Designing spatially heterogeneous strategies for control of virus spread

Y. Wan  S. Roy  A. Saberi

School of Electrical Engineering and Computer Science, Washington State University, USA
E-mail: ywan@eecs.wsu.edu

Abstract: The spread of a virus – whether in a human population, computer network or cell-to-cell – is closely tied to the spatial (graph) topology of the interactions among the possible infectives. The authors study the problem of allocating limited control resources (e.g. quarantine or recovery resources) in these networks in a way that exploits the topological structure, so as to maximise the speed at which the virus is eliminated. For both multi-group and contact-network models for spread, these problems can be abstracted to a particular decentralised control problem for which the goal is to minimise the dominant eigenvalue of a system matrix. Explicit solutions to these problems are provided, using eigenvalue sensitivity ideas together with constrained optimisation methods employing Lagrange multipliers. The proposed design method shows that the optimal strategy is to allocate resources so as to equalise the propagation impact of each network component, as best as possible within the constraints on the resource. Finally, we show that this decentralised control approach can provide significant advantage over a homogeneous control strategy, in the context of a model for SARS transmission in Hong Kong.

Notations

Epidemic model parameters

R_0  basic reproduction ratio
n  number of groups in a multi-group model; number of individuals in contact network model
β_{ji}  for multi-group model, the transmission coefficient from district i to j; for contact network model, the probability that virus spreads from node i spreads to j during one time step
N_i  population in district i in the multi-group model
λ_i(t)  infectiousness of district i
β_r  reference transmission coefficient
f_{ji}  corrective factor specifying transmission coefficient relative to reference
\tilde{f}_{ji}  nominal value for f_{ji} (i.e. before control is applied)
r_i  control variable: scales the transmission coefficients from district i

\xi_i  control variable: scales the transmission coefficients into district i
T_i  the average infectiousness duration of an individual in district i
\tilde{T}  nominal average infectiousness duration without control
\xi_i  control variable: scales the infectious duration in district i
p_i(t)  probability that node i is infected in the contact network model
δ_i  recovery rate of a note i

Key design problem parameters

G  topology matrix
K  gain matrix
D  additive gain matrix
λ_{max}  dominant eigenvalue and its associated right and left eigenvectors
ω_{max}  the optimised parameter values
1 Problem formulation and motivation

The significant impacts of epidemics in recent years highlight the need for controlling virus spread with limited resources [1, 2]. Here, we put forth the perspective that spatially-homogeneous control strategies can allow mitigation of virus spread with sparse resources. Thus, we pose the virus-spreading control problem as a constrained decentralised design task for a dynamic network model and give an analytical methodology to complete the design task. Our design method leads us to the insight that resources (whether vaccination, rapid detection capability, quarantines or other resources) should be allocated so as to equalise the propagation impact of each network component, as best as possible within the constraints on the resource. We apply our method to design spatially-homogeneous controls for Hong Kong’s 2003 severe acute respiratory syndrome (SARS) outbreak [3], which outperform the existing homogeneous controls considered in the literature.

Although our primary focus is on virus-spreading control, this work also constitutes a significant contribution to our ongoing efforts in decentralised controller design and its applications. Recently, researchers in fields such as autonomous vehicle control and sensor networking have recognised the need for decentralised algorithms/controllers that exploit a network’s topological structure (see, e.g. [4]; see also the review article [5]). Concurrently, systems biologists have recognised the significant role played by a network’s graph structure on associated dynamics, in both molecular and population biology (see the review article [6]). Further, in a subset of these efforts, control and/or identification tasks that are based on the network structure are of interest.

From a control-theory standpoint, the existence of stabilising decentralised controllers for networks can be checked using the seminal work of Wang and Davison [7], but the design of practical but high performance controllers remains difficult (whether by numerical methods or explicit analysis). In [8], we posed an optimal decentralised design problem for a simple class of network dynamics and developed a design tool using optimisation machinery together with eigenvalue sensitivity and graph-algebra notions. Our efforts here show that similar tools can be developed for a family of decentralised design problems including some constrained ones. Our results also highlight that particularly simple and structurally insightful design tools can be obtained for certain special classes of network topologies, such as ones with non-negative weights.

This paper is organised as follows. In the remainder of this section, we review two models for computer-virus and biological-virus spread, namely a multi-group model for spatially inhomogeneous populations (Section 1.2), and a contact network model for interactions of individuals (Section 1.3). In turn, we formulate several virus-spread control problems as decentralised design problems. In Section 2, we develop a methodology to solve these decentralised design problems in detail. Finally, in Section 3, we apply our method to synthesise spatially heterogeneous controllers for the SARS outbreak in Hong Kong.

1.1 Epidemic control: brief review

Mathematical epidemiology has a history of more than two centuries. One major focus of the mathematical work on epidemics is the characterisation of the basic reproduction ratio \( R_0 \) (defined as the average number of secondary infections produced during an infected individual’s infectious period, when the individual is introduced into a population where everyone is susceptible [9]). It is well known that for \( R_0 > 1 \) a disease can spread throughout the population and may eventually persist in equilibrium, whereas for \( R_0 < 1 \), the epidemic eventually terminates. The basic reproduction ratio can be computed from models and also found experimentally [3, 9–12].

Control can be viewed as reducing \( R_0 \), and hence stopping the spread of a virus. Common control methods include: (1) vaccination; (2) reduction of local contact rates; (3) shortening of the time between symptom appearance and hospitalisation including improved virus detection; (4) restriction on long-range movement; (5) isolation of symptomatic people and those in contact with them (quarantine) [3, 11, 12]. All these control schemes change one or more parameters of the epidemic model of interest. Hence, by analysing \( R_0 \) or simulating dynamics over a parameter range [3, 11], we can analyse the impact of different control schemes. For example, the strategy of vaccinating newborns in a heterogeneous population (assuming contact rates at only two levels) was studied in [9, 13], in which an age-related model was used.

In the remainder of this section, we give explicit formulations of control-design tasks in the context of two widely studied models.

1.2 Spatial control in a multi-group model: formulation

Spatial interaction structure is critical in epidemics (e.g. SARS [3, 14]). However, there is little work in the literature on designing controls (e.g. isolation and quarantine) that are specialised to the network structure, to optimally mitigate epidemic spread with limited resources. We believe such a systematic design of control parameters at different points in the network can provide guidance for effective epidemic control. Here we review spatially inhomogeneous models for epidemics, and so pose the optimal spread control problem.

Early epidemic modelling assumes homogeneous mixing, that is, any pair of individuals are assumed...
equally likely to interact (or equivalently, the strengths of the interactions are the same). In reality, populations are spatially heterogeneous, and in fact, this spatial structure (social interaction topology) of the population plays an important role in epidemic spread. Usually, multi-group models, also called meta-population models (models in which the population is composed of multiple interacting groups, which internally have homogeneous mixing), are used to represent the spatially inhomogeneous dynamics. Abundant work exists in the literature on constructing and analysing these multi-group models [3, 11, 15–26], for example, Keeling and Rohani [18] establishes the connection of coupling between groups with explicit movement patterns, [25, 26] study the effect of aviation traffic on global epidemics and [3, 11, 20–22] provide case studies of various diseases such as smallpox, measles, SARS and foot-and-mouth disease. Of particular interest to us, [27–29] show how to calculate the epidemic threshold in a heterogeneous population. Colizza et al. [27] calculates the threshold from a reaction–diffusion point of view. In [28, 29], the basic reproduction ratio $R_0$ is shown to be the dominant eigenvalue of the next generation operator, the elements of which are defined as the expected numbers of new infections within a group that are produced by one infective with another group during its infectious period. ($R_0$ is usually real since an interaction network structure is most often non-negative and irreducible. We limit ourselves to the real maximum eigenvalue case in our later design.) The next generation matrix has been used to calculate $R_0$ for various applications [3, 11]. Often, stochastic models for epidemic spread are also used because chance fluctuations can be large, especially in the early stages of an epidemic [3, 12].

