let $\xi \coloneqq (d+2)\rho$.

Lemma 6 *For $t \le \tau \le t+H$ and $u \in B(v, H)$, we have*

$$\mathbb{P}_\Phi(w_\tau^u \neq z_\tau^u) \le \xi^{H+1-\text{dist}^{\text{st}}((v,t),(u,\tau))}\rho. \tag{8}$$

Then, Proposition 4 is straightforward from Lemma 6 combined with a union bound argument (details in Appendix D.2).

Appendix D. Proof details in Section C

In this section, we present the proof details which are omitted in the proof outline.

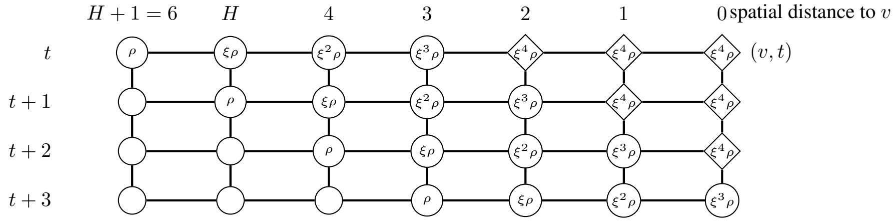

Figure 4: An example of the induction argument of Lemma 6: suppose we have proved Equation (8) for ST nodes (q, t') with $\text{dist}^{\text{st}}((v, t), (q, t')) \geq 3$ (drawn as circles) and then consider ST nodes (u, τ) with $\text{dist}^{\text{st}}((v, t), (u, \tau)) \leq 2$ (drawn as diamonds). By induction hypothesis, each of the $\leq d + 2$ ST neighbors of (u, τ) has probability $\leq \xi^{H+1-3}\rho = \xi^3\rho$ taking different decisions under $\pi(H)$ and $\pi(H + 1)$. By Lemma 5, the probability of (u, τ) taken different decisions under $\pi(H)$ and $\pi(H + 1)$ is $\leq d \cdot \xi^3\rho \cdot \mathbb{P}(z^u_\tau \neq w^u_\tau | z^{\Gamma(u)}_\tau \neq w^{\Gamma(u)}_\tau) + (\xi^3\rho + \xi^3\rho) \cdot \rho \leq \xi^4\rho$, where the last inequality is due to Assumption 1.

D.1. Proof of Lemma 5

When solving for (6), in order to compute the optimal decision vector under policy $\pi(H)$ at time step t, i.e., $\pi_t(H)$, it also needs to compute a "tentative decision rule" from time $t + 1$ to time $t + H$, which we denote them together by w_τ for $t \leq \tau \leq t + H$. Similarly, we denote the optimal decision vector at time t and "tentative decision rule" under policy $\pi(H + 1)$ as z_τ for $t \leq \tau \leq t + H$. In addition, we define $z_{t-1} \coloneqq \pi_{t-1}(H + 1), w_{t-1} \coloneqq \pi_{t-1}(H)$. Note that z_τ and w_τ are random variables, which are measurable with respect to $\sigma(x_0, x_{[\tau-1]}, \Phi_{[\tau]})$. For any node u with $\text{dist}(v, u) > H + 1$, we have defined $z^u_\tau = w^u_\tau = 0$. Thus, we consider node u such that $\text{dist}(v, u) \leq H + 1$. Define the local version of $f_\tau(\cdot)$ function concerning node u,

$$f^u_\tau(x^u; x^{\Gamma(u)}_t, x^u_{t-1}, \Phi_t) \coloneqq \Phi^u_{t-1,t}(x^u_{t-1}, x^u) + \Phi^u_t(x^u) + \sum_{q \in \Gamma(u)} \Phi^{u,q}_t(x^u_t, x^q_t).$$

Let z^{-u}_τ (resp., w^{-u}_τ) denote the actions at all nodes other than u under the vector z_τ (resp., w_τ). Since z^u_τ is the optimal solution at time τ when restricting to the subgraph $B(v, H + 1)$ and using $H + 1$-step look-ahead,

$$f^{H+1}_\tau(z^u_\tau, z^{-u}_\tau; z_{\tau-1}, \Phi_\tau) + V^{H+1}_\tau(z^u_\tau, z^{-u}_\tau) \geq f^{H+1}_\tau(w^u_\tau, z^{-u}_\tau; z_{\tau-1}, \Phi_\tau) + V^{H+1}_\tau(w^u_\tau, z^{-u}_\tau).$$

After rearranging the terms in the above inequality,

$$f^u_\tau(w^u_\tau; z^{\Gamma(u)}_\tau, z^u_{\tau-1}, \Phi_\tau) - f^u_\tau(z^u_\tau; z^{\Gamma(u)}_\tau, z^u_{\tau-1}, \Phi_\tau) \leq V^{H+1}_\tau(z^u_\tau, z^{-u}_\tau) - V^{H+1}_\tau(w^u_\tau, z^{-u}_\tau). \quad (9)$$

Similarly, by optimality of w^u_τ,

$$f^H_\tau(w^u_\tau, w^{-u}_\tau; w_{\tau-1}, \Phi_\tau) + V^H_\tau(w^u_\tau, w^{-u}_\tau) \geq f^H_\tau(z^u_\tau, w^{-u}_\tau; w_{\tau-1}, \Phi_\tau) + V^H_\tau(z^u_\tau, w^{-u}_\tau).$$

Hence, we define the following positive random variable which is effectively the optimality gap between switching from action w^u_τ to z^u_τ under policy $\pi(H)$,

$$\begin{aligned}
\Delta_\tau^u &:= \left[f_\tau^H(w_\tau^u, w_\tau^{-u}; w_{\tau-1}, \Phi_\tau) + V_\tau^H(w_\tau^u, w_\tau^{-u})\right] - \left[f_\tau^H(z_\tau^u, w_\tau^{-u}; w_{\tau-1}, \Phi_\tau) + V_\tau^H(z_\tau^u, w_\tau^{-u})\right] \\
&= \Big[f_\tau^u(w_\tau^u; w_\tau^{\Gamma(u)}, w_{\tau-1}^u, \Phi_\tau) - f_\tau^u(z_\tau^u, w_\tau^{\Gamma(u)}; w_{\tau-1}^u, \Phi_\tau)\Big] + \left[V_\tau^H(w_\tau^u, w_\tau^{-u}) - V_\tau^H(z_\tau^u, w_\tau^{-u})\right]
\end{aligned}$$

