Supplementary Information
Scenario-aware Control of Multi-pathway Spread Processes: Application to Biological Invasions

Prathyush Sambaturu, Manisha Sudhir, Hongze Chen, Anil Vullikanti,
Rangaswamy Muniappan and Abhijin Adiga

1 SIR model

We define the network-based Susceptible-Infectious-Removed (SIR) diffusion process [4]. Each node is in one of the following states: susceptible (**S**), infectious (**I**) or recovered (**R**). Let $S \subseteq V$ denote the initial set of nodes in state **I** that seed the diffusion process at $t = 0$. At any time t, each node u in state **I** infects its susceptible neighbor v with probability proportional to the weight $w(u, v, \lambda, t)$ via the labeled edge (u, v, λ). Then, u transitions to state **R** at time $t + 1$.

2 The multi-pathway model

We focus on developing practical control algorithms for invasive alien species spread that affect important agricultural crops, using the agent-based model of McNitt et al. [5] referred to as the MultiPath model (Figure S1). The dynamics of the MultiPath model follow the Susceptible-Exposed-Infectious (SEI) [4]. Each node is in one of the following states: susceptible (**S**), exposed (**E**), or infectious (**I**). Let $S \subseteq V$ denote the initial set of nodes in state **I** that seed the diffusion process at $t = -1$. At any time t, each node u in state **I** infects its susceptible neighbor v with probability equal to the weight $w(u, v, \lambda, t)$ via the labeled edge (u, v, λ, t). An exposed node transitions to the infectious state after ℓ time steps, where ℓ is the latency period. It corresponds to the time taken for the population of the IAS to build up. A node has two periodic time-varying attributes, suitability $\epsilon(v, t)$ for pest establishment and infectivity $\rho(v, t)$. For the short-distance dispersal, the probability that node v is infected by its neighbor v' within its Moore neighborhood is $w(v', v, \lambda_s, t) = \epsilon(v, t)\big(1 - \exp(-\alpha_s \rho(v', t))\big)$, where λ_s is the edge label corresponding to the pathway and α_s is a tunable pathway parameter. For two nodes v and v' within a group, the probability of within group transmission from v' to v is $w(v', v, \lambda_\ell, t) = \epsilon(v, t)\big(1 - \exp(-\alpha_\ell \rho(v', t))\big)$, where λ_ℓ is the pathway label and α_ℓ is the pathway parameter. For the group-to-group transmission, a directed flow network is defined with groups as nodes and the edge weight for the edge from Q_i to Q_j denoted by F_{ij}. Suppose $g(v) = Q_j$ and $g(v') = Q_i$, then the probability that v is infected by v' through this pathway is $w(v', v, \lambda_{\ell d}, t) = \epsilon(v, t)\big(1 - \exp(-\alpha_{\ell d} F_{ij} \rho(v', t))\big)$, where $\lambda_{\ell d}$ is the pathway label and $\alpha_{\ell d}$ is the pathway parameter. The details of network construction are in McNitt et al. [5].

3 Hardness results

Lemma 1. *The IAScontrol problem is NP-complete even when G is a tree.*

Proof. Our reduction is from a variation of the Unbalanced Graph Cut problem [3]: given a graph $G = (V, E)$, a source node s, and node cost c_v, the goal is to choose a subset $V' \subset V$ such that $\sum_{v \in V'} c_v \leq B$, and $|\{u : u \text{ is reachable from } s \text{ in } G - V'\}|$ is minimized. By modifying the reduction in [3] (who consider the edge version of the problem), it can be shown that the above variation is also NP-hard when G is a tree.

We observe that the MultiPath model generalizes the SI process on a graph, by considering a single pathway in which each node is in a singleton group. We reduce the above variation of the Unbalanced Graph Cut problem to the IAScontrol on a tree, with the model parameters chosen so that the MultiPath model corresponds to the SI process. Taking $T = |V|$, the number of infections is equal to the number of nodes reachable from S in the residual graph after the intervened set of nodes are removed. □

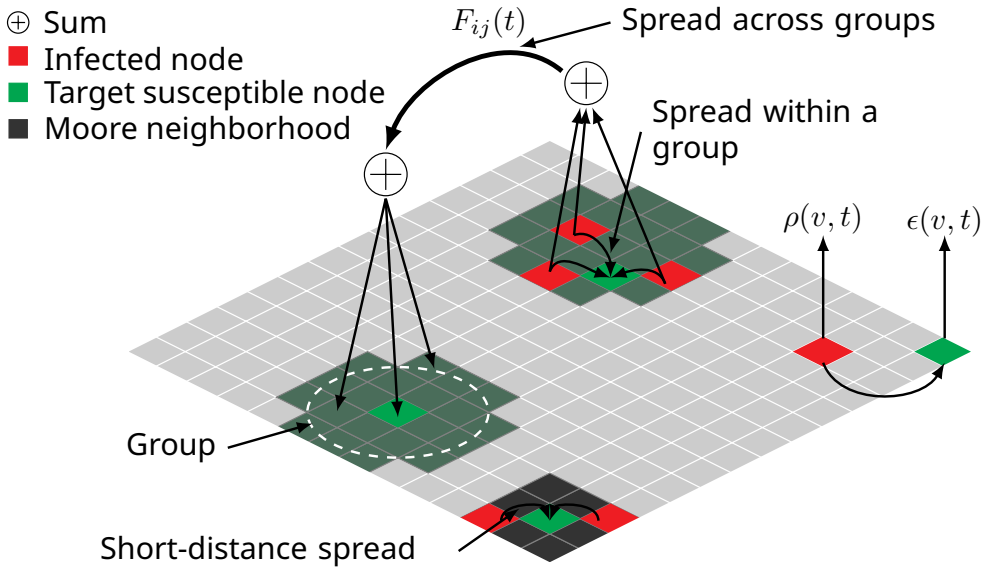


Figure S1: An illustration of the multi-pathway model. [5]

An alternative form of the problem (referred to as the IAScontrolMinBudget problem) is: given a target bound K on the number of infections, choose $\mathcal{Q}^*$ so that $\inf_{\mathrm{T}}(G, \mathrm{S}, \tau_{\mathrm{d}}, \{v \mid g(v) \in \mathcal{Q}^*\}) \leq K$, and $\sum_{q \in \mathcal{Q}^*} c_q$ is minimized.

Lemma 2. *It is NP-hard to approximate the IAScontrolMinBudget to within an $O(2^{\log^{1-o(1)} n/2})$ factor.*

Proof. Our reduction is from the node version of the Label Cut problem [6], which is defined in the following manner: given a graph $G = (V, E)$, a source node s, a sink node t, a label $\ell(v) \in L$ for each node $v \in V$, and a cost c_i for each label $i \in L$, the objective is to choose a subset $L' \subset L$ such that the nodes s and t are disconnected in $G - \{v : \ell(v) \in L'\}$, and $\sum_{i \in L'} c_i$ is minimized. The hardness result of [6] (who consider the edge labeled version of the problem) can be modified to show that the same hardness holds for the node version of Label Cut as well.

We reduce the node version of the Label Cut problem to an instance G' of IAScontrolMinBudget. G' is basically the same as G, with an additional set U of n nodes connected to the node t (so that the total number of nodes is $2n$). We consider the set L to be the groups $\mathcal{Q}$, and choose parameters so that there is a single pathway in the MultiPath model, which corresponds to the SI model. We choose the target number of infections $K = n$. Then, the $\inf_{\mathrm{T}}(G, \mathrm{S}, \tau_{\mathrm{d}}, \{v \mid g(v) \in \mathcal{Q}^*\}) \leq K$ if and only if the node t is disconnected from s when the nodes in $\{v : g(v) \in \mathcal{Q}^*\}$ are removed, else the number of infections will be at least n, since all the nodes in U will be infected if t is infected. Thus, a solution $\mathcal{Q}^*$ to the IAScontrolMinBudget instance on G' corresponds to a min-label s, t cut in G. Since the number of nodes in G' is $2n$, the hardness result follows from the bound of [6]. □

4 Time-expanded graph

4.1 Definition

We represent the MultiPath model as an SIR process (instead of the SEI process) on an auxiliary network called the *time-expanded network*. Let $H_{\mathrm{te}}(V_{\mathrm{te}}, E_{\mathrm{te}})$ denote the time-expanded network corresponding to the multi-pathway model on $G(V, E)$. See Figure 1(c) for a time-expanded graph of G in Figure 1(b). The key

idea is to treat every node u at each time step as a distinct node, i.e., we have $T+1$ copies $\{u_0, \ldots, u_T\}$ of u, where u_i represents the copy of u at time step i. Here, T is the *time horizon* of the spread process. To incorporate the exposed state in the underlying SEI process of the multi-pathway model, we have ℓ additional copies $\{u_{i,0}, \ldots, u_{i,\ell-1}\}$, corresponding to each u_i, where ℓ is the latency period (time to transition from exposed $\mathbf{E}$ to infectious $\mathbf{I}$). The edge set E_{te} consists of exactly the following four types of edges which corresponds to different events in a SEI process: $\mathbf{S} \to \mathbf{E}$, $\mathbf{E} \to \mathbf{E}$, $\mathbf{E} \to \mathbf{I}$, and $\mathbf{I} \to \mathbf{I}$.