Here, we study inhomogeneous (distributed) virus-spread control in a multi-group model for spatial propagation, in particular designing controls that optimally reduce $R_0$. Spatially inhomogeneous models have already been used to study epidemic control; however, little effort has been devoted to inhomogeneous control strategies. For instance, the outbreak of SARS in 2003 aroused a lot of interest in spatial modelling and control [3, 12, 30], because of the geographical patterns observed in the virus spread [3, 12, 14]. Of interest to us, Riley et al. [3] model SARS in Hong Kong using a stochastic multiple-group model, where each group corresponds to a (spatial) district in Hong Kong. The authors identify the basic reproduction ratio $R_0$ for the model and show how homogeneous (identical network-wide) control can be used to reduce $R_0$ to 1. We notice that the early work [17] studies an inhomogeneous control strategy for steady-state behaviour of persistent epidemics in an open population. Our work builds on [17] and introduces an optimal inhomogeneous control of the full epidemic dynamics. In Section 3, we show that our optimal control which exploits network structure can reduce $R_0$ further with the same amount of resource, or equivalently achieve $R_0 = 1$ with less resource.

Let us now formulate the multi-group epidemic model and associated control design problem. Specifically, we will first describe the model in generality and then consider the particular dependence of model parameters on control actions. We note that the model without controls is very similar to the one in [3] and can be viewed as a multi-group SIS model. We consider a multi-group model with $n$ groups, which we refer to as districts since we are primarily interested in the spatial spread of epidemics. We use the notation $N_i$ for the population of each district $i$. The individuals in each district are modelled as transmitting the disease through homogeneous mixing within the district, as well as interaction with individuals in other districts; we model this transmission across groups in an aggregate fashion using an effective coupling (rather than through explicit modelling of individuals’ trajectories); see [19, 31–34] for justification. As in the existing literature [3, 31], we find it convenient to define a reference transmission coefficient and then to define particular transmission coefficients within and between groups relative to this reference. Specifically, we use the notation $\beta$ for the reference transmission coefficient, which identifies a typical rate at which secondary infections would be produced if all contacts were by infectives (and incorporates both the effective contact rate and the average infectiousness of an infected individual [3]). Usually, the reference transmission coefficient is chosen as the transmission rate internal to a prototypical district and assuming that no control actions have yet been taken, although other references can equivalently be used. The actual transmission coefficient $\beta_{ij}$ between district $i$ and district $j$ — i.e. the average rate at which infections in district $i$ would be produced if all contacts with district $j$ were with infectives in that district — is given by $\beta_{ij} = f_{ij}\beta$, where $f_{ij}$ is a corrective factor that specifies the transmission coefficient relative to the reference and is a parameter amenable to control in our model (as we shall specify in the next paragraph). We stress that $f_{ij}$ and $\beta_{ij}$ are also defined within a district, that is, for $i = j$; we note that $f_{ii}$ may differ from 1 in our model, either because of intrinsic differences between the regions or because control actions have been taken. For two different districts ($i \neq j$), $f_{ij}$ captures the relative rate of inhomogeneous mixing as well as any controls that further reduce contact.

Finally, the average duration of the infectious period is denoted as $T_i$ and is also assumed amenable to control. This model can be analysed by tracking the infectiousness $\lambda_i(t)$ of each region $i$, that is, the average total infectiousness of the individuals in the region $i$ (see [3] for details). Briefly, it can be shown that the average rate at which each individual in region $i$ becomes infected is given by $\sum_{j=1}^{n} \beta_{ij}\lambda_i(t)/N_j$, and in turn that rate of change of
\( \lambda_i(\cdot) \) is given by \( \bar{N}_i T_i \sum_{j=1}^{n} \beta_{ij} \lambda_j(\cdot)/\bar{N}_j \) (assuming that the infected population is small compared with the total population). One thus recovers that the next-generation matrix ([3, 28, 29]) is

\[
A = \beta \text{diag}(T, \bar{N}) \times \begin{bmatrix}
  f_{11} & f_{21} & \cdots & f_{n1} \\
  f_{12} & f_{22} & \cdots & f_{n2} \\
  \vdots & \vdots & \ddots & \vdots \\
  f_{1n} & f_{2n} & \cdots & f_{nn}
\end{bmatrix} \left[ \text{diag}(\bar{N}) \right]^{-1}
\]

(1)

Now let us consider applying controls in the context of this model. Specifically, we assume that the corrective factors and infectious period durations have nominal values, namely \( \tilde{f}_{ij} \) and \( \bar{T} \). We consider three sorts of control that deviate from this nominal (see [35] for motivation of these controls in stopping virus spread). (1) We allow the change of the contact rate of individuals in district \( i \) by a factor of \( r_i \in [0, 1] \) which decreases the spread of virus both locally and to spatial neighbours, and hence is modelled as scaling the transmission coefficient or equivalently the corrective factor: \( f_{ij} = r_i \tilde{f}_{ij} \) for all \( i, j \) (including \( j = i \)). Note that \( r_i \) can be reduced by isolation of closely connected and fairly isolated groups such as a school/college or restriction on public assemblage. (2) We allow change of the contact rate of an individual from outside districts to a district \( i \) by a factor of \( c_i \). In this case, \( f_{ij} = c_i f_{ji} \), for all \( i, j \) such that \( i \neq j \). The external contact rate factor \( c_i \in [0, 1] \) can be reduced by prohibition of travel from another district to district \( i \) or similar isolation of arriving travellers for some days (i.e. for a period longer than the incubation period). (3) We allow the average duration of the infectious period of each district \( i \) or \( T_i \) to deviate from \( \bar{T} \) by a factor of \( t_i \), and hence, \( T_i = t_i \bar{T} \). The factor \( t_i \) can be reduced by shortening the time between symptom appearance and hospitalisation in district \( i \), for example, through faster detection of infected individuals. Control measures such as isolation of people who may have contacted an infected individual (i.e. isolation of a neighbourhood with infected people) or isolation of symptomatic people (possibly including some individuals that have false symptoms and are not infected), reduce both \( T_i \) and \( r_i \). The next generation matrix upon specification of the control parameters is

\[
A = \tilde{\beta} \bar{T} \text{diag}(t_i, r_i, \bar{N}_i) \times \left( \text{diag}(\tilde{f}_{ii}) + \text{diag}(c_i) \right) \times \left[ \text{diag}(\bar{N}_i) \right]^{-1}
\]

Let us now formally pose the controller design problem. In doing so, note that one very reasonable performance measure is the dominant eigenvalue of the next generation matrix, which represents the spread rate of the epidemic. Let us say we are interested in designing \( t_i \) and/or \( r_i \). Noting that we can write the next generation matrix as \( A = KG \), where the topology matrix is

\[
G = \tilde{\beta} \bar{T} \text{diag}(\tilde{f}_{ii}) + \beta \bar{T} \text{diag}(\bar{N}_i c_i) \times \left[ \text{diag}(\bar{N}_i) \right]^{-1}
\]

and the gain matrix is \( K = \text{diag}(t_i/r_i) \), we can view the design problem as that of finding a diagonal matrix \( K \) so as to minimise \( \lambda_{\text{max}}(KG) \) [where \( \lambda_{\text{max}}(\cdot) \) denotes the dominant eigenvalue of a matrix], subject to constraints that \( 0 \leq K_i \leq 1 \) and that the \( K_i \) in total exceeds a lower bound \( \Gamma \) (since much resource is needed to make \( K_i \) small). Here is a formal statement.

**Problem 1:** Design diagonal matrix \( K \) such that \( \lambda_{\text{max}}(KG) \) is minimised, where \( K \) is subject to the following constraints.

1. \( \text{tr}(K) = \sum_i K_i \geq \Gamma \),
2. \( 0 \leq K_i \leq 1 \) for all \( i \).

In the case that we restrict long-distance movement (i.e. movement between districts) and so design \( c_i \), we can write the next generation matrix as \( A = D + KG \), where

\[
G = \tilde{\beta} \bar{T} \text{diag}(t_i, r_i, \bar{N}_i) \times \left( \text{diag}(\tilde{f}_{ii}) + \text{diag}(c_i) \right) \times \left[ \text{diag}(\bar{N}_i) \right]^{-1}
\]

\[
K = \text{diag}(c_i) \text{ and } D = \tilde{\beta} \bar{T} \text{diag}(t_i, r_i, \tilde{f}_{ii}).
\]

In this case, we can formulate the design problem as follows.