Let A_τ denote the event such that $w_\tau^u \neq z_\tau^u$ and $w_\tau^{\Gamma(u)} = z_\tau^{\Gamma(u)}$. Then,

$$\begin{aligned}
\Delta_\tau^u \mathbb{I}\{A_\tau\} &= \Big(f_\tau^u(w_\tau^u; w_\tau^{\Gamma(u)}, w_{\tau-1}^u, \Phi_\tau) - f_\tau^u(z_\tau^u, w_\tau^{\Gamma(u)}; w_{\tau-1}^u, \Phi_\tau) + V_\tau^H(w_\tau^u, w_\tau^{-u}) - V_\tau^H(z_\tau^u, w_\tau^{-u})\Big)\, \mathbb{I}\{A_\tau\} \\
&= \Bigg(f_\tau^u(w_\tau^u; z_\tau^{\Gamma(u)}, z_{\tau-1}^u, \Phi_\tau) - f_\tau^u(z_\tau^u; z_\tau^{\Gamma(u)}, z_{\tau-1}^u, \Phi_\tau) \\
&\quad + \Phi_{\tau-1,\tau}^u(w_{\tau-1}^u, w_\tau^u) - \Phi_{\tau-1,\tau}^u(z_{\tau-1}^u, w_\tau^u) - \Phi_{\tau-1,\tau}^u(w_{\tau-1}^u, z_\tau^u) + \Phi_{\tau-1,\tau}^u(z_{\tau-1}^u, z_\tau^u) \\
&\quad + V_\tau^H(w_\tau^u, w_\tau^{-u}) - V_\tau^H(z_\tau^u, w_\tau^{-u})\Bigg)\mathbb{I}\{A_\tau\} \\
&\leq 4c_{\text{time}}\mathbb{I}\{z_{\tau-1}^u \neq w_{\tau-1}^u\} + \Bigg(f_\tau^u(w_\tau^u; z_\tau^{\Gamma(u)}, z_{\tau-1}^u, \Phi_\tau) - f_\tau^u(z_\tau^u; z_\tau^{\Gamma(u)}, z_{\tau-1}^u, \Phi_\tau) \\
&\qquad\qquad + V_\tau^H(w_\tau^u, w_\tau^{-u}) - V_\tau^H(z_\tau^u, w_\tau^{-u})\Bigg)\mathbb{I}\{A_\tau\}.
\end{aligned}$$

Then, by (9), we further have

$$\begin{aligned}
\Delta_\tau^u \mathbb{I}\{A_\tau\} \leq & 4c_{\text{time}}\mathbb{I}\{z_{\tau-1}^u \neq w_{\tau-1}^u\} + \\
& \underbrace{\left(V_\tau^{H+1}(z_\tau^u, z_\tau^{-u}) - V_\tau^{H+1}(w_\tau^u, z_\tau^{-u}) + V_\tau^H(w_\tau^u, w_\tau^{-u}) - V_\tau^H(z_\tau^u, w_\tau^{-u})\right)\mathbb{I}\{A_\tau\}}_{(\natural)}
\end{aligned} \tag{10}$$

Next we expand out the expressions for the approximate value-to-go functions in ($\natural$).

Define the following functions over decision vectors at time $\tau + 1$:

$$\begin{aligned}
g_\tau(x) = g_\tau(x; \Phi_{\tau+1}) &:= f_\tau^{H+1}(x; z_\tau^u, z_\tau^{-u}, \Phi_{\tau+1}) + V_{\tau+1}^{H+1}(x), \\
h_\tau(x) = h_\tau(x; \Phi_{\tau+1})) &:= f_\tau^H(x; w_\tau^u, w_\tau^{-u}, \Phi_{\tau+1}) + V_{\tau+1}^H(x), \\
\delta_\tau(x) = \delta_\tau(x; \Phi_{\tau+1})) &:= \Phi_{\tau,\tau+1}^u(w_\tau^u, x^u) - \Phi_{\tau,\tau+1}^u(z_\tau^u, x^u).
\end{aligned}$$

where we omit their dependency on $\Phi_{\tau+1}$ to simplify the notations. Then, we have

$$\begin{aligned}
\mathrm{RV}_\tau^{H+1}(z_\tau^u, z_\tau^{-u}; \Phi_{\tau+1}) = \max_{x \in \mathcal{A}^{B(v,H+1)}} g_\tau(x), \quad \mathrm{RV}_\tau^{H+1}&(w_\tau^u, z_\tau^{-u}; \Phi_{\tau+1}) \\
&= \max_{x \in \mathcal{A}^{B(v,H+1)}} g_\tau(x) + \delta_\tau(x)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{RV}_\tau^{H}(w_\tau^u, w_\tau^{-u}; \Phi_{\tau+1}) = \max_{x \in \mathcal{A}^{B(v,H)}} h_\tau(x), \quad \mathrm{RV}_\tau^{H}&(z_\tau^u, w_\tau^{-u}; \Phi_{\tau+1}) \\
&= \max_{x \in \mathcal{A}^{B(v,H)}} h_\tau(x) - \delta_\tau(x).
\end{aligned}$$

We again similarly omit the dependency on $\Phi_{\tau+1}$ to simplify the notations and note that $z_{\tau+1}$ is an optimal solution for $\max_{x \in \mathcal{A}^{B(v,H+1)}} g_\tau(x)$ and $w_{\tau+1}$ is an optimal solution for $\max_{x \in \mathcal{A}^{B(v,H)}} h_\tau(x)$. Hence,

$$
\begin{aligned}
(\natural) &= \mathop{\mathbb{E}}_{\Phi_{\tau+1}} [\mathrm{RV}_\tau^{H+1}(z_\tau^u, z_\tau^{-u}) - \mathrm{RV}_\tau^{H+1}(w_\tau^u, z_\tau^{-u}) + \mathrm{RV}_\tau^{H}(w_\tau^v, w_\tau^{-u}) - \mathrm{RV}_\tau^{H}(z_\tau^u, w_\tau^{-u})]\mathbb{I}\{A_\tau\} \\
&\le \mathop{\mathbb{E}}_{\Phi_{\tau+1}} \left[g_\tau(z_{\tau+1}) - (g_\tau(z_{\tau+1}) + \delta_\tau(z_{\tau+1})) + h_\tau(w_{\tau+1}) - (h_\tau(w_{\tau+1}) - \delta_\tau(w_{\tau+1}))\right] \mathbb{I}\{A_\tau\} \\
&= \mathop{\mathbb{E}}_{\Phi_{t+1}} \left[\delta_\tau(w_{\tau+1}) - \delta_\tau(z_{\tau+1})\right] \mathbb{I}\{A_\tau\} \\
&= \mathop{\mathbb{E}}_{\Phi_{t+1}} \left[\Phi_{\tau,\tau+1}^u(w_\tau^u, w_{\tau+1}^u) - \Phi_{\tau,\tau+1}^u(z_\tau^u, w_{\tau+1}^u) - \Phi_{\tau,\tau+1}^u(w_\tau^u, z_{\tau+1}^u) + \Phi_{\tau,\tau+1}^u(z_\tau^u, z_{\tau+1}^u)\right] \mathbb{I}\{A_\tau\} \\
&\le 4c_{\text{time}} \mathop{\mathbb{E}}_{\Phi_{\tau+1}} [\mathbb{I}\{w_{\tau+1}^u \ne z_{\tau+1}^u\}]\mathbb{I}\{A_\tau\},
\end{aligned}
$$

where the last inequality is since when $z_{\tau+1}^u = w_{\tau+1}^u$, the four terms on the RHS cancel out. Hence,

$$
\Delta_\tau^u \mathbb{I}\{A_\tau\} \le 4c_{\text{time}}(\mathbb{I}\{z_{\tau-1}^u \ne w_{\tau-1}^u\} + \mathop{\mathbb{E}}_{\Phi_{\tau+1}} [\mathbb{I}\{w_{\tau+1}^u \ne z_{\tau+1}^u\}]\mathbb{I}\{A_\tau\}).
$$