- $(v_i, u_{i+1,0}, \lambda, i)$, $\forall (v, u, \lambda, i) \in E$ with weight $w(v, u, \lambda, i)$ (captures $\mathbf{S} \to \mathbf{E}$ through pathway λ)
- $(u_{i,r}, u_{i,r+1})$ for $r \in [0, \ell-2]$ (captures $\mathbf{E} \to \mathbf{E}$)
- $(u_{i,\ell-1}, u_{i+\ell})$ (captures $\mathbf{E} \to \mathbf{I}$)
- (u_i, u_{i+1}) (captures $\mathbf{I} \to \mathbf{I}$)

All edges of types other than $\mathbf{S} \to \mathbf{E}$ have weight 1. For the special case of the SI diffusion process ($\ell = 0$), there are no nodes of the form $u_{i,r}$ and it has two edge types, $\mathbf{I} \to \mathbf{I}$ as defined above and $\mathbf{S} \to \mathbf{I}$: $(v_i, u_{i+1,0}, \lambda)$ with weight $w(v, u, , \lambda, i)$.

4.2 Equivalence of MultiPath on G and SIR process on H_{te}

Let $\sigma_G(v,t)$ be the state of a vertex v in G at time t, which can be either $\mathbf{S}$, $\mathbf{E}$, or $\mathbf{I}$. Similarly, let $\sigma_{H_{\text{te}}}(u_i, t)$ (resp. $\sigma_{H_{\text{te}}}(u_{i,r}, t)$) be the state of a vertex u_i (resp. $u_{i,r}$) in H_{te} at time t with $\mathbf{S}$, $\mathbf{I}$, and $\mathbf{R}$ being the possible states. Let $\mathcal{O}_G$ denote a stochastic disease outcome of the SEI model on G – this specifies the state $\sigma_G(v,t)$ for each (v,t), and set of the edges (u, v, λ, t) such that node u infects v at time t through pathway λ. Similarly, let $\mathcal{O}_{H_{\text{te}}}$ denote a disease outcome in the SIR model on H_{te}. We say that $\mathcal{O}_{H_{\text{te}}}$ is consistent with $\mathcal{O}_G$ if: (i) for any u, $\sigma_G(u,i) = \mathbf{I}$ (resp. $\sigma_G(u,i) = \mathbf{S}$) $\iff$ $\sigma_{H_{\text{te}}}(u_i, i) = \mathbf{I}$ (resp. $\sigma_{H_{\text{te}}}(u_i, i) = \mathbf{S}$); (ii) $\sigma_G(u, i+r) = \mathbf{E}$, $r \in [0, \ell-1]$) $\iff$ $\sigma_{H_{\text{te}}}(u_{i,r}, i+r) = \mathbf{I}$; (iii) $\sigma_G(u, i-1) = \sigma_G(u,i) = \mathbf{I}$ $\iff$ $\sigma_{H_{\text{te}}}(u_{i-1}, i) = \mathbf{R}$; and (iv) u infects v on edge (u, v, λ, i) in G at time i $\iff$ node u_{i-1} infects node $v_{i,0}$ at time i on edge (u_{i-1}, v_i, λ). Given $\mathcal{O}_G$, for a time t, let $\mathcal{O}_G(\leq t)$ be a snapshot of $\mathcal{O}_G$ up to time step t. Similarly, $\mathcal{O}_{H_{\text{te}}}(\leq t)$ is a snapshot of $\mathcal{O}_{H_{\text{te}}}$ up to t time steps.

Statement of Theorem 1. *Consider the multi-pathway diffusion process on $G(V,E)$ for T time steps with a latency period $\ell \geq 0$ and the SIR process on the corresponding time-expanded graph $H_{te}(V_{te}, E_{te})$. Then, for any outcome $\mathcal{O}_G$ and a consistent outcome $\mathcal{O}_{H_{te}}$, the probability that $\mathcal{O}_G$ is the outcome in the multi-pathway process on G is equal to the probability that $\mathcal{O}_{H_{te}}$ is the outcome in the SIR process on H_{te}.*

Proof. The proof is by induction on time t. We will assume that $\ell > 0$. The proof for $\ell = 0$ uses the same approach. At $t = 0$, by definition, $\forall v \in V(G)$ we have $\sigma_G(v,0) = \mathbf{I}$ $\iff$ $\sigma_{H_{\text{te}}}(v_0, 0)) = \mathbf{I}$ and $\sigma_G(v,0) \neq \mathbf{E}$. Suppose that $\Pr(\mathcal{O}_G(\leq t)) = \Pr(\mathcal{O}_{H_{\text{te}}}(\leq t))$ for some $t \geq 0$.

For the induction step, we consider all events that can occur at time $t+1$ w.r.t. a node v in $\mathcal{O}_G$ on a case by case basis. It is sufficient to prove that for each such event, the corresponding event in $\mathcal{O}_{H_{\text{te}}}$ for $t+1$ has the same probability.

Case 1. $\mathbf{I} \to \mathbf{I}$ Consider the event where v is in state $\mathbf{I}$ in $\mathcal{O}_G$ at t and it remains in this state at time $t+1$. By model definition, the probability of this event is 1. In $\mathcal{O}_{H_{\text{te}}}$, the corresponding event is, (i) $\sigma_{H_{\text{te}}}(v_t, t) = \mathbf{I}$, (ii) $\sigma_{H_{\text{te}}}(v_{t+1}, t) = \mathbf{S}$, and (iii) $\sigma_{H_{\text{te}}}(v_{t+1}, t+1) = \mathbf{I}$. By induction assumption, (i) and (ii) are true. Since the weight on edge (v_t, v_{t+1}) is 1, this event happens with probability 1.

Case 2. $\mathbf{S} \to \mathbf{S}$ Suppose v is in state $\mathbf{S}$ in $\mathcal{O}_G$ at t and it remains in this state at time $t+1$. For this to occur, v should not be infected through any edge of the form $(v', v, \lambda, t+1)$ at time $t+1$. Let $E(v, \mathcal{O}_G, t+1)$ denote the set of such edges. By model definition, the probability of this event is $\prod_{(v', v, \lambda, t+1) \in E(v, \mathcal{O}_G, t+1)} \big(1 - w(v', v, \lambda, t+1)\big)$. In $\mathcal{O}_{H_{\text{te}}}$, the corresponding event is, (i) $\sigma_{H_{\text{te}}}(v_t, t) = \mathbf{S}$ and (ii) $\sigma_{H_{\text{te}}}(v_{t+1}, t+1) = \mathbf{S}$. By induction assumption, (i) is true. By definition of H_{te}, probability that $v_{t+1,0}$ is infected by the edge $(v'_t, v_{t+1,0}, \lambda, t+1)$

is $w(v', v, \lambda, t+1)$. Thus, the probability of the $\mathbf{S} \to \mathbf{S}$ event is equal to $\prod_{(v',v,\lambda,t+1)\in E(v,\mathcal{O}_G,t+1)} \big(1 - w(v', v, \lambda, t+1)\big)$.

Case 3. $\mathbf{S} \to \mathbf{E}$ Suppose v is in state $\mathbf{S}$ at time t and is infected at time $t+1$ through exactly the edges from $E(v, \mathcal{O}_G, t+1)$. Let E' denote the set of such edges. The probability of this event occurring is $\prod_{(v',v,\lambda,t+1)\in E'} w(v', v, \lambda, t+1)$. The corresponding event in $\mathcal{O}_{H_{\text{te}}}$ is, (i) $\sigma_{H_{\text{te}}}(v_t, t) = \mathbf{I}$, (ii) $\sigma_{H_{\text{te}}}(v_{t+1,0}, t) = \mathbf{S}$, and (iii) $v_{t+1,0}$ is infected at time $t+1$ through edges in the set $E'' = \{(v'_t, v_{t+1,0}, \lambda, t+1) \mid (v', v, \lambda, t+1) \in E'\}$. By induction assumption, (i) and (ii) are true. By definition of H_{te}, probability that $v_{t+1,0}$ is infected by the edge $(v'_t, v_{t+1,0}, \lambda, t+1)$ is $w(v', v, \lambda, t+1)$. Therefore, the probability that $v_{t+1,0}$ is infected through all the edges in E' is $\prod_{(v',v,\lambda,t+1)\in E'} w(v', v, \lambda, t+1)$.