**Problem 2:** Design diagonal matrix \( K \) such that \( \lambda_{\text{max}}(D + KG) \) is minimised, where \( K \) is subject to the constraints that

1. \( \text{tr}(K) \geq \Gamma \),
2. \( 0 \leq K_i \leq 1 \) for all \( i \).

**Remark 1:** We may alternately consider other constraints on \( K \). Two realistic ones are listed below. In this paper, we do not explore these cases in any depth, although our approach of finding an optimal solution can easily be generalised for these cases.
• We may consider a constraint based on resource cost increasing inversely with $K_i$, that is, $\sum 1/K_i \leq \Gamma$. More generally, any cost that is concave with respect to the $K_i$ can be assumed.

• Often, it is natural that the resource cost also scales with $N_i$, or otherwise differs from one district to another. Thus, we may wish to consider constraints of the form $\sum \alpha_i K_i \leq \Gamma$, where $\alpha_i$ is a district-specific scaling factor.

Remark 2: Our focus here, on minimisation of the basic reproductive ratio, is of course only one aspect of epidemic control. One might also consider the design of total epidemic size or duration or pursue design of the steady-state in an open system. Some of these problems are amenable to similar analysis, see, for example, our efforts in [36, 37].

1.3 Inhomogeneous control in a contact network model

Contact network models (also known as agent-based models or automaton models) for virus spread — those in which individuals’ infection states (or state probabilities) are tracked — have been used to model cell-to-cell spread of influenza [38], SARS propagation [39], and computer virus spread [40], among other applications. Contact network models are motivated by the observation that homogeneity usually does not exist in real populations, perhaps not even within small groups. Contact network models are thus appealing in that they can capture the specific network interactions among individuals. It is worth noting that, within the general framework of contact network models, considerable research is focused on special classes of network topologies (e.g. scale-free, small-world, correlated, mesh) [1, 41–45].

A contact network model defined for a general network topology was proposed in [46] (which is motivated by the computer virus application). This paper approximated the epidemic threshold (a threshold on the infection rate to curing rate ratio, above which epidemic occurs, i.e. such that $R_0 = 1$) as the inverse of the dominant eigenvalue of the network’s adjacency matrix, assuming that (at each discrete time step) an infected node infects its adjacent node with a common probability and is cured with a different common probability. However, because the interaction probabilities are identical, this paper does not provide us with insight into topology-based network design, which can potentially lower the network’s vulnerability to virus spread. Our work here seeks topology-exploiting designs, for example, problems of where in the contact network to place limited control resources.

Few works have studied heterogeneous network resource allocation for stopping virus spread. Pastor-Satorras and Vespignani [1] proposed a targeted immunisation strategy (a few nodes with the highest connectivity are immunised) for power-law networks and evaluated its performance using simulation. Wang et al. [2] concluded that selective immunisation (e.g. immunising the upper-level nodes in a tree-like topology or nodes with high connectivity) significantly reduces a network’s vulnerability to virus attack compared with random immunisation. However, this work is also built on simulation, and hence does not provide us with an immunisation strategy that meets a performance requirement or that must operate under particular rigid constraints.

We develop network resource allocation strategies that optimize spread-based performance requirements (e.g. epidemic diminishing rate, number of nodes affected and the total duration of the epidemic), with the motivation that such design will aid in defending networks against virus attacks. In our effort, the network parameters (e.g. the local curing rates and infection rates) from [46] have the flexibility of design. For example, providing a selected set of individuals/nodes with faster detection capabilities and treatment (or, in the case of computer viruses better virus scan softwares) can increase these nodes’ local curing rates. Similarly, providing antibiotics to individuals (equivalently, providing computers with strong firewalls) can safeguard these nodes from common viruses or at least lower the rate of infection from their neighbouring nodes. Each of these control actions is associated with a cost (e.g. financial cost, productivity loss). Thus, it is not realistic to immunise or provide real-time repair to every individual in a network. Instead, we must assign limited control resources to achieve the best performance. Our study indicates how resources can be allocated in a way that appropriately uses the network topology.

To pursue control, we build on the contact network model proposed in [46], with the motivation that this model has already been of interest in studying resource allocation. Our model is a generalisation of [46], in that we allow variation in local curing rates and infection rates throughout the network. As in [46], we model virus spread as a discrete-time dynamics defined on a directed graph. Each node ($i \in 1, \ldots, n$) represents an individual (node) in the network, which may either be infected or susceptible. Each directed edge represents a path along which a virus can spread from one node to another. The branch weight $\beta_{ij}$ represents the probability that virus originating from node $i$ spreads to node $j$ during a time step. Notice that $\beta_{ij}$ increases with the transmission rate from node $i$ to $j$ and the infectiousness of the virus and decreases as protection at node $j$ is increased (e.g. by giving antibiotics to humans or providing firewalls to computers). We set $\beta_{ij} = 0$ if node $j$ is not a neighbour to $i$ (node $i$ cannot transmit an infection to $j$) or node $j$’s protection prevents any infection. An infected node has probability $\delta_i$ to recover at a discrete time step.

We consider two possible control actions in our model. (1) We allow for control that makes a node $j$ less susceptible to
any virus spread. In this case, we assume that the nominal weights are scaled by a constant for all entering branches, that is, the weights become $K_j \beta_{ij}$, $K_j \in [0, 1]$. We note that decreasing $K_j$ from 1 is costly. (2) We allow control of the recovery rate $\delta_j$. We note that increasing the recovery rate is expensive, in that more medicine or quicker hospitalisation is needed (better virus removal programmes or quicker human intervention, respectively, for computer network applications).

Now let us analyse the network's dynamics. Denoting the probability that each node $i$ is infected at time $t$ as $p_i(t)$, we find that the probability the node is infected at time $t + 1$ is

$$p_i[t + 1] = \left(1 - \prod_{j \neq i} (1 - K_j \beta_{ij} p_j[t])\right) + (1 - \delta_i)p_i[t] \quad (3)$$

Assuming $K_j \beta_{ij} p_j[t] \ll 1 \forall i, j, t$ (which is accurate for small time steps), the quantity $1 - \prod_{j \neq i} (1 - K_j \beta_{ij} p_j[t])$ can be well approximated by $\sum_{j \neq i} K_j \beta_{ij} p_j[t]$, and thus we can linearise (3) to obtain the network dynamics

$$P[t + 1] = (D + KG)P[t] \quad (4)$$

where topology matrix

$$G = \begin{bmatrix}
0 & \beta_{12} & \ldots & \beta_{1n} \\
\beta_{21} & 0 & \ldots & \beta_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
\beta_{n1} & \beta_{n2} & \ldots & 0
\end{bmatrix}$$

additive gain matrix $D = \text{diag}(1 - \delta_i)$, gain matrix $K = \text{diag}(K_j)$ and $P[t] = \begin{bmatrix} p_1[t], p_2[t], \ldots, p_n[t] \end{bmatrix}^T$. In the case where only $\delta_i$ are being designed, we find it convenient to use the notation $P[t + 1] = (D + G)P[t]$, since the $K_j$ are fixed.

The diagonal matrices $D$ and $K$ are the ones we are able to design. Our aim is to design $D$ and $K$ so that the network structure best inhibits virus spread, and hence minimises epidemic size. This goal leads us to consider optimisation with respect to a performance measure. The one we consider here is the dominant eigenvalue of $D + KG$ [denoted as $\lambda_{\max}(D + KG)$], as we know that the dominant eigenvalue governs the growth/decay rate of infection.

The matrix $D$ contains the local recovery rate of each node. Increasing $\delta_i$ (or equivalently decreasing $D_i$) can speed up the elimination of a virus, but at higher resource cost. Therefore we aim to design $D$ so that the cost of control (i.e. sum of $\delta_i$) is under a limit, whereas the performance of the design is optimised. Similarly, matrix $K$ represents the virus protection strength for each node. Decreasing $K_j$ for more nodes can also speed up the elimination of a virus. We thus design $K$ so that the cost of control is less than a threshold (i.e. sum of $K_j$ is over a limit), while the performance of the design is optimised.

Let us pose these network design problems formally.

**Problem 3:** Design diagonal matrix $D$, such that $\lambda_{\max}(D + G)$ is minimised, while $D$ satisfies the constraints:
1. $\text{tr}(D) \geq \Gamma$,
2. $0 \leq D_i \leq 1$ for all $i$.