Finally, we have

$$
\begin{aligned}
\mathbb{P}_\Phi(w_\tau^u \ne z_\tau^u, w_\tau^{\Gamma(u)} = z_\tau^{\Gamma(u)}) &\le \mathbb{P}(0 \le \Delta_\tau^u \mathbb{I}\{A_\tau\} \le 4c_{\text{time}}(\mathbb{I}\{z_{\tau-1}^u \ne w_{\tau-1}^u\} + \mathop{\mathbb{E}}_{\Phi_{\tau+1}} [\mathbb{I}\{w_{\tau+1}^u \ne z_{\tau+1}^u\}])) \\
&= \mathbb{P}(w_{\tau-1}^u \ne z_{\tau-1}^u)\mathbb{P}(0 \le \Delta_\tau^u \mathbb{I}\{A_\tau\} \le 4c_{\text{time}}(1 + \mathbb{P}(w_{\tau+1}^u \ne z_{\tau+1}^u | w_{\tau-1}^u \ne z_{\tau-1}^u)) | w_{\tau-1} \ne z_{\tau-1}) \\
&\quad + \mathbb{P}(w_{\tau-1}^u = z_{\tau-1}^u)\mathbb{P}(0 \le \Delta_\tau^u \mathbb{I}\{A_\tau\} \le 4c_{\text{time}} \mathbb{P}(w_{\tau+1}^u \ne z_{\tau+1}^u | w_{\tau-1}^u = z_{\tau-1}^u) | w_{\tau-1} = z_{\tau-1}) \\
&\le \mathbb{P}(w_{\tau-1}^u \ne z_{\tau-1}^u) \cdot g \cdot 4c_{\text{time}}(1 + \mathbb{P}(w_{\tau+1}^u \ne z_{\tau+1}^u | w_{\tau-1}^u \ne z_{\tau-1}^u)) \\
&\quad + \mathbb{P}(w_{\tau-1}^u = z_{\tau-1}^u) \cdot g \cdot 4c_{\text{time}}(\mathbb{P}(w_{\tau+1}^u \ne z_{\tau+1}^u | w_{\tau-1}^u = z_{\tau-1}^u)) \\
&= 4gc_{\text{time}}(\mathbb{P}(w_{\tau-1}^u \ne z_{\tau-1}^u) + \mathbb{P}(w_{\tau+1}^u \ne z_{\tau+1}^u)) \\
&\le (\mathbb{P}(w_{\tau-1}^u \ne z_{\tau-1}^u) + \mathbb{P}(w_{\tau+1}^u \ne z_{\tau+1}^u))\rho
\end{aligned}
$$

where the second last inequality is based on the following observation: conditional on previous decisions, previous reward functions, current interactions and node reward functions at other nodes except u, $\Delta_\tau^u \mathbb{I}\{A_\tau\} \in [0, s]$ for any $s \ge 0$ if and only if $\Phi_\tau^v(w_\tau^u) - \Phi_\tau^v(z_\tau^u)$ is within some length s interval. Moreover, the probability of the above event is upper bounded by g multiplied by s due to the third condition in Assumption 1: for any $a \ne a' \in \mathcal{A}$, $b_1 < b_2$,

$$
\mathbb{P}(\Phi_\tau^v(a) - \Phi_\tau^v(a') \in [b_1, b_2) \mid \Phi_\tau^{\text{inter}}, \{\Phi_\tau^u\}_{u \ne v}) \le g(b_2 - b_1).
$$

D.2. Proof of Lemma 6

We define a new distance metric which is more suitable for the dynamic optimization problem we are interested in. Denote node $v \in V$ at time $t \in [\mathcal{T}]$ as the pair (v, t), which we henceforth call a *ST node*. Define the *ST distance* between two ST nodes (v_1, t_1) and (v_2, t_2) as

$$
\text{dist}^{\text{st}}((v_1, t_1), (v_2, t_2)) = \text{dist}(v_1, v_2) + |t_1 - t_2|.
$$

In particular, if $t_1 = t_2$, then $\text{dist}^{\text{st}}((v_1, t_1), (v_2, t_2)) = \text{dist}(v_1, v_2)$. We also define another parameter:

$$
\xi := (d+2)\rho. \tag{11}
$$

Note that $\xi \leq \frac{1}{2}$ under our assumption that $\rho := 4g(dc_{\text{edge}} + 2c_{\text{time}}) \leq \frac{1}{2(d+2)}$. Recall that $\Gamma(u)$ denotes the set of neighbors of u in G.

Before proving Lemma 6, we present the following claim.

Claim 1. Under the same setting as in Lemma 6, for $t \leq \tau \leq t+H$ and $u \in B(v, H+1)$, we have

$$\mathbb{P}_{\Phi}(w_\tau^u \neq z_\tau^u | w_\tau^{\Gamma(u)}, z_\tau^{\Gamma(u)}) \leq \rho.$$

Proof of Claim 1. Let E_τ denote the event $w_\tau^u \neq z_\tau^u$ given $w_\tau^{\Gamma(u)}, z_\tau^{\Gamma(u)}$. Then, we define the following positive random variable as in the proof of Lemma 5 which is effectively the optimality gap between switching from action w_τ^u to z_τ^u under policy $\pi(H)$,

$$\begin{aligned}\Delta_\tau^u \mathbb{I}\{E_\tau\} &:= \left[f_\tau^H(w_\tau^u, w_\tau^{-u}; w_{\tau-1}, \Phi_\tau) + V_\tau^H(w_\tau^u, w_\tau^{-u})\right] - \left[f_\tau^H(z_\tau^u, w_\tau^{-u}; w_{\tau-1}, \Phi_\tau) + V_\tau^H(z_\tau^u, w_\tau^{-u})\right] \\ &\leq \Phi_\tau^u(w_\tau^u) - \Phi_\tau^u(z_\tau^u) + d \cdot 2c_{\text{edge}} + 2 \cdot 2c_{\text{time}},\end{aligned}$$

where the last inequality is because changing node action at u at time τ affects at most d spatial edges and 2 temporal edges.