Case 4. $\mathbf{E} \to \mathbf{E}$ Consider the event where v is in state $\mathbf{E}$ in $\mathcal{O}_G$ at t and it remains in this state at time $t+1$. By model definition, this can happen only if v was infected at $t-r$ for some $r \in [0, \ell-1]$. Under this assumption, the probability of this event is 1. In $\mathcal{O}_{H_{\text{te}}}$, the corresponding event is, (i) $\sigma_{H_{\text{te}}}(v_{t-r,r}, t) = \mathbf{I}$, (ii) $\sigma_{H_{\text{te}}}(v_{t-r,r}, t) = \mathbf{S}$, and (iii) $\sigma_{H_{\text{te}}}(v_{t-r,r+1}, t+1) = \mathbf{I}$. By induction assumption, (i) and (ii) are true. Since the weight on edge $(v_{t-r,r}, v_{t-r,r+1})$ is 1, this event happens with probability 1 as well.

Case 5. $\mathbf{E} \to \mathbf{I}$ Consider the event where v is in state $\mathbf{E}$ in $\mathcal{O}_G$ at t and transitions to state $\mathbf{I}$ at time $t+1$. By model definition, this can happen only only if v was infected at $t-\ell$. Under this assumption, the probability of this event is 1. In $\mathcal{O}_{H_{\text{te}}}$, the corresponding event is, (i) $\sigma_{H_{\text{te}}}(v_{t-\ell,\ell}, t) = \mathbf{I}$, (ii) $\sigma_{H_{\text{te}}}(v_{t+1}, t) = \mathbf{S}$, and (iii) $\sigma_{H_{\text{te}}}(v_{t+1}, t+1) = \mathbf{I}$. By induction assumption, (i) and (ii) are true. Since the weight on edge $(v_{t-\ell,\ell}, v_{t+1})$ is 1, this event happens with probability 1.

For the special case of SI diffusion process ($\ell = 0$), the proof follows by replacing Case 3 ($\mathbf{S} \to \mathbf{E}$) with ($\mathbf{S} \to \mathbf{I}$) and ignoring Cases 4 and 5. □

5 SpreadBlocking algorithm and performance guarantees

The algorithm outlined in Figure 1(d) is provided below.

Analysis of SpreadBlocking

Let $\{H^1, \ldots, H^M\}$ be the set of M simulation outcomes corresponding to SIR process on H_{te}, where each $H^j = (V_{\text{te}}, E^j_{\text{te}})$, such that $E^j_{\text{te}} \subseteq E_{\text{te}}$. We solve a linear relaxation of the IAScontrol problem, restricted to these samples, and the resulting objective value is guaranteed to be close to the actual expected number of infections. Table S1 summarizes the quantities and variables used in the linear program, referred to as $LP_{\tau_{\text{d}}}$.

Let $\mathcal{Q}' \subseteq \mathcal{Q}$ be any intervention set for τ_{d}. Let $V_{\text{te}}(\mathcal{Q}') = \{v_i, v_{i,r} \in V_{\text{te}} \mid g(v) \in \mathcal{Q}' \text{ and } i \geq \tau_{\text{d}}\}$, be the set of nodes in H^j to which intervention $\mathcal{Q}'$ applies. Let $V(\mathcal{Q}') = \{v \in V \mid v_i, v_{i,r} \in V_{\text{te}}(\mathcal{Q}')\}$ be the set of nodes in G to which intervention $\mathcal{Q}'$ applies. Let $H^j - V_{\text{te}}(\mathcal{Q}')$ denote the subgraph of H^j induced by removing all nodes in $V_{\text{te}}(\mathcal{Q}')$ from H^j. Let $I^j(\mathcal{Q}') = \{v \in V | \ \exists i \text{ s.t. } v_i \text{ or } v_{i,r} \in \mathcal{R}(H^j - V_{\text{te}}(\mathcal{Q}'))\}$ denote the number of infections (nodes still reachable from S_{te} in H^j) in V. Let $I(\mathcal{Q}') = \frac{1}{M}\sum_j I^j(\mathcal{Q}')$ denote the average number of infections in V restricted to the M simulations.

Bicriteria approximation. The hardness in Lemma 2 motivates the notion of bicriteria approximation: we say that a solution $\mathcal{Q}'$ is an (α, β)-approximation if $\inf_{\text{T}}(G, \text{S}, \tau_{\text{d}}, \{v \mid g(v) \in \mathcal{Q}'\}) \leq \alpha \inf_{\text{T}}(G, \text{S}, \tau_{\text{d}}, \{v \mid g(v) \in \mathcal{Q}^*\})$, and $\sum_{q\in\mathcal{Q}'} c_q \leq \beta B$, where $\mathcal{Q}^*$ is an optimal solution with $\sum_{q\in\mathcal{Q}^*} c_q \leq B$.

Let $\hat{\mathcal{Q}}^* = \text{argmin}_{\mathcal{Q}''} I(\mathcal{Q}'')$ be an intervention set that achieves the minimum average number of infections on the simulations. Then, let $I_{opt} = \inf_{\text{T}}(V(\mathcal{Q}^*))$, i.e, the expected number of infections achieved by an optimal solution $\mathcal{Q}^*$ to the given instance of the IAScontrol. We first show that the average number of infections $I(\mathcal{Q}')$ achieved by any intervention set $\mathcal{Q}'$ restricted to the M simulations is close to expected number of infections $\inf_{\text{T}}(\mathcal{Q}')$ for that intervention set.

Algorithm 1: SpreadBlocking algorithm.

Input : $G = (V, E)$, set of sources $S \subseteq V$, budget B, time horizon T, intervention delay τ_{d}
Output : Intervention set $\mathcal{Q}_{\mathrm{SB}} \subseteq \mathcal{Q}$

1 Construct time expanded network H_{te} from G
2 Construct M simulations of the SIR process $\{H^1 = (V_{\mathrm{te}}, E^1_{\mathrm{te}}), \ldots, H^M = (V_{\mathrm{te}}, E^M_{\mathrm{te}})\}$ with $S_{te} = \{u_0 \mid u \in S\}$ as sources on the time-expanded network H_{te} corresponding to SEI process on G.
3 Solve the linear program $LP_{\tau_{\mathrm{d}}}$ defined as follows:

$$
\begin{aligned}
(LP_{\tau_{\mathrm{d}}})\ \min \quad & \frac{1}{M}\sum_j \sum_u z^j_u \\
\forall i < \tau_{\mathrm{d}}, u_i, u_{i,r} \in \mathcal{R}(H^j) \quad : \quad & y^j_{u,i} = 1, y^j_{u,i,r} = 1 \\
\forall u_i, u_{i,r} \in \mathcal{R}(H^j) \quad : \quad & z^j_u \geq y^j_{u,i},\ z^j_u \geq y^j_{u,i,r} \\
\forall (v_{i-1}, u_{i,0}) \in E^j_{\mathrm{te}} : y^j_{u,i,0} \quad \geq \quad & y^j_{v,i-1} - x_{g(u),\tau_{\mathrm{d}}} \\
\forall (u_{i,r}, u_{i,r+1}) \in E^j_{\mathrm{te}} : y^j_{u,i,r+1} \quad \geq \quad & y^j_{u,i,r} - x_{g(u),\tau_{\mathrm{d}}} \\
\forall (u_{i-\ell,\ell-1}, u_i) \in E^j_{\mathrm{te}} : y^j_{u,i} \quad \geq \quad & y^j_{u,i-\ell,\ell-1} - x_{g(u),\tau_{\mathrm{d}}} \\
\forall (u_{i-1}, u_i) \in E^j_{\mathrm{te}} : y^j_{u,i} \quad \geq \quad & y^j_{u,i-1} - x_{g(u),\tau_{\mathrm{d}}} \\
\forall i \geq \tau_{\mathrm{d}}, \forall u_i, u_{i,r} \in \mathcal{R}(H^j) \quad : \quad & y^j_{u,i} \leq 1 - x_{g(u),\tau_{\mathrm{d}}} \\
& y^j_{u,i,r} \leq 1 - x_{g(u),\tau_{\mathrm{d}}} \\
\sum_{Q_q \in \mathcal{Q}} x_{q,\tau_{\mathrm{d}}} \quad \leq \quad & B \\
\text{All variables} \quad \in \quad & [0,1]
\end{aligned}
$$

(Rounding) Let $\mathbf{x}, \mathbf{y}, \mathbf{z}$ be the optimal fraction solution to $LP_{\tau_{\mathrm{d}}}$. Round it to an integral solution X, Y, Z using the following rounding procedure: (i) For each H^j, u_i, set $Y^j_{u,i} = 1$ if $y^j_{u,i} \geq \frac{1}{2}$. Similarly, for each $H^j, u_{i,r}$, set $Y^j_{u,i,r} = 1$ if $y^j_{u,i,r} \geq \frac{1}{2}$. (ii) For each $H^j, u \in V$, set $Z^j_u = 1$ if $z^j_u \geq \frac{1}{2}$. (iii) For each $Q_q \in \mathcal{Q}$, set $X_{q,\tau_{\mathrm{d}}} = 1$ if $x_{q,\tau_{\mathrm{d}}} \geq \frac{1}{2g_m}$ where g_m, such that $g_m \leq k$ is the maximum number of groups associated with the set of nodes on any path in any H^j.