**Problem 4:** Design diagonal matrix $K$, such that $\lambda_{\max}(D + KG)$ is minimised, while $K$ satisfies the constraints:
1. $\text{tr}(K) \geq \Gamma$,
2. $0 \leq K_i \leq 1$ for all $i$.

## 2 Network design

In this section, we address the design problems formulated in Section 1, namely to design diagonal $D$ to minimise $\lambda_{\max}(D + G)$ and to design diagonal $K$ to minimise $\lambda_{\max}(KG)$ and $\lambda_{\max}(D + KG)$, subject to the described constraints. We give methods for finding the optimal resource allocations both for general topologies $G$ and for specific classes of topologies that are common in virus-spreading applications. Our methods turn out to be deeply connected to on-going research on the design of high-performance decentralised controllers for modern networks, so we begin by briefly discussing the connection. We then describe the solution to the $D + G$ case, since the full suite of results is easier to describe/interpret in this case. We further present the design for the design of $\lambda_{\max}(KG)$ and $\lambda_{\max}(D + KG)$, and finally briefly highlight some computational and implementation-related concerns.

### 2.1 Connection to decentralised control

Let us begin by explaining why the three design problems listed above are canonical decentralised controller design problems. Decentralised control – that is, the task of controlling a system with many components each of which has only partial ability to regulate the global dynamics – has been of interest for many years [47–49] and has application in diverse fields such as electric power system control and robotics. The decentralised control of systems comprising simple but highly limited agents that cooperate through sensing/communication has gained a special prominence in recent years, as networks have become ubiquitous and increasing coordination of the network components is critical to achieving desired tasks [4, 5, 7, 8, 50, 51]. Our problems fit within this paradigm for control of modern networks: the control actions in a region or for an individual only directly impact the epidemic spread locally and are constrained (mathematically, only alter the
system matrix through multiplication/addition of a constrained diagonal matrix) and yet must be used in coordination to stop spread globally. We stress that we have not taken a decentralised-control perspective simply for the sake of convenience of analysis or implementation: the design problems that we address are in their essence decentralised, in that control actions (e.g. vaccination) in regions only react to and impact local infected populations; this fundamental constraint is reflected in the diagonal structure of the designable matrices $D$ and $K$.

For modern network tasks such as ours, understanding the role played by the graph topology in permitting stabilisation and high-performance control has been of particular interest [4, 5, 8]. This graph-theoretic viewpoint has highlighted that, although the existence of stabilising decentralised controllers can be checked (see the seminal work of Wang and Davison [48]), the problem of designing high-performance decentralised controllers remains difficult by any means. Our efforts here contribute to the understanding of high-performance decentralised controller design [8, 51].

More formally, the decentralised controller design problem is as follows: for a network with $n$ channels or agents or components, controllers (rules for determining channel inputs or actuations from observations) must be developed, so that the entire network’s evolution meets performance and/or robustness requirements. Although a variety of controller forms and performance requirements are of interest, linear control schemes subject to constraints are especially common, and many performance measures are based on the modes (eigenvalues) governing the dynamics. In such cases, the controller design problem can be abstracted to a linear-algebraic design problem in which, due to decentralisation, the design parameters are contained in diagonal or block-diagonal matrices. We notice that the virus-spreading control problems introduced in Section 2 take this form, in that a linear but constrained decentralised mechanism is used to optimise an eigenvalue-based cost measure; as expected, this problem abstracts to a linear-algebraic one concerned with designing a diagonal gain matrix. It is worth noting that the particular linear algebraic design problems posed here arise in a range of applications, including in autonomous-vehicle control and numerical computation, among others [4, 5, 8, 50].

2.2 Designing $\lambda_{\text{max}}(D + G)$

We address the problem of designing diagonal $D$ such that $D + G$ has minimum dominant eigenvalue, subject to the constraints that $0 \leq D_i \leq L$ and $\tr(D)$ is lower-bounded. For convenience, we refer to an optimum $D$ as $D^*$, the dominant eigenvalue of $D^* + G$ as $\lambda_{\text{max}}^*$, and the corresponding left and right eigenvectors as $\omega_{\text{max}}^*$ and $\varphi_{\text{max}}^*$. Our design method is based on the observation that an optimised topology $D^* + G$ has a very special eigenstructure, based on which we can compute $D^*$ and find the optimal performance.

We begin with the structural result.

**Theorem 1:** Consider a matrix $D + G$, where $D$ is diagonal and $G$ is an $n \times n$ matrix. Consider any $D = D^*$ that minimises the dominant eigenvalue of $D + G$ subject to the constraints (1) $\sum D_i \geq \Gamma$ and (2) $D_i \in [0, L]$, and assuming the dominant eigenvalue of $D + G$ is real and non-repeated. (The theorem can be easily generalised to the case that $D + G$ has real, simple dominant eigenvalues.) The optimising $D^*$ and corresponding eigenvalue/eigenvectors $\lambda_{\text{max}}^*$, $\omega_{\text{max}}^*$ and $\varphi_{\text{max}}^*$ (normalised to unit length) satisfy one of the following two conditions.

1. $\sum D_i^* = \Gamma$. In this case, for each $i$, we either have $0 < D_i^* < L$ and $w_{\text{max}}^* \varphi_{\text{max}}^* = 1$, or we have $D_i^* = L$ or $D_i^* = 0$.

2. $\sum D_i^* > \Gamma$. In this case, for each $i$, we either have $0 < D_i^* < L$ and $w_{\text{max}}^* \varphi_{\text{max}}^* = 0$, or we have $D_i^* = L$ or $D_i^* = 0$.

**Proof:** This result follows from standard theorems on eigenvalue sensitivity [52], as well as theorems on constrained optimisation using Lagrange multipliers [53]. Let us denote $D = d_i^2$, since we require $D_i \geq 0$ for all $i$. The procedure for finding an optimum $D^*$ under constraint is to form the Lagrangian $L = \lambda_{\text{max}}(D + G) + \sum a_i(d_i^2 + m_i^2 - L) - C(\sum d_i^2 - n^2 - \Gamma)$ and set the derivatives of it with respect to all variables (namely $d_i$, $a_i$, $m_i$, $C$ and $n$) to 0 (here, $m_i$ and $n$ are slack variables to transform inequality constraints to equality constraints). This procedure leads to the equations below

$$
d_i^2(\omega_{\text{max}}^* \varphi_{\text{max}}^* + a_i^* - C^*) = 0
$$

$$
d_i^2 + m_i^2 = L
$$

$$
m_i a_i^* = 0
$$

$$
n^2 C^* = 0
$$

$$
\Gamma + n^2 = \sum d_i^2
$$

(5)

Note that the first equation above follows from the eigenvalue sensitivity formula. The two cases in the theorem thus follow automatically from consideration of the variables $n^2$ and $C^*$, one of which must be 0. Specifically, the case where $\sum D_i^* = \Gamma$ follows from setting $n^2$ to 0, while the case where $\sum D_i^* > \Gamma$ follows from setting $C^*$ to 0.

Theorem 1 states that an optimum $D^*$ can be either at or inside the constraint boundaries. When $D^*$ is at the boundary $\sum D_i^* = \Gamma$, each $D_i^*$ falls into one of the three categories: (1) $D_i^* = L$, (2) $D_i^* = 0$ and (3) $w_{\text{max}}^* \varphi_{\text{max}}^* = 1$. When $D^*$ is not at the boundary $\sum D_i^* = \Gamma$, each $D_i^*$ again is at 0 or $L$, or $w_{\text{max}}^* \varphi_{\text{max}}^* = 0$. 

Note that the second equation above follows from the eigenvalue sensitivity formula. The two cases in the theorem thus follow automatically from consideration of the variables $n^2$ and $C^*$, one of which must be 0. Specifically, the case where $\sum D_i^* = \Gamma$ follows from setting $n^2$ to 0, while the case where $\sum D_i^* > \Gamma$ follows from setting $C^*$ to 0. 

Theorem 1 states that an optimum $D^*$ can be either at or inside the constraint boundaries. When $D^*$ is at the boundary $\sum D_i^* = \Gamma$, each $D_i^*$ falls into one of the three categories: (1) $D_i^* = L$, (2) $D_i^* = 0$ and (3) $w_{\text{max}}^* \varphi_{\text{max}}^* = 1$. When $D^*$ is not at the boundary $\sum D_i^* = \Gamma$, each $D_i^*$ again is at 0 or $L$, or $w_{\text{max}}^* \varphi_{\text{max}}^* = 0$. 