Since $\Delta_\tau^u \mathbb{I}\{E_\tau\} \geq 0$, we have the following bound under E_τ,

$$\Phi_\tau^u(w_\tau^u) - \Phi_\tau^u(z_\tau^u) \geq -(d \cdot 2c_{\text{edge}} + 2 \cdot 2c_{\text{time}}).$$

Moreover, since z_τ^u is optimal under $\pi(H+1)$,

$$\begin{aligned}0 &\leq \left[f_\tau^{H+1}(z_\tau^u, z_\tau^{-u}; z_{\tau-1}, \Phi_\tau) + V_\tau^{H+1}(z_\tau^u, z_\tau^{-u})\right] - \left[f_\tau^{H+1}(w_\tau^u, z_\tau^{-u}; z_{\tau-1}, \Phi_\tau) + V_\tau^{H+1}(w_\tau^u, z_\tau^{-u})\right] \\ &\leq \Phi_\tau^u(z_\tau^u) - \Phi_\tau^u(w_\tau^u) + d \cdot 2c_{\text{edge}} + 2 \cdot 2c_{\text{time}},\end{aligned}$$

which leads to

$$\Phi_\tau^u(w_\tau^u) - \Phi_\tau^u(z_\tau^u) \leq (d \cdot 2c_{\text{edge}} + 2 \cdot 2c_{\text{time}}).$$

Combining these two bounds above, we have

$$\begin{aligned}\mathbb{P}(w_\tau^u \neq z_\tau^u | w_\tau^{\Gamma(u)}, z_\tau^{\Gamma(u)}) &\leq \mathbb{P}(-(2dc_{\text{edge}} + 4c_{\text{time}}) \leq \Phi_\tau^u(w_\tau^u) - \Phi_\tau^u(z_\tau^u) \leq \\ &\qquad (2dc_{\text{edge}} + 4c_{\text{time}}) | w_\tau^{\Gamma(u)}, z_\tau^{\Gamma(u)}) \\ &\leq g \cdot 2(2dc_{\text{edge}} + 4c_{\text{time}}) = \rho.\end{aligned}$$

□

Proof of Lemma 6. We prove the lemma by induction on the ST distance. By Claim 1 above, for $0 \leq \tau < H$ and $u \in B(v, H+1)$,

$$\mathbb{P}_{\Phi}(w_\tau^u \neq z_\tau^u) \leq \rho.$$

This serves as the base case for proof of Lemma 6: when (u, τ) satisfies $\text{dist}^{\text{st}}((v,t),(u,\tau)) \geq H+1$, Lemma 6 holds. Suppose that for all $k' > k$ for some $0 \leq k \leq H$, we have that if a node (u, τ) satisfies $\text{dist}^{\text{st}}((v,t),(u,\tau)) \leq k'$, then $\mathbb{P}_{\Phi}(w_\tau^u \neq z_\tau^u) \leq \xi^{H+1-k'}\rho$.

For the inductive step, we consider nodes (u,τ) with $\text{dist}^{\text{st}}((v,t),(u,\tau)) \leq k$ for $0 \leq k \leq H$. For the following, to simply the notations, we write $\mathbb{P}_\Phi$ as $\mathbb{P}$.

$$\begin{aligned}
\mathbb{P}(w_\tau^u \neq z_\tau^u) &= \mathbb{P}(w_\tau^u \neq z_\tau^u, w_\tau^{\Gamma(u)} \neq z_\tau^{\Gamma(u)}) + \mathbb{P}(w_\tau^u \neq z_\tau^u, w_\tau^{\Gamma(u)} = z_\tau^{\Gamma(u)}) \\
&= \mathbb{P}(w_\tau^u \neq z_\tau^u \mid w_\tau^{\Gamma(u)} \neq z_\tau^{\Gamma(u)})\mathbb{P}(w_\tau^{\Gamma(u)} \neq z_\tau^{\Gamma(u)}) + \mathbb{P}(w_\tau^u \neq z_\tau^u, w_\tau^{\Gamma(u)} = z_\tau^{\Gamma(u)}) \\
&\leq \rho(d \cdot \xi^{H-k}\rho) + \mathbb{P}(w_\tau^u \neq z_\tau^u, w_\tau^{\Gamma(u)} = z_\tau^{\Gamma(u)}) \\
&\leq \rho(d \cdot \xi^{H-k}\rho) + \mathbb{P}(w_{\tau-1}^u \neq z_{\tau-1}^u)\rho + \mathbb{P}(w_{\tau+1}^u \neq z_{\tau+1}^u)\rho \\
&\leq \rho(d \cdot \xi^{H-k}\rho) + 2(\xi^{H-k}\rho)\rho \\
&= (d\rho + 2\rho)\xi^{H-k}\rho \\
&\leq \xi^{H+1-k}\rho
\end{aligned}$$

where the first inequality is by induction hypothesis since the spatial neighbors of u has ST distance to (v,t) at most $k+1$ as well as Claim 1; the second inequality is by Lemma 5; the third inequality is again by the induction hypothesis. Hence we complete the induction step. □

Proof [Proof of Proposition 4] We first use Lemma 6 for $(u,\tau) = (v,t)$ and obtain

$$\mathbb{P}(\pi_t^v(H) \neq \pi_t^v(H+1)) \leq \xi^{H+1}\rho.$$

Then, observe that when the locality parameter $H = +\infty$, we obtain the optimal node decision $(x_t^v)^*$. Then we use a union bound over all locality parameters H which is greater than or equal to L.

$$\mathbb{P}(\pi_t^v(L) \neq \pi_t^v(+\infty)) \leq \sum_{H \geq L} \mathbb{P}(\pi_t^v(H) \neq \pi_t^v(H+1)) \leq \sum_{H \geq L} \xi^{H+1}\rho \leq 2\xi^{L+1}\rho \leq \epsilon/(2C)$$

since $\xi \leq 1/2$ and $L = \lfloor \log_2 \frac{4C}{\epsilon} \rfloor$. ■

Appendix E. Proof details for sample approximation

E.1. Bounding the Sampling Loss

In this section, we aim to bound the loss in rewards due to approximating the expected value-to-go function using simulation. The main result of the subsection is given in Proposition 7, which states that Algorithm 1 obtains the same solution as the local policy $\pi(L)$ with high probability.

Proposition 7 *Under the conditions in the Theorem 3, given any $\epsilon > 0$, there exists a function $N = N(\epsilon, d, g, C) = O((\frac{4C}{\epsilon})^{2\log_2 d} g^2 C^4) < \infty$ such that if sample size $n \geq N$, then for any $v \in V$, $t \geq 0$,*

$$\mathbb{P}(\pi_t^v(L) \neq \text{Alg}_t^v) \leq \epsilon/(2C).$$

Proof [Proof of Proposition 7] Suppose $\text{Alg}_t^v \neq \pi_t^v(L)$. By optimality,

$$\phi_t^v(\pi_t^v(L)) - \phi_t^v(\text{Alg}_t^v) + \Delta_t^{-v} := \max_{x_t^{B(v,L)}: x_t^v = \pi_t^v(L)} \left(f_t^L(x_t) + V_t^L(x_t)\right) - \max_{x_t^{B(v,L)}: x_t^v = \text{Alg}_t^v} \left(f_t^L(x_t) + V_t^L(x_t)\right) \geq 0.$$

Similarly,

$$\phi_t^v(\mathrm{Alg}_t^v) - \phi_t^v(\pi_t^v(L)) + \Delta^{-v,n} := \max_{x_t^{B(v,L)}:x_t^v=\mathrm{Alg}_t^v} \left(f_t^L(x_t) + \widehat{V}_t^L(x_t)\right) - \max_{x_t^{B(v,L)}:x_t^v=\pi_t^v(L)} \left(f_t^L(x_t) + \widehat{V}_t^L(x_t)\right) \geq 0.$$

Therefore,

$$-\Delta_t^{-v} \leq \phi_t^v(\pi_t^v(L)) - \phi_t^v(\mathrm{Alg}_t^v) \leq \Delta^{-v,n}.$$

