Controlling Diffusion on Multi-Pathway Spatial Networks: Application to Biological Invasions

---

A

Thesis

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

---

in partial fulfillment

of the requirements for the degree

Master of Science

by

Manisha Sudhir

May 2021

APPROVAL SHEET

This

Thesis

is submitted in partial fulfillment of the requirements
for the degree of

Master of Science

Author: Manisha Sudhir

This Thesis has been read and approved by the examing committee:

Advisor: Anil Vullikanti

Advisor: Abhijin Adiga

Committee Member: Madhav Marathe (Chair)

Committee Member: Henning S. Mortveit

Committee Member:

Committee Member:

Committee Member:

Committee Member:

Accepted for the School of Engineering and Applied Science:

Craig H. Benson, School of Engineering and Applied Science

May 2021

Acknowledgements

I would like to acknowledge Professor Abhijin Adiga and Professor Anil Vullikanti for their invaluable time and efforts spent on mentoring me throughout my Masters. They have taught me all that I know about computational epidemiology as well as taught me the importance of research. I would like to also thank Professor Madhav Marathe for giving me the opportunity to work at the Biocomplexity Institute and Initiative and always encouraging me to learn something new and push my boundaries. I would like to thank Professor Henning S. Mortveit for taking the time to be on my committee. In addition, I would like to thank the staff and facilities at UVA for making my time here invaluable and providing a friendly and approachable learning environment. Finally, I would like to thank my parents for all their love and support.

ABSTRACT

An important challenge in agriculture and food security is the control of invasive alien species (IAS) spread that affect important agricultural crops. However, optimal control of such epidemics is a challenging problem. In this thesis, we consider the problem of controlling a multi-pathway epidemiological process on a temporal network. Our focus is on the problem of group-scale interventions, where the objective is to find an optimal set of regions (or groups of nodes) to intervene at so as to minimize the spread. Such interventions correspond to region-wide management techniques, which are more realistic compared to targeted interventions that are typically studied in network science. In this collaborative work, we designed, implemented and analyzed an algorithm called SPREADBLOCKING for intervention problem. Our method uses sample average approximation technique and a linear relaxation of an integer linear program.

This thesis contributes to the implementation of the simulator, experimental framework and analysis. We implemented the multipathway simulator using vectorization methods, and achieved an order of magnitude speed improvement over the previous version. We integrated the simulator with the intervention algorithm. This involved representing simulation instances, which correspond to Susceptible-Exposed-Infectious (SEI) process on the input network to a Susceptible-Infectious-Recovered (SIR) process on a time-expanded graph. For experimental evaluation of the SPREADBLOCKING algorithm, we implemented popular baselines for comparison of our results. Finally, we conducted experiments to evaluate our intervention algorithm on several real-world networks with respect to budget, introduction scenarios and intervention delays.

Our results show superior performance across model parameters compared to the baselines. We note that early discovery of the IAS and speed of intervention are critical to identify intervention candidates under model uncertainty. We observe that groups with high inflow, even though vulnerable, are not necessarily chosen as candidates for intervention. Across model parameters, we note that performance of group-scale interventions is comparable to individual-based interventions in performance, though the former is more practical from an implementation perspective.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

An important challenge in agriculture and food security is the control of invasive alien species spread (IAS) that affect important agricultural crops. Biological invasions cause disruptions to native ecosystems, and negatively impact health and economy, with annual economic impact of over $120B in the United States alone [25]. At the global scale, the economic impacts of IAS are significant. For example, invasive insect species alone are thought to be responsible for more than 70 billion USD/year in lost ecosystem goods and services, and 6.9 billion USD/year in health costs, far outweighing the economic benefits of these species [5]. Moreover, many IAS affect landscape aesthetics and biodiversity, indirectly affecting tourism and other businesses related to recreation. From researching various sources, it is clear that IAS management and control is an important environmental, social, and economic issue [17]. In this work, we study the spread of a representative pest called *Tuta absoluta* [3], which has been responsible for devastating tomato crops globally.

The spread of IAS has been successfully modeled using network and agent-based methods [32, 8, 23, 21]. In such models, nodes represent spatial regions (e.g., counties), and edges represent flows between regions (e.g., through trade). Node and edge attributes vary in time reflecting seasonal weather patterns and cropping cycles of host crops. Different pathways of spread such as self-mediated, wind, trade, etc. lead to multi-edged networks. Hence, we can say that IAS spread is

a multi-pathway phenomenon driven by various natural and anthropogenic factors [15]. We note that many other natural and social phenomena can be modeled as epidemiological processes on networks [20].

1.1.1 Multi-pathway Model

There are various ways in which a epidemiological processes can be modelled depending on the characteristics of the studied phenomenon. In this thesis the diffusion process used is based on Susceptible-Exposed-Infectious (SEI) process defined in Section 2.1. Some of the other popular diffusion processes include Susceptible-Infectious-Susceptible (SIS) model, Susceptible-Infectious-Removed (SIR) model (defined in Section 2.1).

In this thesis, we focus on the multi-pathway model developed for the spread of IAS in McNitt et al. [21]. The study area is overlaid with a grid and induces the first-level nodes in the network (Figure 2.1). Some nodes in this spatial network belong to groups (called localities), which represent regions of major supply of host crops and demand. A node has two time-varying attributes, suitability $\epsilon(v,t)$ for pest establishment and infectivity $\rho(v,t)$. There are three pathways of spread. Self-mediated dispersal corresponds to a diffusion from one cell to its adjacent cells, local human-mediated dispersal is diffusion within a group (farmer-market interactions), and long-distance dispersal corresponds to diffusion from cells from one group to another (trade). The probability that a node can be infected through a pathway is modeled as a negative exponential function of infectivity and pathway parameters, which can be expressed as edge weights between two nodes.

Accounting for various ways in which dispersal can occur typically results in a complex model. In addition, lack of usable data and systematic modeling methods makes it very hard to validate these models. Also, there are challenges in analyzing such processes. Analysis of even simple diffusion processes on static networks is hard [11]. For complex diffusion processes, one usually would have to rely on in-silico studies for insights into their behavior.

1.2 Controlling Multi-pathway Spread of Invasive Species

Managing invasive species is a major challenge for society. In the case of newly established invaders, rapid action is key for a successful management. Models and algorithms can be used to evaluate various intervention strategies. However, interventions are resource intensive and therefore, are typically applied with budget constraints.

Applying pesticides, setting up pheromone traps and imposing trade embargoes have been the main methods applied to control *T. absoluta*. Such interventions incur huge economic costs, and therefore, designing optimal interventions is a fundamental challenge in agriculture. As in the case of other diffusion processes, controlling diffusion in a multi-pathway model is computationally very challenging. In many ordinary differential equation models, which have been used in the study of spread processes, interventions can be computed optimally (e.g., Medlock and Galvani [22]). However, designing optimal intervention strategies in network-based diffusion models is much harder [4, 27, 29]. Wilder et al. [33] consider optimal interventions in a dynamic population under a continuous-time SIS model. There has been a lot of work in different types of models, e.g., [4, 33, 27, 29]. Prior results do not immediately provide results for the problem considered because the multi-pathway models like those in [21] are different from SIS/SIR models used by the above-mentioned works, and the goal is to minimize the expected number of infections, which is not captured by some of the prior work, e.g., [27, 32, 8].

Typically controlling an outbreak corresponds to removing nodes (or edges) of the network in order to stop or slow down the diffusion process. In an impending epidemic scenario, policy makers have to develop strategies under resource constraints. Therefore, deciding which nodes to intervene at is an important problem that has been well-studied in the context of infectious disease spread and other social phenomena [4, 33, 27]. Here, we focus on developing practical control algorithms in the context of IAS that affect important agricultural crops.

1.3 Group-scale Interventions

This thesis studies group-scale interventions in the multi-pathway epidemiological process described in Section 1.1.1. Throughout this thesis, intervening at a group (or locality) means removing all nodes present in the group. Given resource constraints, we must be able to identify which are the optimal groups to intervene, such that it reduces the total number of accumulated infections in the network. We study group-scale interventions because optimal strategies based on node level characteristics cannot be easily turned into implementable policies; targeted interventions are harder as it is difficult to include every field or cultivated patch of land in the model. In fact, even in infectious disease spread, vaccination policies such as those specified by CDC are at the level of groups (e.g., based on demographics), and almost all the efforts in epidemiology are focused on developing group level strategies, even though this may lead to sub-optimal solutions compared to the node level intervention strategies.

1.4 Contributions

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. [21]. Our contributions are as follows.

1.4.1 Multi-pathway Model and its Implementation

Recall the brief description of the multi-pathway simulator in Section 1.1.1. We implemented an improved and faster version of this model. Firstly, we used the concept of "live-edges", where we first randomly sampled edges consisting of an infectious source node and a susceptible target node with a probability proportional to the edge weight, and then decided which nodes were infected in the current time step. This enabled us to effectively use Pandas (in Python) vectorization methods and leverage DataFrames to store edges, nodes, their attributes and state information. Secondly, for two pathways, we applied aggregation of infections at the group level, which reduced the number of edges to be processed. Here, we leveraged Pandas GroupBy, Join and Map operations to achieve

further speedup. The revised simulator is about 10 times faster for a network of 200 nodes with the same simulation parameters 2.1 compared to the older version. The simulator also used a generalised notion of multi-scale spatial network. The input network for the simulator defines the relationship of nodes at node level as well as "group" level.

1.4.2 Equivalence of Multi-pathway Model to SIR Process on a Time-Expanded Network

To the best of our knowledge there is no work on intervention algorithms for SEI diffusion processes. However, control of SIR diffusion processes is well-studied. To leverage such approaches, we come up with the notion of time-expanded graph of a network, and show that SEI diffusion processes on the multi-pathway network can be provably reduced to an SIR process on the time-expanded network. In Chapter 4 we define the time expanded network and and show this equivalence. The implementation details of representing simulation output as the corresponding subgraph of the time-expanded graph is also described.

1.4.3 Group-scale Intervention Problem and Algorithm

We introduce a new group-scale intervention problem (IAScontrol) to formalize control strategies for IAS. We design algorithm SpreadBlocking for IAScontrol for choosing which groups to intervene, and when, given resource constraints. Our method is a combination of the sample average approximation (SAA) technique, with a linear relaxation of an integer linear program (ILP). We prove rigorous guarantees on its performance. A detailed explanation of these concepts are present in Chapter 5. We also implemented three popular control methods (max. outflow, vulnerability and exhaustive search) to compare the performance of our intervention algorithm with. The Outflow-based method is is used in many works [28, 27, 23]. There are many variants depending on the models used. Here, it is implemented by grouping the total outflows from each locality. In an unweighted graph this would correspond to a degree based heuristic. The vulnerability of a node is the probability that the node will be infected when no interventions are in place subject to certain initial seeding. Candidates for intervention are chosen based on their vulnerability. The exhaustive

baseline corresponds to considering all possible solution sets of size B (budget) in the solution space. This baseline would provide the best solution within the limits of the sample average approximation. More details on the implementation of these methods are present in sections 5.1, 5.2 and 5.3.

1.4.4 Experimental Analysis

We use SpreadBlocking to study real-world networks considered in McNitt et al. [21]. We perform experiments and analyze the algorithm for its effectiveness and solution quality under different values of budget B, intervention delay τ_d and model parameters. We compare the performance of the intervention algorithm with the three baselines. The algorithm's performance is also compared with the targeted intervention case (each node belonging to a distinct group). We analyze the structural properties of groups that appear prominently in the solutions to gain insights into the structural and dynamical properties of the network that influence spread.

1.4.5 Specific Contributions

This thesis contributes to the work in Section 1.4.1, which corresponds to implementing the simulator, and the conversion of simulation instances to equivalent subgraphs of the time-expanded network (Section 1.4.2). In Section 1.4.3, this thesis contributes to the selection and implementation of the algorithms that serve as baselines for the SpreadBlocking algorithm. Experiment design and analysis is another major contribution of this thesis (Section 1.4.4).

Chapter 2

The Multi-pathway model and its Implementation

Spread dynamics of invasive species spread are influenced by habitat suitability and human activities. The spread can occur across a network of nodes that represents the focus region, with linkages (or edges) representing the potential movement between such habitats [9]. Our work uses one such agent-based model to study the multi-pathway spread of invasive alien species. This model was used by Mcnitt et al. [21] to investigate the spread of the South American tomato leafminer, a tomato pest. The model accounts for both self-mediated and human-mediated spread, as well as the propagation mechanisms' spatial heterogeneity, temporal variations, and multi-scale design. The authors demonstrate the significance of trade pathways in the spread of the pest. This model is generic and can be applied to several biological invasion scenarios that include other pathways (for e.g., wind dispersal). The objective of this chapter is to describe the Multi-pathway model and it's implementation. We first explain the concept of this model mathematically in Section 2.1 and Section 2.2 and then describe how this model was improved and implemented using Pandas concepts in Python in Section 2.3, which is a contribution to this thesis.

2.1 Preliminaries

Let $G(V, E)$ be a temporal edge-weighted directed graph. Let the weight of an edge $(u, v) \in E$ at a discrete time step $t = \{0, 1, \ldots, T\}$ be denoted by $w(u, v, t)$, where T is the maximum number of time steps or the *time horizon*. Let $\mathcal{Q} = \{Q_1, Q_2, Q_3, .., Q_k\}$ be a collection of k disjoint subsets of V. Each subset Q_i is a *group*. For a vertex v, let $g(v)$ denote the group it belongs to. Two types of network-based diffusion processes are considered in this paper:

Susceptible-Infectious-Removed (SIR): A node is in Susceptible (**S**) state if it is not yet infected. When a susceptible comes in contact with an infectious node (in state **I**), it can get infected. If it gets infected, then, it transitions from state **S** to state **I**. The node stays in state **I** for exactly one time step. It can infect any of its susceptible neighbors in that time step. In the next time step, it moves to the Removed (**R**) state when it is effectively removed from the network from the perspective of the diffusion process.

Susceptible-Exposed-Infectious (SEI): In this model, a susceptible node when infected, transitions to the exposed state **E** when it is infected but not infectious. It stays in the exposed state for a fixed number of time steps, which is denoted by latency period ℓ. After the latency period, the node transitions to the **I** state.

Let S $\subseteq V$ denote the initial set of nodes in state **I** that seed the diffusion process at $t = 0$. In the SEI process, a node moves from **E** to **I** after ℓ time steps, where ℓ is referred to as latency period.