Note that the second equation above follows from the eigenvalue sensitivity formula. The two cases in the theorem thus follow automatically from consideration of the variables $n^2$ and $C^*$, one of which must be 0. Specifically, the case where $\sum D_i^* = \Gamma$ follows from setting $n^2$ to 0, while the case where $\sum D_i^* > \Gamma$ follows from setting $C^*$ to 0.
We notice that, for any condition in one of the above forms, the number of equations and variables are equal, and so we obtain a possible optimal solution from these equations. Thus, we see that a finite search can be used in theory to find the optimal solution, among these possibilities. However, the number of possible optimal solutions (number of solutions which satisfy one set of equations of this type) grows exponentially with the dimension of the matrix $G$, and so the calculation will be very complicated for even moderate-sized $G$. In the rest of this section, we will show that when $G$ is specially structured – for example, non-negative or symmetric, we can develop more explicit (and hence easier-to-evaluate and interpret) expressions for an optimal solution.

In the following Theorem 2, we show that for a very broad class of topology matrices $G$, an optimising $D$ is one that uses maximum total resource, that is, one for which $\sum D_i = \Gamma$.

**Theorem 2:** Consider $D + G$, where $D$ is diagonal, and $G$ is an $n \times n$ matrix. Assume that the largest eigenvalue of $D + G$ is real and non-repeated for all $D$ such that (1) $\sum D_i \geq \Gamma$ and (2) $D_i \in [0, L]$. Any matrix $D'$ that minimises the dominant eigenvalue of $D + G$ subject to these constraints satisfies $\sum D'_i = \Gamma$, if the left and right eigenvectors of $D + G$ corresponding to the dominant eigenvalue have the same sign patterns for all $D$. Classes of matrices satisfying this condition include (1) irreducible non-negative matrices and (2) diagonally symmetrical matrices (matrices for which there exists diagonal $Q$ such that $Q^{-1}GQ$ is symmetric).

To ease the understanding of those readers not familiar with matrix theory, we note that irreducible matrix is one that cannot be transformed into block upper-triangular matrix by simultaneous row/column permutations. The associated digraph of an irreducible matrix is strongly connected, for example, a path exists between any two nodes in the digraph.

**Proof:** The condition that the left and right eigenvectors of the dominant eigenvalue have the same sign pattern implies the relationship that, for all $i$, $\omega_{\text{max}}/\omega_{\text{max},i} > 0$. Thus, according to the eigenvalue sensitivity theorem, the dominant eigenvalue decreases monotonically with the decrease of $D_i$ for all $i$, since $(\partial \lambda_{\text{max}}(D + G))/\partial D_i = \omega_{\text{max}}/\omega_{\text{max},i}$ is positive. Therefore, an optimal $D'$ is on the boundary $\sum D'_i = \Gamma$. From the Perron–Frobenius theorem, the dominant eigenvalue of any non-negative and irreducible matrix is real and non-repeated, and the left and right eigenvectors associated with the dominant eigenvalues are positive [54], and hence have the same sign pattern. For a diagonally symmetrical $G$, it is easy to check through a similarity transform that the eigenvalues are real and the left and right eigenvectors associated with any eigenvalue are identical related by a positive diagonal scaling and hence have the same sign pattern. Hence, the theorem is proved. 

Theorem 2 guarantees that an optimum $D'$ is located on the boundary $\sum D'_i = \Gamma$, whenever $G$ is an irreducible and non-negative square matrix, and hence simplifies the search for $D'$ when $G$ has this special structure. This simplification is relevant to our applications, because for both the multi-group and contact network models, $G$ is non-negative and (for meaningful interaction topologies) irreducible. In the illustrating example, $G$ is irreducible and non-negative, so we expect that an optimum $D'$ satisfies $\sum D'_i = \Gamma$. In this case, $D'$ can be found by searching only through the first set of possibilities in Theorem 1. This search is formulated in Theorem 3 for some matrices of this sort. Before that, in Lemma 1, we characterise the pattern of the eigenvector associated with the dominant eigenvector under constraints $D_i \in [0, L]$ and $\sum D_i \geq \Gamma$.

**Lemma 1:** Consider the matrix $D + G$, where $D$ is diagonal, and $G$ is an $n \times n$ irreducible non-negative symmetric matrix. Any matrix $D = D'$ minimises the dominant eigenvalue of $D + G$ subject to the constraints $\sum D_i \geq \Gamma$ and $D_i \in [0, L]$ if and only if the components of the eigenvector $\omega_{\text{max}}$ associated with the dominant eigenvalue of $D' + G$ has a special pattern. In particular, it is required that $\omega_{\text{max},i} < \omega_{\text{max},l} < \omega_{\text{max},j}$ for any $i, j, l$ such that $D_i = L$, $0 < D_j < L$ and $D_l = 0$, and $\omega_{\text{max},i}$ are identical for all $i$.

**Proof:** First let us show the necessity. Suppose $\omega_{\text{max}}$ is the eigenvector associated with the dominant eigenvalue of $D' + G$. The conclusion that $\omega_{\text{max},i}(\forall i, s.t. 0 < D_i < L)$ (where we have used the subscript notation to refer to the entries of $\omega_{\text{max}}$ for which $0 < D_i < L$) are the same directly follows from Theorem 1 and the symmetry and non-negativity of $G$. More specifically, when $G$ is a symmetric matrix, so is $D + G$, and thus $\omega_{\text{max},i} = \omega_{\text{max},i}$ for all $i$. Also, from Theorem 1, we know that the eigenvectors $\omega_{\text{max}}$ and $\omega_{\text{max}}^*$ of $D' + G$ satisfy $\omega_{\text{max},i} = \omega_{\text{max},i}$ and $\omega_{\text{max},i}^* = 1$ for each $i$ such that $0 < D_i < L$. Finally, from positivity of $G$, we see that $\omega_{\text{max}}$ has only positive entries. Combining, we recover that $\omega_{\text{max},i}(\forall i, s.t. 0 < D_i < L)$ are identical. Now we need to show that $\omega_{\text{max},i}(\forall i, s.t. D_i = 0) < \omega_{\text{max},i}(\forall i, s.t. 0 < D_i < L) < \omega_{\text{max},i}(\forall i, s.t. D_i = 0)$. Since $D'$ achieves the minimum dominant eigenvalue, decreasing $D_i$ for $i$ such that $D_i = L$ (making them less than $L$) or increasing $D_i$ for $i$ such that $D_i = 0$ (making them larger than $0$) while maintaining $\sum D_i = \Gamma$ should increase $\lambda_{\text{max}}$. Eigenvalue sensitivity naturally leads to the inequality of eigenvector components, since the derivative of $\lambda_{\text{max}}$ with respect to $D_i$ equals $\omega_{\text{max},i}^2$.

For the sufficient condition, we know that if $\omega_{\text{max},i}(\forall i, s.t. D_i = L) < \omega_{\text{max},i}(\forall i, s.t. 0 < D_i < L) < \omega_{\text{max},i}(\forall i, s.t. D_i = 0)$ and $\omega_{\text{max},i}(\forall i, s.t. 0 < D_i < L)$ are the same, the corresponding $D$ achieves a local minimum (from above). It follows from convexity (which can be proved easily using, for example, the Courant–Fisher theorem) that the local minimum is in fact global. 


doi: 10.1049/iet-syb:20070040

© The Institution of Engineering and Technology 2008
Lemma 1 presents the pattern of the eigenvectors associated with the minimised dominant eigenvalue of $D + G$. This allows us to check whether a solution $D$ is optimum by simply evaluating the dominant eigenvectors of $D + G$. Here we define the expression diagonalise($\psi$) as placing the $i$th entry of a vector $\psi$ as the $i$th diagonal entry in a diagonal matrix, for $i = 1, \ldots, n$.

**Theorem 3:** Consider a topology matrix $G$ that is non-negative, irreducible and diagonally symmetrisable. A matrix $D = D^*$ that minimises the dominant eigenvalue of $D + G$ can be found using the following algorithm.