4 $\mathcal{Q}_{\mathrm{SB}} = \{Q_q \mid X_{q,\tau_{\mathrm{d}}} = 1\}$

Lemma 3. *Let the number of groups $|\mathcal{Q}| = k \geq 2$, budget $B \geq 1$, and a positive real number $\epsilon < 1$. If $M \geq (6B+3)\epsilon^{-2}n$ and $\inf_{\mathrm{T}}(V(\mathcal{Q}')) \geq \log k$, with probability at least $1 - \frac{1}{k}$, for any intervention set $\mathcal{Q}' \subseteq \mathcal{Q}$, we have $I(\mathcal{Q}') \in [(1-\epsilon)\inf_{\mathrm{T}}(V(\mathcal{Q}')), (1+\epsilon)\inf_{\mathrm{T}}(V(\mathcal{Q}'))]$.*

Proof. From equivalence in Theorem 1, we have $\mathbb{E}[I(\mathcal{Q}')] = \mathbb{E}[I^j(\mathcal{Q}')] = \inf_{\mathrm{T}}(V(\mathcal{Q}'))$. The $I^j(\mathcal{Q}')$ variables are independent and $\frac{I^j(\mathcal{Q}')}{n} \in [0,1]$ where $|V| = n$ is the number of nodes in G. Using Chernoff bound [1, Theorem 1.1] to $M\frac{I^j(\mathcal{Q}')}{n}$, we have

$$
\Pr\Big(\frac{I^j(\mathcal{Q}')}{n} \notin \Big[(1-\epsilon)\frac{M}{n}\inf_{\mathrm{T}}(V(\mathcal{Q}')), (1+\epsilon)\frac{M}{n}\inf_{\mathrm{T}}(V(\mathcal{Q}'))\Big]\Big) \leq 2\exp(-\frac{\epsilon^2 M}{3n}\inf_{\mathrm{T}}(V(\mathcal{Q}'))).
$$

We have $\inf_{\mathrm{T}}(V(\mathcal{Q}')) \geq \log k$ and $M \geq (6B+3)\epsilon^{-2}n$ from our assumptions. Then, this probability is at most $2e^{-(2B+1)\log k} = \frac{2}{k^{2B+1}}$. The number of possible intervention sets $\mathcal{Q}'$ of size B is at most k^B (as at most k^B needs to be checked). Therefore, for $M = (6B+3)\epsilon^{-2}n$, the probability that there exists an intervention

Term	Definition
M	Number of simulation outcomes
$S_{te} \subseteq V_{\text{te}}$	Fixed set of sources of infection $\forall H^j$
$\mathcal{R}(H^j) \subseteq V_{\text{te}}$	Set of nodes in H^j reachable from S_{te} via a directed path
$x_{q,\tau_{\text{d}}} = 1$	if group $Q_q \in \mathcal{Q}$ is intervened at time-step τ_{d}
$y^j_{u,i} = 1$	if $u_i \in V_{\text{te}}$ is infected in H^j at time-step i (there is a directed from S_{te} to u_i in H^j), i.e., $\sigma_{H_{\text{te}}}(u_i, i) = \mathbf{I}$.
$y^j_{u,i,r} = 1$	if $u_{i,r}$ is infected in H^j at time-step i (there is a directed from S_{te} to $u_{i,r}$ in H^j), i.e, $\sigma_{H_{\text{te}}}(u_{i,r}, i) = \mathbf{I}$
$z^j_u = 1$	if node u_i or $u_{i,r}$ is infected in H^j (corresponds to u being infected within T in G)
g_m	maximum number of groups to which the set of nodes on any path in any H^j belong to. Typically, we expect $g_m << k$

Table S1: Notation for SpreadBlocking algorithm

set $\mathcal{Q}' \subseteq \mathcal{Q}$ such that $I(\mathcal{Q}') \notin [(1-\epsilon)\inf_{\text{T}}(V(\mathcal{Q}')), (1+\epsilon)\inf_{\text{T}}(V(\mathcal{Q}'))]$ is at most $k^B \frac{2}{k^{2B+1}} \leq \frac{1}{k}$ for $k \geq 2$ and $B \geq 1$. □

Let $ILP_{\tau_{\text{d}}}$ denote the integral version of $LP_{\tau_{\text{d}}}$, i.e., with all variables required to be in $\{0, 1\}$.

Lemma 4. *For any H^j, and any node $u_i \in V_{te}$ with $y^j_{u,i} < \frac{1}{2}$ (resp. for $u_{i,r} \in V_{te}$ with $y^j_{u,i,r} < \frac{1}{2}$), rounding in SpreadBlocking algorithm ensures that the node u_i (resp. $u_{i,r}$) is not reachable from S_{te} in $H^j - V_{te}(\mathcal{Q}_{SB})$, where $\mathcal{Q}_{SB}$ is the intervention set computed by the algorithm.*

Proof. Let $\mathcal{P}_{u_i,j}$ be the set of paths from S_{te} to node u_i in H^j. Let $U_{\tau_{\text{d}}}(P)$ denote the set of nodes on path P at distance τ_{d} or more from S_{te}. Let us denote $G_{\tau_{\text{d}}}(P) \subseteq \mathcal{Q}_{\text{SB}}$ to be the groups to which nodes in $U_{\tau_{\text{d}}}(P)$ belong.

We will prove the statement for $u_i \in V_{\text{te}}$ (a similar argument works for $u_{i,r}$ case). Given $y^j_{u,i} < \frac{1}{2}$, by rounding in SpreadBlocking we have $Y^j_{u,i} = 0$ in the integral solution (i.e., u_i is not infected in H^j). The node u_i is uninfected in H^j if and only if for every path $P \in \mathcal{P}_{u_i,j}$ at least one group in $G_{\tau_{\text{d}}}(P)$ is in $\mathcal{Q}_{\text{SB}}$, i.e., u_i is not reachable from S_{te} in $H^j - V(\mathcal{Q}_{\text{SB}})$. This corresponds to the constraint $\sum_{Q_q \in G_{\tau_{\text{d}}}(P)} x_{q,\tau_{\text{d}}} \geq 1 - y^j_{u,i} > \frac{1}{2}$ (this path-based constraint could be obtained from the edge constraints in $LP_{\tau_{\text{d}}}$ by adding all the constraints on the edges on path P). Assume for the sake of contradiction that none of the groups in $Q_q \in G_{\tau_{\text{d}}}(P)$ have $x_{q,\tau_{\text{d}}} \geq \frac{1}{2g_m}$. Then, we have $\sum_{Q_q \in G_{\tau_{\text{d}}}(P)} x_{q,\tau_{\text{d}}} < \frac{1}{2}$ as there could be at most g_m groups for any path where each $Q_q \in G_{\tau_{\text{d}}}(P)$ has $x_{q,\tau_{\text{d}}} < \frac{1}{2g_m}$. This is a contradiction as a feasible solution to $LP_{\tau_{\text{d}}}$ satisfies this constraint. Therefore, for a node $u_i \in H^j$, on each path $P \in \mathcal{P}_{u_i,j}$ there exists some group $Q_q \in G_{\tau_{\text{d}}}(P)$ with $x_{q,\tau_{\text{d}}} \geq \frac{1}{2g_m}$ implying that some $Q_q \in G_{\tau_{\text{d}}}(P)$ is in $\mathcal{Q}_{\text{SB}}$. □

Lemma 5. *Let $\mathcal{Q}_{SB} = \{Q_q \mid X_{q,\tau_d} = 1\}$ be the intervention set computed by SpreadBlocking algorithm, then we have $|\mathcal{Q}_{SB}| \leq 2g_m B$.*

