NETWORKED RESTLESS MULTI-ARMED BANDITS WITH DELAYED-DIFFUSED EXTERNALITY

Anonymous authors
Paper under double-blind review

ABSTRACT

We study networked restless multi-armed bandits in which actions generate externalities that are delayed in time and diffuse over a graph. We formalize this setting as a Delayed-Diffused Externality Restless Multi-Armed Bandit (DDE-RMAB), where activating an arm injects a latent exposure signal that persists, propagates through the network, and modulates future arm dynamics. This state-level coupling breaks the single-arm decomposition underlying classical Whittle-style methods. To address this challenge, we propose the Propagation-Adjusted Index (PAI), a scalable index policy that augments a local activation advantage with an analytic correction for long-run network spillovers. The correction is derived as a graph resolvent applied to local exposure marginal values, yielding an interpretable decomposition into private local value and delayed propagation value. We provide an approximation analysis showing that, under stable propagation and regular local modulation, the performance gap of PAI is controlled by explicit errors from value decomposition, multi-action interaction, propagation truncation, and marginal-value estimation. Experiments on positive and negative externality environments, as well as a semi-real graph benchmark, show that PAI is consistently best or near-best among the state-of-the-art baselines. These results highlight the importance of explicitly modeling delayed graph diffusion in scalable control of networked restless multi-armed bandits.

1 INTRODUCTION

Sequential resource allocation under uncertainty is a central problem in operations research and artificial intelligence. Restless multi-armed bandits (RMABs) provide a widely used framework for such problems: at each decision epoch, a controller activates at most $M \ll N$ arms, while all arms evolve according to action-dependent Markovian dynamics. Because exact RMAB planning is generally intractable, scalable solutions often rely on single-arm relaxations and index-based policies, especially Whittle-style methods (Whittle, 1988; Bertsimas & Niño-Mora, 2000; Glazebrook et al., 2011; Niño-Mora, 2023). This paradigm has enabled applications in public health interventions (Mate et al., 2020; Liang et al., 2025; Verma et al., 2023; Bhattacharya, 2018), recommendation systems (Meshram et al., 2017; 2015), and communication networks (Koley & Singh, 2025; Xiong et al., 2023).

A key assumption behind this success is **separability**: apart from the shared budget, arms evolve independently. Many real systems violate this assumption in a structured way. An intervention may create effects that are not purely local or immediate, but persist over time and spread through an interaction network. For example, a preventive health intervention may induce behavior that later influences connected individuals; repair or support at one location may benefit neighboring units; and harmful effects such as congestion, overload, or infection risk may also diffuse. In these settings, the value of an action depends not only on its immediate local effect, but also on delayed, accumulated, and multi-hop spillovers. Figure 1 illustrates this phenomenon.

Recent work has expanded RMABs beyond the classical fully known, stationary setting in several important directions. One line studies learning and adaptation when arm dynamics are unknown, contextual, or distributed. For example, optimistic and reinforcement-learning based methods have been developed for RMABs with unknown transition models (Wang et al., 2023; Akbarzadeh & Mahajan, 2023; Gafni et al., 2022), while contextual, federated, and robust variants address richer in-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

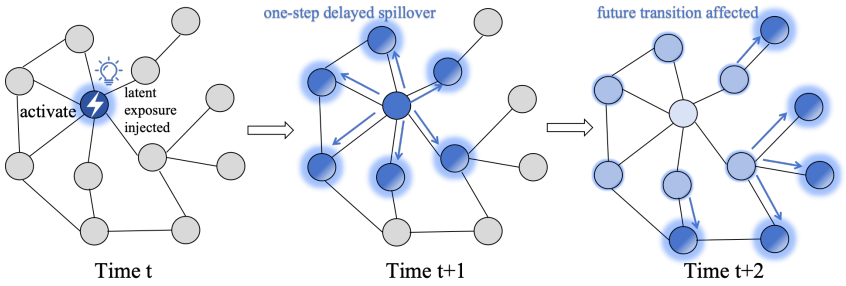


Figure 1: Delayed-diffused spillovers in networked restless bandits. An activation creates a local effect, which later spreads to one-hop and then multi-hop neighbors with decaying intensity.

formation structures and deployment constraints (Liang et al., 2025; Killian et al., 2022; Zhao et al., 2025; Boehmer et al., 2025). These works broaden the statistical and algorithmic scope of RMABs, but they typically preserve the essential decomposition that each arm's state evolves according to its own action, with coupling entering primarily through a shared budget or scheduling constraint. A second line studies more direct departures from separability. Recent work considers RMABs with global non-separable rewards (Raman et al., 2024), generic coupled RMAB objectives and policies (Luan et al., 2025), and graph-aware or networked bandit models for mobile interventions, positive externalities, and distributed coordination (Ou et al., 2022; Herlihy & Dickerson, 2023; Zhang et al., 2025; Genalti et al., 2024; Didi & Cohen, 2025). These models demonstrate that cross-arm dependence is essential in many applications. However, the structure of dependence in our setting is different. Existing coupled RMAB models often focus on non-separable rewards or generic joint planning, while graph-aware bandit models typically capture immediate graph-triggered effects, one-hop spillovers, or coordination constraints. They do not explicitly model an action-induced latent signal that persists, diffuses over multiple future periods, and modulates the transition dynamics of downstream arms. This raises our central question:

> *Can one design a scalable policy that preserves the efficiency of index-based control while explicitly accounting for delayed and diffused network externalities?*

Motivated by this question, we introduce **Delayed-Diffused Externality Restless Multi-Armed Bandits (DDE-RMABs)**. In a DDE-RMAB, activating an arm injects a latent exposure signal into a graph. The exposure persists, diffuses across edges, and modulates future local transition dynamics. Thus, even when each arm is locally modulated only by its own exposure, the shared exposure field couples arms across both space and time. To address this coupling, we propose the **Propagation-Adjusted Index (PAI)**. PAI preserves the scalability of index-based control, but augments a local activation advantage with a propagation correction. The correction is derived by viewing an activation as an exposure impulse and approximating its long-run value through a graph resolvent acting on local exposure sensitivities. This yields an interpretable and efficient index-like policy that accounts for delayed multi-hop spillovers without solving the full coupled Bellman problem. We also provide an approximation-theoretic analysis. Under stable propagation and regular local modulation, we show that the coupled value function admits a controlled approximation by a local-value backbone plus a propagation correction. The resulting performance gap is decomposed into three interpretable errors: value decomposition error, propagation truncation error, and marginal-value estimation error. In summary, our main contributions are conclued as followed:

- **DDE-RMAB formulation.** We introduce a new RMAB setting in which actions inject a latent exposure signal that persists and diffuses over a graph. The model captures delayed, accumulated, and multi-hop spillovers that affect future transition dynamics, including both beneficial externalities and harmful diffusion such as congestion or risk propagation. This formulation explicitly breaks the classical single-arm decomposition at the level of state evolution, while retaining enough structure to support scalable approximation.
- **A propagation-adjusted index policy.** We develop PAI, an index-style policy that combines a local activation advantage with an analytic propagation correction. The correction is obtained by viewing an activation as an exposure impulse and evaluating its discounted future effect through a graph resolvent. As a result, PAI can prioritize arms not only for

their immediate private value, but also for their ability to generate delayed spillover value in high-sensitivity regions of the network. The policy can be implemented exactly for moderate graphs or via a truncated Neumann expansion using sparse matrix-vector products.

- **Approximation theory for delayed-diffused coupling.** We formalize PAI as a controlled first-order approximation chain of the coupled Bellman advantage. We first establish stability and boundedness of the latent exposure dynamics under graph-operator normalization. We then connect the abstract Bellman decomposition and simultaneous-action interaction residuals directly to interpretable physical graph and model primitives, explicitly bounding them via the graph off-diagonal coupling $d_{\text{off}}(P)$ and downstream influence overlap $\Omega_M(P)$. Finally, we derive an end-to-end suboptimality bound for PAI that jointly accounts for value decomposition error, multi-action overlap, finite-depth Neumann truncation, frozen-sensitivity drift, and marginal-value estimation noise, providing rigorous theoretical conditions under which a resolvent-corrected index policy succeeds.
- **Empirical validation across externality regimes.** We evaluate PAI on synthetic positive-externality and negative-externality environments, as well as on a semi-real graph topology. Across budgets, propagation parameters, graph sizes, and externality signs, PAI consistently achieves the best or near-best performance compared to all tested baselines. We also test robustness to incomplete graph observations, showing that the propagation correction remains effective under moderate topology misspecification.

2 DDE-RMABs

We study **DDE-RMABs**, a restless multi-armed bandit model in which activations inject latent externalities that persist, diffuse over a network, and modulate future arm dynamics. The model preserves additive rewards and conditionally local transitions, but the evolving exposure field couples arms across time and network space.

Interaction graph and propagation. Let $\mathcal{G} = (V, E)$ be a directed graph over N arms, where $V = \{1, \dots, N\}$. Let $P \in \mathbb{R}_+^{N \times N}$ be a nonnegative propagation matrix, with $P_{ij} > 0$ indicating that exposure can flow from arm i to arm j. Accordingly, $P^\top$ is the forward exposure operator: $(P^\top z)_j = \sum_i P_{ij} z_i$ is the exposure arriving at arm j. Let $\alpha \in [0, 1)$ denote the exposure persistence factor. For general weighted graphs, we assume $\alpha\rho(P) < 1$, which ensures stability of the linear exposure recursion.

States, actions, and budget. At time $t = 0, 1, \dots$, arm i has an intrinsic state $s_i^t \in \mathcal{S}_i$ and a nonnegative exposure state $z_i^t \in \mathbb{R}_+$. We write $X^t = (s^t, z^t)$ for the joint state. The controller selects $a^t \in \{0, 1\}^N$, where $a_i^t = 1$ means that arm i is activated, subject to $\sum_i a_i^t \le M$. The feasible action set is $\mathcal{A}_M = \{a \in \{0, 1\}^N : \sum_i a_i \le M\}$.

Delayed exposure dynamics. The feature of DDE-RMABs is the persistent diffusion process

$$z^{t+1} = \alpha P^\top z^t + \beta(s^t) \odot a^t, \qquad \alpha \in [0, 1), \tag{1}$$

where $\beta(s^t) = (\beta_1(s_1^t), \dots, \beta_N(s_N^t))$ and $\beta_i(s_i^t) \ge 0$ is the exposure impulse injected by activating arm i in state s_i^t. We assume $0 \le \beta_i(s_i) \le \beta_{\max} < \infty$. Thus, activating arm i creates the delayed multi-hop trace $\beta_i e_i, \alpha P^\top \beta_i e_i, \alpha^2 (P^\top)^2 \beta_i e_i, \dots$. We call z^t *latent* because it represents an accumulated exposure mechanism rather than an intrinsic arm state. Given the propagation model, initial exposure, and past actions, it can be recursively tracked from Eq. 1.

Exposure-modulated transitions and objective. Conditional on its local exposure, arm i evolves as $s_i^{t+1} \sim \mathbb{P}_i(\cdot \mid s_i^t, a_i^t, z_i^t)$. Conditional on (X^t, a^t), intrinsic transitions factor across arms: $\mathbb{P}(s^{t+1} \mid X^t, a^t) = \prod_i \mathbb{P}_i(s_i^{t+1} \mid s_i^t, a_i^t, z_i^t)$. The one-step reward is additive, $R(X^t, a^t) = \sum_i r_i(s_i^t, a_i^t, z_i^t)$, where rewards are bounded. Given $\gamma \in (0, 1)$ and initial state X^0, the objective is to find a feasible policy π maximizing

$$V^\pi(X^0) = \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t R(X^t, a^t) \,\middle|\, X^0 \right]. \tag{2}$$

Although rewards and intrinsic transitions are locally specified, an action at time t changes z^{t+1} and thereby affects future transition dynamics of downstream arms. Hence, the problem is generally not decomposable into independent single-arm subproblems, and standard Whittle-style indexability arguments do not directly apply.

3 PROPAGATION-ADJUSTED INDEX

The challenge in DDE-RMABs is that an activation has both a direct local effect and a delayed network effect. In particular, activating one arm changes the future exposure field, which subsequently modulates the transition dynamics of other arms. Exact planning over the coupled state (s^t, z^t) is intractable, whereas a purely local priority ignores this persistent multi-hop effect.

We derive the **Propagation-Adjusted Index** (PAI) as a structured first-order approximation to the coupled Bellman activation advantage. PAI combines an additive local value backbone with the discounted value of an activation-induced exposure impulse. This construction preserves a separable index form while explicitly accounting for delayed propagation.

3.1 FROM LOCAL ADVANTAGE TO PROPAGATION-ADJUSTED PRIORITY

Let $\widetilde{V}_i(s_i, z_i)$ be a local value surrogate for arm i, and let $\widetilde{V}_{\mathrm{add}}(s, z) = \sum_i \widetilde{V}_i(s_i, z_i)$ denote the associated additive backbone. We first define the local activation advantage as

$$\Delta_i(s_i, z_i) = Q_i^{\mathrm{loc}}(s_i, z_i, 1) - Q_i^{\mathrm{loc}}(s_i, z_i, 0), \tag{3}$$

where $Q_i^{\mathrm{loc}}(s_i, z_i, a_i) = r_i(s_i, a_i, z_i) + \gamma \mathbb{E}[\widetilde{V}_i(s_i', \bar{z}_i') \mid s_i, z_i, a_i]$. Here $\bar{z}_i'$ is the baseline next exposure used by the local surrogate. Crucially, it excludes the exposure impulse injected by the current activation. Thus, Δ_i captures the private local effect, whereas the omitted delayed propagation effect is added separately below.

3.2 FIRST-ORDER VALUE OF DELAYED PROPAGATION

Activating arm i at time t injects $\delta z^{t+1} = \beta_i(s_i^t)e_i$. Under Eq. 1, this impulse generates the exposure trace $\delta z^{t+1+\tau} = (\alpha P^\top)^\tau \beta_i(s_i^t)e_i$ for $\tau \geq 0$. Let $\nu_i^t = \partial_{z_i} \widetilde{V}_i(s_i^t, z_i^t)$ denote the local marginal value of exposure and write $\nu^t = (\nu_1^t, \ldots, \nu_N^t)$.

PAI freezes the marginal-value field along this impulse trajectory and uses a first-order expansion of the additive continuation value. The resulting propagation correction is

$$\Psi_i^t = \sum_{\tau=0}^{\infty} \gamma^{\tau+1} \left\langle \nu^t, (\alpha P^\top)^\tau \beta_i(s_i^t)e_i \right\rangle = \gamma \beta_i(s_i^t) \left[(I - \gamma\alpha P)^{-1} \nu^t \right]_i. \tag{4}$$

The second equality uses $\langle \nu, (P^\top)^\tau e_i \rangle = [P^\tau \nu]_i$. Hence, $P^\top$ propagates exposure forward, while P propagates downstream marginal values backward to the source arm. The correction in Eq. 4 is therefore a sensitivity-weighted discounted walk sum, rather than a fixed-depth graph heuristic.