2.2 Multi-pathway Model for IAS Spread

We will describe briefly the model developed in McNitt et al. [21]. It is illustrated in Figure 2.1. The study region is divided into cells, which correspond to the set of nodes V of the spatial network. The nodes are partitioned into groups (called localities in Mcnitt et al. [21]), which represent regions of major supply of host crops and demand. It is possible that some nodes do not belong to any locality. There are three pathways of spread. Self-mediated dispersal corresponds to diffusion from one

cell to its adjacent cells, local human-mediated dispersal is diffusion within a group farmer-market interactions), and long-distance dispersal corresponds to diffusion from cells from one group to another (trade).

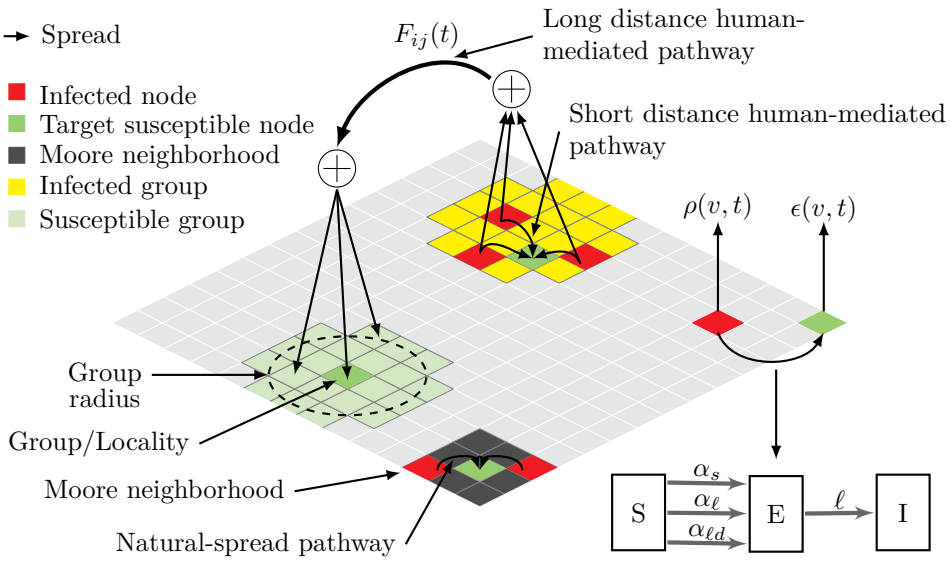

Figure 2.1: An illustration of the multi-pathway model [21].

A node has two periodic time-varying attributes, suitability $\epsilon(v,t)$ for pest establishment and infectivity $\rho(v,t)$. Here, each period unit corresponds to a month in the year. The probability that a node can be infected through a pathway is modeled as a negative exponential function of infectivity and pathway parameters, which can be expressed as edge weights between two nodes. For the short-distance dispersal, the probability that node v is infected by its neighbor v' within its Moore neighborhood M_v (where the range of the Moore neighbourhood is denoted by r_{M}) is:

$$p_s(v,t) = \epsilon(v,t)\Big(1 - \exp\Big(-\alpha_s \sum_{v' \in \mathrm{M}_v(r_{\mathrm{M}})} \rho(v',t)\Big)\Big). \tag{2.1}$$

Expression (2.1) can be disaggregated into source-target probability terms. There are two implicit assumptions that have been made: (i) the probability that a neighbor of v, say v' does not infect v at time t is $e^{-\alpha_s \rho(v',t)}$, and (ii) each neighbor independently infects v with this probability. Under these assumptions, we can rearrange the expression for eq. (2.1) as follows:

$$p_s(v,t) = \epsilon(v,t)\Big(1 - \prod_{v' \in \mathrm{M}_v(r_\mathrm{M})} e^{-\alpha_s \rho(v',t)}\Big). \tag{2.2}$$

The probability depends on the suitability of the cell $\epsilon(v,t)$, infestation level of each neighbouring cell in the Moore neighbourhood with range r_M, $\rho(v',t)$, and the scaling factor, α_s, the pathway parameter. To state more clearly, $e^{-\alpha_s \rho(v',t)}$ is the edge weight of node v to v'. Here, $1 - e^{-\alpha_s \rho(v',t)}$ is the probability that node v' will infect node v. Node v' can be infected by any of the incoming edge weights and hence, Equation 2.2 refers to the probability that at least one of the v' infects v. The same concept can be followed for the rest of the equations. In Equation 2.2, since it is a short distance pathway v' is a node/cell neighbour within the Moore neighbourhood of v.

Local human-mediated dispersal is modeled as the spread within a locality. Every cell v is influenced by cells in its locality L based on their infectiousness:

$$p_\ell(v,t) = \epsilon(v,t)\Big(1 - \exp\Big(-\alpha_\ell \sum_{v' \in L} \rho(v',t)\Big)\Big), \tag{2.3}$$

where α_ℓ is the scaling factor. This equation can be written as:

$$p_\ell(v,t) = \epsilon(v,t)\Big(1 - \prod_{v' \in L} e^{-\alpha_\ell \rho(v',t)}\Big). \tag{2.4}$$

Long-distance human-mediated dispersal corresponds to spread through trade between localities. We define long distance edge-weighted flows F_{ij} as flow of production from locality i to j. The probability of spread is directly proportional to (i) the trade flow F_{ij} from locality i to j and (ii) total infectiousness of the locality, which is just the sum of infectiousness of cells belonging to that locality. Suppose cell v belongs to locality i. Then, the probability of cell v being infected due to this pathway is given by:

$$p_{\ell d}(v,t) = \epsilon(v,t)\Big(1 - \exp\Big(-\alpha_{\ell d} \sum_{j \neq i} \sum_{v' \in L(j)} F_{ji} \rho(v',t)\Big)\Big), \tag{2.5}$$

where $\alpha_{\ell d}$ is the pathway scaling factor. This equation can similarly be rewritten as:

$$p_{\ell d}(v,t) = \epsilon(v,t)\left(1 - \prod_{j \neq i} e^{-\alpha_\ell F_{ji} \sum_{v' \in L(j)} \rho(v',t)}\right) \tag{2.6}$$

2.3 The Multi-pathway Simulator

The simulator in McNitt et al. [21] was improved in terms of speed and capacity for handling larger networks. The performance analysis was done and the results are in Table 2.1. The previous implementation of the multi-pathway model used naive Python constructs. For every infected node per time step, the neighbouring nodes were infected with corresponding edge probabilities. On being infected/exposed, these neighbouring nodes would be respectively added to an "infected array" and an "exposed array". All the nodes in each array would be updated to their respective states in a for loop. Although this method only performs operations on infected nodes at any given time step, it does not scale well with larger networks having a higher degree of infection spread. This method is only beneficial if only a small number of nodes are infected in the network and the infection does not spread. If we have a large network with a very large spread in infections, sequentially checking each infected node and performing sub-operations every simulation step can be quite slow. The new version of the simulator leverages the vectorization and Groupby features of Pandas to significantly improve the performance of the revised version of the simulator.

The basic data structures of Pandas – DataFrames and series – are based on arrays. The built-in Pandas functions are carefully designed to operate on entire arrays, instead of sequentially on individual values. Vectorization is the process of executing operations on entire arrays by leveraging efficient under-the-hood functions. Pandas includes a large collection of vectorized functions for mathematical operations, aggregations and string functions. These functions are optimized to operate specifically on Pandas series and DataFrames. Therefore, it is recommended that vectorized Pandas functions be used wherever possible.

Live edges. In the revised simulator we operate in terms of "edges" instead of "node" operations. We check if an edge is "live" by checking the following:

1. The source state being infectious.
2. Target being suitable for infection.
3. Target being susceptible.
4. The probability of the source being able to infect the target is greater than the randomly generated probability value. (The probability that the source will be able to infect the target are calculated by the pathway equations.)

If all the above holds true then the edge is considered to be a live edge. Any target node that participates in a live-edge, i.e, an end point of a live edge is considered to be exposed at this time interval. Each of these above said operations/checks can be performed on the entire DataFrame at once instead of sequentially in a for loop. All the live-edges at each simulation step are collected at once by Pandas group-by operations and added to the time-expanded output.

Aggregation of infections and inter-level edges. We aggregate the infections for each locality/group by summing up the infections of its constituent cells. This is done by grouping cells by locality first. Pandas enables us to easily aggregate information at the locality level by its "Groupby" operation. These values can further be used when calculating edge probabilities of infection through the pathway equations explained in this section. This drastically lets us reduce the number of edge

Table 2.1: Simulator timings compared with the previous version [21] of the simulator. Parameters used for evaluation: Time steps: 24, simulation runs: 10, start month: 5, Moore range: 1, suitability threshold: 0, latency period: 2.

Network	Nodes	Simulator[Mcnitt et al.]	Simulator(improved)
BD	211	1.5 minutes	7 seconds
PH	673	5.16 minutes	9 seconds
ID	3296	15.84 minutes	11 seconds
VN	503	4.30 minutes	7 seconds
TH	738	2.95 minutes	8 seconds

operations (while calculating edge probabilities of infection in local human mediated pathway and long distance human-mediated dispersal) on the network.

These pandas operations improve the speed of the simulator and thereby provide the capability for handling large networks with larger spread in infections. Although one might argue that the space complexity increases as we are storing states of every node and not just the infected nodes. There is always the general notion of a trade-off between time and space complexity. Reducing the time-complexity would make a huge impact in terms of studying results and performing experiments on large networks such as the USA and regional networks. We compared the two versions of the simulator. Results are in Table 2.1.

Chapter 3

Problem Definition

3.1 Preliminaries

Here, we define the group-scale intervention problem IAScontrol for the SEI process on a network. Intervening at a group means removing all nodes in a group. The intervention is performed τ_{d} time steps after revealing the source. In the SEI process, a node moves from **E** to **I** after ℓ time steps, where ℓ is referred to as latency period. Here, we would like to clarify that the interventions are non-adaptive, i.e., the decision to intervene is not made by observing the system state at time $\tau_{\mathrm{d}} - 1$ or before. Instead, it is based on the expected state of the system at τ_{d}. The reason for introducing the delay parameter is to study the negative effects of delaying interventions. Suppose $V' \subseteq V$ is the set of nodes intervened at τ_{d}, then, let $\inf_{\mathrm{T}}(G, \mathrm{S}, \tau_{\mathrm{d}}, V')$ (we can drop $G, \mathrm{S}, \tau_{\mathrm{d}}$ when context is clear) denote the expected number of nodes exposed at a time horizon T due to SEI diffusion with source nodes S when intervention is applied at nodes in V' at time τ_{d}. Note that unlike SIR process, the steady state for an SEI process is all nodes reachable from S becoming exposed. Therefore, the intervention problem is relevant only when the time horizon T is finite. We will now define the problem and provide an example of the intervention process.

3.2 IASCONTROL

The IAScontrol problem is formally defined below.

Definition 1 (IAScontrol problem)**.**

Instance. Given a temporal edge-weighted directed graph $G(V, E)$, a partition of the vertex set into groups $\mathcal{Q}$, source nodes $\mathrm{S} \subseteq V$, SEI diffusion process on G with transmission probabilities equal to edge weights, budget B, intervention delay τ_{d} and time horizon T.

Goal. Find a set of groups $\mathcal{Q}^* \subseteq \mathcal{Q}$ such that $|\mathcal{Q}^*| \leq B$ and the expected number of infections $\inf_{\mathrm{T}}(G, \mathrm{S}, \tau_{\mathrm{d}}, \{v \mid g(v) \in \mathcal{Q}^*\})$ is minimized. Where, for a vertex v, let $g(v)$ denote the group it belongs to.

3.3 Example

Consider the example in Figure 3.1. There are 4 groups denoted by G1, G2, G3 and G4. G1 has one outflow to G4 and G2 has 3 outflows to G1, G3 and G4. Let v_s be the seed node at time step 0. Let the probability with which a node can infect its neighbouring nodes through the three pathways be 1. Let us also assume that Moore range of infection is 1 and delay equal to 0. At $t = 1$, node v_s infects all its neighbouring nodes (due to short-distance pathway/natural spread). At $t = 2$, the infection spreads through entire G1 (due to short distance human-mediated spread) and G4 (due to long-distance human-mediated spread). Let us now assume that the budget set by the policy maker to intervene at a group is 1. By intuition, if we intervene at G2 because it has the highest number of outflows (degree based intervention - a popular method of intervention) at $t = 2$ then the infection count does not reduce and it does not affect the infection spread in our example network by much. Hence, *a node with high number of outflows is not necessarily an ideal candidate for intervention.* In example 3.2 let us assume the same setting as the previous example. We demonstrate through this example that intervening at G1 at T=1 will prevent spread through G1 and rapid spread to G4 and the rest of the network in future time steps. Hence, *discovery of an infected node early can be beneficial.*

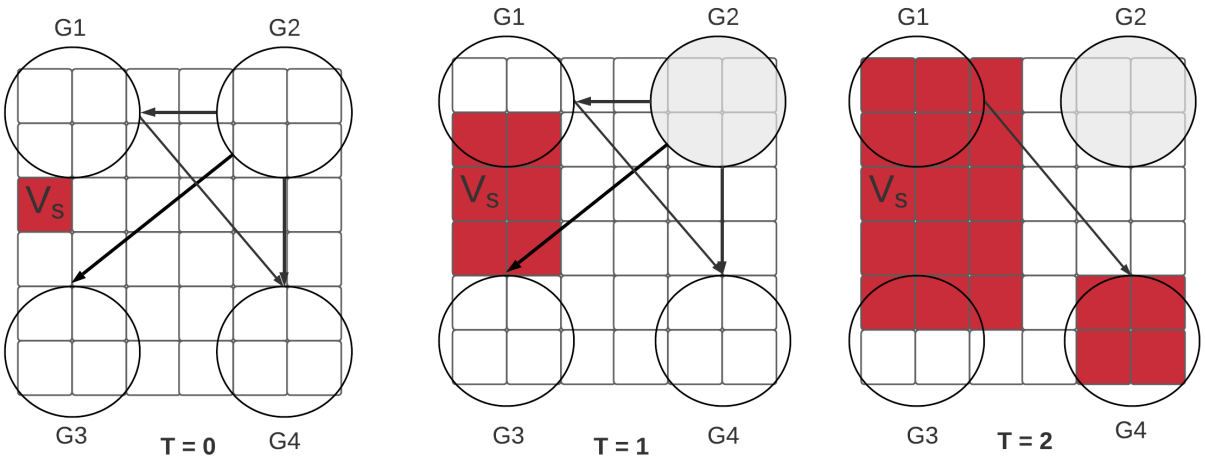

Figure 3.1: Example 1: A node with high number of outflows is not necessarily the ideal candidate for intervention.