Proof. The rounding procedure in SpreadBlocking scales each $x_{q,\tau_{\text{d}}}$ variable by a factor at most $2g_m$. Therefore, $|\mathcal{Q}_{\text{SB}}| = \sum_{Q_q \in \mathcal{Q}} X_{q,\tau_{\text{d}}} \leq \sum_{Q_q \in \mathcal{Q}} 2g_m x_{q,\tau_{\text{d}}} \leq 2g_m B$. The first inequality follows from the rounding and the second inequality follows from the budget constraint in $LP_{\tau_{\text{d}}}$. □

6 Computing g_m

Any simulation instance $H(V_H, E_H)$ can be viewed as a directed graph with $(u, v) \in E_H$ if and only if u infects v. Note that H is a directed acyclic graph (DAG). A topological ordering of the nodes is an ordering that satisfies the following condition: for any two nodes $(u, v) \in E_H$, $u < v$ in the ordering. We recall that every

node v in H has a corresponding node in G. Let $V_H^{\max} \subseteq V_H$ be the set of maximal elements. These are nodes that do not infect any other node. Let $g(v)$ denote the group to which its corresponding node in G belongs to. If v does not belong to any group, then $g(v) = \varnothing$. For any path P, let *pathgroups* $\text{pg}(P) = \{g(v) \neq \varnothing \mid v \in P\}$ denote the set of groups corresponding to the nodes of P. Let $\mathcal{P}_g(v)$ denote the set of $\text{pg}(P)$ for all paths P that start from some seed node and end at v. Note that $\mathcal{P}_g(v)$ contains only the distinct pathgroups. Algorithm 2 computes $\mathcal{P}_g(v)$ for all v. Then, it computes g_m as the size of the largest set in $\cup_{v \in V_H^{\max}} \mathcal{P}_g(v)$.

Algorithm 2: Computing $g_m(H)$ for a simulation cascade H, the maximum number of groups in any directed path from the set of sources S to a terminal in H.

Input : Simulation cascade $H(V_H, E_H)$
Output : $g_m(H)$

1 Let $\langle v_1, v_2, \ldots, v_n \rangle$ denote a topologically sorted order of the nodes of H.
2 $\forall v \in V_H$, let $\mathcal{P}_g(v) = \big\{\{\}\big\}$. `// Each` $\mathcal{P}_g(v)$ `is initialized to a set containing an empty set.`
3 **for** $i = 1, \ldots, n$ **do**
4 | Let $\mathcal{P}_g(v_i) = \cup_{u \text{ for } (u,v_i) \in V_H} \mathcal{P}_g(v_i)$ `//` v_i `inherits the pathgroups from its in-neighbors.`
5 | **if** $g(v_i) \neq \varnothing$ **then**
6 | | For each $A \in \mathcal{P}_g(v_i)$, $A \leftarrow A \cup \{g(v_i)\}$ `// The individual pathgroups are updated by the group of` v_i.
7 | **end**
8 **end**
9 $g_m = \max\{|A| \mid A \in \cup_{v \in V_H^{\max}} \mathcal{P}_g(v)\}$.

Validity. Firstly, we show that the Lines 4-7 indeed compute $\mathcal{P}_g(v)$ by the following inductive argument based on the topological ordering of nodes. **Base case:** Lines 4-7 compute $\mathcal{P}_g(v)$ when v is a seed node (or a minimal node in the ordering). As Line 4 does not change $\mathcal{P}_g(v)$ since in this case v does not have any predecessors and therefore, Line 6 simply adds $\{g(v)\}$ setting $\mathcal{P}_g(v) = \big\{\{g(v)\}\big\}$ if $g(v) \neq \varnothing$. **Assumption:** At iteration i, the topological ordering ensures that all in-neighbors of v_i are processed before it is processed. Hence, we will assume that for each in-neighbor of v_i, $\mathcal{P}_g(u)$ is correctly computed by this iteration. **Inductive step.** Any path P from a seed node to v_i is of the form $P'uv$ where u is some in-neighbor of v_i. Note that $\text{pg}(P'u) \in \mathcal{P}_g(u)$ by the inductive assumption. $\text{pg}(P) = \text{pg}(P'u) \cup \{g(v_i)\}$ if $g(v_i) \neq \varnothing$, else $\text{pg}(P) = \text{pg}(P'u)$. Lines 4-7 compute this for all in-neighbors. Finally, we note that it is enough to consider only the terminal nodes for computing g_m since for every $(u, v) \in E_H$ for each $A \in \mathcal{P}_g(u)$, there exists $A' \in \mathcal{P}_g(v)$ such that $A \subseteq A'$. Hence, proved.

Complexity. Note that the total number of groups is k and time horizon is T. Let $k' \leq \min(T, k)$. For any node v, $|\mathcal{P}_g(v)| \leq 2^{k'}$ as in the worst case all possible subsets of the k' groups are possible. Line 4 takes $O(d_{in}(v)2^{2k'})$ computations where d_{in} is the in-degree of v (d_{in} set union operations). Line 6 requires $O(2^{2k'})$ computations. Since $\sum_{v \in V_H} d_{in} = |E_H|$, the total time complexity is $O(|E_H|2^{2k'})$.

7 Multi-scenario robust optimization

Algorithm 3: Robust optimization for multi-scenario intervention.

Input : $G = (V, E)$, set of sources $S \subseteq V$, budget B, time horizon T, intervention delay τ_{d}
Output : intervention set $\mathcal{Q}_{\mathrm{SB}} \subseteq \mathcal{Q}$

Construct time expanded network H_{te} from G Construct M simulations of the SIR process $\{H^{C,1} = (V_{\mathrm{te}}, E_{\mathrm{te}}^{C,1}), \ldots, H^{C,M_C} = (V_{\mathrm{te}}, E_{\mathrm{te}}^{C,M_C})\}$ for scenario $C \in \mathcal{I}$, such that $\sum_C M_C = M$, on the time-expanded network H_{te} corresponding to SEI process on G.

Solve the linear program $LP_{\tau_{\mathrm{d}}}$ defined as follows:

$$
\begin{aligned}
(LP_{\tau_{\mathrm{d}}})\ \min \max_{C \in \mathcal{I}} \quad & \frac{1}{M_C} \sum_j \sum_u z_u^{C,j} \\
\forall C, \forall i < \tau_{\mathrm{d}}, u_i, u_{i,r} \in \mathcal{R}(H^{C,j}) \quad : \quad & y_{u,i}^{C,j} = 1, y_{u,i,r}^{C,j} = 1 \\
\forall C, \forall u_i, u_{i,r} \in \mathcal{R}(H^{C,j}) \quad : \quad & z_u^{C,j} \geq y_{u,i}^{C,j},\ z_u^{C,j} \geq y_{u,i,r}^{C,j} \\
\forall C, \forall (v_{i-1}, u_{i,0}) \in E_{\mathrm{te}}^{C,j} : y_{u,i,0}^{C,j} \quad \geq \quad & y_{v,i-1}^{C,j} - x_{g(u),\tau_{\mathrm{d}}} \\
\forall C, \forall (u_{i,r}, u_{i,r+1}) \in E_{\mathrm{te}}^{C,j} : y_{u,i,r+1}^{C,j} \quad \geq \quad & y_{u,i,r}^{C,j} - x_{g(u),\tau_{\mathrm{d}}} \\
\forall C, \forall (u_{i-\ell,\ell-1}, u_i) \in E_{\mathrm{te}}^{C,j} : y_{u,i}^{C,j} \quad \geq \quad & y_{u,i-\ell,\ell-1}^{C,j} - x_{g(u),\tau_{\mathrm{d}}} \\
\forall C, \forall (u_{i-1}, u_i) \in E_{\mathrm{te}}^{C,j} : y_{u,i}^{C,j} \quad \geq \quad & y_{u,i-1}^{C,j} - x_{g(u),\tau_{\mathrm{d}}} \\
\forall C, \forall i \geq \tau_{\mathrm{d}}, \forall u_i, u_{i,r} \in \mathcal{R}(H^{C,j}) \quad : \quad & y_{u,i}^{C,j} \leq 1 - x_{g(u),\tau_{\mathrm{d}}} \\
& y_{u,i,r}^{C,j} \leq 1 - x_{g(u),\tau_{\mathrm{d}}} \\
\sum_{Q_q \in \mathcal{Q}} x_{q,\tau_{\mathrm{d}}} \quad \leq \quad & B \\
\text{All variables} \quad \in \quad & [0, 1]
\end{aligned}
$$

(Rounding) Let $\mathbf{x}, \mathbf{y}, \mathbf{z}$ be the optimal fraction solution to $LP_{\tau_{\mathrm{d}}}$. Round it to an integral solution X, Y, Z using the following rounding procedure:

(i) For each $C \in \mathcal{S}$, H^j, u_i, set $Y_{u,i}^{C,j} = 1$ if $y_{u,i}^{C,j} \geq \frac{1}{2}$. Similarly, for each $H^{C,j}, u_{i,r}$, set $Y_{u,i,r}^{C,j} = 1$ if $y_{u,i,r}^{C,j} \geq \frac{1}{2}$.