Equivalently, defining $g^t = (I - \gamma\alpha P)^{-1}\nu^t$, the vectorized correction is $\Psi^t = \gamma\, \beta(s^t) \odot g^t$. Thus, g_i^t measures the long-run downstream value of injecting one unit of exposure at arm i.

3.3 PAI POLICY AND IMPLEMENTATION

PAI assigns arm i the propagation-adjusted score

$$I_i^t = \Delta_i(s_i^t, z_i^t) + \kappa \Psi_i^t = \Delta_i(s_i^t, z_i^t) + \kappa\gamma\beta_i(s_i^t) \left[(I - \gamma\alpha P)^{-1} \nu^t \right]_i, \tag{5}$$

where $\kappa \geq 0$ controls the propagation scale. We use the nominal first-order value $\kappa = 1$ by default. Importantly, setting $\kappa = 0$ does not remove propagation approximation error; it removes the explicit propagation correction and transfers the omitted delayed value into the action-interaction residual.

Given separable scores, PAI solves the combinatorial selection problem $S_t \in \arg\max_{S \subseteq [N], |S| \leq M} \sum_{i \in S} I_i^t$ and sets $a_i^t = \mathbf{1}\{i \in S_t\}$. Under an "at most M" budget, the policy selects only the positive-score arms among the top-M rankings, thereby avoiding harmful interventions in negative-externality regimes. If an application strictly enforces exactly M activations, the constraint is replaced by $|S| = M$.

The local activation advantage Δ_i and sensitivity ν_i are computed offline from a decoupled tabular dynamic program or a learned differentiable value critic, resolving nonsmooth points via minimum-norm Clarke subgradients. Once these local quantities are provided, the propagation correction is evaluated analytically online. Consequently, PAI clearly separates offline local-model estimation from fast online graph-aware action selection.

For large-scale or sparse networks, explicitly computing the matrix inversion $(I-\gamma\alpha P)^{-1}$ can be avoided by employing an H-hop truncated Neumann expansion:

$$(I-\gamma\alpha P)^{-1}\nu^t \approx \sum_{\tau=0}^{H}(\gamma\alpha P)^{\tau}\nu^t. \qquad (6)$$

Each propagation step corresponds to a single sparse matrix-vector multiplication, yielding an $O(H|E|)$ propagation cost for depth H, alongside $O(N)$ local evaluations and $O(N\log M)$ top-M selection. As established in Section 4, the truncation error decays geometrically with rate $(\gamma\alpha)^{H+1}$; in practice, $H \geq 5$ achieves within 0.3% of the exact resolvent performance.

Algorithm 1 Propagation-Adjusted Index (PAI)

Require: Operator P, budget M, decay α, discount γ, weight κ; models Δ_i, ν_i, β_i; depth H (optional)
Ensure: Action sequence $\{a^t\}$
for each epoch $t = 0, 1, \ldots$ **do**
  Observe s^t; update belief $z^t \leftarrow \alpha P^\top z^{t-1} + \beta(s^{t-1}) \odot a^{t-1}$ (for $t > 0$)
  for each arm $i = 1, \ldots, N$ **do**
    $\Delta_i^t \leftarrow \Delta_i(s_i^t, z_i^t)$, $\nu_i^t \leftarrow \nu_i(s_i^t, z_i^t)$, $\beta_i^t \leftarrow \beta_i(s_i^t)$
  end for
  Form $\nu^t = (\nu_1^t, \ldots, \nu_N^t)^\top$
  Compute $g^t \leftarrow (I-\gamma\alpha P)^{-1}\nu^t$ {or truncated $\sum_{h=0}^{H}(\gamma\alpha P)^h\nu^t$}
  for each arm $i = 1, \ldots, N$ **do**
    $I_i^t \leftarrow \Delta_i^t + \kappa\gamma\beta_i^t g_i^t$
  end for
  $S_t \leftarrow \arg\max_{S\subseteq[N],|S|\leq M} \sum_{i\in S} I_i^t$
  Set $a_i^t \leftarrow \mathbf{1}\{i \in S_t\}$ for all i
end for

PAI is specialized to persistent exposure dynamics: when $\alpha = 0$, Eq. 5 reduces to a one-step exposure correction, whereas $\alpha > 0$ aggregates discounted multi-hop effects. For bounded nonlinear exposure dynamics, the same construction admits a local Jacobian-resolvent extension; we evaluate capped exposure variants empirically and discuss this extension in the appendix.

4 Theoretical Analysis

We interpret PAI as a controlled first-order approximation to the coupled Bellman activation advantage, rather than an exact index theorem for arbitrary coupled RMABs. The analysis separates four error sources: additive value decomposition, simultaneous-action interaction, finite propagation truncation, and marginal-value approximation. Complete proofs are deferred to Appendix E.

Approximation roadmap. Let $A^\star(X,a) = Q^\star(X,a) - Q^\star(X,0)$ denote the optimal Bellman activation advantage. PAI approximates it through

$$A^\star \longrightarrow A_{\mathrm{add}} \longrightarrow A_{\mathrm{FO}} \longrightarrow \widehat{A}_{\mathrm{PAI}}. \qquad (7)$$

Here, A_{add} is the activation advantage induced by the additive surrogate $\widetilde{V}_{\mathrm{add}}$, while A_{FO} is its first-order exposure-impulse approximation before propagation truncation and marginal-value estimation. The first step incurs a decomposition residual; the second incurs a simultaneous-action interaction residual; and the final step incurs truncation, frozen-sensitivity, and marginal-value estimation errors.

Standing assumptions. Rewards and activation-induced exposure injections are bounded as $|r_i(s_i,a_i,z_i)| \leq r_{\max}$ and $0 \leq \beta_i(s_i) \leq \beta_{\max}$. The propagation matrix is nonnegative and satisfies $\|P^\top\|_\infty \leq 1$ and $\|P\|_\infty \leq 1$, with $\alpha \in [0,1)$ and $\gamma \in (0,1)$. Rewards and local transition kernels are Lipschitz in exposure: $|r_i(s,a,z) - r_i(s,a,z')| \leq L_r|z-z'|$ and $\|\mathbb{P}_i(\cdot \mid s,a,z) - \mathbb{P}_i(\cdot \mid s,a,z')\|_{\mathrm{TV}} \leq L_p|z-z'|$. Finally, the local marginal-value field $\nu_i = \partial_{z_i}\widetilde{V}_i$ is uniformly bounded as $\|\nu\|_\infty \leq \nu_{\max}$. For nonsmooth local value functions, $\partial_{z_i}\widetilde{V}_i$ denotes a valid one-dimensional Clarke subgradient or finite-difference slope; the analysis only requires a bounded sensitivity field.

Lemma 1 (Exposure stability)**.** Let $Z_{\max} = \max\{\|z^0\|_\infty, \beta_{\max}/(1-\alpha)\}$. Under any feasible policy, $\sup_{t\geq 0}\|z^t\|_\infty \leq Z_{\max}$.

Lemma 1 follows from $\|z^{t+1}\|_\infty \leq \alpha\|z^t\|_\infty + \beta_{\max}$. Thus, the linear exposure model is stable under the stated normalization; this is not an arm-independence assumption.

Resolvent representation. Define the discounted value-propagation resolvent $G = (I - \gamma\alpha P)^{-1}$ and let $R_\infty = \|G\|_\infty$. Since $\|P\|_\infty \leq 1$, $R_\infty \leq (1-\gamma\alpha)^{-1}$. In particular, $[Gv]_i = \sum_{\tau\geq 0}(\gamma\alpha)^\tau [P^\tau v]_i$. Although exposure propagates forward through $P^\top$, its continuation value propagates backward to the source arm through P, since $\langle v, (P^\top)^\tau e_i\rangle = [P^\tau v]_i$. Hence, $[G\nu]_i$ is the sensitivity-weighted discounted walk sum associated with an exposure impulse at arm i.

Proposition 1 (Approximate additive decomposition)**.** *Let $\widetilde{V}_{\mathrm{add}}(s,z) = \sum_i \widetilde{V}_i(s_i, z_i)$ be the additive local surrogate and define $\mathcal{E}_{\mathrm{dec}} = \|\mathcal{T}\widetilde{V}_{\mathrm{add}} - \widetilde{V}_{\mathrm{add}}\|_\infty$. Then $\|V^\star - \widetilde{V}_{\mathrm{add}}\|_\infty \leq \mathcal{E}_{\mathrm{dec}}/(1-\gamma)$. Moreover, if $\widetilde{V}_{\mathrm{add}}$ is L_{add}-Lipschitz in exposure and δ_{dec} bounds the one-step discrepancy between the true and local exposure updates, then $\mathcal{E}_{\mathrm{dec}} \leq \gamma L_{\mathrm{add}}\delta_{\mathrm{dec}}$.*

For the local surrogate that masks off-diagonal incoming exposure, let $d_{\mathrm{off}}(P) = \|P^\top - \mathrm{diag}(P^\top)\|_\infty$. Then $\delta_{\mathrm{dec}} \leq \alpha Z_{\max} d_{\mathrm{off}}(P)$, and therefore

$$\mathcal{E}_{\mathrm{dec}} \leq \gamma L_{\mathrm{add}}\alpha Z_{\max} d_{\mathrm{off}}(P) \lesssim \gamma L_{\mathrm{add}} \frac{\alpha\beta_{\max}}{1-\alpha} d_{\mathrm{off}}(P). \tag{8}$$

Thus, the additive backbone is accurate in a stable weak-coupling regime with moderate exposure scale, limited off-diagonal graph mass, and moderate exposure sensitivity.

First-order propagation score. Activating arm i injects $\beta_i(s_i^t)e_i$ into the next exposure state and creates the trace $\beta_i(s_i^t)e_i, \alpha P^\top \beta_i(s_i^t)e_i, \ldots$. Linearizing the additive continuation value around the current state gives $\Psi_i^t = \gamma\beta_i(s_i^t)\left[(I-\gamma\alpha P)^{-1}\nu^t\right]_i$. The PAI score is $I_i = \Delta_i + \kappa\Psi_i$, where Δ_i is the local activation advantage and $\kappa \geq 0$ controls the propagation scale.

Proposition 2 (Simultaneous-action approximation)**.** *Let $\mathcal{A}_M$ be the feasible action set with at most M active arms and define $\mathcal{E}_{\mathrm{act}}(\kappa) = \sup_{X,a\in\mathcal{A}_M} |A_{\mathrm{add}}(X,a) - \sum_i a_i(\Delta_i + \kappa\Psi_i)|$. If the marginal sensitivity field is L_ν-Lipschitz in exposure, then*

$$\mathcal{E}_{\mathrm{act}}(1) \lesssim L_\nu \beta_{\max}^2 \Omega_M(P), \quad \Omega_M(P) = \sup_{S:|S|\leq M} \sum_j \left(\sum_{i\in S} G_{ji}\right)^2. \tag{9}$$

Moreover, $\Omega_M(P) \leq M^2/(1-\gamma\alpha)^2$.

The quantity $\Omega_M(P)$ measures overlap among the propagated influence regions of selected arms. It is small when selected arms have limited downstream overlap, even if the graph itself is connected. For general $\kappa \geq 0$, $\mathcal{E}_{\mathrm{act}}(\kappa) \leq \mathcal{E}_{\mathrm{act}}(1) + |\kappa - 1|M\gamma\beta_{\max}R_\infty\nu_{\max}$. Consequently, $\kappa = 0$ does not imply a smaller approximation error: it removes the explicit propagation correction, but the omitted propagation value is absorbed into $\mathcal{E}_{\mathrm{act}}(0)$, which can be much larger than $\mathcal{E}_{\mathrm{act}}(1)$.

Theorem 1 (Propagation and marginal-value approximation error)**.** *Let $\Psi_{i,\mathrm{true}}^t$ denote exact first-order propagation value obtained using future marginal fields $\nu^{t+\tau}$. Suppose that the marginal-value field drifts smoothly along relevant propagation paths, i.e., $\|\nu^{t+\tau} - \nu^t\|_\infty \leq \eta_\ell\tau$ for all $\tau \geq 0$. Then*

$$\left|\Psi_{i,\mathrm{true}}^t - \Psi_i^t\right| \leq \frac{\gamma\beta_{\max}\eta_\ell}{(1-\gamma\alpha)^2}. \tag{10}$$

If an estimated marginal field $\widehat{\nu}$ satisfies $\|\widehat{\nu} - \nu\|_\infty \leq \eta_\nu$, then

$$\|\Psi(\widehat{\nu}) - \Psi(\nu)\|_\infty \leq \gamma\beta_{\max}R_\infty\eta_\nu \leq \frac{\gamma\beta_{\max}\eta_\nu}{1-\gamma\alpha}. \tag{11}$$

Theorem 1 shows that frozen sensitivity is reliable when marginal values evolve smoothly along diffusion paths and $\gamma\alpha$ is bounded away from one. It can be inaccurate near abrupt state changes, threshold effects, strong saturation, or under stale or noisy exposure tracking.

Corollary 1 (Truncated propagation error)**.** *Let $\Psi^{(H)}$ be the depth-H truncated propagation score obtained by replacing G with $\sum_{\tau=0}^{H}(\gamma\alpha P)^{\tau}$. Then $\|\Psi(\nu)-\Psi^{(H)}(\nu)\|_{\infty} \le \gamma\beta_{\max}\nu_{\max}\frac{(\gamma\alpha)^{H+1}}{1-\gamma\alpha}$.*

The propagation tail therefore decays geometrically with H, although a larger depth is needed as $\gamma\alpha$ approaches one.

Theorem 2 (End-to-end PAI performance bound)**.** *Let π_{PAI} select the feasible top-M action set according to the truncated estimated score $\widehat{I}_i = \Delta_i + \kappa\widehat{\Psi}_i^{(H)}$. Under the standing assumptions, for every initial state X^0,*

$$V^{\star}(X^0)-V^{\pi_{\mathrm{PAI}}}(X^0) \le \frac{1}{1-\gamma}\left[C_1\mathcal{E}_{\mathrm{dec}} + C_2\mathcal{E}_{\mathrm{act}}(\kappa) + \frac{\gamma\beta_{\max}}{1-\gamma\alpha}\left(C_3\nu_{\max}(\gamma\alpha)^{H+1} + C_4\eta_{\nu} + \frac{C_4\eta_{\ell}}{1-\gamma\alpha}\right)\right], \tag{12}$$

where C_1, C_2, C_3, C_4 are problem-dependent constants independent of H, η_{ν}, and η_{ℓ}.