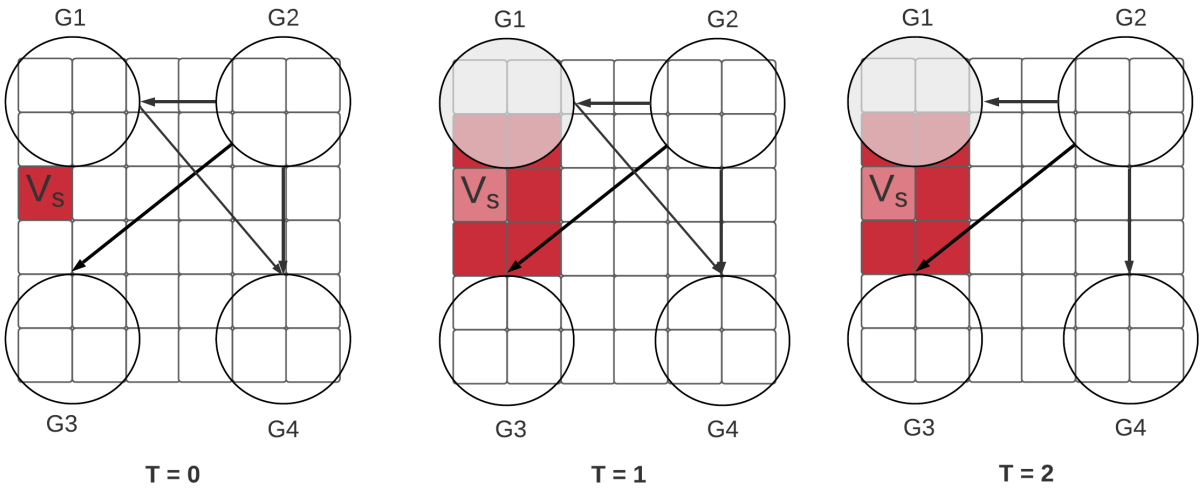

Figure 3.2: Example 2: Discovery of infected node early can be beneficial. Intervening at G1 at T=1 prevents further spread in G1 and G4.

Chapter 4

Time expanded network

In this section, we use the concept of time-expanded networks to represent multi-pathway model as an SEI process on another network called the time-expanded network. In Section 4.1 we define the time expanded network and in Section 4.2 we show that the SIR process on a network is "equivalent" to the SEI process on a time-expanded network with an example. This thesis contributes to the interfacing of the output of the simulator as a time expanded network with the intervention algorithm, which will be described later in this chapter.

4.1 The Time-expanded Network

Let $H_{\text{te}}(V_{\text{te}}, E_{\text{te}})$ be the time-expanded network corresponding to the SEI process on $G(V, E)$. 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. To incorporate the exposed state in SEI process, we have additional copies of a node at each time step, where for a latency period $\ell \geq 1$ and each time step i, we have ℓ additional copies given by $\{u_{i,0}, \ldots, u_{i,\ell-1}\}$ of u. The edge set E_{te} consists of exactly the following four types of edges which corresponds to different events in a SEI process:

- $(v_i, u_{i+1,0})$, $\forall (v, u) \in E$ with weight $w(v, u, i)$ (captures $\mathbf{S} \to \mathbf{E}$)
- $(u_{i,r}, u_{i,r+1})$ 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 $\mathbf{S} \to \mathbf{I}$: $(v_i, u_{i+1,0})$ with weight $w(v, u, i)$. An example of the time-expanded network is shown in Figure 4.2. This is the time-expanded network of the graph G ($a \to b \to c$) shown in Figure 4.1. Here the latency period is 2. The subscripts of each node denote the time-step copy of the node. For example 'a_0', the 0 indicates the copy of the node at time-step 0. The subscripts of the black nodes indicate the time-step and the latency count. For example $a_{0,0}$, the first 0 indicates the node a at time-step 0 and the second 0 in the subscript indicates the first latency period.

4.2 Equivalence

We state equivalence by showing that the SEI process on G is *equivalent* to the SIR process on the time-expanded network H_{te}. This is formally explained below. 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 $\mathcal{O}_G$ (example Figure 4.1) 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 edge-time tuples $((u, v), t)$ such that node u infects v at time t. Similarly, let $\mathcal{O}_{H_{\text{te}}}$ (example Figure 4.2) denote a disease outcome in the SIR model on H_{te}. We say $\mathcal{O}_{H_{\text{te}}}$ is consistent with $\mathcal{O}_G$ if: (i) node u has $\sigma_G(u, 0) = \mathbf{I}$ in $\mathcal{O}_G$ (resp. $\sigma_G(u, 0) = \mathbf{S}$) $\iff$ node u_0 has $\sigma_{H_{\text{te}}}(u_0, 0) = \mathbf{I}$ (resp. $\sigma_{H_{\text{te}}}(u_0, 0) = \mathbf{S}$). (ii) node u has $\sigma_G(u, i) = \mathbf{X}$ in $\mathcal{O}_G$ $\iff$ node u_i has $\sigma_{H_{\text{te}}}(u_i, i) = \mathbf{X}$ in $\mathcal{O}_{H_{\text{te}}}$ for $\mathbf{X} \in \{\mathbf{S}, \mathbf{I}\}$. (iii) node u has $\sigma_G(u, i + r) = \mathbf{E}$ in $\mathcal{O}_G$ for the $(r + 1)^{th}$ time step (where $r \in [0, \ell - 1]$) of latency period $\iff$ node $u_{i,r}$ has $\sigma_{H_{\text{te}}}(u_{i,r}, i + r) = \mathbf{I}$. (iv) node u has $\sigma_G(u, i) = \mathbf{I}$ s.t. $\sigma_G(u, i - 1) = \mathbf{I}$ if and only if $\sigma_{H_{\text{te}}}(u_i, i) = \mathbf{R}$. (v) infection spreads on an edge $(u, v) \in \mathbf{E}$ at time $t = i$ in $\mathcal{O}_G$ $\iff$ node u_{i-1} infects node $v_{i,0}$ at $t = i$ in $\mathcal{O}_{H_{\text{te}}}$
Given $\mathcal{O}_G$, for a time t, let $\mathcal{O}_G(t)$ be a snapshot of $\mathcal{O}_G$ up to time step t. Similarly, $\mathcal{O}_{H_{\text{te}}}(t)$ is a snapshot of $\mathcal{O}_G$ for t time steps.

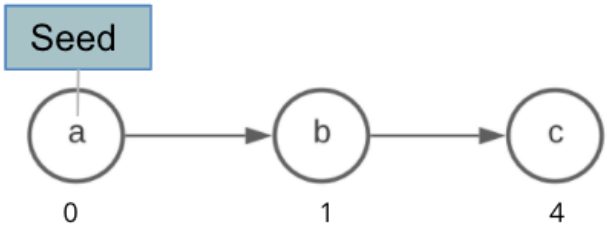

Figure 4.1: An example of network $\mathcal{O}_G$ with time-steps at which the nodes become infected.

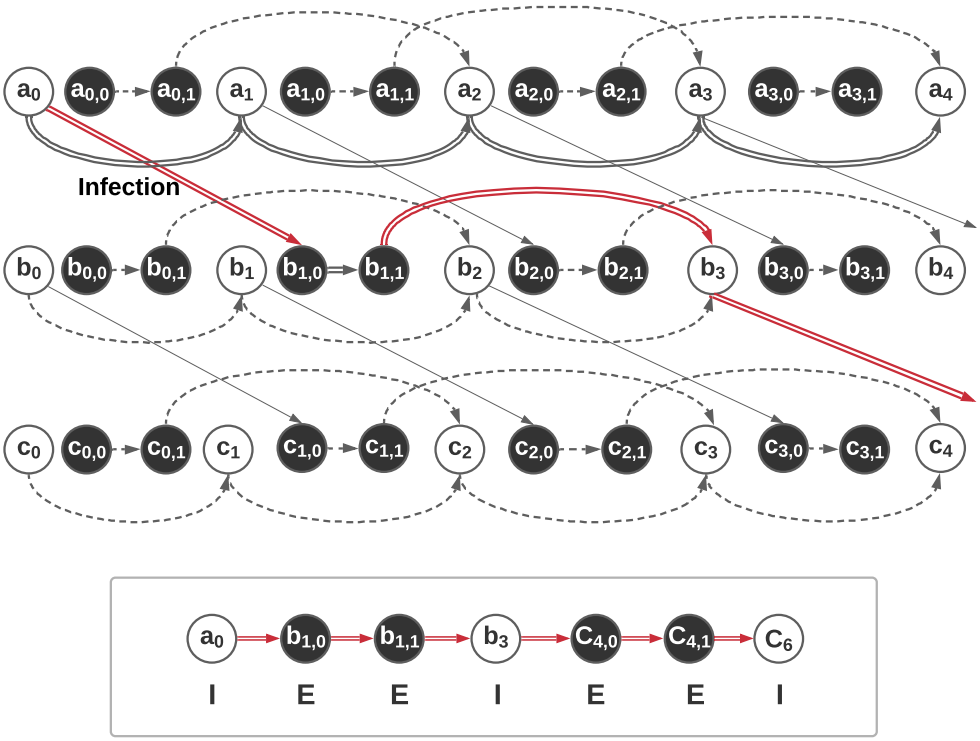

Figure 4.2: A snapshot of the time expanded graph $\mathcal{O}_{H_{\text{te}}}$ corresponding to $\mathcal{O}_G(t)$ for latency period $\ell = 2$. (Bottom) Shows the infection path with the state of each node at each time-step.

Theorem 2. *Consider the SEI 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 $H_{te}(V_{te}, E_{te})$ with the following initial conditions: At time 0, $\forall v \in V(G)$, $\Pr(\sigma_G(v, 0) = \mathbf{I}) = \Pr(\sigma_{H_{te}}(v_0, 0) = \mathbf{I})$, $\Pr(\sigma_G(v, 0) = \mathbf{E}) = 0$, $\Pr(\sigma_{H_{te}}(v_0, 0) = \mathbf{R}) = 0$ and the remaining nodes in H_{te} are in state $\mathbf{S}$. Then, for any outcome $\mathcal{O}_G$ and a consistent outcome $\mathcal{O}_{H_{te}}$, we have $\Pr[\mathcal{O}_G$ is the outcome in the SEI process on $G] = \Pr[\mathcal{O}_{H_{te}}$ is the outcome in the SIR process on $H_{te}]$.*

The proof was contributed by the collaborators.

In the above example (Figure 4.1 and Figure 4.2), the probability of node a being infected at time-step 0 in our original graph ($\mathcal{O}_G$) is equivalent to the probability of node a being infected at time-step 0 (a_0) in our time expanded graph $\mathcal{O}_{H_{\text{te}}}$. In $\mathcal{O}_G$ in the example (Figure 4.1), node a is infected at time-step 0, b is infected at time step 1 and c is infected at time-step 4. This infection spread is equivalent to the infection in $\mathcal{O}_{H_{\text{te}}}$ (Figure 4.2). In $\mathcal{O}_{H_{\text{te}}}$ each node is treated as a separate node. The path of infection is denoted by the red double arrows and the black double arrows show that the node is reachable from the source node of infection.

Node a_0 is infectious at time-step 0 since it is the seed node. Then $b_{1,0}$ and $b_{1,1}$ transition to the exposed (**E**) state at time step 1 and then transitions into infectious (**I**) state (b_3) at time-step 3. $C_{4,0}$ and $C_{4,1}$ then go to the exposed state (**E**) before being infectious (**I**) at time-step 6 (C_6) and this process continues. The nodes go to the exposed (**E**) state before being infectious for two time-steps because the latency period (ℓ) is equal to two in this example. Figure (4.1 bottom) shows the representation of the infection path sub-graph of $\mathcal{O}_{H_{\text{te}}}$ with each node's respective states shown. This is just an example of one such path of infection.

4.3 Implementation

The multi-pathway simulator's (Section 2.3) output was translated into a time-expanded network so that it could be interfaced with the intervention algorithm (Chapter 5). Each simulation outcome was mapped to its respective transitions. The events ($\mathbf{S} \rightarrow \mathbf{E}$,$\mathbf{E} \rightarrow \mathbf{E}$, $\mathbf{E} \rightarrow \mathbf{I}$, $\mathbf{I} \rightarrow \mathbf{I}$) were detected in each simulation outcome for each node depending on the transition states of the nodes. These 'live-edges' were collected at each sample and added to the output. The live-edge end points correspond to graph G, they were converted to edges of graph H_{te} by recording the following information:

- **Simulation step** or sample associated with the event.
- **Source node** corresponds to the node that affected the target node's state change.

- **Target node** gives the target node's information.

- **Source time-step and target time-step** is the time-step at which the source node and target nodes got its current states. This information is useful in terms of recording it as $\mathcal{O}_{H_{\text{te}}}$. For example, in Figure 4.2 the transition $a_{1,0}$ to $b_{1,0}$ corresponds to the transition from $\mathbf{S} \rightarrow \mathbf{E}$, source time-step and target time-step record the first subscript in the variables (a, b) which is the time-step information. Similarly in $\mathbf{E} \rightarrow \mathbf{E}$ which is ($b_{1,0}$ to $b_{1,1}$) records the first subscript information from one black node to another black node. $\mathbf{E} \rightarrow \mathbf{I}$ corresponds to $b_{1,1}$ to b_3 which again records the first subscript (or time-step) information from the black node to a white node transition in the example. $\mathbf{I} \rightarrow \mathbf{I}$ corresponds to a transition from a white node to another white node. In this way source and target time-steps retain information that corresponds to the information related to the first subscript of the nodes in $\mathcal{O}_{H_{\text{te}}}$.

- **Source index and target index** information correspond to the latency period information in the time-expanded graph or the second subscript information of each node in the example (Figure 4.2).

- **Pathway** through which the transition occurred (short-distance, short-distance human-mediated and long-distance human-mediated pathway).