(ii) For each C, and for each $H^{C,j}, u \in V$, set $Z_u^{C,j} = 1$ if $z_u^{C,j} \geq \frac{1}{2}$.

(iii) For each $Q_q \in \mathcal{Q}$, set $X_{q,\tau_{\mathrm{d}}} = 1$ if $x_{q,\tau_{\mathrm{d}}} \geq \frac{1}{2g_m}$ where g_m, such that $g_m \leq k$ is the maximum number of groups associated with the set of nodes on any path in any H^j.

return $\mathcal{Q}_{\mathrm{SB}} = \{Q_q \mid X_{q,\tau_{\mathrm{d}}} = 1\}$

8 Additional details for experiments and results

The table of networks is provided in Table S2.

Computation environment. The simulator [5] and the intervention algorithm are implemented using the Python programming language. We used the Gurobi software [2] to implement and solve the GroupInt algorithm. All experiments in this paper are performed on an HPC system that runs Linux $x86\_64$ operating system where each compute node contains 40 threads with a total of 320GB memory.

Implementation details and computation time. The results shown above were generated through the use of a two-part pipeline: first, simulating infection cascades M times per sample set, then running the

Table S2: List of networks used and their attributes.

net.	name	nodes	edges	groups	gp. edges
BD	Bangladesh	211	6846	7	141
ID	Indonesia	3296	110640	35	2181
PH	Philippines	673	20108	16	450
TH	Thailand	738	27666	5	48
VN	Vietnam	503	16746	15	426

intervention algorithm using the generated cascades on different combinations of budgets and intervention delays. The pipeline was run for different values of M and different sample sets were run in parallel through the use of a high-performance computing cluster.

The multi-pathway simulator was run as a simple, single-threaded process; its computation time is directly proportional to the number of simulations M and the network size/complexity. The SpreadBlocking algorithm involves building an ILP model using a simulated cascade, then solving it using the Gurobi Optimizer. Therefore, the size of the ILP and the time needed to build it depend on M and the cascade size (which in turn depends on network size, model parameters, and the time horizon). Accordingly, we have configured the Gurobi Optimizer to utilize a variable number of threads per linear program (up to 20), as a function of cascade size, number of simulations M, and intervention delay. The optimizer also uses different optimization methods depending on the number of allotted threads: barrier method for one thread, and simultaneous barrier, dual simplex, and primal simplex if given more.

Due to numerous variables in play as well as multiple solvers running in parallel, the computation time of the intervention phase can be difficult to predict in advance. Sometimes an LP model is solved quickly by primal simplex, but in other cases, the complex structure of the model forces the barrier method to run to completion. In theory, the required time should increase as the complexity of the ILP instance increases, such as when the intervention delay is high or when the budget is close to half of the total number of groups in the network, but in reality the running times for our computations were highly variable. For example, for the BD network with $M = 250$, a delay $\tau_{\mathrm{d}} = 6$ and a budget $B = 3$ takes about 5-6 minutes to solve, whereas $\tau_{\mathrm{d}} = 3$ and $B = 3$ takes 10-30 minutes, due to the solver using twice as many threads for the former compared to the latter. Among all ILP instances, the longest solve time was roughly 54 minutes, though this is a conspicuous outlier: all but six instances were able to finish solving in under 30 minutes.

References

[1] Devdatt P. Dubhashi and Alessandro Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms.* Cambridge University Press, 2009.

[2] LLC Gurobi Optimization. Gurobi optimizer reference manual, 2021.

[3] Ara Hayrapetyan, David Kempe, Martin Pál, and Zoya Svitkina. Unbalanced graph cuts. In *Proceedings of the 13th Annual European Conference on Algorithms*, ESA'05, page 191–202, Berlin, Heidelberg, 2005. Springer-Verlag.

[4] Madhav Marathe and Anil Kumar S Vullikanti. Computational epidemiology. *Communications of the ACM*, 56(7):88–96, 2013.

[5] Joseph McNitt, Young Yun Chungbaek, Henning Mortveit, Madhav Marathe, Mateus R Campos, Nicolas Desneux, Thierry Brévault, Rangaswamy Muniappan, and Abhijin Adiga. Assessing the multi-pathway threat from an invasive agricultural pest: *Tuta absoluta* in Asia. *Proceedings of the Royal Society B*, 286(1913):20191159, 2019.

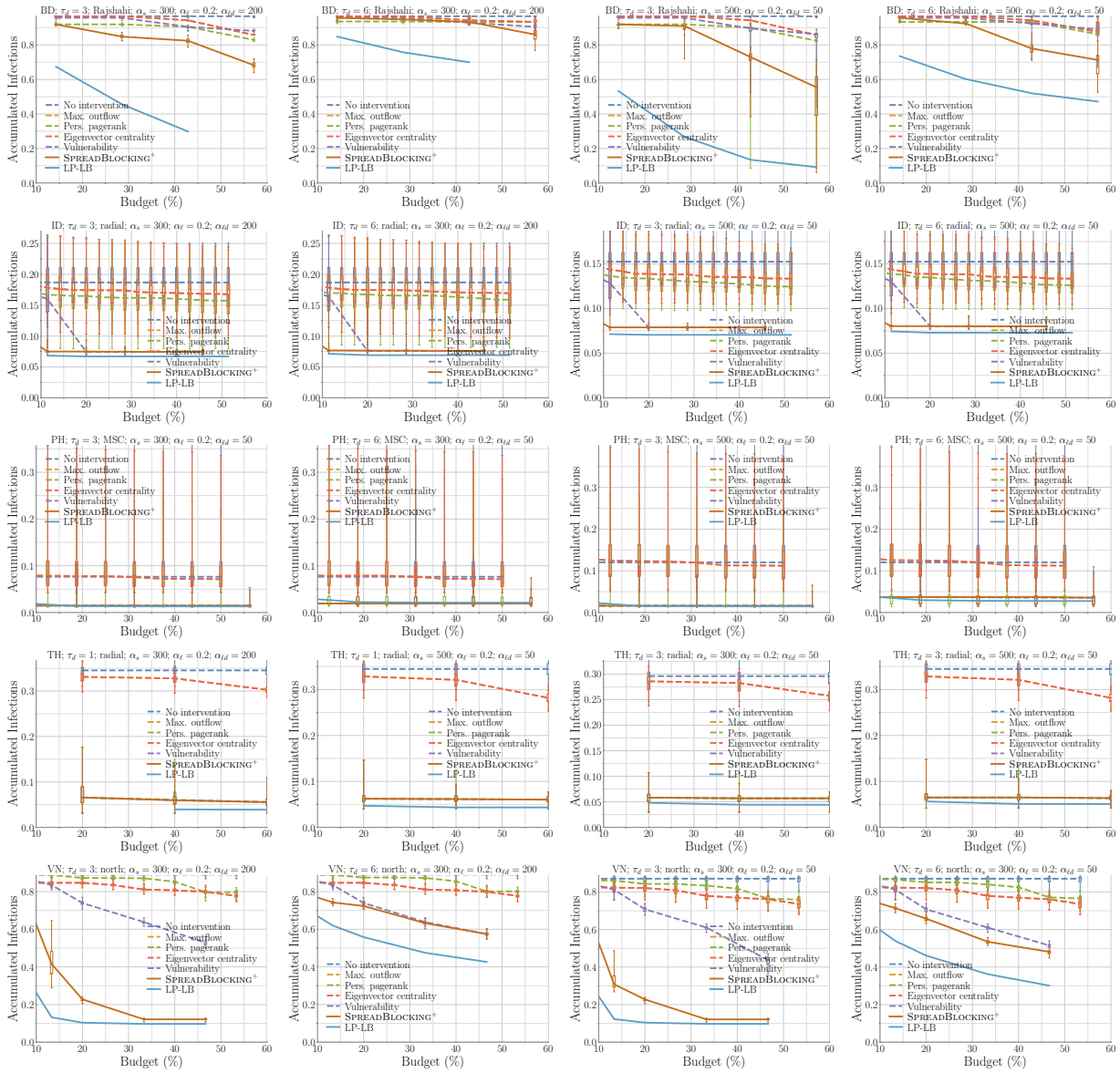


Figure S2: Comparison with baselines. Accumulated infections vs. budget. This is continued from Figure 2.