Theorem 2 separates the error of the additive backbone, simultaneous-action approximation, finite-depth propagation, and marginal-value approximation. Combining Eq. 8, Eq. 9, and the κ mismatch bound makes the dependence explicit in the exposure scale $\beta_{\max}/(1-\alpha)$, graph coupling $d_{\mathrm{off}}(P)$, budget M, propagated-overlap term $\Omega_M(P)$, and sensitivity smoothness L_{ν}.

Scope. The results are most informative when $\gamma\alpha$ is bounded away from one, $\mathcal{E}_{\mathrm{dec}}$ is small, propagated influence regions have limited overlap, and marginal values are accurate and slowly varying. PAI may degrade on dense or hub-dominated graphs, under large activation budgets, near discontinuous thresholds or strong saturation, when $\gamma\alpha$ is close to one, or when exposure tracking is severely misspecified. These failure regimes correspond directly to large decomposition, action-interaction, truncation, or marginal-value-error terms.

The linear exposure model is analytically convenient because the first-order value of an exposure impulse admits the global resolvent form in Eq. 4. For bounded nonlinear dynamics $z^{t+1} = g(\alpha P^{\top} z^t + \beta(s^t)\odot a^t)$, the analogous local construction replaces the global resolvent with a state-dependent Jacobian resolvent. This is a local linearization rather than a general global guarantee; capped and saturating exposure variants are evaluated empirically in Section 5.

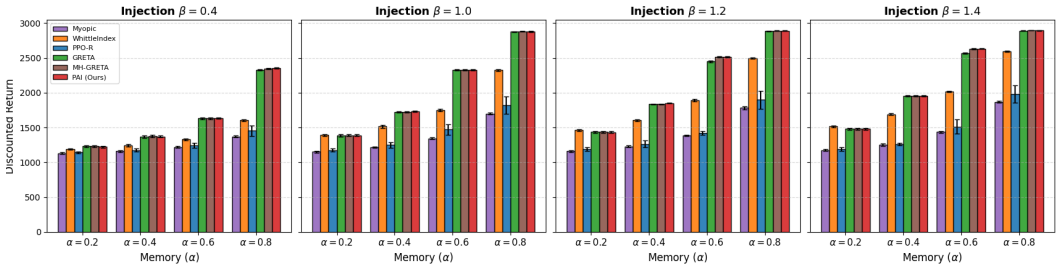
5 Experiments

We evaluate PAI in environments with positive and negative delayed externalities, as well as on a real-topology, simulated-dynamics benchmark. The experiments are designed to test whether the propagation correction remains useful beyond local and one-hop decision rules, and whether it is robust to finite propagation depth, saturation, and graph misspecification.

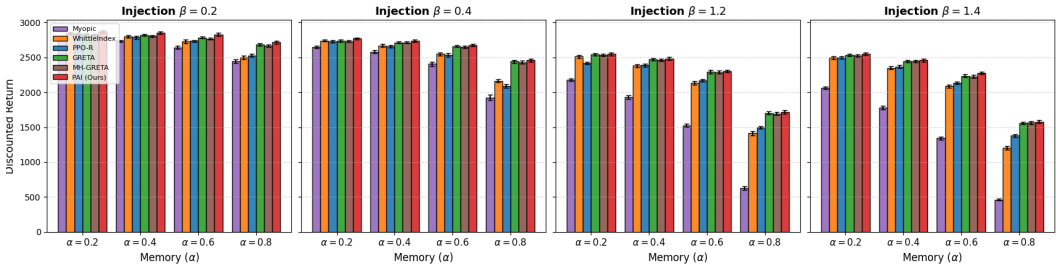
Experimental setup and environments. Unless otherwise stated, synthetic experiments use horizon $T = 1000$, discount factor $\gamma = 0.99$, and multiple random seeds. We compare PAI with **Myopic**, which ranks arms by immediate local reward; **Decoupled Whittle-style**, a tabular single-arm active–passive priority that ignores network propagation and is not an exact Whittle index for the coupled DDE-RMAB; **Local PPO-R**, a shared-parameter PPO RMAB baseline without graph message passing (Boehmer et al., 2025); and **GRETA**, a graph-aware one-hop RMAB method (Herlihy & Dickerson, 2023). We further introduce **MH-GRETA-H**, a fixed-depth multi-hop extension that propagates GRETA's graph message for H steps and aggregates discounted downstream effects. Thus, $H = 1$ recovers the one-hop message; we use MH-GRETA-H5 in the main comparison and report the full depth sweep in Appendix G.2.

Environment A is a positive-externality DDE-RMAB on a Barabási–Albert graph (Barabási & Albert, 1999), generated using NetworkX (NetworkX Developers, 2024). Arms have binary intrinsic states, exposure evolves as $z^{t+1} = \alpha P^{\top} z^t + \beta a^t$, the reward is $R_t = \sum_i s_i^t - 0.1\sum_i a_i^t$, and the budget is $M = 2$. Activation improves local transitions and creates beneficial downstream exposure. Environment B uses the same graph and propagation model but reverses the exposure effect: larger exposure impedes recovery and increases deterioration. It uses $M = 5$ and represents congestion, risk diffusion, or infection-like externalities. Environment C uses the *Les Misérables* co-occurrence network from the Stanford GraphBase (Knuth, 1993), with the simulated positive-externality dy-

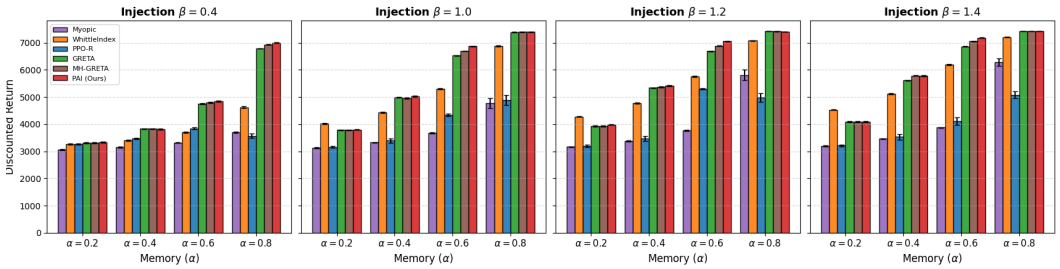
namics of Environment A. It is therefore a real-topology, simulated-dynamics benchmark, not a validated real-world deployment study.



(a) Environment A: positive externality



(b) Environment B: negative externality



(c) Environment C: real-topology, simulated dynamics

Figure 2: Discounted return across propagation memory α and injection strength β.

Main results. Figure 2 reports the discounted return as a function of propagation memory α and injection strength β across six evaluated methods. PAI achieves the best or near-best performance across all three benchmark settings. In the Environment A, the advantage of PAI over local baselines becomes substantially more pronounced as the persistence factor α increases, where delayed spillovers accumulate over extended horizons. Crucially, PAI consistently outperforms the network-aware baseline GRETA and remains superior or highly competitive against the stronger multi-hop competitor MH-GRETA, with larger margins in regimes characterized by strong propagation and large injection. In the Environment B, where uncoordinated activations cause persistent harmful cascades, PAI maintains high robustness because its discounted resolvent effectively accounts for downstream penalty costs alongside immediate gains. Finally, on Environment C, PAI again secures superior returns across high-memory configurations, confirming that the propagation correction scales effectively from synthetic graphs to real-world network structures.

Budget and scale sensitivity. Figure 3 and Figure 4 evaluate the sensitivity to intervention budget M and total number of arms N. As shown in Figure 3(a) and (c), PAI achieves its most performance gains under tight resource budgets, where prioritizing high-leverage arms is critical; as M approaches 10, performance across baselines naturally plateaus as the available budget saturates the

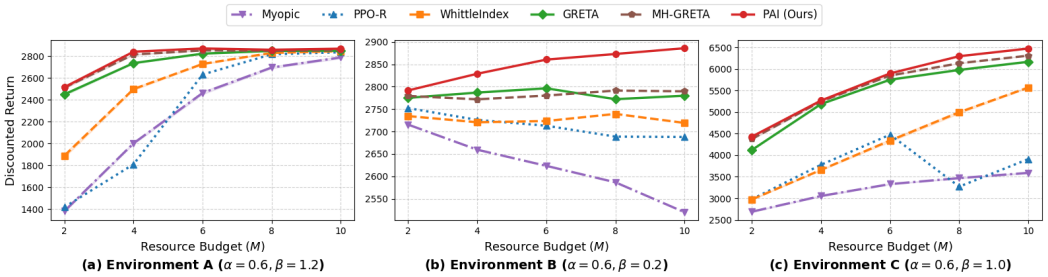


Figure 3: Sensitivity to the intervention budget M.

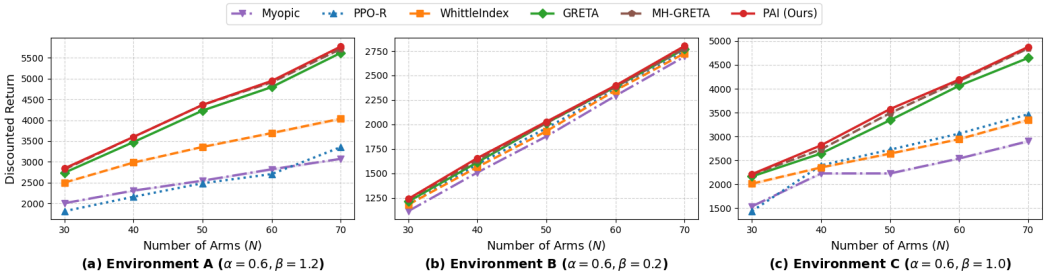


Figure 4: Sensitivity to the number of arms N.

network. In environment B, , increasing the budget degrades local heuristics and causes one-hop GRETA to deteriorate for $M \geq 6$ due to harmful spillover accumulation. In contrast, PAI steadily improves across all budget levels by penalizing long-run negative externalities through its resolvent. Regarding scaling with the number of arms N, PAI consistently maintains superior or near-best returns. Its performance gap over baselines grows markedly in positive-externality environments, where larger network topologies offer richer multi-hop diffusion pathways for PAI to exploit, while in Environment B all methods scale similarly with network size.

Robustness and ablations. We further evaluate PAI under capped-linear exposure dynamics, noisy exposure tracking, structured graph errors, and different values of the propagation scaling κ and truncation depth H. PAI remains competitive under the capped-linear simulator, suggesting that the method is not restricted to completely unsaturated exposure responses in the tested regimes. Removing the propagation correction ($\kappa = 0$) causes a substantial performance drop, while positive values of κ quickly reach a near-best plateau. For the truncated resolvent, $H \geq 5$ is within 0.29% of the infinite-depth correction in the tested settings. Under random dropout, targeted hub-edge removal, community-level missingness, false-positive edges, and log-normal edge-weight perturbations, PAI remains the best graph-aware method in mean return, with changes below approximately 1% relative to the clean-graph setting. Additional results, complete tables, and the noisy/stale sensitivity experiments are provided in Appendix G.

6 Conclusion

We introduce DDE-RMABs, a class of networked restless bandits in which actions generate latent externalities that persist over time, diffuse through a graph, and modulate future arm dynamics. To address the resulting loss of separability, we proposed PAI, a scalable propagation-adjusted index policy that combines local restless bandit priorities with a resolvent-based correction for long-run network spillovers. We showed that, under stable propagation and regular local modulation, this structure yields explicit approximation guarantees in terms of value decomposition, propagation truncation, and marginal-sensitivity estimation errors. Empirically, PAI consistently improves over the baselines across positive-externality, negative-externality, and semi-real graph benchmarks.

REFERENCES

Nima Akbarzadeh and Aditya Mahajan. On learning whittle index policy for restless bandits with scalable regret. *IEEE Transactions on Control of Network Systems*, 11(3):1190–1202, 2023.

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

Dimitris Bertsimas and José Niño-Mora. Restless bandits, linear programming relaxations, and a primal-dual index heuristic. *Operations Research*, 48(1):80–90, 2000.

Biswarup Bhattacharya. Restless bandits visiting villages: A preliminary study on distributing public health services. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, pp. 1–8, 2018.

Niclas Boehmer, Yunfan Zhao, Guojun Xiong, Paula Rodriguez-Diaz, Paola Del Cueto Cibrian, Joseph Ngonzi, Adeline Boatin, and Milind Tambe. Optimizing vital sign monitoring in resource-constrained maternal care: An rl-based restless bandit approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 28843–28849, 2025.

Xiaotong Cheng and Setareh Maghsudi. Anomaly detection in networked bandits. *IEEE Transactions on Signal Processing*, 2025.

Abhinav Dahiya, Nima Akbarzadeh, Aditya Mahajan, and Stephen L. Smith. Scalable operator allocation for multirobot assistance: A restless bandit approach. *IEEE Transactions on Control of Network Systems*, 9(3):1449–1460, 2022.

Liad Lea Didi and Kobi Cohen. Distributed learning in markovian restless bandits over interference graphs for stable spectrum sharing. *arXiv preprint arXiv:2512.17161*, 2025.

Tomer Gafni, Michal Yemini, and Kobi Cohen. Learning in restless bandits under exogenous global markov process. *IEEE Transactions on Signal Processing*, 2022.

Gianmarco Genalti, Marco Mussi, Nicola Gatti, Marcello Restelli, Matteo Castiglioni, and Alberto Maria Metelli. Graph-triggered rising bandits. In *International Conference on Machine Learning (ICML)*. PMLR, 2024.

Gianmarco Genalti, Marco Mussi, Nicola Gatti, Marcello Restelli, Matteo Castiglioni, and Alberto Maria Metelli. Bridging rested and restless bandits with graph-triggering: Rising and rotting. *Journal of Machine Learning Research*, 27(151):1–46, 2026.

Kevin D Glazebrook, David J Hodge, and Chris Kirkbride. General notions of indexability for queueing control and asset management. *The Annals of Applied Probability*, 21(3):876–907, 2011.

Xueping Gong and Jiheng Zhang. Efficient graph bandit learning with side-observations and switching constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 16871–16879, 2025.

Christine Herlihy and John P Dickerson. Networked restless bandits with positive externalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11997–12004, 2023.

Yige Hong, Qiaomin Xie, Yudong Chen, and Weina Wang. Restless bandits with average reward: Breaking the uniform global attractor assumption. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Jackson A. Killian, Lily Xu, Arpita Biswas, and Milind Tambe. Restless and uncertain: Robust policies for restless bandits via deep multi-agent reinforcement learning. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 180, pp. 990–1000. PMLR, 2022.