- **Level 1 intervention** node information provides information on locality/group the source node (in the transition) belongs to. This column information is useful if we are intervening at the locality level.

- **Level 0 intervention** provides source node information. In $\mathbf{E} \rightarrow \mathbf{E}$, $\mathbf{E} \rightarrow \mathbf{I}$, $\mathbf{I} \rightarrow \mathbf{I}$ events it is assigned -1. In $\mathbf{S} \rightarrow \mathbf{E}$ events it is the source node of the transition. This column information is useful if we are intervening at the cell/node level.

Our simulator has two modes for producing the output, the first mode being level 1 mapping (at locality level) and the second mode being level 0 mapping (at node level). The second mode is required to interface the simulator with the SpreadBlocking algorithm. Hence, we need to

convert group–cell edges to cell–cell edges in the case of human-mediated dispersal pathways. This increases the time complexity of the simulator.

Chapter 5

Intervention Algorithms

In this section we will first discuss popular methods to compare our intervention algorithm against (Sections 5.1, 5.2, and 5.3) and then describe the Algorithm SpreadBlocking in Section 5.4.

5.1 Outflow-Based Heuristic

In this method, given the budget B, we choose top B groups with the highest outflows in the group-to-group network. In an unweighted graph this would correspond to a degree based heuristic. Traditionally, this method is used in many works [28, 27]. For example, Nopsa et al. [23] identify important nodes in grain networks, where nodes correspond to large storage cites. Preciado et al. [26] consider the problem of containing an epidemic outbreak in a weighted, directed contact network within a given budget and explore degree weighted based edges as a comparison to study their problem. McNitt et al. [21] also consider high outflow nodes for intervention. Typically, the rationale for this baseline is that higher the outflow from a node, the higher the probability that it infects its network neighbors and therefore, spreading the infection throughout the network. We implement this baseline by grouping the edges (locality to locality edges) by the source locality and calculating the total sum of the weights. This signifies the total outflows from each locality. The localities are then ranked from highest to lowest in outflows. The top B groups are chosen given budget B. The sum of these weights are not time-dependent and weren't summed based on month the weighted outflow took place. Also, we note that this is invariant to the seeding scenario as well.

5.2 Vulnerability of a Group

Vulnerability of a node corresponds to the probability that it will get infected when no interventions are in place subject to certain initial seeding. In most works, the initial seeding corresponds to randomly infecting a fixed number of nodes in the population. The empirical vulnerability is computed by observing every other node and recording the number of times each node is infected. The nodes infected the most number of times were chosen as the most vulnerable nodes. Vulnerability corresponds to the introduction of random seeding scenarios by which we can study the groups that are highly likely to be infected. Typically these random seeding scenarios are used to study vulnerability but in our case this randomness isn't very useful. Policy makers tend to have an idea of possible introduction scenarios. This can be through prior invasion data, monitoring country borders, transport hubs, etc. Hence, in our case the seeding scenarios have a prior, where cells that are source nodes (seed nodes) are set to a probability infection of 1. Also, we note that since the IAS spread considered in this work is an SEI process, as the time horizon T tends to infinity, all nodes reachable from the seed node will be infected. Since our interest is in the short-term predictions and control of IAS spread, we limited our time horizon to one year or 12 simulation time-steps. By running simulations for these many time-steps under given seeding scenarios, we compute the vulnerability of each cell to be the probability that it will be infected by T time-steps. For this baseline, we group the probability of infection for all the cells in each locality until the 12th time-step and rank them by the highest to lowest vulnerable groups. We then choose the top B groups, where B corresponds to budget.

5.3 Exhaustive Search

This baseline corresponds to considering all possible solutions of size B in the solution space. Since the number of solutions is exponential in B, this can be implemented only in cases where the number of groups is small. For larger networks and budgets, reproducing an exhaustive solution set can be quite computationally intensive and often times not feasible. We perform the exhaustive baseline method for networks with at most 10 groups.

5.4 The SpreadBlocking Algorithm

SpreadBlocking (Algorithm 1) is based on the sample average approximation (SAA) technique from stochastic optimization. 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. We use the following quantities and variables in the linear program, referred to as $LP_{\tau_{\text{d}}}$. This algorithm was designed and implemented by other members in our collaborative effort. In this thesis, we will describe it briefly outlining its validity and performance guarantees.

Table 5.1: Summary of notation used.

$S_{te} \subseteq V_{\text{te}}$	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)
B	#groups that can be intervened at time-step $\tau_{\text{d}} \geq 1$ (budget)

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. Let $\hat{\mathcal{Q}}_{opt} = \text{argmin}_{\mathcal{Q}''} I(\mathcal{Q}'')$ 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.

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 corresponding to SEI process on G (as described in Section 5)
3: Solve the linear program $LP_{\tau_{\mathrm{d}}}$ defined as follows:

$$
\begin{aligned}
(LP_{\tau_{\mathrm{d}}})\ \min \quad & \frac{1}{M}\sum_j \sum_u z^j_u \\
\forall i < \tau_{\mathrm{d}}, u_i, u_{i,r} \in \mathcal{R}(H^j) \quad : \quad & y^j_{u,i} = 1, y^j_{u,i,r} = 1 \text{ - - > (1)} \\
\forall u_i, u_{i,r} \in \mathcal{R}(H^j) \quad : \quad & z^j_u \geq y^j_{u,i},\ z^j_u \geq y^j_{u,i,r} \text{ - - > (2)} \\
\forall (v_{i-1}, u_{i,0}) \in E^j_{\mathrm{te}} : y^j_{u,i,0} \quad \geq \quad & y^j_{v,i-1} - x_{g(u),\tau_{\mathrm{d}}} \text{ - - > (3)} \\
\forall (u_{i,r}, u_{i,r+1}) \in E^j_{\mathrm{te}} : y^j_{u,i,r+1} \quad \geq \quad & y^j_{u,i,r} - x_{g(u),\tau_{\mathrm{d}}} \text{ - - > (4)} \\
\forall (u_{i-\ell,\ell-1}, u_i) \in E^j_{\mathrm{te}} : y^j_{u,i} \quad \geq \quad & y^j_{u,i-\ell,\ell-1} - x_{g(u),\tau_{\mathrm{d}}} \text{ - - > (5)} \\
\forall (u_{i-1}, u_i) \in E^j_{\mathrm{te}} : y^j_{u,i} \quad \geq \quad & y^j_{u,i-1} - x_{g(u),\tau_{\mathrm{d}}} \text{ - - > (6)} \\
\forall i \geq \tau_{\mathrm{d}}, \forall u_i, u_{i,r} \in \mathcal{R}(H^j) \quad : \quad & y^j_{u,i} \leq 1 - x_{g(u),\tau_{\mathrm{d}}} \text{ - - > (7)} \\
& y^j_{u,i,r} \leq 1 - x_{g(u),\tau_{\mathrm{d}}} \\
\sum_{Q_q \in \mathcal{Q}} x_{q,\tau_{\mathrm{d}}} \quad \leq \quad & B \\
\text{All variables} \quad \in \quad & [0,1]
\end{aligned}
$$

4: (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 $x_{q,\tau_{\mathrm{d}}} \geq \frac{1}{2k}$ where $|\mathcal{Q}| = k$.

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

5.5 Algorithm Approach

- **Step 1:** Construct the time-expanded network from G. Similar to the example shown in Figure 4.1.

- **Step 2:** Run M simulations of the SIR process and create the subgraphs corresponding to each simulation instance of the time-expanded network. This output is given by the simulator mentioned in Section 2.3. An example of H, can be seen in the Figure 5.2. Let us assume

each group consists of one node for simplicity. If one of the nodes is reachable from the source (a_0) then it is part of the H subgraph. For example, the double lines (in red) in the graph indicate that the node is reachable from the seed node (a_0) and is the path of infection spread. Similarly, we have M such H subgraphs. E_{te} are the live edges in simulation H_1 (j=1). An example of a live edge here is the edge between a_0 and $b_{1,0}$. In the example shown in 5.1 we only have a time horizon (T) of four. All the double lines in the graph are reachable nodes from source node a_0.

Any edge that is part of a path from the source to a node is reachable. Live edges are those that are part of a path made of only live edges from the source to a node. An example of an edge that is not live is a_1 to b_2.

- **Step 3:** Now that the concepts of live-edge and reachability are discussed, let us look into solving the linear program $LP_{\tau_{\text{d}}}$ in Step 3. The objective of the LP is to minimise $\frac{1}{M}\sum_j \sum_u z_u^j$. The z variables are indicators of whether the nodes they represent are reachable from the seed nodes. In Figure 5.1, if any node is reachable from the source it will be infected, and $z_{u,j} = 1$ if the nodes aren't reachable from the source $z_{u,j}$=0. All the nodes connected by the red double-lined path of the figure have values of $z_{u,j}$ being 1. For example, the variable corresponding to $b_{1,0}$, $z_{b,1}$ is equal to 1 since $b_{1,0}$ is reachable from a_0. Therefore, the sum in the objective corresponds to total infections across all simulations. Including the factor $1/M$ in the objective makes this expression the sum of empirical probabilities of each node being infected.

  Constraint 1: If the u_i (the white nodes in the example) are reachable from the source and $u_{i,r}$ (the black nodes in the example are reachable from the source) in the sample H, their values are set to 1. If at least one of u_i or $u_{i,r}$ are infected then we say that u is infected in the sample.

  Constraint 2: For every time-step less than τ_{d}, any node reachable from the source will be infected. Suppose we are intervening at time-step 3 ($\tau_{\text{d}} = 3$). If any node appears before that and it is reachable from the source then it is infected. Any black node or white

node (Figure 5.1) in a simulation is going to be infected if it is reachable from the source and if the i (white node) and/or $i + r$ (black node) value is less than the intervention time. If $b_{2,0}$ is infected, and then vaccinated before time-step 4 then b_4 is not taken into consideration because $b_{2,0}$ is already vaccinated by then. If node b_2 is infected then everything in the path b_0, b_1, b_2 before it will be infected and hence the corresponding z variables will have value 1. If node b_2 is not infected, then the value is 0. However, there is a possibility that it can be infected by some other path at some point of time i then everything in its path can be 0/1. To summarize:

- If node v is intervened, then, v may or may not be infected.
- If v is not infected then v may or may not be infected (as it can be infected from another edge).
- There are four scenarios for any edge from u to v in sample H1:

  Case 1: u is infected, v is not intervened at.

  Case 2: u is infected, v is intervened at.

  Case 3: u is not infected, v is not intervened at

  Case 4: u is not infected, v is intervened at

Constraint 3: This constraint corresponds to the S to E process in the graph edge. Note that $x_{q,\tau_{\mathrm{d}}} = 1$ means if group $Q_q \in \mathcal{Q}$ is intervened at time-step τ_{d}.

Constraint 4: This constraint corresponds to the I to E process in the graph edge.

Constraint 5: This constraint corresponds to the E to I process in the graph edge.

Constraint 6: This constraint corresponds to the I to I process in the graph edge. Example: b_2 to b_3. If b_2 is infected then b_3 will also continue to be infected unless b_3 is intervened.

Constraint 7: When a group that u belongs to is intervened then it shouldn't be infected else it may or may not be infected in the future time-steps if the group that u belongs to hasn't been intervened at.

- **Step 4:** This constraint corresponds to the summed values being rounded to 0 or 1.

- **Step 5:** Returns the optimal solution set for intervention.

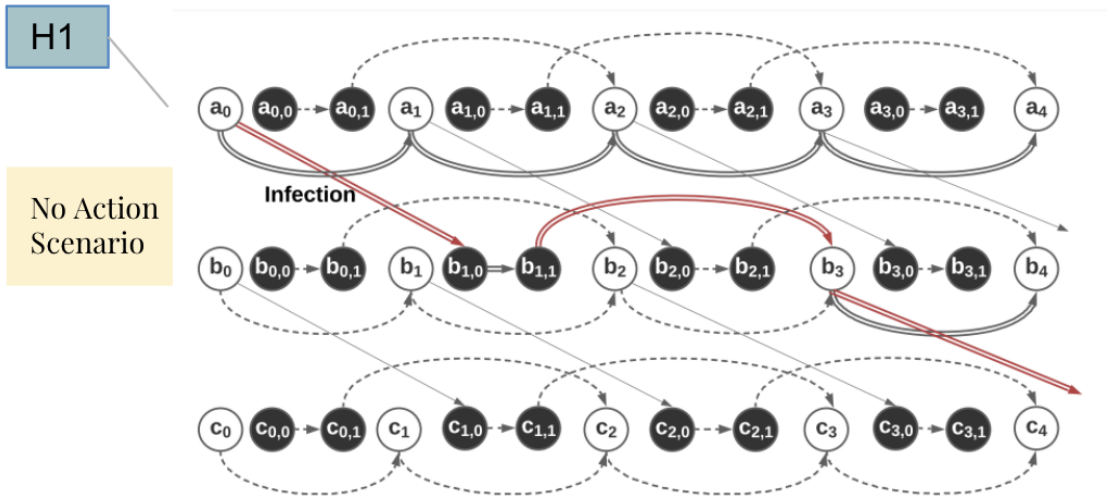

Figure 5.1: No action: A snapshot of the time-expanded graph $\mathcal{O}_{H_{\text{te}}}$ corresponding to $\mathcal{O}_G(t)$ for latency period $\ell = 2$.

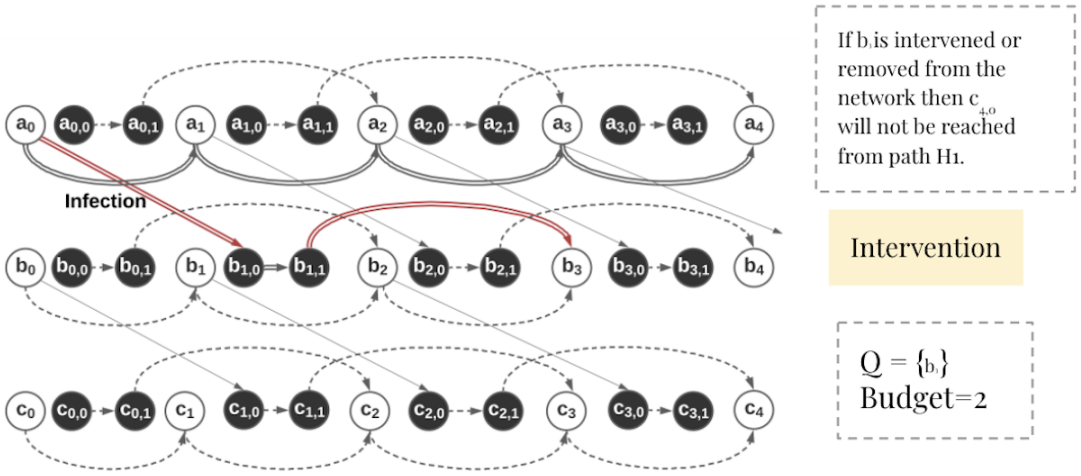

Figure 5.2: Interventions in the scenario depicted in Figure 5.1.