[6] Peng Zhang, Jin-Yi Cai, Lin-Qing Tang, and Wen-Bo Zhao. Approximation and hardness results for label cut and related problems. *Journal of Combinatorial Optimization*, 2009.

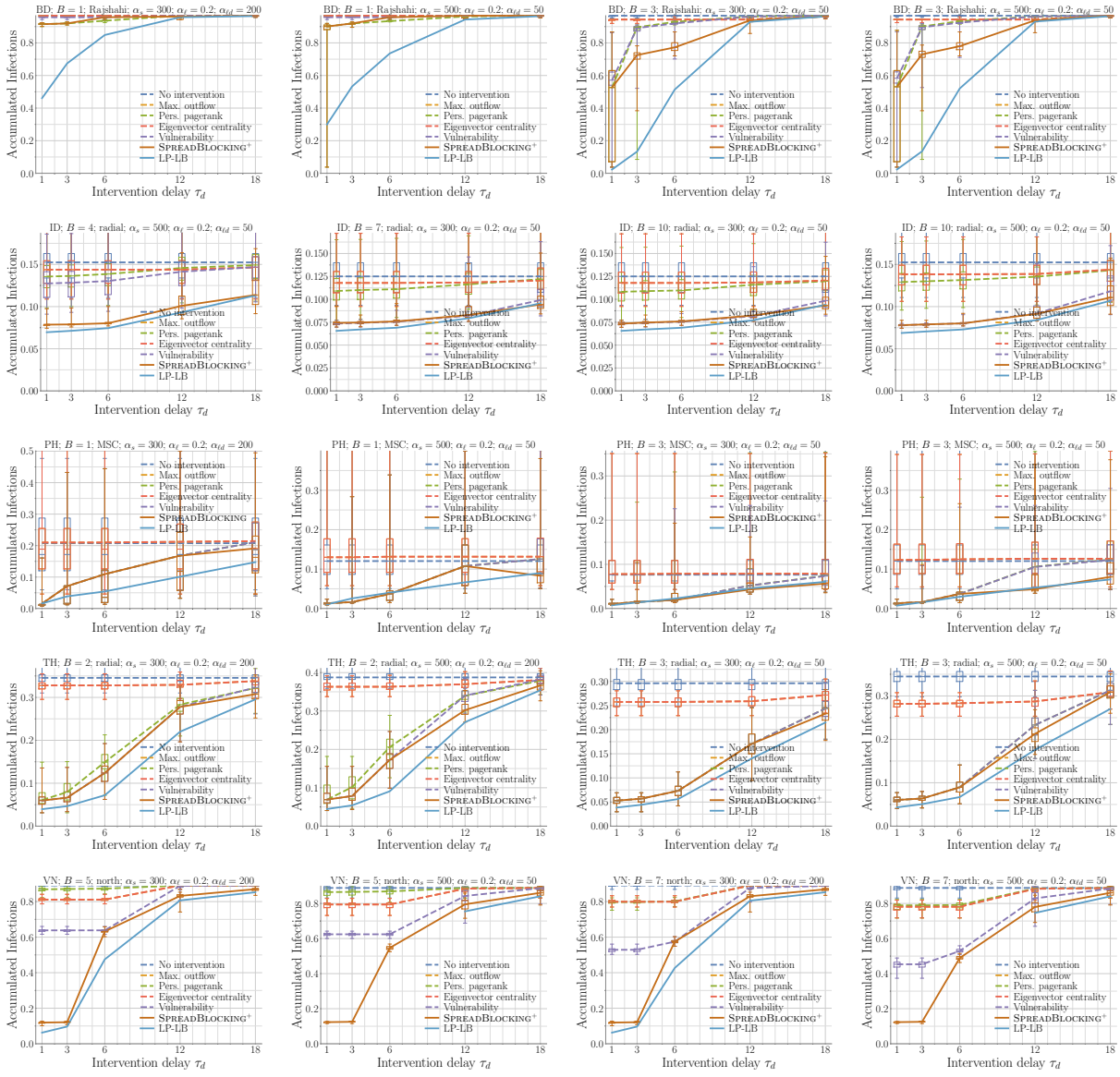


Figure S3: Comparison with baselines. Accumulated infections vs. intervention delay.

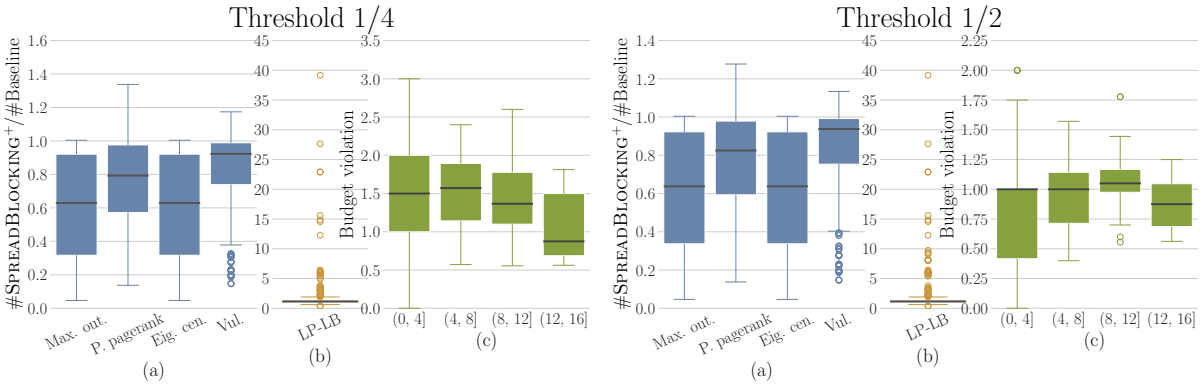


Figure S4: Evaluating the results with respect to the number of infections and budget violation. This is continued from Figure 3 in the main paper.

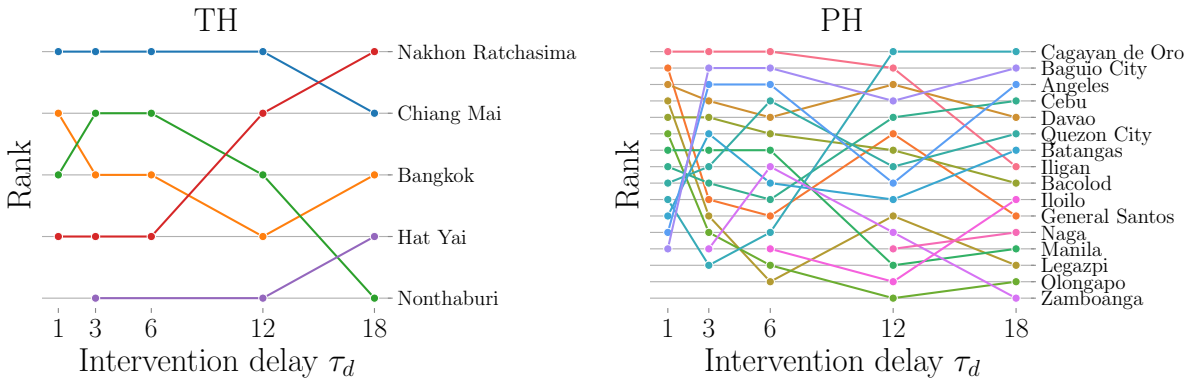


Figure S5: Ranking of groups based on their frequency of occurrence in solution for varying τ_{d} across model parameters.

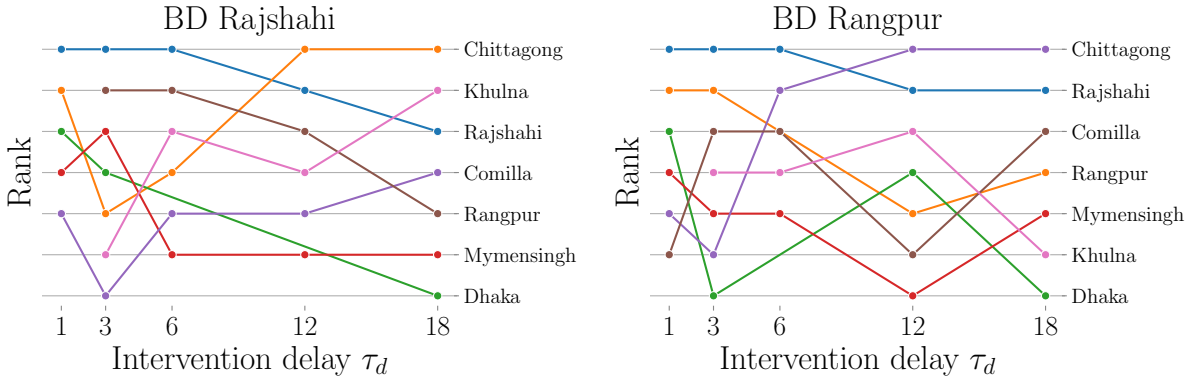


Figure S6: Solution sets under different seeding scenarios for BD. Continued from Figure 5.