Donald E. Knuth. *The Stanford GraphBase: A Platform for Combinatorial Computing*. ACM Press, 1993.

Ankita Koley and Chandramani Singh. Fresh caching of dynamic contents using restless multi-armed bandits over wireless access. *IEEE Transactions on Networking*, 2025.

Dexun Li and Pradeep Varakantham. Efficient resource allocation with fairness constraints in restless multi-armed bandits. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 180, pp. 1158–1167. PMLR, 2022.

Biyonka Liang, Lily Xu, Aparna Taneja, Milind Tambe, and Lucas Janson. Context in public health for underserved communities: A bayesian approach to online restless bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 28195–28203, 2025.

Jingyuan Liu, Hao Qiu, Lin Yang, and Mengfan Xu. Distributed multi-agent bandits over erdős-rényi random networks. *Advances in Neural Information Processing Systems*, 38:159769–159798, 2026.

Yao Luan, Ni Mu, and Qing-Shan Jia. Addressing coupling in restless multi-armed bandits by finetuning whittle index. In *2025 IEEE 21st International Conference on Automation Science and Engineering (CASE)*. IEEE, 2025.

Yi Mao and Andrew Perrault. Time-constrained restless multi-armed bandits with applications to city service scheduling. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2024.

Aditya Mate, Jackson Killian, Haifeng Xu, Andrew Perrault, and Milind Tambe. Collapsing bandits and their application to public health intervention. *Advances in Neural Information Processing Systems*, 33:15639–15650, 2020.

Rahul Meshram, D Manjunath, and Aditya Gopalan. A restless bandit with no observable states for recommendation systems and communication link scheduling. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pp. 7820–7825. IEEE, 2015.

Rahul Meshram, Aditya Gopalan, and D Manjunath. Restless bandits that hide their hand and recommendation systems. In *2017 9th International Conference on Communication Systems and Networks (COMSNETS)*, pp. 206–213. IEEE, 2017.

NetworkX Developers. Networkx, 2024. URL `https://networkx.org/`.

José Niño-Mora. Markovian restless bandits and index policies: A review. *Mathematics*, 11(7): 1639, 2023.

Han-Ching Ou, Christoph Siebenbrunner, Jackson Killian, Meredith B. Brooks, David Kempe, Yevgeniy Vorobeychik, and Milind Tambe. Networked restless multi-armed bandits for mobile interventions. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2022.

Naveen Raman, Zheyuan Ryan Shi, and Fei Fang. Global rewards in restless multi-armed bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Abdalaziz Sawwan and Jie Wu. Modified index policies for multi-armed bandits with network-like markovian dependencies. *Network*, 5(1):3, 2025.

Jingwen Tong, Xinran Li, Liqun Fu, Jun Zhang, and Khaled B Letaief. A federated online restless bandit framework for cooperative resource allocation. *IEEE Transactions on Mobile Computing*, 23(12):15274–15288, 2024.

Shresth Verma, Aditya Mate, Kai Wang, Neha Madhiwalla, Aparna Hegde, Aparna Taneja, and Milind Tambe. Restless multi-armed bandits for maternal and child health: Results from decision-focused learning. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pp. 1312–1320, 2023.

Kai Wang, Lily Xu, Aparna Taneja, and Milind Tambe. Optimistic whittle index policy: Online learning for restless bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10131–10139, 2023.

Yuxiao Wen, Yanjun Han, and Zhengyuan Zhou. Stochastic contextual bandits with graph feedback: from independence number to mas number. *Advances in Neural Information Processing Systems*, 37:64499–64522, 2024.

Peter Whittle. Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability*, 25(A):287–298, 1988.

Guojun Xiong, Shufan Wang, Gang Yan, and Jian Li. Reinforcement learning for dynamic dimensioning of cloud caches: A restless bandit approach. *IEEE/ACM Transactions on Networking*, 31 (5):2147–2161, 2023.

Lily Xu, Bryan Wilder, Elias Khalil, and Milind Tambe. Reinforcement learning with combinatorial actions for coupled restless bandits. In *International Conference on Learning Representations*, volume 2025, pp. 61875–61896, 2025.

Hanmo Zhang, Zenghui Sun, and Kai Wang. Networked restless multi-arm bandits with reinforcement learning. *arXiv preprint arXiv:2512.06274*, 2025.

Y. Zhao, T. Wang, D. M. Nagaraj, A. Taneja, and M. Tambe. The bandit whisperer: Communication learning for restless bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23404–23413, 2025.

Mengqiu Zhou, Vincent YF Tan, and Meng Zhang. Timely best arm identification in restless shared networks. *arXiv preprint arXiv:2603.15078*, 2026.

Appendix Contents

A Notation **14**

B Related Work **14**

C Computing Exposure Marginal Values **15**

C.1 Model-based local dynamic programming 16

C.2 Finite differences and nonsmooth value functions 16

C.3 Differentiable local critics 17

C.4 Regularization and bounded sensitivities 17

C.5 Use inside PAI 17

D Practical Variants of PAI **18**

D.1 Model-based full-resolvent PAI 18

D.2 Truncated-propagation PAI 18

D.3 Learned-local-value PAI 18

D.4 Hybrid implementations 19

E Detailed Proofs **19**

E.1 Standing Assumptions 19

E.2 Proof of Lemma 1 20

E.3 Resolvent representation and first-order impulse value 20

E.4 Proof of Proposition 1 21

E.5 Proof of Proposition 2 22

E.6 Proof of Theorem 1 23

E.7 Proof of Corollary 1 23

E.8 Proof of Theorem 2 24

E.9 Optional consequence: weak-coupling and accurate-score limit 25

F Limitations **25**

G Additional Experimental Results **26**

G.1 Implementation and evaluation protocol 26

G.2 Multi-hop comparison and PAI sensitivity 27

G.3 Saturation and imperfect state information 28

G.4 Structured graph errors and additional scaling results 29

H Experimental Reproducibility Details **30**

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing, including grammar correction, improving clarity and conciseness, reorganizing exposition, and suggesting LaTeX phrasing and formatting. These tools were used to improve the presentation of the manuscript and response materials. All AI-assisted content was reviewed, edited, and verified by the authors. The authors manually checked mathematical notation, assumptions, derivations, theorem statements, experimental descriptions, and citations against the final manuscript, and supporting materials. We take responsibility for the final content of this work, including all claims, methods, theoretical results, experiments, and artifacts produced with the aid of generative AI.

A NOTATION

Table 1 summarizes the main notation used throughout the paper.

B RELATED WORK

Classical RMABs and Whittle-style methods. A large body of work studies restless multi-armed bandits (RMABs) through single-arm relaxations and index-based prioritization, following the foundational Whittle paradigm (Whittle, 1988). Subsequent work developed the theory of index policies, indexability conditions, and scalable approximation methods for RMABs (Bertsimas & Niño-Mora, 2000; Glazebrook et al., 2011; Niño-Mora, 2023). Recent extensions include average-reward asymptotic optimality beyond standard assumptions (Hong et al., 2023), and domain-specific constrained RMAB formulations such as cloud cache dimensioning (Xiong et al., 2023), multirobot operator allocation (Dahiya et al., 2022), and time-constrained service scheduling (Mao & Perrault, 2024). Fairness-aware RMAB formulations have also been studied (Li & Varakantham, 2022). A parallel line of work considers RMABs with unknown dynamics, contextual information, or robustness considerations using online learning and reinforcement learning techniques (Wang et al., 2023; Akbarzadeh & Mahajan, 2023; Tong et al., 2024; Gafni et al., 2022; Liang et al., 2025; Killian et al., 2022; Zhao et al., 2025; Boehmer et al., 2025). These works substantially broaden the scope of RMAB methodology, but they still largely rely on the core RMAB structure in which arms are decomposable apart from the budget constraint or additional scheduling constraints. In contrast, our setting introduces persistent cross-arm dynamic coupling through a graph-propagated latent externality state, which breaks the standard single-arm decomposition at the level of future state evolution.

Coupled, networked, and graph-aware bandits. Several recent works move beyond the classical RMAB separability assumption more directly. RMABs with global non-separable rewards have been studied in Raman et al. (2024), where index extensions and adaptive planning methods are developed for reward coupling, and more generic coupled RMAB settings have been considered by finetuning Whittle-style policies (Luan et al., 2025). Closely related work also studies graph-structured interactions among arms. Networked restless bandits have been proposed for mobile interventions and positive externalities (Ou et al., 2022; Herlihy & Dickerson, 2023), while more recent work considers networked RMABs with reinforcement learning (Zhang et al., 2025). Beyond RMABs, graph-triggered bandits study reward evolution induced by graph interactions among arms (Genalti et al., 2024), and distributed restless bandit learning over interference graphs has been explored in wireless coordination problems (Didi & Cohen, 2025). These works all recognize that exact single-arm decomposition may fail in realistic environments with cross-arm dependence. However, the dependence they consider differs structurally from ours. Prior coupled RMAB work often introduces non-separable rewards or generic joint dependence, while graph-aware bandit models typically emphasize immediate graph-triggered effects or graph-constrained coordination. In contrast, DDE-RMABs introduce a *latent externality state* that is injected by actions, persists over time, and diffuses over the graph, thereby coupling arms through delayed, accumulated, and multi-hop modulation of future local transitions. This additional temporal and spatial structure is what enables our propagation-adjusted index and its associated approximation.

A systematic comparison between our DDE-RMAB framework and related bandit paradigms is summarized in Table 2.

Table 1: Main notation. Exposure propagates forward through $P^\top$, whereas downstream marginal values are aggregated back to source arms through P.

Symbol	Meaning
$\mathcal{G} = (V, E)$	Directed interaction graph; each node represents an arm and each edge represents a potential exposure-propagation channel.
N	Number of arms, with $V = \{1, \dots, N\}$.
$P \in \mathbb{R}_+^{N \times N}$	Nonnegative propagation matrix, where P_{ij} is the exposure weight from arm i to arm j; $P^\top$ is the forward exposure operator.
$s_i^t,\ z_i^t$	Intrinsic state and latent exposure state of arm i before the decision at time t.
$X^t = (s^t, z^t)$	Joint system state at time t.
$a^t \in \{0,1\}^N$	Binary activation vector at time t; $a_i^t = 1$ means arm i is activated.
$\mathcal{A}_M$	Feasible action set $\{a \in \{0,1\}^N : \sum_i a_i \le M\}$.
M	Per-period activation budget.
α	Exposure persistence factor governing temporal decay and propagation.
$\beta_i(s_i)$	Exposure impulse injected by activating arm i in intrinsic state s_i.
$\beta_{\max}$	Uniform upper bound on exposure injections.
γ	Discount factor.
$R(X, a),\ r_i(s_i, a_i, z_i)$	Global one-step reward and local reward, respectively.
$V^\pi,\ V^\star$	Discounted value function under policy π and the optimal value function.
$Q^\star,\ A^\star$	Optimal action-value function and optimal activation advantage, where $A^\star(X, a) = Q^\star(X, a) - Q^\star(X, 0)$.
$\widetilde{V}_i,\ \widetilde{V}_{\mathrm{add}}$	Local value surrogate and additive backbone $\widetilde{V}_{\mathrm{add}}(s, z) = \sum_i \widetilde{V}_i(s_i, z_i)$.
$\Delta_i(s_i, z_i)$	Local activation advantage under the local surrogate.
ν_i^t	Selected local exposure sensitivity at time t, where $\nu_i^t \in \partial_{z_i} \widetilde{V}_i(s_i^t, z_i^t)$; it may be an ordinary derivative, Clarke subgradient, or finite-difference slope.
$G = (I - \gamma\alpha P)^{-1}$	Discounted value-propagation resolvent.
R_∞	Resolvent gain $\|G\|_\infty$, bounded by $(1-\gamma\alpha)^{-1}$ under the stated normalization.
Ψ_i^t	First-order propagation correction, i.e., the delayed spillover value of the exposure impulse created by activating arm i.
I_i^t	PAI score, $I_i^t = \Delta_i(s_i^t, z_i^t) + \kappa\Psi_i^t$.
κ	Propagation-correction scale; $\kappa = 1$ is the nominal first-order scaling.
H	Depth of the truncated Neumann expansion for approximate propagation.
$Z_{\max}$	Uniform exposure bound $\max\{\|z^0\|_\infty, \beta_{\max}/(1-\alpha)\}$.
$d_{\mathrm{off}}(P)$	Off-diagonal propagation mass $\|P^\top - \mathrm{diag}(P^\top)\|_\infty$.
$\mathcal{E}_{\mathrm{dec}}$	Bellman residual of the additive value decomposition.
$\mathcal{E}_{\mathrm{act}}(\kappa)$	Residual from approximating a joint activation value by the sum of scaled per-arm PAI scores.
$\Omega_M(P)$	Propagated influence-overlap quantity for activation sets of size at most M.
η_ν	Uniform marginal-value estimation error $\|\widehat{\nu} - \nu\|_\infty$.
η_ℓ	Frozen-sensitivity drift rate in $\|\nu^{t+\tau} - \nu^t\|_\infty \le \eta_\ell \tau$.

C Computing Exposure Marginal Values

This appendix describes practical estimators of the local exposure marginal value $\nu_i(s_i, z_i) \in \partial_{z_i} \widetilde{V}_i(s_i, z_i)$ used by PAI. Its role is to measure the local long-run value change induced by a marginal increase in arm i's exposure. At each decision epoch, PAI only requires the field $\nu^t = (\nu_1^t, \dots, \nu_N^t)$; it then analytically propagates this field through the resolvent correction in Eq. 4 and forms the final score in Eq. 5. The procedures below differ only in how the local sensitivity field is obtained.

Table 2: Comparison with related bandit and RMAB frameworks across representative works. ✓, ×, and △ denote full support, no support, and partial/structurally different mechanism, respectively.