5.6 Guarantees of the algorithm

1. For any H^j, and any node $u_i \in V_{\text{te}}$ with $y^j_{u,i} < \frac{1}{2}$ (resp. for $u_{i,r} \in V_{\text{te}}$ 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 $H^j - V_{\text{te}}(\mathcal{Q}_{\text{SB}})$, where $\mathcal{Q}_{\text{SB}}$ is the intervention set computed by the algorithm.

2. Solution returned by our current approach is not optimal but we guarantee that it is closer to the optimal solution. The algorithms is at most a fixed constant times worse than that of the optimal solution. Informally, Theorem 3 states that with high probability $(1 - 1/k)$, the total accumulated infections obtained with our solution is at most 6 times the accumulated infections obtained by the optimal solution. To put it in naive terms, if a node belonging a group is saved (i.e. not infected) then the algorithm guarantees that the intervention set will contain groups to intervene at that ensure that all paths leading to that node will not be able to infect that node.

3. The number interventions we perform may violate the budget constraint. However, violation factor is at most the maximum number of distinct groups that a path from any two vertices contains.

Theorem 3. *Let $M \geq 24nk \log k$. Let $\mathcal{Q}_{SB}$ be the intervention set computed by* SpreadBlocking *algorithm. Then with probability $1 - \frac{1}{k}$, $\inf_{\text{T}}(V(\mathcal{Q}_{SB})) \leq 6 \inf_{\text{T}}(V(\mathcal{Q}^*))$ where $\mathcal{Q}^* \subseteq \mathcal{Q}$ is the optimal solution for the given instance of IAScontrol, and $|\mathcal{Q}_{SB}| \leq 2mB$. Here, m is the maximum number of distinct groups represented in a path between any two vertices.*

Chapter 6

Experiments and Results

We conducted experiments on several real-world networks and addressed the following questions:

1. **Assessment of the algorithm:** How does the algorithm fare w.r.t. popular heuristics for intervention?

2. **Effect of budget, intervention delay, model parameters and seeding scenarios:** How do solution sets as well as efficacy differ under different scenarios?

3. **Comparison with targeted intervention:** How do the results compare with the traditional targeted intervention case (each node belongs to a distinct group)?

4. **Structural properties of the solution set:** What are the attributes of groups that appear prominently in the solution set? What insights does this provide us into the structural and dynamical properties of the network that influence spread?

5. **Seeding Scenarios:** When multiple sources of infection are introduced/combination of seeding scenarios are implemented, how do the intervention solutions compare?

6.1 Data sets

Table 6.1 lists all networks incorporated in our analysis. These networks were constructed by McNitt et al. ([21]) and are publicly available. There are several versions of the networks depending

on the gravity model parameters. We used the values 2 and 500 for the distance function exponent and cut-off respectively. These are among the best model parameters obtained after calibration in their work (again available in their supplementary material). Each network has groups containing on an average 20–30 nodes capturing key urban and producing areas. For most countries, a significant portion of the nodes do not belong to any group. However, these nodes together cover less than 20% of the total production and population in each country.

Table 6.1: List of networks used for experiments and their attributes.

network	name	nodes	edges	groups	group 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

6.2 Experimental Setup

We modified the available multi-pathway model implementation [21] suitably to generate simulation outcomes in the appropriate format for the IAScontrol algorithm. The range of values for each parameter of the multi-pathway model were chosen to cover the best models with highest fit to ground truth. We used a full factorial design for each network with the pathway parameters and its values listed in Table 6.2.

Table 6.2: Multi-pathway Parameters.

parameter	values
α_s	[300,500]
α_ℓ	[0,0.2]
$\alpha_{\ell d}$	[50,200]
r_{M}	[1]
Start month	5
Number of Simulations	100
Time-steps (T)	24

6.2.1 Implementation and Computational Resources

We used Python 3.7 and its libraries to implement the multi-pathway simulator and Sqlite to store and analyze our data from our experiments. We used the Gurobi software [12] to implement and solve the GroupInt algorithm. All experiments in this thesis was performed on an HPC system that runs Linux $x86\_64$ operating system with a memory of 100GB.

6.3 Performance Evaluation

We compared SpreadBlocking to the different heuristics defined earlier for increasing values of budget and intervention delay. In each case the average accumulated infections was used as the metric for evaluation. This is the sum of the probabilities of node being infected by time T, where T is the time horizon. The results for the networks are in Figures 6.1 and 6.2. We observe consistent superior performance of SpreadBlocking across networks and model parameters, budget and intervention delay. For small networks, we also see that SpreadBlocking performs close to exhaustive search algorithm (Figure 6.1, BD network). Since SpreadBlocking is a bi-criteria approximation algorithm, the solution provided can violate the budget constraints (Table 6.3). We observed this phenomenon in experiments as well. Therefore, to compare with the other methods, we considered the budget for which the solution was provided, not the budget that was provided in the input specification.

In Table 6.3, we compare the intervention benefit obtained with the solution for the relaxed ILP ($LP_{\tau_{\mathrm{d}}}$) of SpreadBlocking. The approximation factor with respective objective of IAScontrol is computed using the objective of $LP_{\tau_{\mathrm{d}}}$, which is a lower bound on the optimal. We note that SpreadBlocking has much better approximation guarantees in practice.

Table 6.3: Performance of SpreadBlocking algorithm. Budget violation is the ratio of budget used by SpreadBlocking to B, whereas "obj. approximation factor" is the ration of objective value of SpreadBlocking to that of the LP_{τ_d}. Each entry correspond to the {min, avg., max} values of budget violation (and obj. approximation factor) computed over all runs on the network.

network	budget violation	obj. approximation factor
BD	[1.00,1.57,5.00]	[1.00,1.03,1.47]
ID	[1.00,1.11,1.57]	[1.00,1.00,1.01]
PH	[1.00,1.57,3.00]	[1.00,1.00,1.05]
VN	[1.00,1.85,4.00]	[1.00,1.01,1.07]

6.4 Analysis of Solution Sets

6.4.1 Assessment of the Algorithm with Popular Baselines

We analyzed the solution sets obtained under model uncertainty for different (B, τ_d) pairs. We notice that the performance of SpreadBlocking proves to be consistently superior to the other baselines and comparable to the exhaustive baseline (BD). In Figure 6.1 for the Bangladesh network we observe that the SpreadBlocking algorithm performs almost as well as the exhaustive baseline for a budget of foue. As intervention delay increases some groups become more prominent for intervention and may contribute to reducing the overall number infections, hence at intervention delay of six for the BD network is not as close to the exhaustive intervention. But as budget increases the SpreadBlocking tends to do as well as the exhaustive intervention case. In the PH network we observe that the max outflow baseline is very close to the SpreadBlocking. This is because intervening at groups with higher weighted outflows prove more beneficial for the PH network. We also observe that the vulnerability baseline consistently performs poorly (next to the no intervention case) across all networks. Production areas which are very close to high-consumption localities (large urban areas) are particularly vulnerable. Because local production typically does not satisfy demands of such localities, they have high inflows from other production areas and possibly from other countries. As a result, these localities are quickly infected. Once introduced to such localities, farmer–market interactions (local human-mediated dispersal) can facilitate the introduction of the pest to nearby production regions where it can establish [21]. They may take awhile to establish (at

a higher delay) before really proliferating and infecting other nodes at a faster rate. Hence its role in the spread throughout the entire country is quite limited.

In the VN network we see that SpreadBlocking does significantly better than the other baselines at a lower budget, with higher budget scenarios the other baselines catch-up. For the ID network max outflow and SpreadBlocking do significantly better than the vulnerability and no intervention case.

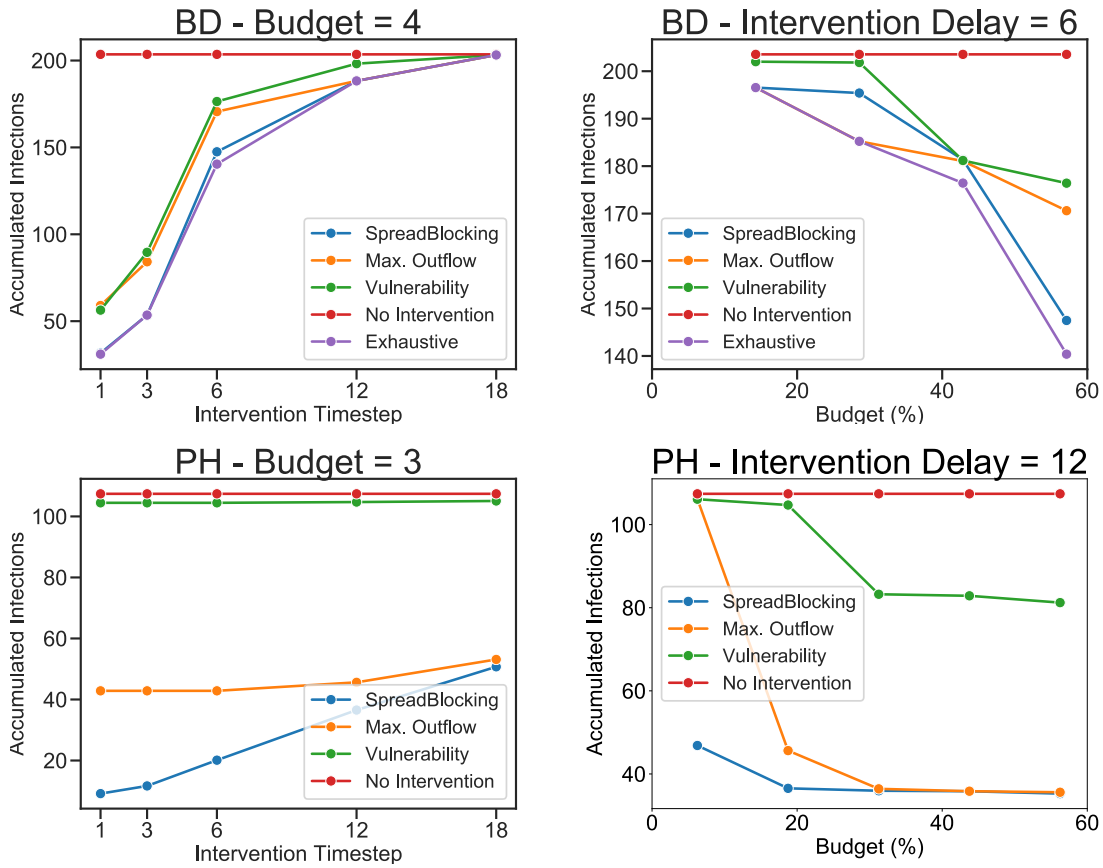

Figure 6.1: Comparison of algorithm with respect to budget and intervention delay for the parameter set: $\alpha_s \in 500$, $\alpha_\ell \in 0.2$, $\alpha_{\ell d} \in 200$, Moore range $r_{\mathrm{M}} = 1$, start month= 5, countries: BD, PH.

6.4.2 Effect of Budget, Intervention Delay, Model Parameters and Seeding Scenarios

For a given (B, τ_{d}) pair, we gathered all solutions from the algorithm (including those that violated the budget constraint) for different model parameters and ordered the groups by their frequency of

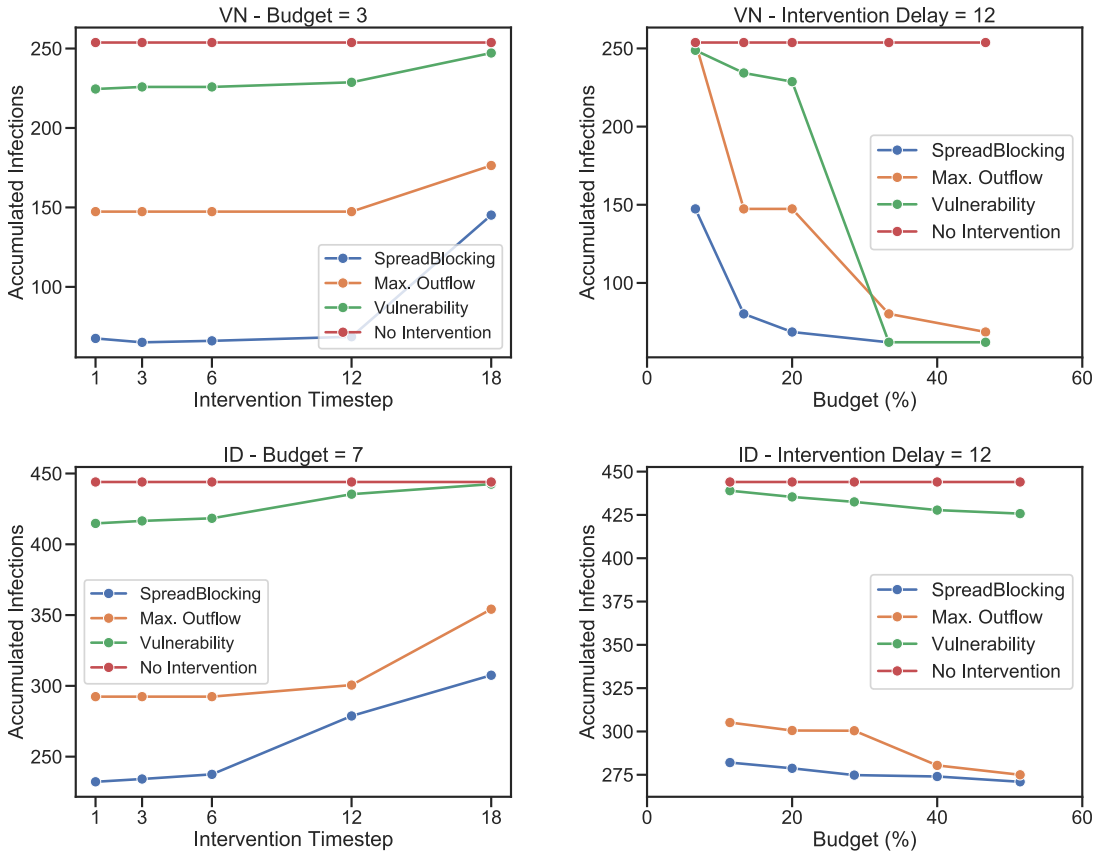

Figure 6.2: Comparison of algorithm with respect to budget and intervention delay for the parameter set: $\alpha_s \in 500$, $\alpha_\ell \in 0.2$, $\alpha_{\ell d} \in 200$, Moore range $r_{\mathrm{M}} = 1$, start month= 5, countries: VN, ID.

occurrence in the solution space. Figure 6.7 ranks the groups by the frequency of their occurrence in the solutions for a given τ_{d}. The results are network and seeding scenario dependent. In some networks, some groups are consistently picked in the solutions for increasing τ_{d}. These are major production areas which are infected in the beginning of the invasion. Since this is an SEI process, they continue to influence the spread. In Figure 6.7 (VN), we observe that Haiphong and Hanoi are consistently picked and are at the same rank irrespective of delay, this is because the Haiphong-Hanoi region are big sources of production (even surplus) in the northern Vietnam region [21]. This prominence is also seen in the contour plot in Figure 6.3, by the consistent dark blue shades in the map.