1. Find diagonal matrix $Q$ such that $Q^{-1}GQ$ is symmetric. We denote $Q^{-1}GQ$ as $G$. (See [55] for a method for checking if a matrix is symmetrisable and if so how the diagonal matrix can be chosen to achieve symmetry.)

2. Choose some gains $D_i$ as 0 and some other gains $D_i$ as 1 and rearrange the rows and columns of $D + G$ in such a way that those $D_i$ fixed at 0 and 1 are in the lower right corner. The permuted matrix can be decomposed as

$$
\begin{bmatrix}
\hat{G}_{11} + D_{A1} & \hat{G}_{12} \\
\hat{G}_{21} & \hat{G}_{22} + D_{B}
\end{bmatrix}
$$

where the matrix $D_A$ is unknown, and the matrix $D_B$ has entries fixed at 0 or 1.

3. Solve the equation $-1^T\hat{G}_{11}1 - 1^T\hat{G}_{12}(\lambda I - \hat{G}_{22} - D_B)^{-1}\hat{G}_{21}1 + 1^T\lambda 1 = \Gamma - \text{tr}(D_B)$ for $\lambda$. This can be done through a simple numerical procedure, such as a gradient search. (This is a simple procedure because the optimisation is with respect to a single variable. Standard optimisation tools can solve this problem in minimal time.)

4. Calculate $D_A$ using $D_A = \text{diagonalise}(-\hat{G}_{11}1 - \hat{G}_{12}(\lambda I - \hat{G}_{22} - D_B)^{-1}\hat{G}_{21}1 + \lambda I)$. If $0 \leq D_{A1} \leq LI$, and the pattern of eigenvector associated with the dominant eigenvalue of $D + G$ follows Lemma 1, an optimum $D$ is $\text{diag}(D_A, D_B)$. Otherwise, go to step (2) until a solution is achieved.

**Proof:** We know from similarity that the eigenvalues of $D + G$ are same as that of $D + \hat{G}$, where $\hat{G} = Q^{-1}GQ$ and $Q$ is the positive diagonal matrix such that $G$ is symmetric. Hence, we can without loss of generality, consider designing $D^*$ to minimise $D^* + \hat{G}$. Since all $v_{\text{max},i}$ of $D^* + G$ whose corresponding $D_i$ are not fixed at 0 or 1 are equal (let us normalise them to 1), we know an optimum $D$ satisfies

$$
\begin{bmatrix}
\hat{G}_{11} + D_{A1} & \hat{G}_{12} \\
\hat{G}_{21} & \hat{G}_{22} + D_{B}
\end{bmatrix}
\begin{bmatrix}
1 \\
v
\end{bmatrix} = \lambda
\begin{bmatrix}
1 \\
v
\end{bmatrix}
$$

where $D_B$ contains the entries in $D$ that are 0 and 1. $D_A$ must be found, and $\hat{G}_{11}, \hat{G}_{12}, \hat{G}_{21}$, and $\hat{G}_{22}$ are submatrices of $\hat{G}$ (upon appropriate permutation). A little bit of algebra leads to the solution of $\lambda, D$ and $v$. If the pattern of the eigenvector associated with the dominant eigenvalue of $D + G$ follows Lemma 1, and the $D$ matrix is within the constraints $0 \leq D \leq LI$, we have found a global optimum solution according to Lemma 1.

Although we have presented an algorithm for diagonally symmetrisable $G$, a slightly more complicated algorithm exists for all $G$ that are non-negative matrices; we omit this algorithm here in the interest of space. We also note that the above algorithm may be computationally intensive, in that the steps may have to be repeated up to $3^n$ times to find the optimum. For some topology matrices $G$, the calculation of $D^*$ can further be simplified. We will describe the procedure to calculate $D^*$ under these circumstances in Theorem 4. As a preliminary step, let us first explicitly compute an optimal $D$ when the constraints on individual $D_i$ are relaxed.

**Lemma 2:** Consider the matrix $D + G$, where $D$ is diagonal, and $G$ is an $n \times n$ symmetric and non-negative irreducible matrix. $D$ that minimises the dominant eigenvalue of $D + G$ subject only to the constraint $\sum D_i \geq \Gamma$ can be found as follows. First find $\bar{\lambda}_{\text{max}} = (1/n)(1 + \sum \bar{G}_{ii})$ and then find $\bar{D}_i = \lambda_{\text{max}} - \sum_{i \neq j} \bar{G}_{ij}$.

**Proof:** First, we notice all entries of $\bar{v}_{\text{max}}$ are identical, based on Lemma 1 and the fact that here only the constraint $\sum D_i \geq \Gamma$ is enforced. Thus, an optimizing $\bar{D}$ and eigenvalue satisfy $(\bar{D} + \bar{G})1 = \bar{\lambda}_{\text{max}}1$. Therefore, the $\bar{D}_i$’s make the row sums of $D + G$ equal. $\bar{D}_i$ also satisfies $\sum \bar{D}_i = \Gamma$, since $G$ and thus $\bar{D} + G$ is symmetric. A little bit of algebra leads to the expressions for $\bar{\lambda}_{\text{max}}$ and $\bar{D}_i$.

Lemma 2 states that, without the constraints that $D_i \in [0, L]$, an optimum $D$ (denoted as $D^*$) is the one that equalises the row sum of $D + G$, that is, resource is allocated to each part of the network so as to make all their impacts identical. However, when the individual $D_i$ are constrained, sometimes an optimum $D$ cannot be reached. Building on Lemma 2, Theorem 4 illustrates an easy way to find an optimum $D$ under several circumstances.

**Theorem 4:** Consider a matrix $D + G$, where $D$ is diagonal and $G$ is an $n \times n$ non-negative, irreducible and diagonally symmetrisable matrix. We can find $D = D^*$ that minimises the dominant eigenvalue of $D + G$ subject to the constraints (1) $\sum D_i \geq \Gamma$ and (2) $D_i \in [0, L]$ using the following algorithm.

1. Find diagonal $Q$ such that $Q^{-1}GQ$ is symmetric. We name $Q^{-1}GQ$ as $G$.

2. Calculate $\lambda$ and $\bar{D}_i$ following Lemma 2, and check if $0 \leq \bar{D}_i \leq L$ for all $i$. Denote the set of indices $i$ such that $\bar{D}_i > L$ as $L^+$ and the set such that $\bar{D}_i < 0$ as $L^-$. If
\( L^+ = \phi \) and \( L^- = \phi \) (i.e. both sets are empty), we have 
\( \lambda_{\text{max}} = \lambda \) and \( D_i = D_i^* \).

3. If \( L^+ \neq \phi \) and \( L^- = \phi \), we set \( D_i = L \) for \( i \in L^+ \).

That is, \( D_i \) can be calculated by first applying: \( D_i = L \) for \( i \in L^+ \), \( \nu_{\text{max},1,i} = 1 \) for \( i \notin L^+ \) and \( \sum D_i = \Gamma \) and solving the above equations for \( D_i \) as in Theorem 3. If \( D_i \leq L \forall i \), we have \( D^* = D \). Otherwise, we must recursively reduce those \( D_i \) such that \( D_i > L \) to \( L \) and recompute \( D \) until all \( D_i \leq L \).

4. If \( L^+ \neq \phi \) and \( L^- \neq \phi \), we set \( D_i = 0 \) for \( i \in L^- \).

That is, \( D_i \) can be calculated by first applying: \( D_i = 0 \) for \( i \in L^- \), \( \nu_{\text{max},1,i} = 1 \) for \( i \notin L^- \) and \( \sum D_i = \Gamma \) and solving the above equations for \( D_i \) as in Theorem 3. If \( D_i \geq 0 \forall i \), we have \( D^* = D \). Otherwise, we must recursively increase those \( D_i \) such that \( D_i < 0 \) to 0 and recompute \( D_i \) until all \( D_i \geq 0 \).

**Proof:** This theorem considers three cases. The simplest case is when \( L^+ = \phi \) and \( L^- = \phi \). In this case, obviously \( D^* = D \) is an optimum solution, as shown in Lemma 2. The cases when one of \( L^+ , L^- \) is \( \phi \) and the other is not are similar; hence, we only consider the case that \( L^+ = \phi \) and \( L^+ \neq \phi \). In this case, an optimal solution has the special feature that \( D_i^* = L \) for \( i \in L^+ \). This proof is complicated; however, the idea is as follows.