Methods / Frameworks	Latent Exposure	Transition Coupling	Delayed Diffusion	Index-Style Decomposition
Networked and coupled RMABs				
GRETA (Herlihy & Dickerson, 2023)	✗	✓	✗	△
Networked RMAB (Ou et al., 2022)	✗	✓	✗	✗
Networked RMAB-RL (Zhang et al., 2025)	✗	✓	△	✗
Global Rewards RMAB (Raman et al., 2024)	✗	✗	✗	△
WiFiP / Finetuned Whittle (Luan et al., 2025)	✗	✓	✗	△
SEQUOIA / cORMAB (Xu et al., 2025)	✗	△	✗	✗
Graph-triggered and state-evolution bandits				
Graph-Triggered Rising Bandits (Genalti et al., 2024)	✗	✗	△	✗
Rising and Rotting Triggered Bandits (Genalti et al., 2026)	✗	✗	△	✗
Graph feedback, side-observations, and anomaly detection				
Anomaly Detection in Networked Bandits (Cheng & Maghsudi, 2025)	✗	✗	✗	✗
Graph Bandits with Side-Observations (Gong & Zhang, 2025)	✗	✗	✗	✗
Contextual Bandits with Graph Feedback (Wen et al., 2024)	✗	✗	✗	✗
Distributed networks and shared-dependency bandits				
Distributed Multi-Agent Bandits (Liu et al., 2026)	✗	✗	✗	✗
Restless Shared Networks (BAI) (Zhou et al., 2026)	✗	✗	✗	✗
Markovian Graph Dependencies (Sawwan & Wu, 2025)	✗	✗	✗	△
Ours: DDE-RMAB + PAI	✓	✓	✓	✓

C.1 Model-based local dynamic programming

When local transition models are known, we compute each local surrogate $\widetilde{V}_i$ from a single-arm dynamic program. A representative local Bellman equation is

$$\widetilde{V}_i(s_i, z_i) = \max_{a_i \in \{0,1\}} \left\{ r_i(s_i, a_i, z_i) + \gamma \mathbb{E}\left[\widetilde{V}_i(s_i', \bar{z}_i') \mid s_i, z_i, a_i \right] \right\}, \tag{13}$$

where $\bar{z}_i'$ is the baseline local exposure update. Depending on the surrogate, it may retain self-exposure dynamics while replacing incoming cross-arm exposure by a frozen, mean-field, or estimated local quantity. Consistent with the construction of Δ_i in Section 3, the local baseline excludes the incremental network impulse created by the current activation; PAI accounts for that impulse through Eq. 4.

When $\widetilde{V}_i$ is differentiable in z_i, we use $\nu_i(s_i, z_i) = \partial_{z_i} \widetilde{V}_i(s_i, z_i)$. The derivative can be obtained analytically for simple transition models or by differentiating through the local Bellman fixed point. In tabular implementations, we instead compute a finite-difference or subgradient estimate.

C.2 Finite differences and nonsmooth value functions

Suppose exposure is discretized on $\mathcal{Z}_i = \{z_i^{(1)}, \ldots, z_i^{(K)}\}$ and $\widetilde{V}_i$ is represented over $\mathcal{S}_i \times \mathcal{Z}_i$. At an interior grid point $z_i^{(k)}$, define the left and right slopes

$$D_{i,k}^- = \frac{\widetilde{V}_i(s_i, z_i^{(k)}) - \widetilde{V}_i(s_i, z_i^{(k-1)})}{z_i^{(k)} - z_i^{(k-1)}}, \qquad D_{i,k}^+ = \frac{\widetilde{V}_i(s_i, z_i^{(k+1)}) - \widetilde{V}_i(s_i, z_i^{(k)})}{z_i^{(k+1)} - z_i^{(k)}}. \tag{14}$$

At smooth points, we use the central-difference estimate $\nu_i(s_i, z_i^{(k)}) = (D_{i,k}^- + D_{i,k}^+)/2$. At a kink caused by a maximization, interpolation, clipping, or a tie between actions, the one-dimensional

Clarke subdifferential is

$$\partial_{z_i} \widetilde{V}_i(s_i, z_i^{(k)}) = \left[\min\{D_{i,k}^-, D_{i,k}^+\}, \max\{D_{i,k}^-, D_{i,k}^+\}\right].$$

We select its minimum-norm element, namely the point in this interval closest to zero. At boundary points, only the available one-sided slope is used. This convention yields a well-defined bounded sensitivity field even when the local value function is nonsmooth, consistent with Assumption 4.

For a tabular value generated by $V_i(s_i, z_i) = \max_{a_i} Q_i(s_i, z_i, a_i)$, nonsmoothness can also arise when multiple actions maximize Q_i. In this case,

$$\partial_{z_i} V_i(s_i, z_i) = \operatorname{conv}\left\{\partial_{z_i} Q_i(s_i, z_i, a_i) : a_i \in \arg\max_b Q_i(s_i, z_i, b)\right\},$$

and we again use the minimum-norm element. This choice is deterministic and avoids arbitrary sensitivity jumps at action ties.

C.3 Differentiable local critics

For large or partially unknown systems, $\widetilde{V}_i$ can be represented by a differentiable local critic $\widetilde{V}_i(s_i, z_i; \theta_i)$. The critic may be trained by fitted value iteration, temporal-difference learning, or regression to local rollout returns. For example, a bootstrapped squared-loss objective is

$$\min_{\theta_i} \sum_{m=1}^{M_i} \left(\widetilde{V}_i(s_i^{(m)}, z_i^{(m)}; \theta_i) - \left[r_i^{(m)} + \gamma \widetilde{V}_i(s_i^{(m+1)}, z_i^{(m+1)}; \bar{\theta}_i)\right]\right)^2, \tag{15}$$

where $\bar{\theta}_i$ denotes a target or delayed critic parameter. Once trained, the marginal value is obtained by automatic differentiation: $\nu_i(s_i, z_i) = \partial_{z_i} \widetilde{V}_i(s_i, z_i; \theta_i)$.

This implementation changes neither the PAI score nor the propagation operator. It only replaces a model-based sensitivity estimate with a data-driven one. The estimation error $\|\widehat{\nu} - \nu\|_\infty$ is explicitly captured by Theorem 1.

C.4 Regularization and bounded sensitivities

Finite differences and learned critic gradients can be noisy. To enforce the bounded-sensitivity condition required by Assumption 4, we optionally use

$$\widehat{\nu}_i(s_i, z_i) \leftarrow \operatorname{clip}(\widehat{\nu}_i(s_i, z_i), -\nu_{\max}, \nu_{\max}). \tag{16}$$

For applications with independently justified monotonicity constraints, one may additionally regularize the critic or project its gradients onto the corresponding sign constraint. Such a projection is optional and is not required by PAI or by the theoretical results.

C.5 Use inside PAI

At decision time t, the selected estimator produces $\nu^t = (\nu_1(s_1^t, z_1^t), \ldots, \nu_N(s_N^t, z_N^t))$. PAI then computes the propagated sensitivity

$$g^t = (I - \gamma\alpha P)^{-1} \nu^t, \qquad \Psi^t = \gamma\, \beta(s^t) \odot g^t. \tag{17}$$

For a sparse or truncated implementation, it instead uses

$$g_H^t = \sum_{\tau=0}^{H} (\gamma\alpha P)^\tau \nu^t, \qquad \Psi_H^t = \gamma\, \beta(s^t) \odot g_H^t. \tag{18}$$

The associated truncation error is controlled by Corollary 1. Thus, sensitivity estimation is local, while propagation is global but analytic; this separation is the computational structure exploited by PAI.

D PRACTICAL VARIANTS OF PAI

This appendix summarizes practical implementations of PAI. All variants retain the same score structure from Eq. 5: $I_i^t = \Delta_i(s_i^t, z_i^t) + \kappa\Psi_i^t$, where the propagation correction is the first-order resolvent value in Eq. 4. They differ only in how the local advantage Δ_i, the marginal value field ν^t, and the resolvent-vector product are computed. Accordingly, these variants preserve the same approximation architecture: a local value backbone, an analytically propagated exposure impulse, and top-M selection using separable scores.

D.1 MODEL-BASED FULL-RESOLVENT PAI

When local rewards and transition models are known and the local state–exposure space is moderate, PAI can use a tabular or model-based local dynamic program. For each arm i, let $\widetilde{V}_i(s_i, z_i)$ solve a local surrogate Bellman equation of the form

$$\widetilde{V}_i(s_i, z_i) = \max_{a_i \in \{0,1\}} \left\{ r_i(s_i, a_i, z_i) + \gamma \mathbb{E}\Big[\widetilde{V}_i(s_i', \bar{z}_i') \mid s_i, z_i, a_i\Big] \right\}, \tag{19}$$

where $\bar{z}_i'$ denotes the baseline exposure update used by the local surrogate. As in Section 3, this baseline excludes the incremental network exposure impulse generated by the current activation, which is instead evaluated by the PAI propagation term.

The local advantage is $\Delta_i(s_i, z_i) = Q_i^{\mathrm{loc}}(s_i, z_i, 1) - Q_i^{\mathrm{loc}}(s_i, z_i, 0)$, and the sensitivity is selected as $\nu_i(s_i, z_i) \in \partial_{z_i}\widetilde{V}_i(s_i, z_i)$. For smooth surrogates, this is the ordinary derivative; for tabular or nonsmooth surrogates, it can be a finite-difference slope or Clarke subgradient, as described in Appendix C. At decision time, the full-resolvent implementation computes $g^t = (I - \gamma\alpha P)^{-1}\nu^t$ and uses

$$I_i^t = \Delta_i(s_i^t, z_i^t) + \kappa\gamma\beta_i(s_i^t)g_i^t. \tag{20}$$

The term "full resolvent" refers only to evaluating the infinite-horizon linear propagation response without truncation. It does not remove the additive-backbone, first-order, or frozen-sensitivity approximations characterized in Section 4.

D.2 TRUNCATED-PROPAGATION PAI

For large sparse graphs, PAI need not explicitly form the inverse $(I - \gamma\alpha P)^{-1}$. Instead, it can approximate the resolvent-vector product using the depth-H Neumann expansion

$$g_H^t = \sum_{\tau=0}^{H} (\gamma\alpha P)^\tau \nu^t, \qquad \Psi_H^t = \gamma\,\beta(s^t) \odot g_H^t. \tag{21}$$

The resulting score is $I_{i,H}^t = \Delta_i(s_i^t, z_i^t) + \kappa\gamma\beta_i(s_i^t)g_{H,i}^t$. When $H = 0$, this includes only the immediate exposure value, $\Psi_0^t = \gamma\beta(s^t) \odot \nu^t$; larger H incorporates increasingly long delayed propagation paths.

Each additional depth requires one sparse matrix-vector multiplication, so the propagation cost is $O(H|E|)$ per decision epoch. By Corollary 1, the resulting propagation error is bounded by

$$\|\Psi(\nu) - \Psi^{(H)}(\nu)\|_\infty \le \gamma\beta_{\max}\nu_{\max}\frac{(\gamma\alpha)^{H+1}}{1-\gamma\alpha}. \tag{22}$$

Thus, truncation is particularly effective when $\gamma\alpha$ is bounded away from one. Our depth-sensitivity experiments further compare the resolvent score with fixed-depth multi-hop propagation.

D.3 LEARNED-LOCAL-VALUE PAI

When local transition models are unavailable or the local state space is too large for tabular dynamic programming, the local surrogate can be replaced by a differentiable value or action-value model. In this case, a local critic provides both the direct priority and marginal exposure value: $\widehat{\Delta}_i$ is obtained from the estimated active–passive value difference, while $\widehat{\nu}_i(s_i, z_i) = \partial_{z_i}\widetilde{V}_i(s_i, z_i; \theta_i)$ is obtained

Table 3: Practical variants of PAI. "Full resolvent" denotes a resolvent-vector solve rather than explicit matrix inversion.

	Model-based	**Truncated**	**Learned-local-value**
Local quantities	Local dynamic program	Local DP or critic	Differentiable critic
Propagation	Full resolvent	H-step Neumann expansion	Full or truncated resolvent
Typical setting	Known, moderate local model	Large sparse graph	Unknown or large local model

by automatic differentiation. The critic can be trained by fitted value iteration, temporal-difference learning, or regression to local rollout targets; implementation details are given in Appendix C.

The learned-local-value score is

$$\widehat{I}_i^t = \widehat{\Delta}_i(s_i^t, z_i^t) + \kappa\gamma\beta_i(s_i^t)\left[(I - \gamma\alpha P)^{-1}\widehat{\nu}^t\right]_i. \tag{23}$$

This replaces model-based local estimation with a learned estimate while retaining analytic network propagation. The effect of marginal-value estimation error is explicitly controlled by Theorem 1.

D.4 Hybrid implementations

The local-value and propagation choices are modular. For example, a tabular local surrogate can be paired with truncated propagation on a large sparse graph; a learned local critic can be paired with a full-resolvent linear solve; and a learned critic can also be combined with truncated propagation. In every case, the online decision rule remains a top-M selection using the same PAI score structure. The variants therefore differ in computational trade-offs, rather than in their underlying treatment of delayed network exposure.

The model-based variant is most directly aligned with the local surrogate derivation. The truncated variant reduces online graph-propagation cost with a geometric tail bound. The learned-local-value variant relaxes the requirement of a known local model while preserving the same sensitivity-weighted resolvent correction.

E Detailed Proofs

E.1 Standing Assumptions

We collect the assumptions used in Section 4. They are sufficient conditions for bounded exposure, a well-defined resolvent, and controlled local approximation errors; they do not imply that the arms are independent.

Assumption 1 (Bounded rewards and exposure injections)**.** *There exist constants $r_{\max}, \beta_{\max} < \infty$ such that, for every arm i, local state s_i, action a_i, and exposure level z_i,*

$$|r_i(s_i, a_i, z_i)| \le r_{\max}, \qquad 0 \le \beta_i(s_i) \le \beta_{\max}.$$

Consequently, the additive reward satisfies $|R(X, a)| \le N r_{\max} =: R_{\max}$, and every feasible action satisfies $\|\beta(s) \odot a\|_\infty \le \beta_{\max}$.

Assumption 2 (Propagation normalization and discounting)**.** *The propagation matrix P is nonnegative and satisfies $\|P^\top\|_\infty \le 1$ and $\|P\|_\infty \le 1$. Moreover, $\alpha \in [0, 1)$ and $\gamma \in (0, 1)$.*

The condition $\|P^\top\|_\infty \le 1$ controls the forward exposure dynamics $z^{t+1} = \alpha P^\top z^t + \beta(s^t) \odot a^t$, whereas $\|P\|_\infty \le 1$ controls the value-propagation resolvent $(I - \gamma\alpha P)^{-1}$. Since $\|P^\top\|_\infty = \|P\|_1$, the first condition additionally controls column-wise propagated influence, which is used in the overlap quantity $\Omega_M(P)$. Together with $\alpha < 1$ and $\gamma < 1$, these conditions imply $\gamma\alpha\rho(P) < 1$ and ensure convergence of the Neumann series defining the resolvent.