Similarly, in PH, Bandang and Bukittinggi have the second and third highest outflows and are also highly weighted outflows which is why they remain highly ranked even as time delay increases.

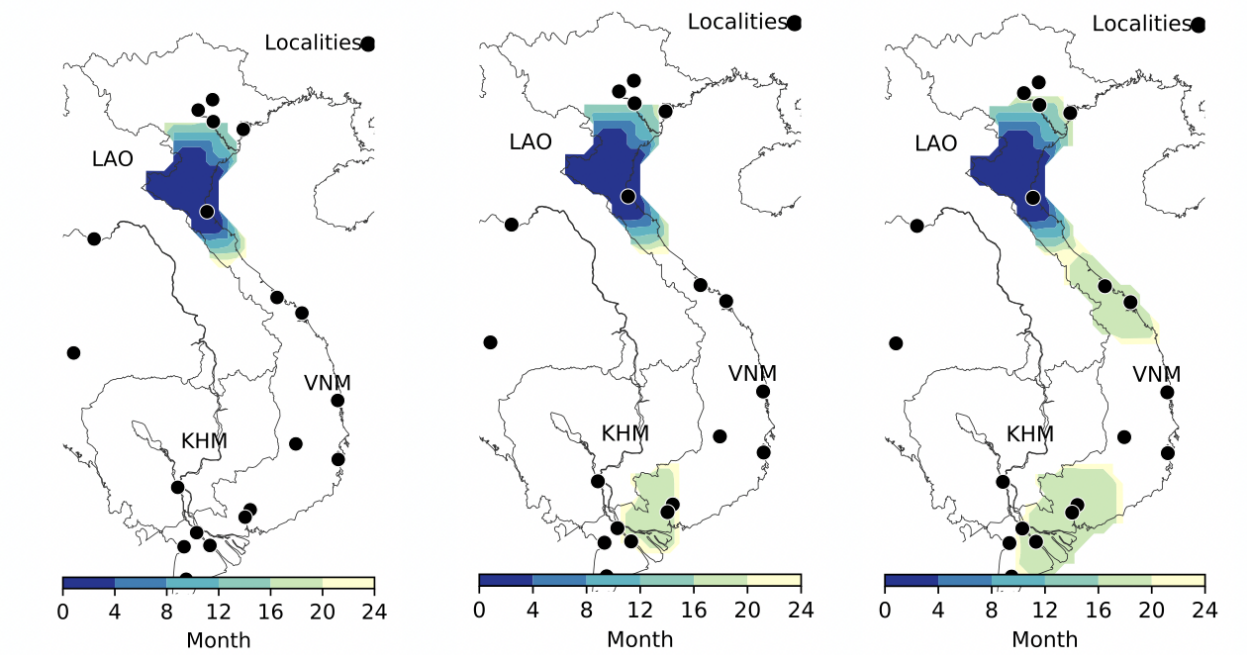

Figure 6.3: Study of varying model parameters. Model Parameters: (LHS) $\alpha_s \in 300$, $\alpha_\ell \in 0$, $\alpha_{\ell d} \in 50$, (MID) $\alpha_s \in 400$, $\alpha_\ell \in 0.1$, $\alpha_{\ell d} \in 100$, (RHS) $\alpha_s \in 500$, $\alpha_\ell \in 0.2$, $\alpha_{\ell d} \in 200$; Moore range $r_{\mathrm{M}} = 1$, start month= 5, country: VN.

Cagayan de Oro and Manila are major production centers [21], which is why they are extremely significant with a lower delay and lose significance as delay increases as infection spreads from these localities to other localities with high outflows.

Hence, *different localities become more prominent with increase in delay, as the infection spreads to cells close to them.* This indicates that as τ_{d} increases, the variability in the solutions increases, this can also be seen in BD, more exapmles of BD are explained in Section 6.4.2.3. Our results consistently indicate that *early discovery of the IAS and speed of intervention are critical to obtain stable intervention solutions under model uncertainty.* These results vary depending on the seeding scenario, location of major hubs and production sources and time delay.

From the contour map in VN (Figure 6.3), we also observe that *different models lead to different spread patterns in Vietnam.* With lower values for pathway parameters, the infection doesn't spread to south but with higher values, the infection starts spreading to the south. Once, the infection spreads to the south, localities geographically located very close to each other (almost in clusters) tend to infect each other very quickly, this pattern can also be seen in Indonesia in Figure 6.7-ID

where with increase in time delay there is high variation in the solution. This is due to the nodes being infected in the south rapidly once the infection spreads. Hence, it is better to intervene before the infection spreads to south in the early time delays.

6.4.2.1 Analysis of Start Month of Infection

With varying start months of infection spread, infection spread patterns change as well. From Figure 6.4 we can see that infection spreads very rapidly in start month 3, this is possibly due to the infection spreading during peak growing season. Hence, we can conclude that *timing of interventions is crucial depending on the peak growing season.* For example, intervening early is crucial at start month 3 due to peak production, growing and trade season, but not as crucial when the start month is 7.

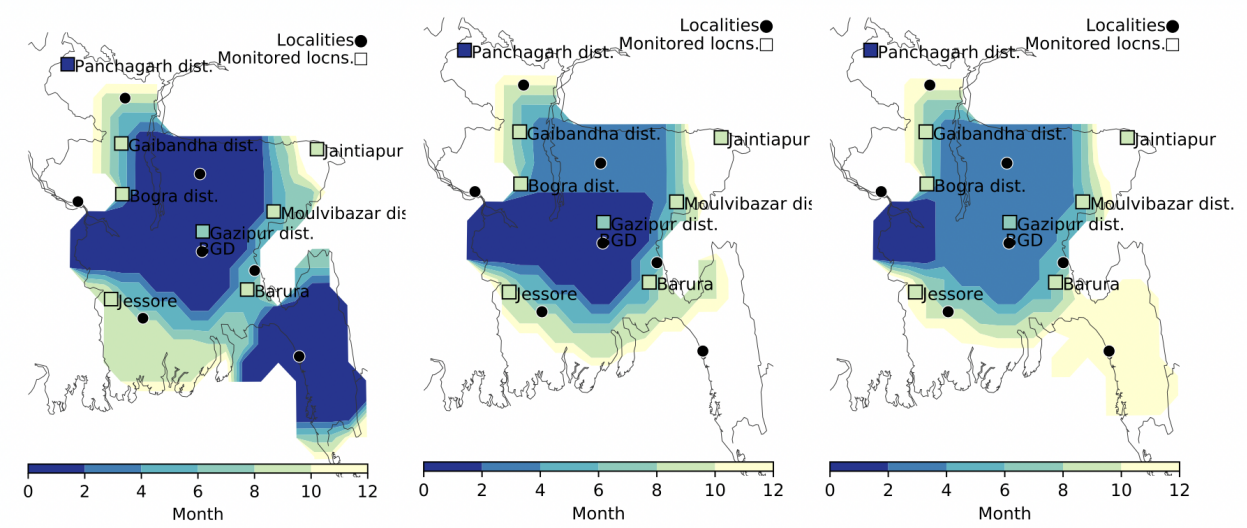

Figure 6.4: Study of intervention delays at start month (LHS) 3, (MID) 5, (RHS) 7. Model Parameters: (a)$\alpha_s \in 300$, $\alpha_\ell \in 0.1$, $\alpha_{\ell d} \in 100$, Moore range $r_{\mathrm{M}} = 1$.

6.4.2.2 Analysis of Inflows/Outflows of Localities

We also analyzed the frequency of occurrence with respect to node attributes inflow and outflow (Figure 6.5). These graphs study the structural properties of the groups that appear in the solution set. The Y-axis is total weighted accumulated outflows from a group and the X-axis is the total

weighted inflows pertaining to the group. From these graphs we can see that with change in intervention delay, certain groups become more prominent while other groups become less prominent. For example, in the first figure it may be more beneficial to intervene at groups in the seeded area and its nearby localities but as delay increases, (delay 18) other localities become much more prominent/beneficial to intervene at and the prominence of the groups switches (change in size of the blue dots in the image). The results of this analysis confirm the analysis made in Section 6.4.2.

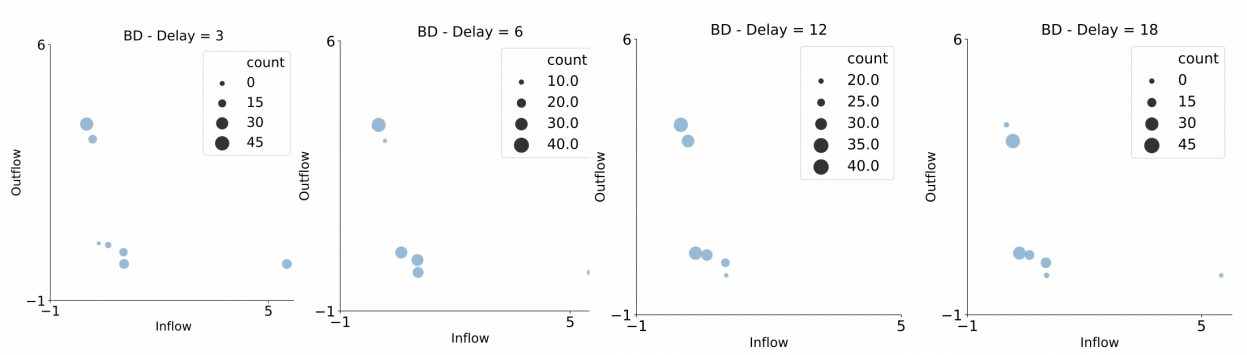

Figure 6.5: Analyzing the rank with respect to node attributes for BD.

6.4.2.3 Analysis of Seeding Scenarios

For the country-specific studies, the seeded location was decided based on the analysis in Mcnitt et al. ([21]) of possible points of entry through different pathways. Extra seeding scenario's for the countries Vietnam, Thailand and Bangladesh were added based on cases that were not seen/studied before. In each case, we chose those cells or groups as seeds which are at high risk of invasion. These include cells at the border of the neighboring infested country or groups which have high influx of travellers or trading activity with other countries. For example, in Bangladesh we seeded areas in the North (Rangpur) and in the East (Rajshahi) of the country. Figure 6.6 has the ranking plots and the contour plots (for the no intervention case) for both the seeding scenarios. The darker shades of blue in the contour plots indicate the spread of infection at earlier time delays and lighter as the time delay increases. Rajshahi and Rangpur are major production areas in Bangladesh with high weighted outflows to neighbouring localities. From figure 6.6 we can see that *ranking of groups is dependent on the seeding scenario and intervention delay.* In the

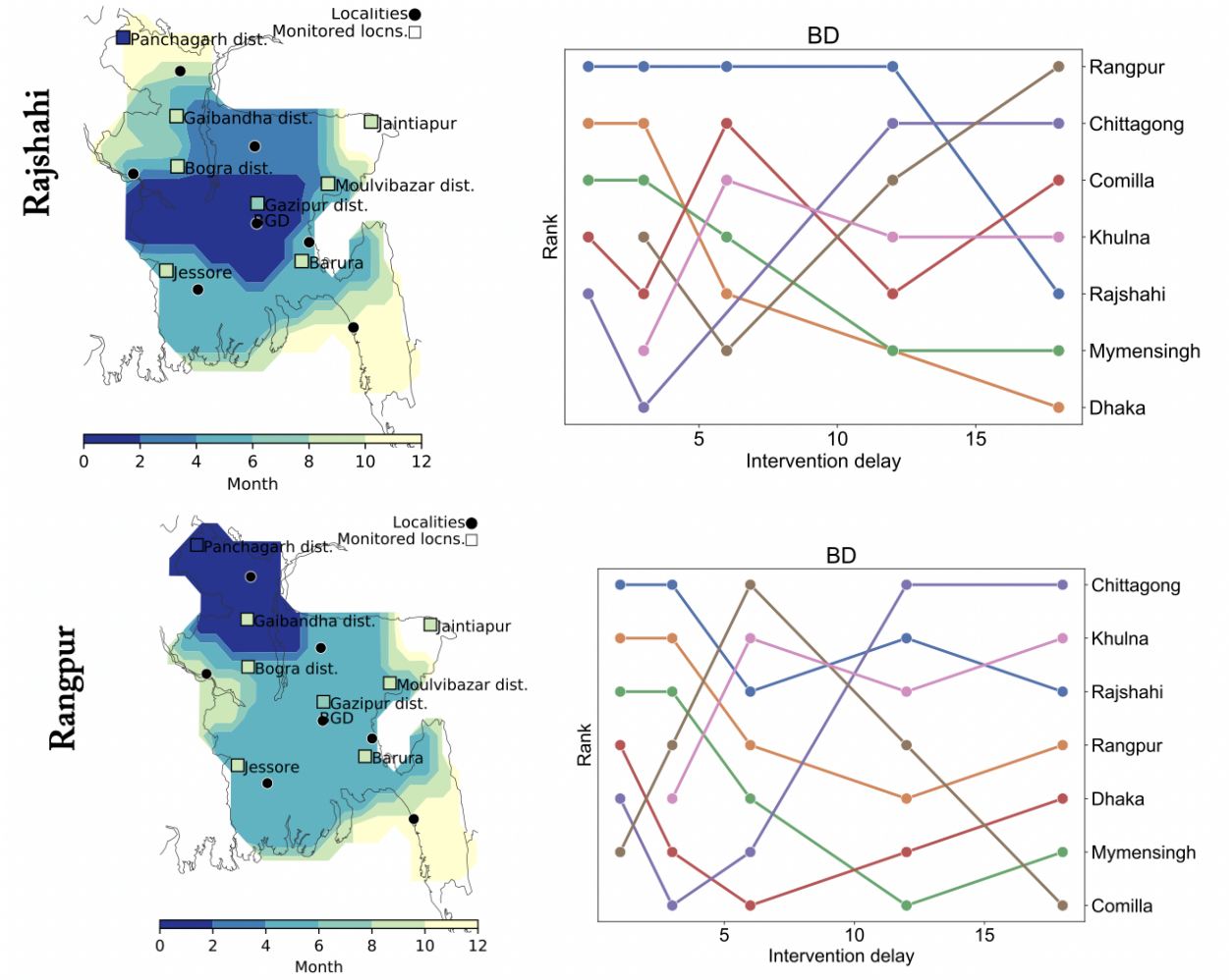

Figure 6.6: Performance rank of groups based on their frequency of occurrence in solution sets for a τ_d across various model parameters and seeding scenarios for BD alongside the no intervention map.