By the third condition in the Assumption 1, we have that

$$\begin{aligned}\mathbb{P}\left(\mathrm{Alg}_t^v(L) \neq \pi_t^v(L)\right) &\leq \mathbb{P}\left(-\Delta_t^{-v} \leq \Phi_t^v(\pi_t^v(L)) - \Phi_t^v(\mathrm{Alg}_t^v) \leq \Delta^{-v,n}\right) \\ &\leq g\,\mathbb{E}\left[\Delta^{-v,n} + \Delta_t^{-v}\right].\end{aligned}$$

By definitions of Δ_t^{-v} and $\Delta^{-v,n}$ above, we have

$$\begin{aligned}\Delta^{-v,n} + \Delta_t^{-v} &\leq \max_{x_t^{B(v,L)}:x_t^v=\pi_t^v(L)} \left|V_t^L(x_t) - \widehat{V}_t^L(x_t)\right| + \max_{x_t^{B(v,L)}:x_t^v=\mathrm{Alg}_t^v} \left|V_t^L(x_t) - \widehat{V}_t^L(x_t)\right| \\ &\leq 2\max_{x_t}\left|V_t^L(x_t) - \widehat{V}_t^L(x_t)\right|.\end{aligned}$$

Therefore, it suffices to show that

$$\mathbb{E}\left[\max_{x_t^{B(v,L)}} \left|V_t^L(x_t) - \widehat{V}_t^L(x_t)\right|\right] \leq \frac{\epsilon}{4gC}.$$

Therefore, it suffices to show that

$$\mathbb{E}\left[\max_{x_t^{B(v,L)}} \left(V_t^L(x_t) - \widehat{V}_t^{L,n}(x_t)\right)\right] \leq \frac{\epsilon}{4gC} \text{ and } \mathbb{E}\left[\max_{x_t^{B(v,L)}} \left(\widehat{V}_t^{L,n}(x_t) - V_t^L(x_t)\right)\right] \leq \frac{\epsilon}{4gC}.$$

By definition, given any $x_{t+L} \in \mathcal{A}^{B(v,L)}$, $\widehat{V}_{t+L}^{L,n}(x_{t+L}) = V_{t+L}^L(x_{t+L}) = 0$. Hence,

$$\mathbb{E}[\max x_{t+L}(\widehat{V}_{t+L}^{L,n}(x_{t+L}) - V_{t+L}^L(x_{t+L}))] = 0.$$

Next, for $t \leq \tau \leq t+L-1$, we derive a recursive relation between $\mathbb{E}[\max x_\tau(\widehat{V}_\tau^{L,n}(x_\tau) - V_\tau^L(x_\tau))]$ and $\mathbb{E}[\max x_{\tau+1}(\widehat{V}_{\tau+1}^L(x_{\tau+1}) - V_{\tau+1}^L(x_{\tau+1}))]$. Given any $x_\tau \in \mathcal{A}^{B(v,L)}$

$$\begin{aligned}\widehat{V}_\tau^{L,n}(x_\tau) - V_\tau^L(x_\tau) &= \frac{1}{n}\sum_{s=1}^n \max_{x_{\tau+1}}(f_{\tau+1}^L(x_{\tau+1};x_\tau,\phi_{\tau+1}^{(s)}) + \widehat{V}_{\tau+1}^{L,n}(x_{\tau+1})) - \mathbb{E}[\max_{x_{\tau+1}}(f_{\tau+1}^L(x_{\tau+1};x_\tau,\Phi_{\tau+1}) + V_{\tau+1}^L(x_{\tau+1}))] \\ &= \frac{1}{n}\sum_{s=1}^n \max_{x_{\tau+1}}(f_{\tau+1}^L(x_{\tau+1};x_\tau,\phi_{\tau+1}^{(s)}) + \widehat{V}_{\tau+1}^{L,n}(x_{\tau+1})) - \frac{1}{n}\sum_{s=1}^n \max_{x_{\tau+1}}(f_{\tau+1}^L(x_{\tau+1};x_\tau,\phi_{\tau+1}^{(s)}) + V_{\tau+1}^L(x_{\tau+1})) \\ &+ \frac{1}{n}\sum_{s=1}^n \max_{x_{\tau+1}}(f_{\tau+1}^L(x_{\tau+1};x_\tau,\phi_{\tau+1}^{(s)}) + V_{\tau+1}^L(x_{\tau+1})) - \mathbb{E}[\max_{x_{\tau+1}}(f_{\tau+1}^L(x_{\tau+1};x_\tau,\Phi_{\tau+1}) + V_{\tau+1}^L(x_{\tau+1}))] \\ &\leq \max_{x_{\tau+1}}(\widehat{V}_{\tau+1}^{L,n}(x_{\tau+1}) - V_{\tau+1}^L(x_{\tau+1})) + \frac{1}{n}\sum_{s=1}^n \max_{x_{\tau+1}}(f_{\tau+1}^L(x_{\tau+1};x_\tau,\phi_{\tau+1}^{(s)}) + V_{\tau+1}^L(x_{\tau+1})) \\ &- \mathbb{E}[\max_{x_{\tau+1}}(f_{\tau+1}^L(x_{\tau+1};x_\tau,\Phi_{\tau+1}) + V_{\tau+1}^L(x_{\tau+1}))]\end{aligned}$$

We now bound the expectation of later two terms on the right-hand side of the above inequality. Let $Y^{(s)} := \max_{x_{\tau+1}}(f^L_{\tau+1}(x_{\tau+1}; x_\tau, \phi^{(s)}_{\tau+1}) + V^L_{\tau+1}(x_{\tau+1}))$. Then, $\{Y^{(s)}\}_{1\leq s\leq n}$ are independent random variables with expectation $\mu := \mathbb{E}[\max_{x_{\tau+1}}(f^L_{\tau+1}(x_{\tau+1}; x_\tau, \Phi_{\tau+1}) + V^L_{\tau+1}(x_{\tau+1}))]$. Note that

$$\mathbb{E}[\frac{1}{n}\sum_{s=1}^{n} Y^{(s)} - \mu] = \frac{1}{n}\,\mathbb{E}[\sum_{s=1}^{n} Y^{(s)} - n\mu] \leq \frac{1}{n}\,\mathbb{E}\,|\sum_{s=1}^{n} Y^{(s)} - n\mu|$$

Let $\widetilde{m}$ denote the number of nodes in $B(v, L)$. $\widetilde{m} = 1 + d + \cdots + d^L \leq \frac{d^{L+1}}{d-1}$. Since $\widetilde{m}d/2$ is the maximum number of edges in $B(v, L)$ by the handshaking lemma, we have that

$$Y^{(s)} = \mathrm{RV}^L_\tau(x_\tau; \phi_{\tau+1}(s)) \geq -L\widetilde{m}C_{\text{node}} - L\widetilde{m}\frac{d}{2}c_{\text{edge}} - L\widetilde{m}c_{\text{time}} =: \text{lb},$$

where this lower bound is achieved when all nodes receives the worst possible individual reward $-C_{\text{node}}$, the temporal interactions are $-c_{\text{time}}$ and the edge interactions are $-c_{\text{edge}}$. Similarly we obtain the upper bound,

$$Y^{(s)} = \mathrm{RV}^L_\tau(x_\tau; \phi_{\tau+1}(s)) \leq L\widetilde{m}C_{\text{node}} + L\widetilde{m}\frac{d}{2}c_{\text{edge}} + L\widetilde{m}c_{\text{time}} =: \text{ub}.$$