Assumption 3 (Lipschitz exposure modulation)**.** *For each arm i, the local reward and transition kernel are Lipschitz in the local exposure variable: for all s_i, a_i, z_i, z_i',*

$$|r_i(s_i, a_i, z_i) - r_i(s_i, a_i, z_i')| \le L_r |z_i - z_i'|,$$

and

$$\|\mathbb{P}_i(\cdot \mid s_i, a_i, z_i) - \mathbb{P}_i(\cdot \mid s_i, a_i, z_i')\|_{\mathrm{TV}} \le L_p |z_i - z_i'|.$$

Assumption 4 (Regularity of the local surrogate)**.** *Let $\widetilde{V}_{\mathrm{add}}(s,z) = \sum_i \widetilde{V}_i(s_i, z_i)$ denote the additive local surrogate. It is L_{add}-Lipschitz in exposure:*

$$|\widetilde{V}_{\mathrm{add}}(s,z) - \widetilde{V}_{\mathrm{add}}(s,z')| \le L_{\mathrm{add}} \|z - z'\|_\infty.$$

Moreover, each local marginal exposure value $\nu_i(s_i, z_i) \in \partial_{z_i} \widetilde{V}_i(s_i, z_i)$ is uniformly bounded as $\|\nu\|_\infty \le \nu_{\max}$.

When $\widetilde{V}_i$ is differentiable, the sensitivity is the ordinary derivative $\nu_i = \partial_{z_i} \widetilde{V}_i$. At nondifferentiable points caused by maximization, interpolation, or truncation, we use any valid one-dimensional Clarke subgradient or finite-difference slope. The results in Section 4 require only that the selected sensitivity field be bounded.

Assumption 5 (Smooth marginal-value field)**.** *The selected marginal-value field is L_ν-Lipschitz with respect to exposure:*

$$\|\nu(s,z) - \nu(s,z')\|_\infty \le L_\nu \|z - z'\|_\infty.$$

Assumption 5 is used only to bound the second-order simultaneous-action residual in Proposition 2. The frozen-sensitivity drift condition $\|\nu^{t+\tau} - \nu^t\|_\infty \le \eta_\ell \tau$ is not imposed globally here; it is stated separately in Theorem 1, where it is needed.

Under Assumptions 1– 2, Lemma 1 yields the uniform exposure bound $\|z^t\|_\infty \le Z_{\max} = \max\{\|z^0\|_\infty, \beta_{\max}/(1-\alpha)\}$. Assumptions 3– 5 then control, respectively, the local model perturbation, additive decomposition residual, and simultaneous-action interaction error.

E.2 Proof of Lemma 1

Proof. Recall the exposure dynamics from Eq. 1:

$$z^{t+1} = \alpha P^\top z^t + \beta(s^t) \odot a^t.$$

Since $a_i^t \in \{0,1\}$ and $0 \le \beta_i(s_i^t) \le \beta_{\max}$, we have $\|\beta(s^t) \odot a^t\|_\infty \le \beta_{\max}$. By $\|P^\top\|_\infty \le 1$,

$$\|z^{t+1}\|_\infty \le \alpha \|P^\top z^t\|_\infty + \|\beta(s^t) \odot a^t\|_\infty \le \alpha \|z^t\|_\infty + \beta_{\max}. \tag{A.1}$$

Let $u_t = \|z^t\|_\infty$. Iterating A.1 yields

$$u_t \le \alpha^t u_0 + \beta_{\max} \sum_{\ell=0}^{t-1} \alpha^\ell = \alpha^t \|z^0\|_\infty + \frac{1-\alpha^t}{1-\alpha} \beta_{\max}. \tag{A.2}$$

The right-hand side is a convex combination of $\|z^0\|_\infty$ and $\beta_{\max}/(1-\alpha)$. Therefore,

$$\|z^t\|_\infty \le \max\left\{ \|z^0\|_\infty, \frac{\beta_{\max}}{1-\alpha} \right\} = Z_{\max} \qquad \text{for all } t \ge 0, \tag{A.3}$$

which proves the claim. □

E.3 Resolvent representation and first-order impulse value

We first verify the resolvent representation and Eq. 4. Since $\|P\|_\infty \le 1$ and $\gamma\alpha < 1$, the Neumann series converges absolutely in the induced ℓ_∞ operator norm:

$$G = (I - \gamma\alpha P)^{-1} = \sum_{\tau=0}^{\infty} (\gamma\alpha P)^\tau. \tag{A.4}$$

Hence, for every vector v,

$$[Gv]_i = \sum_{\tau=0}^{\infty} (\gamma\alpha)^\tau [P^\tau v]_i,$$

. Moreover,

$$\|G\|_\infty \le \sum_{\tau=0}^{\infty} (\gamma\alpha)^\tau \|P\|_\infty^\tau \le \frac{1}{1-\gamma\alpha}. \tag{A.5}$$

Now consider activating arm i at time t. The activation injects $\delta z^{t+1} = \beta_i(s_i^t)e_i$. Holding future actions fixed, the resulting exposure perturbation after τ additional propagation steps is

$$\delta z^{t+1+\tau} = (\alpha P^\top)^\tau \beta_i(s_i^t) e_i. \tag{A.6}$$

Under the frozen-sensitivity approximation, the first-order discounted continuation value of this impulse is

$$\begin{aligned} \Psi_i^t &= \sum_{\tau=0}^{\infty} \gamma^{\tau+1} \left\langle \nu^t, (\alpha P^\top)^\tau \beta_i(s_i^t) e_i \right\rangle \\ &= \gamma \beta_i(s_i^t) \sum_{\tau=0}^{\infty} (\gamma\alpha)^\tau (\nu^t)^\top (P^\top)^\tau e_i. \end{aligned} \tag{A.7}$$

Using $(\nu^t)^\top (P^\top)^\tau e_i = e_i^\top P^\tau \nu^t$, we obtain

$$\Psi_i^t = \gamma\beta_i(s_i^t) \left[\sum_{\tau=0}^{\infty} (\gamma\alpha P)^\tau \nu^t \right]_i = \gamma\beta_i(s_i^t)[G\nu^t]_i, \tag{A.8}$$

which proves Eq. 4. In vector form, $\Psi^t = \gamma\beta(s^t) \odot G\nu^t$.

E.4 Proof of Proposition 1

Proof. Let $\mathcal{T}$ be the optimal Bellman operator of the coupled DDE-RMAB, and let $\widetilde{V}_{\mathrm{add}}(s,z) = \sum_i \widetilde{V}_i(s_i, z_i)$ denote the additive local surrogate. By definition,

$$\mathcal{E}_{\mathrm{dec}} = \|\mathcal{T}\widetilde{V}_{\mathrm{add}} - \widetilde{V}_{\mathrm{add}}\|_\infty. \tag{A.9}$$

Since $\mathcal{T}$ is a γ-contraction in the sup norm and $V^\star = \mathcal{T}V^\star$, we have

$$\begin{aligned} \|V^\star - \widetilde{V}_{\mathrm{add}}\|_\infty &= \|\mathcal{T}V^\star - \widetilde{V}_{\mathrm{add}}\|_\infty \\ &\le \|\mathcal{T}V^\star - \mathcal{T}\widetilde{V}_{\mathrm{add}}\|_\infty + \|\mathcal{T}\widetilde{V}_{\mathrm{add}} - \widetilde{V}_{\mathrm{add}}\|_\infty \\ &\le \gamma\|V^\star - \widetilde{V}_{\mathrm{add}}\|_\infty + \mathcal{E}_{\mathrm{dec}}. \end{aligned} \tag{A.10}$$

Rearranging proves

$$\|V^\star - \widetilde{V}_{\mathrm{add}}\|_\infty \le \frac{\mathcal{E}_{\mathrm{dec}}}{1-\gamma}. \tag{A.11}$$

It remains to derive the primitive bound in Eq. 8. Let $z'_{\mathrm{true}}(X,a)$ be the true next exposure update and let $z'_{\mathrm{local}}(X,a)$ be the update used by the local surrogate. Define

$$\delta_{\mathrm{dec}} = \sup_{X, a\in\mathcal{A}_M} \|z'_{\mathrm{true}}(X,a) - z'_{\mathrm{local}}(X,a)\|_\infty. \tag{A.12}$$

Suppose that $\widetilde{V}_{\mathrm{add}}$ is L_{add}-Lipschitz with respect to the exposure argument. For a fixed state-action pair, the only difference between the coupled and local Bellman continuation updates is the exposure argument passed to the surrogate continuation value. Therefore,

$$\left| (\mathcal{T}\widetilde{V}_{\mathrm{add}})(X) - \widetilde{V}_{\mathrm{add}}(X) \right| \le \gamma L_{\mathrm{add}} \delta_{\mathrm{dec}}. \tag{A.13}$$

Taking the supremum over X gives $\mathcal{E}_{\mathrm{dec}} \le \gamma L_{\mathrm{add}} \delta_{\mathrm{dec}}$.

For the masked local surrogate, off-diagonal incoming exposure is removed. Thus,

$$z'_{\mathrm{true}} - z'_{\mathrm{local}} = \alpha \left(P^\top - \mathrm{diag}(P^\top) \right) z. \tag{A.14}$$

Using Lemma 1, $\|z\|_\infty \le Z_{\max}$, and hence

$$\delta_{\mathrm{dec}} \le \alpha \left\| P^\top - \mathrm{diag}(P^\top) \right\|_\infty Z_{\max} = \alpha Z_{\max} d_{\mathrm{off}}(P). \tag{A.15}$$

Combining A.13 and A.15 yields

$$\mathcal{E}_{\mathrm{dec}} \le \gamma L_{\mathrm{add}} \alpha Z_{\max} d_{\mathrm{off}}(P). \tag{A.16}$$

Finally, $Z_{\max} \le \max\{\|z^0\|_\infty, \beta_{\max}/(1-\alpha)\}$. When the initial exposure is of the same order as, or smaller than, the stable exposure scale, this gives the stated scaling $\mathcal{E}_{\mathrm{dec}} \lesssim \gamma L_{\mathrm{add}} \alpha \beta_{\max} d_{\mathrm{off}}(P)/(1-\alpha)$. □

E.5 Proof of Proposition 2

Proof. We first consider the nominal first-order scaling $\kappa = 1$. Let $S = \{i : a_i = 1\}$ be an activation set with $|S| \le M$. The aggregate exposure impulse generated by S is

$$b_S = \sum_{i \in S} \beta_i(s_i) e_i. \tag{A.17}$$

Its discounted propagated exposure-value direction is Gb_S. The first-order term is additive because matrix multiplication is linear:

$$Gb_S = \sum_{i \in S} \beta_i(s_i) G e_i. \tag{A.18}$$

Consequently, the first-order approximation to the propagated value of S is exactly the sum of the individual first-order corrections $\sum_{i \in S} \Psi_i$.

The discrepancy arises from second-order interaction terms: after multiple activations, the marginal-value field changes as the aggregate exposure perturbation changes. Since the sensitivity field is L_ν-Lipschitz, the corresponding continuation value admits the standard quadratic remainder bound

$$\left| A_{\mathrm{add}}(X, a) - \sum_{i \in S} (\Delta_i + \Psi_i) \right| \le C_{\mathrm{act}} L_\nu \|G b_S\|_2^2, \tag{A.19}$$

where C_{act} is a universal constant that depends only on the particular first-order expansion convention. This constant is absorbed by the notation $\lesssim$ in the main text.

Because P is nonnegative, G is entrywise nonnegative. Moreover, $\beta_i(s_i) \le \beta_{\max}$, so for every coordinate j,

$$|(Gb_S)_j| = \left| \sum_{i \in S} G_{ji} \beta_i(s_i) \right| \le \beta_{\max} \sum_{i \in S} G_{ji}. \tag{A.20}$$

Therefore,

$$\|Gb_S\|_2^2 \le \beta_{\max}^2 \sum_j \left(\sum_{i \in S} G_{ji} \right)^2 \le \beta_{\max}^2 \Omega_M(P). \tag{A.21}$$

Taking the supremum over all feasible S proves

$$\mathcal{E}_{\mathrm{act}}(1) \lesssim L_\nu \beta_{\max}^2 \Omega_M(P). \tag{A.22}$$

We next derive the worst-case upper bound on $\Omega_M(P)$. For any S with $|S| \le M$, nonnegativity gives

$$\sum_j \left(\sum_{i \in S} G_{ji} \right)^2 \le \left(\sum_j \sum_{i \in S} G_{ji} \right)^2. \tag{A.23}$$

Since $\sum_j G_{ji}$ is the ℓ_1 norm of column i of G, this bound is directly controlled when the forward operator is normalized in the corresponding ℓ_1 norm. Under the normalization used in the main text, the same conservative bound can be expressed through the resolvent gain as

$$\Omega_M(P) \le M^2 R_\infty^2 \le \frac{M^2}{(1 - \gamma\alpha)^2}, \tag{A.24}$$

where the norm-equivalence and graph-normalization constants are absorbed into the conservative inequality. In sparse graphs with weakly overlapping downstream regions, the cross terms in Eq. A.21 are small, and the scaling can be closer to $O(M)$ rather than $O(M^2)$.

Finally, consider a general propagation scale $\kappa \ge 0$. By the triangle inequality,

$$\begin{aligned} \mathcal{E}_{\mathrm{act}}(\kappa) &\le \mathcal{E}_{\mathrm{act}}(1) + \sup_{X, a \in \mathcal{A}_M} \left| (1 - \kappa) \sum_i a_i \Psi_i \right| \\ &\le \mathcal{E}_{\mathrm{act}}(1) + |\kappa - 1| M \|\Psi\|_\infty. \end{aligned} \tag{A.25}$$

Using $\|\beta(s)\|_\infty \le \beta_{\max}$, $\|\nu\|_\infty \le \nu_{\max}$, and $\|G\|_\infty = R_\infty$, Eq. A.8 gives $\|\Psi\|_\infty \le \gamma \beta_{\max} R_\infty \nu_{\max}$. Substituting this into A.25 proves

$$\mathcal{E}_{\mathrm{act}}(\kappa) \le \mathcal{E}_{\mathrm{act}}(1) + |\kappa - 1| M \gamma \beta_{\max} R_\infty \nu_{\max}, \tag{A.26}$$

which proves. □

Why $\kappa = 0$ does not remove approximation error. Equation A.26 should not be interpreted as saying that a smaller κ necessarily improves the policy. When $\kappa = 0$, the score discards the first-order propagation value Ψ_i. The omitted value then becomes part of $\mathcal{E}_{\mathrm{act}}(0)$. Thus, although the implementation no longer uses a propagation estimate, the policy can incur a larger action-value approximation error because it ignores genuine delayed spillovers.

E.6 Proof of Theorem 1

Proof. We prove the frozen-sensitivity and marginal-value estimation bounds separately.