first seeding scenario, seven percent of the seeded nodes belong to Rajshahi, which is why in the initial intervention delays it remains to be prominent in the ranking. With the increase in delay the prominence of Rashahi reduces and Rangpur dramatically increases. This is because Rajshahi has now infected Rangpur and it becomes prominent because it is a large production area and likely to infect neighbouring localities fast, this could also be because there is evidence of vegetable flow from Rangpur region to other regions particularly during winter [21]. Hence, intervening at Rangpur at a later time delay would be more beneficial than intervening at Rajshahi even though it was prominently the infector in the early time steps. Similarly we can see the same in the Rangpur seeding scenario where majority of the nodes seeded belonged to Rangpur. But here we see the

opposite scenario where Rajshahi loses significance. We can thus say that, *with increasing delay, production areas affected later become important to intervene at.* Another interesting point is that Mymensing starts off ranking high in both seeding scenarios but also loses significance as time delay increases. This is due to Mymensingh being positioned very close to the seeded and production areas and hence one of the first few localities to get infected. With increase in delay other nodes become more important to intervene. *Seeded localities can lose significance over time but this is scenario specific to Bangladesh.*

6.4.3 Comparison with Targeted Intervention

The objective here is to assess the more practical and realistic group-scale intervention with a better performing but difficult-to-implement (in real life) individual-based interventions. In Figure 6.8, we compare the two for one country. Since each group on an average has around 20 nodes, for comparison sake, we have expressed the results for the group-scale intervention in terms of number of nodes intervened at (# groups $\times$ avg. nodes per group in the network). Across model parameters, we note that performance of group-scale interventions is comparable to individual-based interventions. This is because much of the production and population is centered around these groups. The number and size of localities is controlled by population threshold and locality radius parameters [21].

6.5 Computation Time and Scalability

SpreadBlocking algorithm scales well for all networks considered in the paper. However, for certain instances of BD network, it can take longer ($\approx$ 15 minutes). This could be due to the solution space of the instances of BD network. The main bottleneck is solving the linear program, but additional pruning techniques can help reduce the variables in the linear program, thereby speeding up the algorithm.

6.6 Ongoing work

Effect of number of input simulation instances on the solution quality and computation time. Currently we have a fixed number of simulation instances and we would like to further perform experiments to explore our solution set and impact of it on the results. In particular, we would like to know how many simulation instances are required for the solution set to stabilize with respect to number of instances. This would also depend on network structure and size.

Robust optimization under different seeding scenarios. Given a temporal edge-weighted directed graph $G(V, E)$, a partition of the vertex set into groups Q, and a collection of sets of source nodes $T \subseteq 2^V$, and finite set of source nodes $S \subseteq T$ as defined by the policy maker, the SEI diffusion process on G with transmission probabilities equal to edge weights, budget B, intervention delay τ_{d} and time horizon T. The *goal* is to find a set of groups $Q* \subseteq Q$ such that $|Q*| \leq B$ and maximum of the expected number of infections $\inf_{\mathrm{T}}(G, \mathrm{S}, s, \tau_{\mathrm{d}}, v | g(v) \in Q*)$ across the seeding scenarios is minimised.

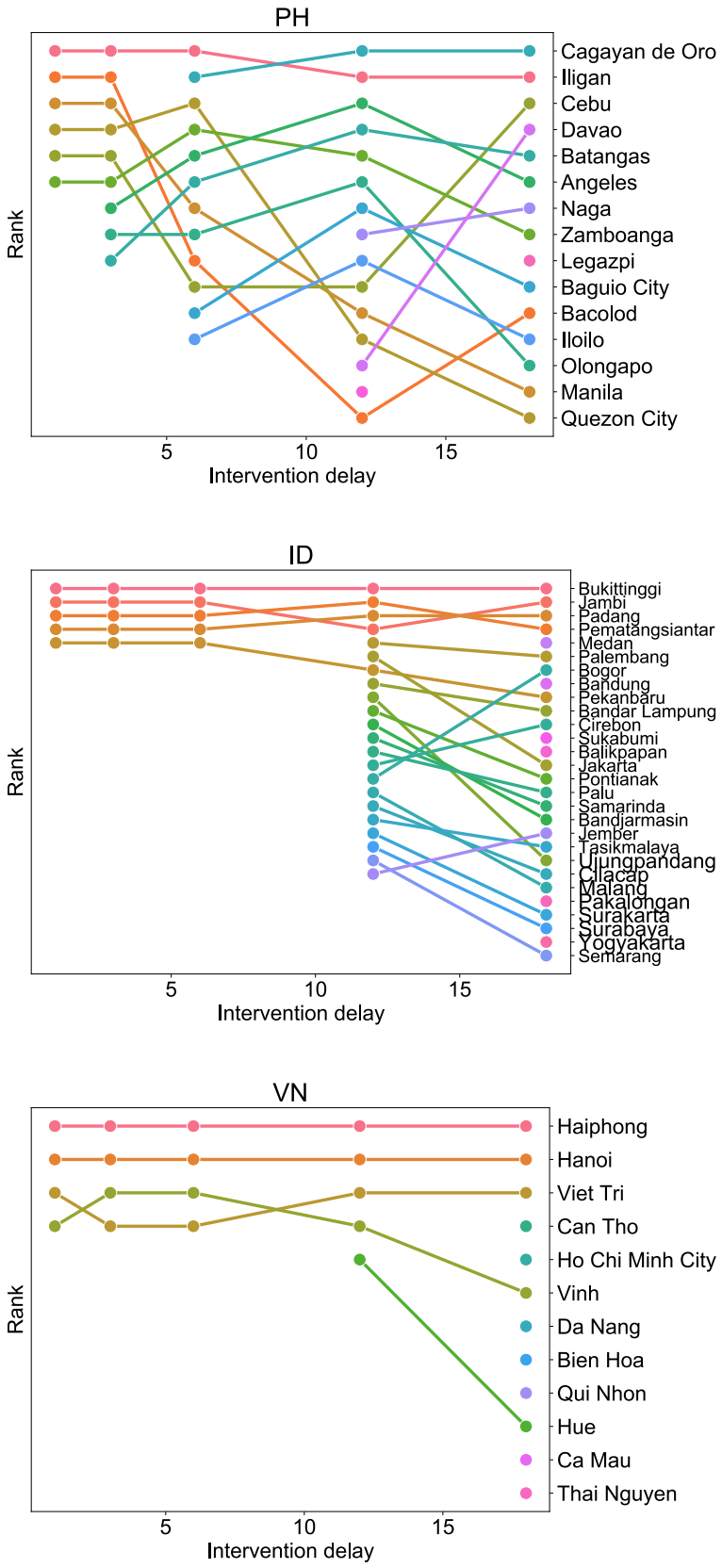

Figure 6.7: Performance rank of groups based on their frequency of occurrence in solution sets for a τ_{d} across various model parameters, countries: PH, ID, VN.

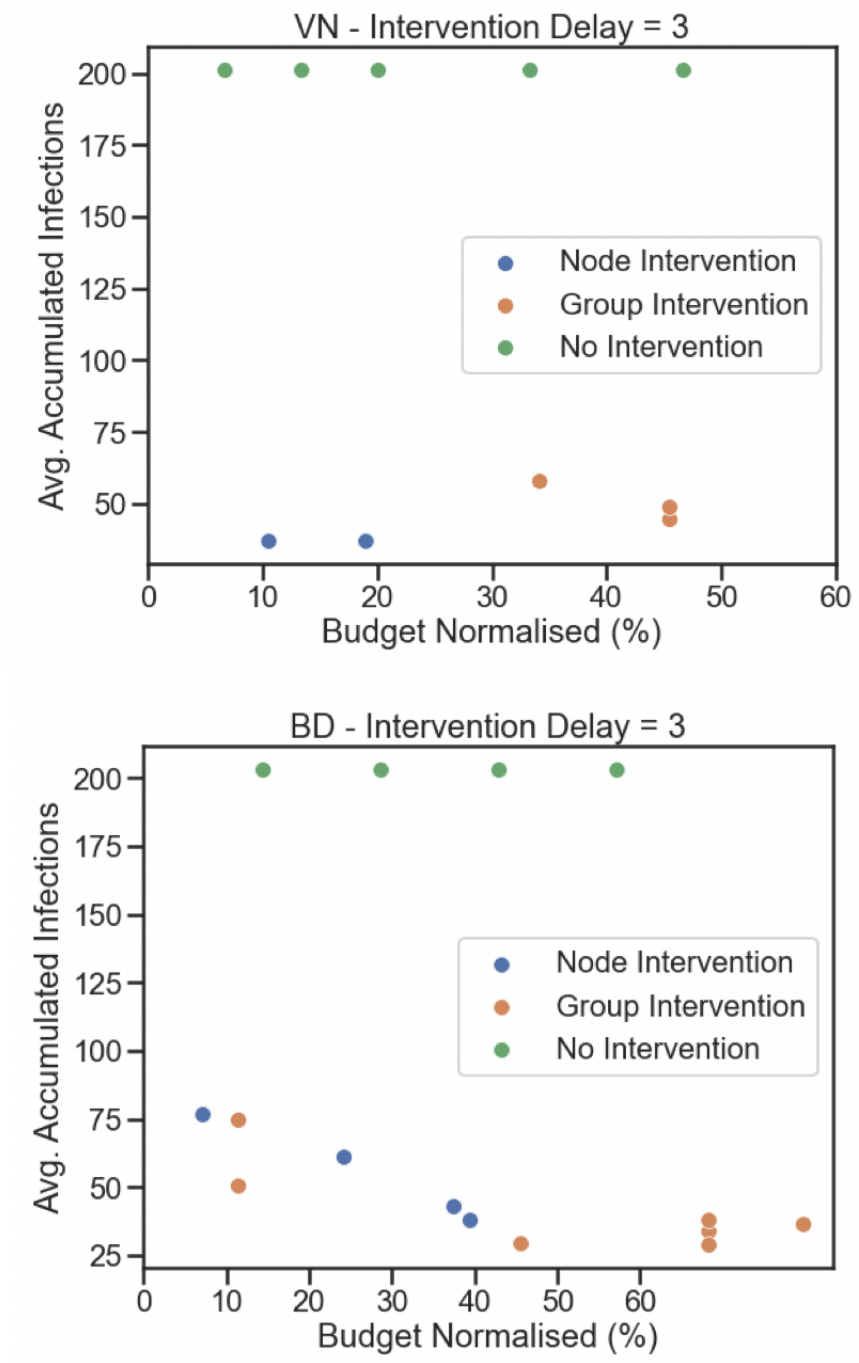

Figure 6.8: Comparison of group-based and individual-based interventions for the parameter set: $\alpha_s \in 300$, $\alpha_\ell \in 0.2$, $\alpha_{\ell d} \in 50$, Moore range $r_{\mathrm{M}} = 1$, start month $= 5$.

Chapter 7

Discussion

7.1 Agent-Based Models in IAS Spread

Agent-based models (ABMs) are made up of autonomous, interacting computational objects called agents that are often located in space and time [14]. Agents can differ in terms of different properties, such as being similar or having several unique properties; there can be millions of agents or a small number of agents [2]. Depending on the context and modeling space, some agents may have rule-based behavior (e.g., backward induction on an extensive game form) [16] or more complex behaviours (e.g., based on heuristics derived from cognitive psychology or neuroscience). Agents in an ABM receive feedback from their environment and react by taking actions.

In the recent years, ABM-based approaches are being increasingly applied to model the spread of invasive alien species (IAS). They range in complexity from simple cellular automata models [11] to multi-model frameworks [6] that include network, phenology and bioeconomic models. Ercsey-Ravasz et al. [8] identify influential nodes in the international food trade network using a dynamic food flux model. Sutrave et al. [31] use a network model to compare several strategies for selecting optimal sentinel plots for monitoring pathogens. Xing et al. [34] evaluate global networks of cropland connectivity for key vegetatively propagated crops important for food security in the tropics. For each crop, potential movement between geographic location pairs is evaluated using a gravity model, with associated uncertainty quantification.

Here, we use the model of McNitt et al. [21], who develop a discrete-time SEI diffusion process over a multi-pathway spatial network to model the spread of the pest in the South-East-Asia region. Unlike SIR models, SEI models suppose that a susceptible individual first goes through a latent (exposed) period before becoming infectious. Many realistic diseases are modeled this way [18]. Although some insight has already been gained from adaptations of network-theoretic models, the real challenge is to understand the epidemiologically-important characteristics of real trade networks. The spread of these infectious diseases are quite complex to model [13] as they can be transmitted through multiple pathways such as trade, human-mediation, packages, etc. [21].

Modeling the spread of diseases through a generalised multi-pathway multi-scale network model has the advantage of representing the propagation of the spread of disease through "node" level (immediate transmission from one host crop to another) and through "group" level (trade or human mediation from one region to another). Mcnitt et al. [21] explicitly consider multiple pathways of introduction and spread for *T. absoluta*. Earlier modeling efforts for *T. absoluta* have only accounted for ecological aspects and self-mediated spread [7, 11].