Since $\text{lb} \leq Y^{(s)} \leq \text{ub}$,

$$\begin{aligned}\frac{1}{n}\,\mathbb{E}[|\sum_{s=1}^{n} Y^{(s)} - n\mu|] &= \frac{1}{n}\int_{t\geq 0} \mathbb{P}(|\sum_{s=1}^{n} Y^{(s)} - n\mu| \geq t)dt \\ &\leq \frac{2}{n}\int_{t\geq 0} e^{-\frac{2t^2}{n(\text{ub}-\text{lb})^2}}dt \\ &= \frac{1}{\sqrt{n}}(\text{ub} - \text{lb})\int_{x\geq 0} e^{-\frac{x^2}{2}}dx \\ &= \frac{\sqrt{2\pi}}{2\sqrt{n}}(\text{ub} - \text{lb})\end{aligned}$$

where the first equality is by the property of expectation of non-negative random variables, the first inequality is by Hoeffding's inequality, the second equality is by change of variables, and the last equality is since $\frac{1}{\sqrt{2\pi}}\int_{x\in\mathbb{R}} e^{-\frac{x^2}{2}}dx = 1$.

Taking expectation of $\max_{x_\tau}(\widehat{V}^{L,n}_\tau(x_\tau) - V^L_\tau(x_\tau))$, we have

$$\begin{aligned}\mathbb{E}[\max_{x_\tau}(\widehat{V}^{L,n}_\tau(x_\tau) - V^L_\tau(x_\tau))] &\leq \mathbb{E}[\max_{x_{\tau+1}}(\widehat{V}^L_{\tau+1}(x_{\tau+1}) - V^L_{\tau+1}(x_{\tau+1}))] + \frac{\sqrt{2\pi}}{2\sqrt{n}}(\text{ub} - \text{lb}) \\ &\leq \mathbb{E}[\max_{x_{\tau+1}}(\widehat{V}^L_{\tau+1}(x_{\tau+1}) - V^L_{\tau+1}(x_{\tau+1}))] + \frac{\sqrt{2\pi}}{2\sqrt{n}}L\widetilde{m}C\end{aligned} \tag{12}$$

where the last inequality is due to $C := 2C_{\text{node}} + 2dc_{\text{edge}} + 4c_{\text{time}} \geq 2C_{\text{node}} + dc_{\text{edge}} + 2c_{\text{time}}$.

Applying Equation (12) L times for $t \leq \tau \leq t + L - 1$, we have

$$\mathbb{E}[\max_{x_t}(\widehat{V}^L_t(x_t) - V^L_t(x_t))] \leq \frac{\sqrt{2\pi}}{2\sqrt{n}}L^2\widetilde{m}C \leq \frac{\epsilon}{4gC},$$

for $n \geq N(\epsilon, d, g, C) = \frac{8\pi g^2 C^4 L^4 \widetilde{m}^2}{\epsilon^2}$. Under the condition in the Theorem 3, let $L = \lfloor \log_2 \frac{4C}{\epsilon} \rfloor$. Then,

$$\begin{aligned} N(\epsilon, d, g, C) &= \frac{8\pi g^2 C^4}{\epsilon^2} (\log_2 \frac{4C}{\epsilon})^4 (\frac{4C}{\epsilon})^{2\log_2 d} (\frac{d}{d-1})^2 \\ &= O((\frac{4C}{\epsilon})^{2\log_2 d} g^2 C^4) \end{aligned} \tag{13}$$

where $O(\cdot)$ notation omits logarithmic factors. Similarly, for such $n \geq N(\epsilon, d, g, C)$, we have

$$\mathbb{E}[\max_{x_t}(V_t^L(x_t) - \widehat{V}_t^L(x_t))] \leq \frac{\sqrt{2\pi}}{2\sqrt{n}} L^2 \widetilde{m} C \leq \frac{\epsilon}{4gC}.$$

All together, we have

$$\mathbb{E}\left[\max_{x_t^{B(v,L)}} \left(V_t^L(x_t) - \widehat{V}_t^{L,n}(x_t)\right)\right] \leq \frac{\epsilon}{4gC} \text{ and } \mathbb{E}\left[\max_{x_t^{B(v,L)}} \left(\widehat{V}_t^{L,n}(x_t) - V_t^L(x_t)\right)\right] \leq \frac{\epsilon}{4gC}$$

as desired.

■

Appendix F. Bounding the total loss

We now show Algorithm 1 achieves a near-optimal total reward by establishing Theorem 3.
Proof [Proof of Theorem 3] As a result of Proposition 4 and Proposition 7, we have that for all $v \in V$, and $t \in [\mathcal{T}]$,

$$\mathbb{P}(\text{Alg}_t^v \neq (x_t^v)^*) \leq \mathbb{P}(\pi_t^v(L) \neq (x_t^v)^*) + \mathbb{P}(\text{Alg}_t^v \neq \pi_t^v(L)) \leq \epsilon / C.$$

Recall the definition of C in Equation (3). Since the largest possible change in total rewards when switching from one node action to another is upper bounded by C,

$$\begin{aligned} |\mathcal{R}(ALG) - \mathcal{R}^*| &\leq \sum_{v \in V, t \in [\mathcal{T}]} \mathbb{P}(\text{Alg}_t^v \neq (x_t^v)^*) C \\ &\leq \epsilon \cdot |V| \mathcal{T}. \end{aligned}$$

■

Appendix G. Computation Efficiency

Next we look at the computation requirement of Algorithm 1. In terms of sampling, Algorithm 1 needs to simulate $n = O((\frac{1}{\epsilon})^{2\log_2 d})$ samples from each of the reward functions $\{\Phi_{t+1}, \Phi_{t+2}, \cdots, \Phi_{t+L}\}$. We illustrate these sample paths in Figure 5. The following proposition shows the computation requirement of Algorithm 1 for deciding the action of node v at each time step t.

Proposition 8 *The computational requirement of Algorithm 1 is $O(|V|\mathcal{T}e^{\text{poly}(\frac{1}{\epsilon})})$, where the model parameters d, g, C and $|\mathcal{A}|$ are constants in the $O(\cdot)$ notation.*

To establish Proposition 8, we prove the following lemma on computation requirement for any L-local ($L \geq 1$) algorithm. Then, the computation needed in Proposition 8 is by letting $L = \lfloor \log_2 \frac{4C}{\epsilon} \rfloor$ and $n = N(\epsilon, d, g, C)$ defined in Equation (13). We defer its proof after introducing the following Lemma.

Lemma 9 *The computation requirement for* Alg_t^v *for* $v \in V$ *and* $t \in [\mathcal{T}]$ *under Algorithm 1 is* $LK^2 n$ *where* n *is the sample size and* $K = |\mathcal{A}|^{d^L}$ *is an upper bound on number of decision vectors to enumerate over for the optimization problem in* (5).