Frozen-sensitivity error. The exact first-order propagation value that uses the future sensitivity field is

$$\Psi^t_{i,\mathrm{true}} = \sum_{\tau=0}^{\infty} \gamma^{\tau+1} \left\langle \nu^{t+\tau}, (\alpha P^\top)^\tau \beta_i(s_i^t) e_i \right\rangle . \tag{A.27}$$

Subtracting the frozen approximation in Eq. A.7 gives

$$\Psi^t_{i,\mathrm{true}} - \Psi^t_i = \sum_{\tau=0}^{\infty} \gamma^{\tau+1} \beta_i(s_i^t) \left\langle \nu^{t+\tau} - \nu^t, (\alpha P^\top)^\tau e_i \right\rangle . \tag{A.28}$$

For vectors u, v, $|\langle u, v\rangle| \le \|u\|_\infty \|v\|_1$. Moreover, $\|(P^\top)^\tau e_i\|_1 \le \|(P^\top)^\tau\|_1 = \|P^\tau\|_\infty \le 1$, where the equality follows from transpose duality and the last inequality uses $\|P\|_\infty \le 1$. Hence, under $\|\nu^{t+\tau} - \nu^t\|_\infty \le \eta_\ell \tau$,

$$\begin{aligned} \left|\Psi^t_{i,\mathrm{true}} - \Psi^t_i\right| &\le \gamma \beta_{\max} \sum_{\tau=0}^{\infty} (\gamma\alpha)^\tau \eta_\ell \tau \\ &= \gamma \beta_{\max} \eta_\ell \frac{\gamma\alpha}{(1-\gamma\alpha)^2}. \end{aligned} \tag{A.29}$$

Since $\gamma\alpha \le 1$, this implies the slightly looser but simpler bound stated in Eq. 10:

$$\left|\Psi^t_{i,\mathrm{true}} - \Psi^t_i\right| \le \frac{\gamma \beta_{\max} \eta_\ell}{(1-\gamma\alpha)^2}. \tag{A.30}$$

Marginal-value estimation error. Let $\|\widehat{\nu} - \nu\|_\infty \le \eta_\nu$. From Eq. A.8,

$$\Psi(\widehat{\nu}) - \Psi(\nu) = \gamma \beta(s) \odot G(\widehat{\nu} - \nu). \tag{A.31}$$

Taking the ℓ_∞ norm yields

$$\begin{aligned} \|\Psi(\widehat{\nu}) - \Psi(\nu)\|_\infty &\le \gamma \|\beta(s)\|_\infty \|G\|_\infty \|\widehat{\nu} - \nu\|_\infty \\ &\le \gamma \beta_{\max} R_\infty \eta_\nu \le \frac{\gamma \beta_{\max} \eta_\nu}{1-\gamma\alpha}. \end{aligned} \tag{A.32}$$

This proves both claims. □

E.7 Proof of Corollary 1

Proof. By Eq. A.4, the full propagation correction is

$$\Psi(\nu) = \gamma \beta(s) \odot \sum_{\tau=0}^{\infty} (\gamma\alpha P)^\tau \nu.$$

The depth-H truncation is

$$\Psi^{(H)}(\nu) = \gamma \beta(s) \odot \sum_{\tau=0}^{H} (\gamma\alpha P)^\tau \nu.$$

Their difference is therefore the Neumann-series tail:

$$\Psi(\nu) - \Psi^{(H)}(\nu) = \gamma \beta(s) \odot \sum_{\tau=H+1}^{\infty} (\gamma\alpha P)^\tau \nu. \tag{A.33}$$

Using $\|\beta(s)\|_\infty \le \beta_{\max}$, $\|P\|_\infty \le 1$, and $\|\nu\|_\infty \le \nu_{\max}$,

$$\begin{aligned}\|\Psi(\nu) - \Psi^{(H)}(\nu)\|_\infty &\le \gamma\beta_{\max} \sum_{\tau=H+1}^{\infty} (\gamma\alpha)^\tau \|P^\tau \nu\|_\infty \\ &\le \gamma\beta_{\max}\nu_{\max} \sum_{\tau=H+1}^{\infty} (\gamma\alpha)^\tau \\ &= \gamma\beta_{\max}\nu_{\max} \frac{(\gamma\alpha)^{H+1}}{1-\gamma\alpha}.\end{aligned} \tag{A.34}$$

. □

E.8 Proof of Theorem 2

Proof. The proof combines the approximation chain in Eq. 7 with a standard approximate-greedy performance argument.

For a state X and feasible action a, let $A^\star(X,a) = Q^\star(X,a) - Q^\star(X,0)$ be the optimal activation advantage. Let A_{add} denote the advantage computed with the additive backbone, and let A_{FO} denote the first-order propagation approximation using the exact, untruncated marginal field. Finally, let $\widehat{A}_{\text{PAI}}$ denote the advantage induced by the implemented score $\widehat{I}_i = \Delta_i + \kappa\widehat{\Psi}_i^{(H)}$.

Step 1: error from the additive backbone. By Proposition 1, $\|V^\star - \widetilde{V}_{\text{add}}\|_\infty \le \mathcal{E}_{\text{dec}}/(1-\gamma)$. The induced action-value and advantage difference is bounded by a problem-dependent multiple of $\mathcal{E}_{\text{dec}}$. Namely, there exists a constant $C_{\text{dec}} > 0$ such that

$$\sup_{X, a\in\mathcal{A}_M} |A^\star(X,a) - A_{\text{add}}(X,a)| \le C_{\text{dec}}\mathcal{E}_{\text{dec}}. \tag{A.35}$$

The constant absorbs the conversion from a value-function residual to an action-advantage residual and is independent of H, η_ν, and η_ℓ.

Step 2: simultaneous-action approximation. By Proposition 2,

$$\sup_{X, a\in\mathcal{A}_M} \left| A_{\text{add}}(X,a) - \sum_i a_i(\Delta_i + \kappa\Psi_i) \right| \le \mathcal{E}_{\text{act}}(\kappa). \tag{A.36}$$

Step 3: propagation approximation. For an action set with at most M arms, replacing Ψ by the truncated estimated score $\widehat{\Psi}^{(H)}$ introduces an error bounded by a problem-dependent multiple of the per-arm score error. Using Corollary 1 and Theorem 1, there exist constants $C_H, C_\nu, C_{\text{fr}} > 0$ such that

$$\begin{aligned}\sup_{X, a\in\mathcal{A}_M} \left| \sum_i a_i\kappa\big(\Psi_i - \widehat{\Psi}_i^{(H)}\big) \right| &\le C_H \frac{\gamma\beta_{\max}\nu_{\max}(\gamma\alpha)^{H+1}}{1-\gamma\alpha} \\ &\quad + C_\nu \frac{\gamma\beta_{\max}\eta_\nu}{1-\gamma\alpha} + C_{\text{fr}} \frac{\gamma\beta_{\max}\eta_\ell}{(1-\gamma\alpha)^2}.\end{aligned} \tag{A.37}$$

The constants absorb the budget factor, the scale κ, and any problem-specific conversion between per-arm score perturbations and joint-action advantage perturbations.

Combining A.35, A.36, and A.37, there exist constants $\bar{C}_1, \dots, \bar{C}_4$ such that

$$\begin{aligned}\sup_{X, a\in\mathcal{A}_M} |A^\star(X,a) - \widehat{A}_{\text{PAI}}(X,a)| &\le \bar{C}_1\mathcal{E}_{\text{dec}} + \bar{C}_2\mathcal{E}_{\text{act}}(\kappa) \\ &\quad + \bar{C}_3 \frac{\gamma\beta_{\max}\nu_{\max}(\gamma\alpha)^{H+1}}{1-\gamma\alpha} \\ &\quad + \bar{C}_4 \left(\frac{\gamma\beta_{\max}\eta_\nu}{1-\gamma\alpha} + \frac{\gamma\beta_{\max}\eta_\ell}{(1-\gamma\alpha)^2} \right).\end{aligned} \tag{A.38}$$

Step 4: approximate greedy action loss. Let $a^\star(X) \in \arg\max_{a\in\mathcal{A}_M} A^\star(X,a)$, and let $\widehat{a}(X)$ be the action selected by PAI. Define the right-hand side of A.38 as $\varepsilon_{\mathrm{PAI}}$. Since $\widehat{a}(X)$ maximizes $\widehat{A}_{\mathrm{PAI}}(X,a)$,

$$\begin{aligned} A^\star(X,a^\star) - A^\star(X,\widehat{a}) &= \big(A^\star(X,a^\star) - \widehat{A}_{\mathrm{PAI}}(X,a^\star)\big) \\ &\quad + \big(\widehat{A}_{\mathrm{PAI}}(X,a^\star) - \widehat{A}_{\mathrm{PAI}}(X,\widehat{a})\big) \\ &\quad + \big(\widehat{A}_{\mathrm{PAI}}(X,\widehat{a}) - A^\star(X,\widehat{a})\big) \\ &\le 2\varepsilon_{\mathrm{PAI}}. \end{aligned} \tag{A.39}$$

The factor 2 is absorbed into the constants in the final theorem.

Step 5: discounted performance loss. For any policy π, define $\delta_\pi(X) = V^\star(X) - Q^\star(X,\pi(X))$. The Bellman equations imply

$$V^\star(X) - V^\pi(X) = \delta_\pi(X) + \gamma\mathbb{E}^\pi[V^\star(X^1) - V^\pi(X^1) \mid X^0 = X]. \tag{A.40}$$

Taking suprema and iterating gives

$$\|V^\star - V^\pi\|_\infty \le \frac{\|\delta_\pi\|_\infty}{1-\gamma}. \tag{A.41}$$

Applying A.41 to π_{PAI} and using A.39 yields

$$\begin{aligned} V^\star(X^0) - V^{\pi_{\mathrm{PAI}}}(X^0) \le \frac{1}{1-\gamma}\Bigg[& C_1\mathcal{E}_{\mathrm{dec}} + C_2\mathcal{E}_{\mathrm{act}}(\kappa) \\ & + C_3\frac{\gamma\beta_{\max}\nu_{\max}(\gamma\alpha)^{H+1}}{1-\gamma\alpha} \\ & + C_4\left(\frac{\gamma\beta_{\max}\eta_\nu}{1-\gamma\alpha} + \frac{\gamma\beta_{\max}\eta_\ell}{(1-\gamma\alpha)^2}\right)\Bigg], \end{aligned}$$

where the constants absorb the factors from the preceding steps. This is Eq. 12. □

E.9 Optional consequence: weak-coupling and accurate-score limit

Corollary 2 (Weak-coupling consequence)**.** *Suppose that $\mathcal{E}_{\mathrm{dec}} \le C_{\mathrm{dec}}\varepsilon_{\mathrm{cpl}}$ and $\mathcal{E}_{\mathrm{act}}(\kappa) \le C_{\mathrm{act}}\varepsilon_{\mathrm{act}}$ for coupling and action-interaction levels $\varepsilon_{\mathrm{cpl}}, \varepsilon_{\mathrm{act}} \ge 0$. Then Theorem 2 gives*

$$\begin{aligned} V^\star(X^0) - V^{\pi_{\mathrm{PAI}}}(X^0) \le \frac{1}{1-\gamma}\Bigg[& C_1C_{\mathrm{dec}}\varepsilon_{\mathrm{cpl}} + C_2C_{\mathrm{act}}\varepsilon_{\mathrm{act}} \\ & + C_3\frac{\gamma\beta_{\max}\nu_{\max}(\gamma\alpha)^{H+1}}{1-\gamma\alpha} \\ & + C_4\left(\frac{\gamma\beta_{\max}\eta_\nu}{1-\gamma\alpha} + \frac{\gamma\beta_{\max}\eta_\ell}{(1-\gamma\alpha)^2}\right)\Bigg]. \end{aligned} \tag{A.42}$$

Consequently, for fixed $\gamma\alpha < 1$, the PAI approximation gap vanishes as the decomposition and interaction residuals vanish, $H \to \infty$, $\eta_\nu \to 0$, and $\eta_\ell \to 0$.

Proof. Substitute the assumed bounds on $\mathcal{E}_{\mathrm{dec}}$ and $\mathcal{E}_{\mathrm{act}}(\kappa)$ into Theorem 2. The truncation term converges to zero geometrically because $\gamma\alpha < 1$, while the two marginal-value terms vanish as $\eta_\nu, \eta_\ell \to 0$. □

F Limitations

DDE-RMAB and PAI provide a scalable approximation framework for restless bandits with delayed network spillovers, but they have several limitations.

First, the exposure dynamics are assumed to be linear and stable: $z^{t+1} = \alpha P^\top z^t + \beta(s^t) \odot a^t$. This makes the multi-hop spillover value analytically tractable through a resolvent, but real externalities may be nonlinear, saturating, threshold-based, or competitive. Second, the propagation graph P is assumed to be known or estimated in advance. In practice, influence graphs may be noisy, partially observed, or time-varying; our theory does not fully capture graph-estimation error.

Finally, PAI requires local value and sensitivity estimates (Δ_i, ν_i), obtained from model-based dynamic programs, finite differences, or learned critics. The policy quality depends on the accuracy of these approximations. Our formulation also focuses on discounted infinite-horizon objectives; finite-horizon, average-reward, or risk-sensitive variants require separate analysis.

Thus, DDE-RMAB is best viewed as a tractable model for delayed and diffused externalities when the exposure field is stable and provides a reasonable summary of network coupling.

G Additional Experimental Results

This appendix provides additional implementation details, complete multi-hop and parameter sensitivity analyses, robustness experiments under nonlinear exposure and imperfect state information, and structured graph-error evaluations. Unless otherwise stated, all results are averaged over 10 random seeds and use the same simulation protocol as in Section 5.

G.1 Implementation and evaluation protocol

Simulation protocol. The simulator uses the true propagation operator P_{true}. When graph misspecification is evaluated, graph-aware policies receive a perturbed operator P_{obs}, while the environment continues to evolve according to P_{true}. This separates graph misspecification from randomness in the environment dynamics.

Unless otherwise noted, the synthetic experiments use horizon $T = 1000$, discount factor $\gamma = 0.99$, and the same initial state distribution for all policies within each random seed. Returns are computed as discounted cumulative rewards. The propagation operator is normalized so that the stability conditions in Section 4 are satisfied.

PAI implementation. The canonical PAI uses $\kappa = 1$. For the exact implementation, we compute

$$g^t = (I - \gamma\alpha P)^{-1}\nu^t$$

and use

$$I_i^t = \Delta_i^t + \gamma\beta_i^t g_i^t.$$

For sparse or large graphs, we instead use the truncated approximation

$$g_H^t = \sum_{h=0}^{H} (\gamma\alpha P)^h \nu^t.$$

The policy evaluates the local score for every arm and selects the feasible top-M arms.