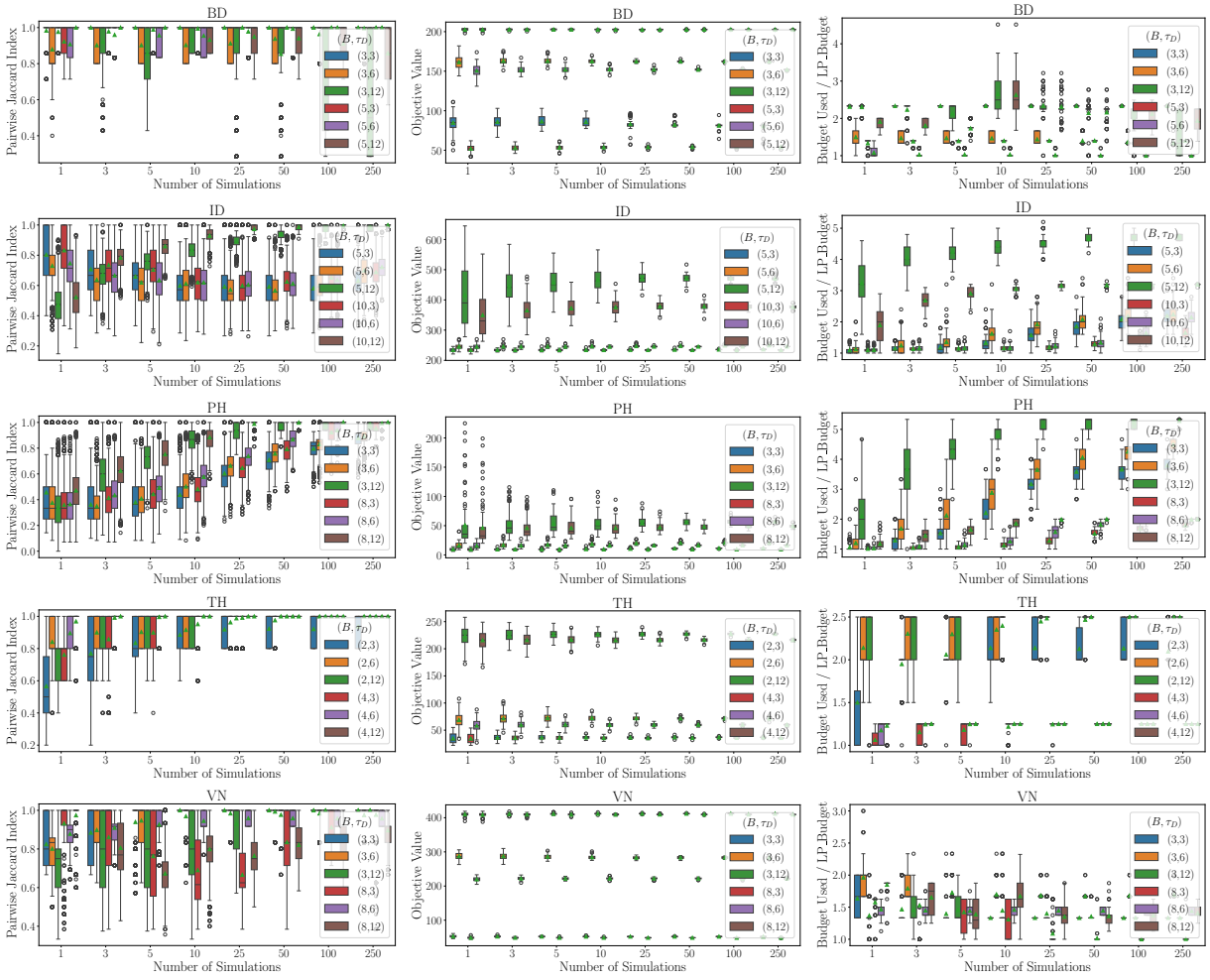
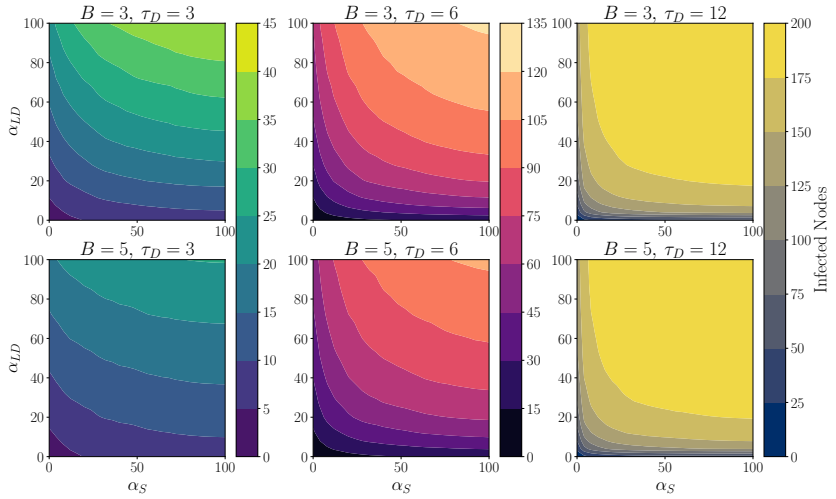
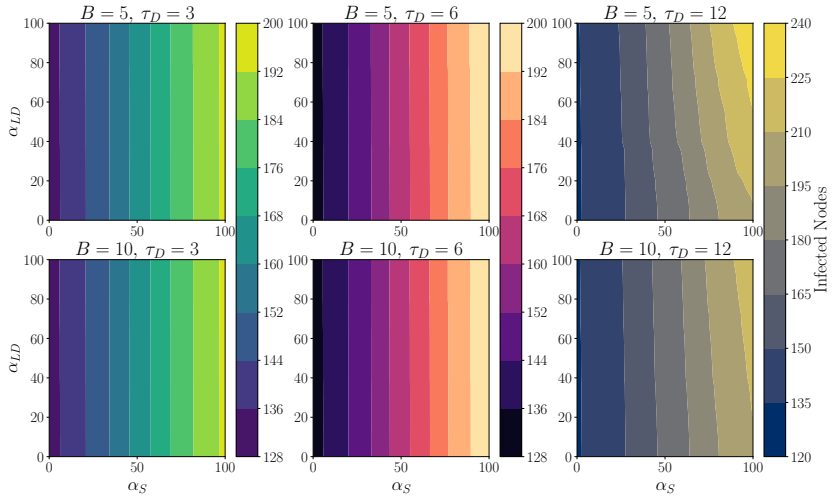


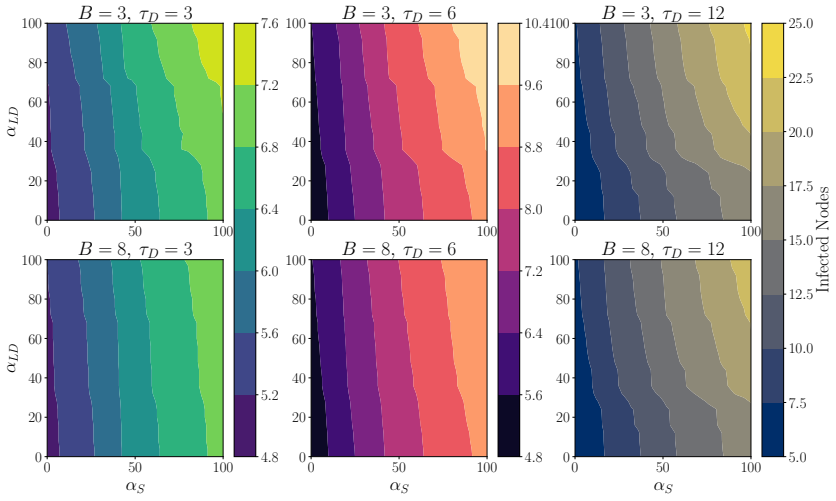
Figure S7: Solution stability with respect to number of simulation instances M for three networks. For each M, we considered 100 replicates, and therefore, we have 100 corresponding solutions. The first column corresponds to Jaccard index between each pair of solutions for increasing M. The second column corresponds to objective value while the third column corresponds to budget violation.



(a) BD



(b) ID



(c) PH

Figure S8: The influence of the different pathways on the efficacy of intervention solutions. Continued from Figure 7.