Let us say an unconstrained optimum \((\hat{D})\) yields \( m \) elements in \( L^+ \) \((m \in \hat{D}) \) greater than 1, and let us consider an optimal solution when these \( \hat{D} \) are constrained to \( L \) (while the other entries are left unconstrained for now). We will prove that the eigenvector components associated with these \( m \) elements are smaller than the eigenvector components associated with all the other elements \((\nu_{\text{max},1,i} < \nu_{\text{max},1,L^+})\), and hence prove that this optimum is in fact a global one according to Lemma 1. We show this using induction. Moreover, if the solution obtained by fixing \( D_j = L \) for \( i \in L^+ \) still has some \( D_j > L \), we will repeat the above process to suppress those \( D_j > L \) at \( L \).

**Basis:** Let us use \( Z = \hat{D} + \hat{G} \) for the matrix that produces the unconstrained global optimum, call its dominant eigenvalue \( \hat{\lambda} \) and assume \((\text{wlog})\) that the first \( m \) entries of \( \hat{D} \) are too big \((e.g. \text{larger than } L)\). Let us first consider decreasing the gain in entry 1 to unity while optimising the other gains (right now, we do not apply constraints to these gains). Here we denote \( D_j \) as gain in entry \( i \). Then the matrix of interest changes from \( Z \) to 

\[
Z + \begin{pmatrix}
-\alpha & \Delta D_2 & \cdots & \Delta D_n \\
\end{pmatrix}
\]

where \( \alpha > 0 \) and \( \sum_{i=2}^{n} \Delta D_i = \alpha \). We would like to prove two statements: first, the eigenvector component corresponding to entry 1 is less than the components associated with all the other entries \((e.g. \nu_{\text{max},1} < \nu_{\text{max},1,i})\); second, the gain corresponding to entries from 2 to \( m \) are still at least \( L \).

Recalling Theorem 1, we know that all \( \nu_{\text{max},1,i} \) are the same, say \( c_1 \). For convenience, let us call \( \nu_{\text{max},1} = \nu_1 \). Hence, the eigenvector associated with the dominant eigenvalue \((\text{denoted by } \lambda)\) for the new matrix is \( \nu = [c_1, \nu_1, \cdots, c_1] \). The Courant–Fischer theorem \([56]\) states that \( \lambda = \max_{\nu} \nu^T Z \nu \) and \( \lambda = \max_{\nu} \nu^T (Z + \Delta \nu) \). \( \lambda \) can be written as \( \lambda = \max_{\nu_{\text{max},1}} (\nu^T Z \nu - \nu_1 c_1 + c_1 \alpha) \). When we decrease the gain in entry 1, we know that the maximum eigenvalue of the new matrix, say \( \lambda_1 \), is larger than \( \lambda \). And we also know \( \nu_1^T Z \nu < \lambda_1 \), since \( \nu \) is not the eigenvector for \( \lambda \). Thus, from the expression for \( \lambda_1 \), the first entry in the eigenvector for the new matrix is smaller than the rest of the entries \((i.e. \nu_1 < c_1)\). Thus, the first statement is proved.

For the second statement, we prove it as follows. Since \( \nu_1 < c_1 \), the eigenvector associated with the optimal gain changes from \( 1_n \) to \( 1_n + \nu_1 \begin{bmatrix} 1^n \end{bmatrix} \), where we know \( \nu_1 < 0 < \epsilon \) (since only one gain has been moved so far).

Further, we know that maximum eigenvalue of the new matrix, say \( \lambda_1 \), is larger than \( \lambda \). Plugging into the eigenvector equation and doing some algebra, we finally obtain the following

\[
(\lambda - \lambda_1)1_{n-1} + \epsilon(\lambda_1 - Z_{2,n,1}1 - Z_{2,n,1}\nu_1/c) = (1 + \epsilon) \begin{bmatrix} \Delta D_2 \\ \vdots \\ \Delta D_n \end{bmatrix}
\]

However, since \( \lambda > \lambda_1, \psi < 0 < \epsilon \) and \( Z \) is a non-negative matrix, we recover that all entries in the expression on the left are positive, and hence the change in gains \( \begin{bmatrix} \Delta D_2 \\ \vdots \\ \Delta D_n \end{bmatrix} \) must be all positive and so each other’s gain strictly increases.

Thus, we see that the gain corresponding to entries from 2 to \( m \) are still at least \( L \).

**Induction:** Suppose that we bring any \( l \) of the \( D_j \) to \( L \) \((l < m)\). Let us assume first that the eigenvector components corresponding to the \( l \) entries are less than the components associated with all the other entries \((e.g. \nu_{\text{max},1,l} < \nu_{\text{max},1,i})\); second, that the gain corresponding to entries other than these \( l \) ones are still at least \( L \). Let us show that after we bring another \((l + 1)\) \( D_j \) to \( L \), we still have the appropriate eigenvector component majorisation and condition on the gains.

Let us consider bringing \( l + 1 \) of the \( \hat{D}_j \) from the unconstrained optimum s.t. \( i \in L^+ \) to \( L \). We can do this using two steps. The first step is to move all but one of the
offending gains to $L$ ($l$ gains), and the second step is to bring the last gain to $L$. First note that this is possible, since after the first step, all other gains remain greater than $L$ by assumption. Without loss of generality, for notational convenience, let us assume that we first bring the first $l \bar{D}_i$, s.t. $i \in L^-$ to $L$, and then move $D_{i+1}$. Denote the dominant eigenvalue after the first step as $\lambda_1$ and the one that after bringing the last gain to $L$ as $\lambda_2$. Again applying the Courant–Fischer theorem, $\lambda_1 = \max_l \alpha^l(D + G)v$ and $\lambda_2 = \max_l \alpha^l(D + G + \Delta)v$, where diagonal matrix $\Delta$ corresponding to changing $D_{i+1}$ to $L$ has entries as follows: $\Delta_{i+1/l+1} = -b (b > 0), \Delta_{i,j} = 0$ for $i \in [1, \ldots, l]$ and $\Delta = \Delta D_{i}$ for $i \in [l+2, \ldots, n]$. We also have $\sum_{i=l+1}^{n} \Delta D_{i} = b$. With a argument similar to that given in the basis argument, we can show that the $(l+1)$th eigenvector component is less than the (identical) components after position $l+1$. Repeating this argument with each $D_i$ out of the $l+1$ possibilities set to $L$ last, we can reach the conclusion that the eigenvector components corresponding to all the $l+1$ entries are less than the rest common entries.

The proof that the remaining gains $D_i$ increase (and hence that they remain larger than $L$ if they were originally larger than $L$ without the constraints) after bringing these $l+1$ gains $D_i$ to $L$ is based on the knowledge that the eigenvector components corresponding to all the $l+1$ entries are less than the remaining (identical) entries. This can be proved formally in a fashion very similar to the case where a single gain is moved, which we have addressed in the basis step of the induction. Thus the details are omitted.

In case some other gains exceed $L$ in the process, these can be reduced in the same fashion.

For the case that $L^- \neq \phi$ and $L^+ = \phi$, the proof is analogous to the case $L^- = \phi$ and $L^+ \neq \phi$ that we have proved here, and hence it is omitted. \hfill \Box

Theorem 4 provides an easy way to calculate the diagonal matrix $D$ that minimises the dominant eigenvalue of $D + G$ for a diagonally symmetrizable and non-negative $G$. We can first calculate an optimum $D$ without the individual constraints on $D_i$, that is, if every $D_i$ satisfies its constraint, we have found an optimum. Otherwise, if $D_i$ that violate their constraints are either all larger than $L$ or all less than 0, the actual optimal $D_i$ for positions where constraints are violated are equal to boundary values. This allows us to quickly locate the $D_i$’s at the boundary rather than to try all combinations of $K$ at boundary to find an optimal solution. In fact, at most $n$ cases (rather than $3^n$) need to be considered. If $D$ has entries less than 0 and greater than $L$ at the same time, we must fall back on Theorem 3, that is, search through the possible combinations of $D_i$ at the boundaries.