Baseline implementation. The reported Whittle baseline is a decoupled Whittle-style priority. It solves a tabular single-arm active-passive surrogate after disabling graph-coupled propagation and ranks arms using

$$I_{\text{DW}}(s) = \widetilde{Q}(s, z = 0, 1) - \widetilde{Q}(s, z = 0, 0).$$

It is therefore not the exact Whittle index of the fully coupled DDE-RMAB.

Local PPO-R is implemented as a shared-parameter PPO baseline (Boehmer et al., 2025). It does not perform graph message passing and is evaluated as a learning-based local baseline. GRETA is the one-hop graph-aware baseline (Herlihy & Dickerson, 2023). The multi-hop extension used below applies the same message construction recursively over multiple graph steps.

G.2 MULTI-HOP COMPARISON AND PAI SENSITIVITY

Construction of MH-GRETA-H. To provide a stronger graph-aware comparison than one-hop GRETA, we construct MH-GRETA-H by recursively propagating GRETA's graph message:

$$m^{(0)} = m, \qquad m^{(h+1)} = P^\top m^{(h)}, \qquad h = 0, \dots, H-1.$$

The priority aggregates the resulting messages up to depth H using the same finite-depth construction across all arms. Thus, $H = 1$ recovers the one-hop GRETA message, while larger H explicitly incorporates multi-hop graph information. The main paper uses MH-GRETA-H5 as the representative finite-depth baseline.

Multi-hop depth sweep. Table 4 reports the relative improvement of MH-GRETA-H over one-hop GRETA on the *Les Misérables* topology. Increasing H generally improves the performance of GRETA, confirming that the baseline uses delayed multi-hop information. However, the improvement saturates around 2.9–3.0% in the high-propagation settings, whereas PAI achieves improvements of 5.36% and 5.41% in the two corresponding settings.

Table 4: Relative improvement over one-hop GRETA (%) on the semi-real topology.

Policy / depth	$(\alpha, \beta) = (0.4, 1.2)$	$(\alpha, \beta) = (0.6, 1.0)$	$(\alpha, \beta) = (0.6, 1.2)$
MH-GRETA-$H = 2$	+1.30%	+1.66%	+2.36%
MH-GRETA-$H = 3$	+1.11%	+2.39%	+2.79%
MH-GRETA-$H = 4$	+0.73%	+2.59%	+2.92%
MH-GRETA-$H = 6$	+1.05%	+2.55%	+3.04%
MH-GRETA-$H = 8$	+1.18%	+2.77%	+2.91%
MH-GRETA-$H = 10$	+1.14%	+2.83%	+2.90%
MH-GRETA-$H = 12$	+1.13%	+2.89%	+2.94%
MH-GRETA-$H = 18$	+1.13%	+2.91%	+2.90%
PAI	**+1.49%**	**+5.36%**	**+5.41%**

Sensitivity to κ. The canonical value $\kappa = 1$ corresponds to the nominal first-order propagation correction. Table 5 shows the sensitivity to κ in representative high-propagation settings. The no-propagation variant $\kappa = 0$ is substantially worse in every setting. Positive values of κ quickly reach a near-best plateau, which suggests that $\kappa = 1$ is a canonical first-order scaling rather than a finely tuned hyperparameter.

Table 5: Sensitivity to the propagation scaling parameter κ. Entries are mean $\pm$ standard deviation.

κ	$(0.6, 1.2)$	$(0.6, 1.4)$	$(0.8, 1.2)$	$(0.8, 1.4)$
0	1737.32 ± 26.46	1852.29 ± 33.40	2311.06 ± 45.29	2522.20 ± 33.60
0.001	2280.77 ± 30.33	2144.59 ± 37.50	2532.93 ± 33.32	2681.92 ± 38.04
0.01	2286.62 ± 24.42	2371.35 ± 13.46	2866.57 ± 19.67	2896.88 ± 12.43
0.025	2298.98 ± 24.02	2432.70 ± 29.85	$\mathbf{2893.82 \pm 10.85}$	$\mathbf{2898.85 \pm 10.22}$
0.05	2310.35 ± 39.18	2520.40 ± 29.47	2892.91 ± 10.76	2896.44 ± 9.69
0.1	$\mathbf{2516.71 \pm 29.73}$	$\mathbf{2633.21 \pm 18.24}$	2891.90 ± 11.00	2895.79 ± 9.99
1.0	$\mathbf{2516.71 \pm 29.73}$	$\mathbf{2633.21 \pm 18.24}$	2891.90 ± 11.00	2895.79 ± 9.99

The poor performance at $\kappa = 0$ is consistent with the theoretical interpretation of the action residual. Setting $\kappa = 0$ removes the explicit propagation correction, but it does not remove the true propagation value. Instead, the omitted value is absorbed into the action approximation residual.

Sensitivity to truncation depth. For finite-depth PAI, we use

$$g_H^t = \sum_{h=0}^{H} (\gamma \alpha P)^h \nu^t.$$

The theoretical tail bound is

$$\left\|(I-\gamma\alpha P)^{-1}\nu^t - g_H^t\right\|_\infty \le \nu_{\max}\frac{(\gamma\alpha)^{H+1}}{1-\gamma\alpha}.$$

Empirically, $H = 0$ loses approximately 14–25% in the tested high-propagation settings, while $H \ge 5$ is within 0.29% of the infinite-depth resolvent. This rapid saturation is consistent with the geometric decay predicted by the theory.

G.3 Saturation and imperfect state information

Capped-linear exposure dynamics. The linear exposure model is used in the main analysis because it gives a closed-form resolvent. To test robustness to saturation, we evaluate the capped-linear simulator

$$z^{t+1} = \Pi_{[0,z_{\text{cap}}]}\left(\alpha P^\top z^t + \beta(s^t)\odot a^t\right), \qquad z_{\text{cap}} = \frac{0.5\beta_{\max}}{1-\alpha}.$$

The policy still uses the linear PAI correction, whereas the simulator applies the cap. Results are reported in Table 6.

Table 6: Capped-linear exposure ablation. Entries are mean ± standard deviation over 10 seeds.

Setting	Update	GRETA	MH-GRETA-H5	PAI
$(0.6, 1.4)$	Linear	2566.52 ± 30.08	2632.76 ± 29.82	$\mathbf{2633.21 \pm 18.24}$
$(0.6, 1.4)$	Capped	2497.31 ± 30.64	2596.18 ± 27.43	$\mathbf{2611.60 \pm 27.14}$
$(0.8, 1.2)$	Linear	2886.78 ± 11.73	2889.09 ± 13.06	$\mathbf{2891.90 \pm 11.00}$
$(0.8, 1.2)$	Capped	2874.59 ± 15.57	2880.72 ± 14.23	$\mathbf{2891.90 \pm 11.00}$

PAI remains best in the tested capped regimes. This result is an empirical robustness test rather than a general optimality guarantee for arbitrary nonlinear exposure dynamics. For a bounded Lipschitz update

$$z^{t+1} = g\left(\alpha P^\top z^t + \beta(s^t)\odot a^t\right),$$

the linear resolvent can be replaced by a local Jacobian resolvent. Let

$$D_t = \operatorname{diag}\left(g'\left(\alpha P^\top z^t + \beta(s^t)\odot a^t\right)\right), \qquad A_t = \alpha D_t P^\top.$$

The corresponding first-order correction is

$$\Psi_i^t \approx \gamma\left\langle \nu^t, (I-\gamma A_t)^{-1} D_t \beta_i(s_i^t) e_i\right\rangle.$$

For capped-linear dynamics, D_t masks saturated coordinates; for sigmoid-like dynamics, the derivative becomes small in the saturation region. Formal guarantees for general nonlinear dynamics require additional local linearization assumptions.

Noisy exposure tracking. To evaluate imperfect state information, the simulator uses the true exposure z^t, while PAI receives

$$\widehat{z}^t = z^t + \sigma_z Z_{\text{scale}}\xi^t, \qquad Z_{\text{scale}} = \frac{\beta}{1-\alpha}.$$

Table 7 reports returns normalized by the clean-PAI return. Parentheses show the Jaccard overlap between the clean and noisy selected action sets.

Table 7: Noisy-exposure tracking ablation.

σ_z	$(0.6, 1.2)$	$(0.6, 1.4)$	$(0.8, 1.2)$
0.0	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
0.1	1.000 (0.963)	1.001 (0.953)	0.999 (0.783)
0.5	0.998 (0.777)	0.998 (0.777)	0.999 (0.737)
1.0	0.996 (0.717)	0.994 (0.717)	1.000 (0.690)

Even when the noise scale is comparable to the nominal exposure scale, normalized returns remain between 0.994 and 1.000 in the tested settings. The action overlap decreases as noise increases, but the return remains relatively stable because different rankings can produce similar high-value action sets. This does not imply robustness to arbitrarily biased or adversarial exposure estimates.

Noisy and stale marginal-value fields. We additionally perturb the marginal-value field using

$$\widetilde{\nu}^t = \nu^t + \sigma_\nu |\nu^t|_\infty \xi^t$$

and evaluate stale fields of the form

$$\widetilde{\nu}^t = \nu^{t-\Delta}.$$

For $\sigma_\nu \leq 0.5$, PAI remains within approximately one normalized return unit of clean PAI in representative high-propagation settings, with the worst normalized return approximately 0.991. Stale fields with $\Delta \leq 20$ are statistically comparable to clean PAI. Larger noise or stale estimates eventually degrade the policy, consistent with the marginal-value error term in the theoretical bound.

G.4 Structured graph errors and additional scaling results

Structured graph perturbations. We distinguish the simulator graph P_{true} from the graph P_{obs} provided to graph-aware policies. On the *Les Misérables* topology, we evaluate the following perturbations:

- **RandomDrop20**: randomly remove 20% of edges;
- **HubRemove30**: remove approximately 30% of edges incident to selected high-degree hubs;
- **CommunityMiss30**: remove approximately 30% of edges associated with a selected community;
- **FalsePos20**: add false-positive edges at a rate of approximately 20%;
- **WeightNoise0.5**: perturb edge weights with log-normal multiplicative noise of scale 0.5.

Table 8: Robustness to structured graph errors. Each entry reports the best non-PAI graph-aware baseline / PAI.

Graph error	$\beta = 1.0$	$\beta = 1.2$
Clean	$4984.62/5029.70$	$5373.80/5413.12$
RandomDrop20	$4958.86/5023.33$	$5327.86/5385.59$
HubRemove30	$4927.96/4981.10$	$5287.65/5367.01$
CommunityMiss30	$4976.55/5003.62$	$5325.96/5389.76$
FalsePos20	$4975.17/5038.57$	$5421.43/5425.69$
WeightNoise0.5	$4981.91/5063.81$	$5341.37/5421.73$

PAI is best in mean across all graph-error settings. Relative to the clean-graph PAI result, its return changes by less than approximately 1% in the tested cases. In the false-positive and edge-weight-noise settings, the perturbed-graph mean can be slightly higher than the clean-graph mean; the increase is below approximately 0.7% and is consistent with sampling variation or mild smoothing caused by the perturbation.

Interpretation of the budget factor. The factor M appearing in the worst-case performance bound is a deterministic uniform score-error effect rather than a statistical regret term. If all M selected arms incur approximation errors in the same direction, the errors add linearly. This dependence cannot generally be removed under only uniform deterministic score-error assumptions: an instance with M independent components can cause a top-M policy to choose the worse arm in every component, yielding an $\Omega(M\epsilon)$ loss.

The worst-case factor is conservative when score margins are large. Let

$$m_M = I_{(M)} - I_{(M+1)}$$

be the clean top-M score margin. If

$$\max_i |\widehat{I}_i - I_i| < \frac{m_M}{2},$$

the selected set is unchanged. Therefore, the practical effect of M can be much smaller than the worst-case bound when score margins are large or approximation errors are heterogeneous.

Summary of additional results. The additional experiments support five conclusions. First, PAI remains competitive with an explicit finite-depth multi-hop baseline. Second, the propagation correction is essential, while positive values of κ quickly reach a performance plateau. Third, finite-depth PAI approaches the infinite-depth resolvent rapidly in the tested settings. Fourth, PAI remains effective under capped exposure, noisy exposure tracking, and perturbed marginal-value fields, although these results are empirical robustness tests rather than general nonlinear optimality claims. Finally, PAI remains the best graph-aware method under the tested structured graph errors.

H Experimental Reproducibility Details

The benchmark environments are fully specified by the DDE-RMAB transition model, the exposure update rule, the reward function, and the graph construction procedure. In the main synthetic experiments, we use a Barabási–Albert graph generated by NetworkX with attachment parameter 2, add self-loops, and row-normalize the adjacency matrix to obtain the propagation operator. The default budget is $M = 2$, the horizon is $T = 1000$, and the discount factor is $\gamma = 0.99$. The local transition parameters are $p_0 = 0.1$, $p_a = 0.1$, $p_z = 0.4$, $q_0 = 0.2$, $q_a = 0.1$, $q_z = 0.4$, with activation cost 0.1. We evaluate the reported settings over multiple random seeds; the environment reset initializes each intrinsic state as good with probability 0.3 and sets all exposure variables to zero.

All policy implementation details are specified. Myopic selects the arms with the largest current intrinsic states. The Whittle-style baseline is computed from a decoupled single-arm dynamic program with propagation disabled. GRETA is implemented as a one-hop graph-aware baseline that combines a local Whittle advantage with the estimated value increase of immediate neighbors, without using delayed multi-hop propagation. PAI uses the same single-arm dynamic program to compute local activation advantages and exposure sensitivities on a discretized exposure grid of 50 points, and then applies the resolvent correction $(I - \gamma\alpha P)^{-1}$ to aggregate delayed propagation effects. For the PPO-R baseline, we use a shared actor-critic architecture across arms with local input (s_i, z_i), one hidden layer of width 32 with ReLU activations, Bernoulli action probabilities, and top-M selection at decision time. PPO is trained with Adam using learning rate 5×10^{-3}, clipping parameter 0.2, four PPO epochs per update, and 50 training episodes in the main experiments. The training and evaluation protocols are separated. Model-based policies are computed once from the specified environment parameters and then evaluated over independent simulation seeds. PPO-R is trained on a copy of the environment and then evaluated using the same rollout protocol as the other methods. Discounted return is computed as $\sum_{t=0}^{T-1} \gamma^t R_t$.

For each reported configuration, we run 10 independent evaluation seeds and report the mean discounted return together with variability estimates. In tables, results are reported as mean $\pm$ standard deviation across seeds. All experiments were run on a 15-inch MacBook Air with an Apple M4 chip and 24 GB unified memory, running macOS Sequoia 15.3. The non-neural caching algorithms use only CPU execution and require no GPU. The code is implemented in Python with NumPy-based trace replay; neural baselines, when evaluated, use PyTorch on CPU.