Proof [Proof of Lemma 9] We show this by induction. Let a_τ denote the amount of computation needed to compute $\widehat{V}_\tau^{L,n}(\cdot)$. Since $\widehat{V}_{t+L}^{L,n}(\cdot) = 0$, $a_{t+L} = 0$. Suppose now we obtain $\widehat{V}_\tau^{L,n}(\cdot)$ function with computational effort a_τ. Given a decision vector $x_{\tau-1}$ and a realization ϕ_τ, the optimal x_τ can be solved by enumerating all possible decision vectors, whose cardinality is at most K. Under the assumption that $\{\Phi_t\}_t$ are independent, we can use the same estimation for $\widehat{V}_\tau^{L,n}(\cdot)$ for different realizations of $\Phi_{\tau-1}$. This implies:

$$a_{\tau-1} = K \cdot n \cdot K + a_\tau,$$

where the first K is the number of possible decision vectors $x_{\tau-1}$, n is the number of samples, and the second K is the computation needed for enumeration. Hence, we have $a_t = LK^2 n = L|\mathcal{A}|^{2d^L} n$. ■

Proof [Proof of Proposition 8] Let $L = \lfloor \log_2 \frac{4C}{\epsilon} \rfloor$. Then, we have

$$|\mathcal{A}|^{2d^L} \leq \mathrm{e}^{2 \ln |\mathcal{A}| \cdot (\frac{4C}{\epsilon})^{\log_2 d}}.$$

Moreover, we let the sample size $n = N(\epsilon, d, g, C) = O((\frac{4C}{\epsilon})^{2\log_2 d} g^2 C^4)$ defined in Equation (13). With d, g, C and $|\mathcal{A}|$ as constants, by Lemma 9, the computation needed for Algorithm 1 is upper bounded by

$$L|\mathcal{A}|^{2d^L} n = O(\mathrm{e}^{2 \ln |\mathcal{A}| \cdot (\frac{4C}{\epsilon})^{\log_2 d}} (\frac{1}{\epsilon})^{2 \log_2 d}) = O(\mathrm{e}^{\mathrm{poly}(\frac{1}{\epsilon})}).$$

■

Appendix H. Interactions must be small to have correlation decay

In this section, we construct a sequence of static (i.e., single period) decision networks indexed by graph degree d with $c_{\text{edge}} = \Theta(1/d)$ such that there is no near-optimal local algorithm for these networks. The decision networks we construct satisfy all parts of Assumption 1 *except* the small-interaction requirement $4g(dc_{\text{edge}} + 2c_{\text{time}}) \leq \frac{1}{2(d+2)}$. Thus, our construction justifies the need for the upper bound on the strength of the interactions in Assumption 1. Admittedly, there is some gap between our assumption $c_{\text{edge}} \leq \Theta(1/d^2)$, and the scale $c_{\text{edge}} = \Theta(1/d)$ at which we show here that long-range correlations arise. In comparison, previous work [21] also assumed $c_{\text{edge}} \leq \Theta(1/d^2)$ to obtain correlation decay in a static random decision network. In our dynamic setting, we have the same scaling to ensure no long-range correlation.

Definition 10 *A* d*-regular graph* $G = (V, E)$ *is an* γ-edge expander *for* $\gamma \in (0, 1)$ *if for any* $S \subseteq V$ *such that* $|S| \leq |V|/2$*, the number of edges between* S *and* $V \backslash S$ *(the "cut size") is at least* $|S| d\gamma$*, i.e.,*

$$\mathrm{cut}(S) := |\{(i, j) \in E : i \in S, j \in V \backslash S\}| \geq |S| d\gamma\,.$$

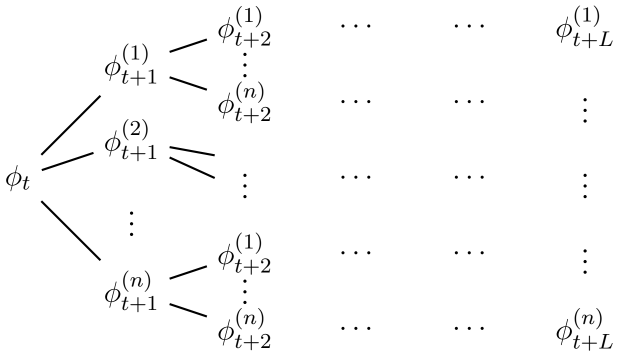

Figure 5: Approximating for value-to-go estimation via sample averages. Each node in this tree is a vector consisting of reward functions in $B(v, L)$.

Construction. Fix degree $d \geq 35$. It is well known that there exists $m_1 < \infty$, such that for any even integer m with $m > m_1$, there is a d-regular graph with m nodes that is a $\frac{1}{3}$-edge expander [18, 35]. In fact, a random d-regular graph has this property asymptotically almost surely (a.a.s.).[1] That is, let G be uniformly drawn from random d-regular graphs with m nodes where $m \geq m_1$, then G is a $\frac{1}{3}$-edge expander almost surely. We define a static random decision network (G, Φ) with action set $\{0, 1\}$ as follows:

- Node rewards: The node rewards $\Phi^v(1)$ are i.i.d. from Uniform$[-1, 1]$; and $\Phi^v(0) = 0$.
- Edge rewards: The edge rewards are "ferromagnetic":

$$\Phi^{u,v}(x^u, x^v) := \begin{cases} c_{\text{edge}} & \text{if } x^u = x^v \\ 0 & \text{otherwise}, \end{cases}$$

  where $c_{\text{edge}} := 6/d$.

Since the constructed decision network is static, there are no temporal interactions in our construction and hence we omit x_0.

The following claim shows that there does not exist near-optimal local algorithms when the small interaction condition in Assumption 1 does not hold.

Claim 2. *For the decision network (G, Φ) uniformly drawn from $d-$regular random graph with m nodes, the optimal action vector is either all 1s or all 0s a.s. Each of these possibilities arises with probability $1/2$. In particular, the optimal solution has long-range correlations. In particular, any L-local algorithm which treats the possible node actions 1 and 0 symmetrically achieves expected payoff at least $m/3$ below the optimal.*

Proof [Proof of Claim 2.] Consider any action vector x such that the majority of actions is 1. We show that the payoff of x is less than the payoff of all 1s. Let S be the set of nodes where x takes action 0. Since G is a $\frac{1}{3}$-edge expander a.s., cut$(S) \geq |S|d/3$. It follows that the total edge

[1] Random d-regular graphs are "almost Ramanujam", i.e., the absolute value of the second largest eigenvalue of their adjacency matrix is bounded above by $2\sqrt{d-1} + \epsilon$ a.a.s. as proved in [18]. The claimed edge expansion property then follows, e.g., using [35, Theorem 4.14].

rewards under x is at least $c_{\text{edge}}|S|d/3 = 2|S|$ smaller than that under all 1s. On the other hand the difference between the total node rewards under x and that under all 1s is $\sum_{v \in S} \Phi^v(1) \geq -|S|$ since $\Phi^v(1) \in [-1, 1]$, i.e., the total rewards under x is at least $2|S| - |S| = |S|$ smaller than the total rewards under all 1s. Similarly, one can show that for any action vector x such that the majority of actions is 0, the total reward under x is at least $m - |S|$ smaller than the total reward under all 0s. It follows that the optimal solution is either all 1s or 0s. Moreover, the optimal solution is all 1s if $\sum_{v \in V} \phi^v(1) \geq 0$ and all 0s otherwise. Since the distribution of i.i.d Uniform distributiion $[-1, 1]$ is symmetric, each of these above possibilities arises with probability $\frac{1}{2}$.