Characteristics of diseases and pests may differ in infectivity, transmissibility, host crop distribution, and extent of community spread. But those characteristics can easily be integrated (parameterized) by controlling the model parameters and including appropriate spread kernels in our simulator.

7.2 Control in Epidemiological Models

Managing invasive species is a major challenge for society. In the case of newly established invaders, rapid action is key for a successful management. However, interventions are resource intensive and many times limited in availability. Therefore, typically, the goal is to optimally intervene within a given budget constraint (B). In addition, there is potentially a huge cost incurred when there is a delay in discovering a biological invasion (much like in the case of infectious diseases such as COVID-19). A question that arises (answered through our experiments) is what is the effect of budget, intervention delay, model parameters and seeding scenarios on the solution set? How does the solution set change as these parameters change? In the real world scenario, intervening/providing

vaccinations can be quite expensive and limited to resources, hence every aspect of the solution set needs to be studied carefully before making a decision to intervene.

A simple example of intervention is one in which the nodes with the highest degrees (largest numbers of neighbors) are removed one by one until the budget is exhausted [28]. We use a similar baseline to compare our algorithm (SpreadBlocking). Vaccination depending on susceptible size is another example. In [19], authors initialize node scores with their degree values, recalculate a specific immunized node's score based on its local knowledge, and then substitute the specific immunized node with its non-immunized higher-score neighbor. Another common strategy used is page rank. Authors in [30] use an individual's movement based vaccination (IMV) strategy, where individuals are vaccinated based on their movement behaviours. This strategy is used on an SIR model. There are also papers that focus on "edge" removal rather than "node" removal [24].

In ODE models, interventions can be computed optimally, e.g., Medlock and Galvani [22]. However, optimizing individual-based interventions in network SEIR models is much harder [4, 27, 29]. Wilder et al. [33] consider optimal interventions in a dynamic population under a continuous-time SIS model.

7.3 Group Based Interventions

Since targeted immunization of specific nodes is harder to implement, optimal strategies based on node level characteristics, such as the various methods described above, cannot easily be converted into implementable policies. For example, Sutrave et al. [31] identify important counties to monitor and intervene at. We note that the similar intervention policies are developed in the case of infectious disease epidemiology as well. For example, CDC vaccine policies are at the group level (e.g., based on demographics), even though this can result in sub-optimal solutions compared to node level intervention strategies.

We build on the work of Sambaturu et al.([29]), who use the SAA approach for the simple SIR model; specifically, we extend their approach to the IAS model by considering the process on a time-expanded network, and searching for group interventions. We note that group-scale vaccination has been studied in the context of infectious disease spread and other socio-technical phenomena,

e.g., Zhang et al. [35]. Even though the work of Zhang et al. considers node removals in SIR diffusion processes, there is a key difference in the problem formulation, which makes it hard to compare it with our approach. The budget in their work corresponds to total number of individuals who can be vaccinated, not number of groups. There, the objective is to find an optimal allocation of vaccines to each group, while in our case, the objective is to select the best groups to intervene. Once intervened, all individuals in our group will be vaccinated. Another important distinction is that Zhang et al. (like the degree-based algorithm) does not account for the seeding scenario. In a temporal graph setting, Gauvin et al. ([10]) rank sub-graphs using tensor decomposition, and thus identify important groups to intervene at.

Through our experiments we find that group-based interventions are comparable to individual-based interventions. This is partly because most of the production happens within a locality or in the vicinity of a locality (Figure S1 in [21]). This implies that much of the susceptible cells are near these localities or groups. Therefore, any solution set of individual-based intervention has a majority of cells belonging to groups. Therefore most of the spread will occur near the locality/groups. However, group-based interventions need not show the same efficacy if this is not the case. Secondly, we note that from Figure 6.8 that the node-based intervention is slightly inferior to group-based intervention in some cases. This is because of differences in the actual and empirical probabilities of infection and the fact that SpreadBlocking does not guarantee a solution that minimizes accumulated infections.

7.4 Discussion on Group Size and Number of Groups

Spread of infections from one locality to another can be due to many factors such as start month, pathway parameter, neighbouring localities, etc. In various epidemiological modeling scenarios groups are considered to be a certain section of the population dependent on factors such as age [1], temporal graphs of interaction or actions [10], etc. In modeling spread of diseases of the tomato leafminer pest, McNitt et al. [21] consider the boundaries of a group or locality to be modelled through existing data. The authors define a locality/groups as centers of consumption and production. From the perspective of consumption, the authors selected cities with population

greater than a certain population threshold in the entire study region. The number and size of localities is controlled by population threshold and locality radius parameters. The authors chose 250,000 as the threshold for the model with the main criterion for the choice being coverage of population and knowledge of major wholesale markets. Then major production centers were added if their population did not meet the threshold [21]. A natural question is how modeling decisions such as number of groups or size of the group are varied. Suppose the network has many more localities participating in the transmission of the pest. On the outset, it is attractive as it has the potential to provide intervention solutions at a higher spatial resolution. However, the main limitation in having a high-resolution model is the unavailability of data to calibrate and validate the model, which makes it harder to justify implementation of the given solution. A large number of groups also increases computation time while simulating this infection spread. At the same time, if the size of each group is increased, it would lead to lower accuracy in simulation results. Also, it might not be feasible to intervene at a group if it is too large (like a state or a province). Hence, careful consideration of the size of each group and total number of groups accurately is important.

Chapter 8

Summary and Future Work

In this work, we developed a simulation-based group-scale intervention algorithm SpreadBlocking for controlling the spread in a multi-pathway model. We applied it to study the spread of invasive species. Our results show superior performance under uncertainty in model parameters compared to popular baselines. Motivated by this, one possible direction of study is to design robust optimization algorithms that account for uncertainty in introduction scenarios. Also, multi-stage interventions are relevant for the IAS domain since in most scenarios, the country is not prepared for the invasion and therefore, limited resources are available for immediate control. Scalability to large networks with large number of groups is a challenge. Pruning techniques can be explored to reduce the number of variables and constraints in the integer linear program.

Bibliography

[1] Melissa K et al. Andrew. Age and frailty in covid-19 vaccine development. *The Lancet*, 396(11), 2020.

[2] Steven C. Bankes. Agent-based modeling: A revolution? *Proceedings of the National Academy of Sciences*, 99(suppl 3):7199–7200, 2002.

[3] Antonio Biondi, Raul Narciso C Guedes, Fang-Hao Wan, and Nicolas Desneux. Ecology, worldwide spread, and management of the invasive south american tomato pinworm, *tuta absoluta*: past, present, and future. *Annual Review of Entomology*, 63:239–258, 2018.

[4] Christian Borgs, Jennifer Chayes, Ayalvadi Ganesh, and Amin Saberi. How to distribute antidote to control epidemics. *Random Structures & Algorithms*, 37(2):204–222, 2010.

[5] Corey Bradshaw, Boris Leroy, Céline Bellard, David Roiz, Céline Albert, Alice Fournier, Morgane Barbet-Massin, Jean-Michel Salles, Frederic Simard, and Franck Courchamp. Massive yet grossly underestimated global costs of invasive insects. *Nature Communications*, 7, 10 2016.

[6] LR Carrasco, JD Mumford, A MacLeod, T Harwood, Giselher Grabenweger, AW Leach, JD Knight, and RHA Baker. Unveiling human-assisted dispersal mechanisms in invasive alien insects: integration of spatial stochastic simulation and phenology models. *Ecological Modelling*, 221(17):2068–2075, 2010.

[7] Nicolas Desneux, Eric Wajnberg, Kris Wyckhuys, Giovanni Burgio, Salvatore Arpaia, Consuelo Narváez-Vasquez, Joel Gonzalez-Cabrera, Diana Ruescas, Elisabeth Tabone, Jacques Frandon, Jeannine Pizzol, Christine Poncet, Tomas Cabello, and Alberto Urbaneja. Biological invasion

of european tomato crops by *Tuta absoluta*: Ecology, geographic expansion and prospects for biological control. *Journal of Pest Science*, 83:197–215, 08 2010.

[8] Mária Ercsey-Ravasz, Zoltán Toroczkai, Zoltán Lakner, and József Baranyi. Complexity of the international agro-food trade network and its impact on food safety. *PloS one*, 7(5):e37810, 2012.

[9] Joseph Ferrari, Evan Preisser, and Matthew Fitzpatrick. Modeling the spread of invasive species using dynamic network models. *Biological Invasions*, 16:949–960, 04 2014.

[10] Laetitia Gauvin, André Panisson, Alain Barrat, and Ciro Cattuto. Revealing latent factors of temporal networks for mesoscale intervention in epidemic spread. *arXiv preprint arXiv:1501.02758*, 2015.

[11] Ritter YA Guimapi, Samira A Mohamed, George O Okeyo, Frank T Ndjomatchoua, Sunday Ekesi, and Henri EZ Tonnang. Modeling the risk of invasion and spread of *Tuta absoluta* in Africa. *Ecological Complexity*, 28:77–93, 2016.

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

[13] Hans Heesterbeek, Roy M. Anderson, Viggo Andreasen, Shweta Bansal, Daniela De Angelis, Chris Dye, Ken T. D. Eames, W. John Edmunds, Simon D. W. Frost, Sebastian Funk, T. Deirdre Hollingsworth, Thomas House, Valerie Isham, Petra Klepac, Justin Lessler, James O. Lloyd-Smith, C. Jessica E. Metcalf, Denis Mollison, Lorenzo Pellis, Juliet R. C. Pulliam, Mick G. Roberts, and Cecile Viboud. Modeling infectious disease dynamics in the complex landscape of global health. *Science*, 347(6227), 2015.

[14] Miller JH Holland JH. Artificial adaptive agents in economic-theory. *American Economic Review*, 1991.

[15] Philip E Hulme. Trade, transport and trouble: managing invasive species pathways in an era of globalization. *Journal of Applied Ecology*, 46(1):10–18, 2009.

[16] Scott E. Page John H. Miller. *Complex Adaptive Systems: An Introduction to Computational Models of Social Life*. Princeton University Press, 2009.

[17] Brian Alan Johnson, André Derek Mader, Rajarshi Dasgupta, and Pankaj Kumar. Citizen science and invasive alien species: An analysis of citizen science initiatives using information and communications technology (ict) to collect invasive alien species observations. *Global Ecology and Conservation*, 21:e00812, 2020.

[18] Kwang Ik Kim and Zhigui Lin. Asymptotic behavior of an SEI epidemic model with diffusion. *Mathematical and Computer Modelling*, 47(11):1314–1322, 2008.

[19] Yang Liu, Yong Deng, and Bo Wei. Local immunization strategy based on the scores of nodes. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 26(1):013106, 2016.

[20] M. Marathe and A. Vullikanti. Computational epidemiology. *Communications of the ACM*, 56(7):88–96, 2013.

[21] 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.

[22] Jan Medlock and Alison P Galvani. Optimizing influenza vaccine distribution. *Science*, 325(5948):1705–1708, 2009.

[23] John F Hernandez Nopsa, Gregory J Daglish, David W Hagstrum, John F Leslie, Thomas W Phillips, Caterina Scoglio, Sara Thomas-Sharma, Gimme H Walter, and Karen A Garrett. Ecological networks in stored grain: Key postharvest nodes for emerging pests, pathogens, and mycotoxins. *BioScience*, page biv122, 2015.

[24] Jorge M. Pacheco, Sven Van Segbroeck, and Francisco C. Santos. *Disease Spreading in Time-Evolving Networked Communities*, pages 291–316. Springer Singapore, Singapore, 2017.

[25] David Pimentel, Rodolfo Zuniga, and Doug Morrison. Update on the environmental and economic costs associated with alien-invasive species in the United States. *Ecological economics*, 52(3):273–288, 2005.

[26] V. M. Preciado, M. Zargham, and D. Sun. A convex framework to control spreading processes in directed networks. In *2014 48th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6, 2014.

[27] Sudip Saha, Abhijin Adiga, B Aditya Prakash, and Anil Kumar S Vullikanti. Approximation algorithms for reducing the spectral radius to control epidemic spread. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 568–576. SIAM, 2015.

[28] Marcel Salathé and James H. Jones. Dynamics and control of diseases in networks with community structure. *PLOS Computational Biology*, 6(4):1–11, 04 2010.

[29] Prathyush Sambaturu, Bijaya Adhikari, B Aditya Prakash, Srinivasan Venkatramanan, and Anil Vullikanti. Designing effective and practical interventions to contain epidemics. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1187–1195, 2020.

[30] Md Shahzamal, Bernard Mans, Frank de Hoog, Dean Paini, and Raja Jurdak. Vaccination strategies on dynamic networks with indirect transmission links and limited contact information. *PLOS ONE*, 15(11):e0241612, Nov 2020.

[31] Sweta Sutrave, Caterina Scoglio, Scott A Isard, JM Shawn Hutchinson, and Karen A Garrett. Identifying highly connected counties compensates for resource limitations when evaluating national spread of an invasive pathogen. *PLoS One*, 7(6):e37793, 2012.

[32] Andrew J Tatem. The worldwide airline network and the dispersal of exotic species: 2007–2010. *Ecography*, 32(1):94–102, 2009.

[33] Bryan Wilder, Sze-Chuan Suen, and Milind Tambe. Preventing infectious disease in dynamic populations under uncertainty. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[34] Yanru Xing, John F Hernandez Nopsa, Kelsey F Andersen, Jorge L Andrade-Piedra, Fenton D Beed, Guy Blomme, Mónica Carvajal-Yepes, Danny L Coyne, Wilmer J Cuellar, Gregory A Forbes, Jan F Kreuze, Jürgen Kroschel, P Lava Kumar, James P Legg, Monica Parker, Elmar Schulte-Geldermann, Kalpana Sharma, and Karen A Garrett. Global Cropland Connectivity: A Risk Factor for Invasion and Saturation by Emerging Pathogens and Pests. *BioScience*, 70(9):744–758, 07 2020.

[35] Yao Zhang, Abhijin Adiga, Sudip Saha, Anil Vullikanti, and B Aditya Prakash. Near-optimal algorithms for controlling propagation at group scale on networks. *IEEE Transactions on Knowledge and Data Engineering*, 28(12):3339–3352, 2016.