Now consider any given L and any L-local algorithm which treats 1 and 0 symmetrically. By symmetry, each node decision is a priori equally likely to be 1 or 0. By symmetry, each node decision is a priori equally likely to be 1 or 0. It follows that in a large network, about half the decisions will be 1 and the other half will be 0 under the L-local algorithm. Formally speaking, the expected number of 1s is $m/2$, and the variance in the total number of 1s is $\text{Var}[\sum_{v \in V} \mathbb{I}\{x^v = 1\}] = \sum_{v \in V} \text{Var}[\mathbb{I}\{x^v = 1\}] = \sum_{v \in V} (\frac{1}{2})^2 = \frac{1}{4}m$.

Hence for any $m > 250$, we know by Chebyshev's inequality that with probability at least 0.9, the number of 1s will be in the range $|S| \in (0.4m, 0.6m)$, i.e., the payoff will be at least $0.4m$ below the optimal (see the previous paragraph) Combining, the local algorithm suffers expected payoff loss at least $0.9 \times 0.4m \geq m/3$.

■

Appendix I. Missing Details for Section 4

In the following, we explain in details the simulation environment of our experiment. There are 5 dynamic decision networks parameterized by interaction strength c for both the spatial and temporal dimensions, with $c = 0.1, 0.2, 0.3, 0.4$, and 0.5. These decision networks share all other components, which we list below.

- *Graph G*: Using the *NetworkX* package in Python, we randomly generate[2] a 3-regular graph with 500 vertices.
- *Time horizon $\mathcal{T}$:* To simplify the simulation and reduce the computational effort, we set the time horizon to be 2 for all decision networks.
- *Interaction function*: Both the spatial interaction and the temporal interaction are *ferromagnetic*, meaning that agreeing actions incur a bonus of c, whereas disagreeing actions result in no reward.
- *Action set $\mathcal{A}$*: We assume a binary action set – that is, for all $v \in V$ and $t \in [\mathcal{T}]$, $x_t^v \in \{0, 1\}$, where action zero is viewed as the default action, meaning that $\Phi_t^v(0) = 0$.
- *Node reward*: The random node rewards, for both time periods, when taking action 1 are assumed to i.i.d. and follow the uniform distribution on $[-1, 1]$.

[2] statistics source: https://networkx.org/documentation/stable/reference/generated/networkx.generators.random_graphs.random_regular_graph.html

We sample $n_1 = 10$ instances for each decision network, where these instances differ in terms of the realized node rewards at the first time period. Having multiple realizations allow us to compute the confidence intervals for the performance of our algorithm. We denote by $\{\phi_1^{v,(i)}\}_{v\in V}$ the realized first-period node rewards (when taking action 1) in the i-th instance, where each $\phi_1^{v,(i)}$ is sampled according to the node reward distribution, i.e., uniformly from $[-1, 1]$. Note that the realized node rewards $\{\phi_1^{v,(i)}\}_{v\in V}$ for each $i \in [n_1]$ are shared by all the i-th instances of all decision networks.

To remove the loss in rewards due to sampling, we control the variability in the second period node rewards. That is, we pre-generate two independent sets of samples of node rewards for the second time period. The first set contains $n_{2,est} = 100$ samples, which are used to compute solutions at the first time period; and the second set contains $n_{2,eval} = 30$ samples, which are used to estimate the total payoff under the solutions computed using the first set of samples. We denote the node rewards when taking action 1 by $\{\phi_{2,est}^{v,(i)}\}_{v\in V}$ for the i-th sample in the first set and by $\{\phi_{2,eval}^{v,(i)}\}_{v\in V}$ for the i-th sample in the second set.

For each instance, we compute several solutions, with one obtained by solving the global optimization problem, and the others obtained by our local algorithms with different locality parameters. To solve the network optimization problem, either globally or locally, we write a *Mixed Integer Program* (MIP) and solve it through *Gurobi* [22]. The decision variables of the MIP are:

- node actions for the first time period: $\{x_1^v\}_{v\in V}$;
- disagreement indicator of neighboring nodes for $t = 1$: $\{y_1^e\}_{e\in E}$;
- node actions for the second time period for each sample j: $\{x_2^{v,(j)}\}_{v\in V, j\in[n_{2,est}]}$;
- disagreement indicator of neighboring nodes for $t = 2$ for each sample j: $\{y_2^{e,(j)}\}_{e\in E, j\in[n_{2,est}]}$;
- temporal disagreement indicator for each node for each sample j: $\{y^{v,(j)}\}_{v\in V, j\in[n_{2,est}]}$.

And the formulation of our MIP is given below.

$$\begin{aligned}
\max\quad & \sum_{v\in V} \phi_1^{v,(i)} \cdot x_1^v + \sum_{e=(u,v)\in E} c\cdot(1-y_1^e) + \\
& \frac{1}{n_{2,est}} \sum_{j\in[n_{2,est}]} \left[\sum_{v\in V} \phi_{2,est}^{v,(j)} \cdot x_2^{v,(j)} + \sum_{e\in E} c\cdot(1-y_2^{e,(j)}) + \sum_{v\in V} c\cdot(1-y^{v,(j)}) \right] \\
\text{s.t.}\quad & y_1^e \geq x_1^u - x_1^v \qquad \forall\, e=(u,v)\in E \\
& y_1^e \geq x_1^v - x_1^u \qquad \forall\, e=(u,v)\in E \\
& y_2^{e,(j)} \geq x_2^{u,(j)} - x_2^{v,(j)} \qquad \forall\, e=(u,v)\in E,\ j\in[n_{2,est}] \\
& y_2^{e,(j)} \geq x_2^{v,(j)} - x_2^{u,(j)} \qquad \forall\, e=(u,v)\in E,\ j\in[n_{2,est}] \\
& y^{v,(j)} \geq x_1^v - x_2^{v,(j)} \qquad \forall\, v\in V,\ j\in[n_{2,est}] \\
& y^{v,(j)} \geq x_2^{v,(j)} - x_1^v \qquad \forall\, v\in V,\ j\in[n_{2,est}] \\
& x_1^v, y_1^e, x_2^{v,(j)}, y_2^{e,(j)}, y^{v,(j)} \in \{0,1\} \qquad \forall\, v\in V,\ e\in E,\ j\in[n_{2,est}]
\end{aligned}$$

Note that V and E are either nodes and edges of the entire graph when solving for the global optimal solution, or nodes and edges of a local graph when solving for the solution using our local algorithm. Although the MIP is given for obtaining a first time period solution, a similar MIP can be used to estimate the payoff of a given first time period solution, where we take $\{x_1^v\}_{v\in V}$ as given and replace rewards $\{\phi_{2,est}^{v,(j)}\}$ with $\{\phi_{2,eval}^{v,(j)}\}$.