### 2.2.1 Interpretations of the optimal design $D^* + G$

In this section, we have discussed designing a matrix $D$ to minimise the dominant eigenvalue of $D + G$ subject to constraints. This design specifically allows us to allocate limited repair resources to a contact network, so as to best fight against the spread of a virus. From this viewpoint, it is instructive to study the structure of an optimising $D = D^*$, and hence the structure of $D^* + G$.

The structure of an optimising $D$ is highly dependent on the structure of the matrix $G$, which describes the connection topology of the network. The theorems give us the insight that, for a symmetric $G$, the matrix $D$ should be chosen to best equalise the row sums of $D + G$, within the permitted constraints. In terms of resource allocation, this means that placing the most resource (whether sensing capabilities, vaccinations, quarantine or other resources) at the nodes that have strong connections best prevents virus spread. This makes sense since these nodes have the strongest potential to spread the virus throughout the network if they are infected and similarly to heal the network when they are healthy. Eliminating viruses at these nodes as soon as possible can quickly quench the spread. In case the individual constraints prevent placing enough resource at a node, nearby nodes are provided with extra resources to prevent spread. It is interesting to note that our optimal control strategy in general does not shatter the contact network into multiple components (which has been shown in [57] to be infeasible because a very large number of vertices, even of high degrees, must be removed to break the network), but instead sufficiently impedes the spread of the virus (in a probabilistic sense) to reduce the basic reproduction ratio below 1 with limited resource. An interesting future direction may be to extend our design to the case where resources are allocated to many nodes (individuals) at once, so as to reflect the concept of placing epidemic sensing capacities at locations visited by many individuals as proposed in [57].

It is worth noting that this design is suitable for repair resource allocation before the break-out of a virus or real time during a virus. In other words, this design is robust to the initial location of the virus. This is useful even when real-time allocation of resources after the start of an epidemic is not possible, or when it is hard to locate and respond to infected nodes network-wide in an epidemic. When the initially affected nodes are known, the design can be improved further using this additional information. We leave this improvement to later work.

We recall that our analysis of the contact-network model is based on a linear approximation, and so it is worthwhile to briefly consider the dynamics of the original nonlinear model upon application of our design strategy. A (continuous-time) Markov-process formulation of the nonlinear dynamic model [58] makes clear that, in fact, epidemics always eventually die out in models such as ours, and so the time until die-off is of particular interest. Depending on the dominant eigenvalue of the next-generation operator, the die-off may
either be rapid (e.g. logarithmic in the size of the network) or the virus may be essentially persistent (requiring a die-off time that is exponential in network size) [58]. Let us now consider applying our repair-resource design to networks of increasing size. Assuming that the network topology is symmetric and that enough repair resources are provided to ensure that eigenvalues of the next-generation operator are uniformly within a bounded set strictly inside the unit circle, we can replicate the majorisation-based argument of [58] to prove that the die-off time is logarithmic in network size.

Example: We illustrate calculation of \( D = D^* \) to minimise the dominant eigenvalue of \( D + G \) for the non-negative and symmetric topology matrix

\[
G = \begin{bmatrix}
0 & 1 & 1 \\
1 & 4 & 8 \\
1 & 8 & 16 \\
\end{bmatrix}
\]

subject to the constraints \( D_i \in [0, 1] \) and \( \sum D_j \geq \Gamma \) (Fig. 1). We consider the following two cases.

1. \( \Gamma = 1.5 \). In this case, we have \( D_1^* = 0.4167, D_2^* = 0.4792, D_3^* = 0.6042 \), \( \lambda^* = 0.7917 \) and \( \psi_{\text{max}} = [1 \ 1 \ 1]^T \).

2. \( \Gamma = 2.9 \). In this case, we find that \( D_1^* = 0.928, D_2^* = 0.972 \) and \( D_3^* = 1 \), which yields \( \lambda^* = 1.2661 \) and \( \psi_{\text{max}} = [1 \ 1 \ 0.7047]^T \). We notice that the sum of each row of \( D^* + G \) is not identical in this case and that \( D_1 \) is increased by a larger fraction than \( D_2 \) to make up for the resource deficiency at node 3.

A brief discussion of the procedure for finding the optimal in the example is worthwhile. When \( \Gamma = 1.5 \), the \( D_i \) obtained following step (2) in Theorem 4 are feasible. Hence, this single step finds an optimum \( D \). When \( \Gamma = 2.9 \), \( D_1 \) obtained following step (2) is larger than 1, whereas \( D_1 \) and \( D_3 \) are smaller than 1. This satisfies the condition of step (3). Hence, by fixing \( D_3 \) at 1, and following the calculation in Theorem 3, we find an optimum \( D^* \).

![Figure 1](image.png)  
*Figure 1  Illustration of a three-node network topology*

### 2.3 Designing \( \lambda_{\text{max}}(KG) \) and \( \lambda_{\text{max}}(D + KG) \)

In this section, we address the other two design problems needed for virus-spreading control, design of diagonal \( K \) to minimise the dominant eigenvalue of \( KG \) and \( D + KG \) (\( D \) diagonal), respectively.

The results are similar to those for the \( D + G \) case, so the proofs are omitted. Here, we denote an optimum gain \( \tilde{K}^* \) that minimises the dominant eigenvalue of \( KG \) (or \( D + KG \)) subject to the constraints (1) \( \sum K_i \geq \Gamma \) and (2) \( K_i \in [0, L] \), assuming the dominant eigenvalue of \( KG \) is real and non-repeated. The optimising \( \tilde{K}^* \) and the corresponding left and right eigenvectors \( \psi_{\text{max}}^* \) and \( \varphi_{\text{max}}^* \) satisfy one of the following conditions.

1. \( \sum K_i^* \geq \Gamma \). In this case, for each \( i \) we either have \( 0 < K_i^* < L \) and \( \lambda^*_{\text{max}}K_i^{*\gamma} \psi_{\text{max}}^*, \varphi_{\text{max}}^* = 1 \) or we have \( K_i^* = L \) or \( K_i^* = 0 \).

2. \( \sum K_i^* > \Gamma \). In this case, for each \( i \) we either have \( 0 < K_i^* < L \) and \( \psi_{\text{max}}^*, \varphi_{\text{max}}^* = 0 \) or we have \( K_i^* = L \) or \( K_i^* = 0 \).

An optimum \( K \) can be found through a search algorithm, as stated in Theorem 6. Theorem 6 is for the \( KG \) case, and Theorem 7 is for the \( D + KG \) case.

**Theorem 6:** Consider a topology matrix \( G \) that is non-negative, irreducible and diagonally symmetrisable. The \( K = K^* \) that minimises the dominant eigenvalue of \( KG \) subject to the constraints (1) \( \sum K_i \geq \Gamma \) and (2) \( K_i \in [0, L] \) can be found using the following algorithm.

1. Find diagonal \( Q \) such that \( Q^{-1} GQ \) is symmetric. We denote \( Q^{-1} GQ \) as \( \tilde{G} \).
2. Set some \( K_i \) to 0 and some \( K_i \) to \( L \), remove the rows and columns of \( \tilde{G} \) corresponding to the \( K_i \) chosen as 0 and rearrange the remaining rows and columns of \( \tilde{G} \) in such a way that those \( K_i \) fixed at \( L \) are at the lower right corner. The resulting matrix can then be written as

\[
\begin{bmatrix}
K_{11} & 0 \\
0 & \bar{K}
\end{bmatrix}
\]
3. The eigenvalue $\lambda$ satisfies $\lambda^T(\hat{G}_{11} - \hat{G}_{12}(L\hat{G}_{22} - \lambda I)^{-1}L\hat{G}_{21})^{-1} = \Gamma - \text{trace}(L)$, and can be found through a simple numerical procedure. $K_A$ can be found as $K_A = \text{diagonalise}(\lambda^T(\hat{G}_{11} - \hat{G}_{12}(L\hat{G}_{22} - \lambda I)^{-1}L\hat{G}_{21})^{-1})$. Check whether $K$ is feasible.

4. Repeat with different $K_i$ set to their boundary values until a global minimum $\lambda$ is found.

**Theorem 7:** Consider a topology matrix $G$ that is non-negative, irreducible and diagonally symmetrisable. The $K = K^*$ that minimises the dominant eigenvalue of $D + KG$ subject to the constraints (1) $\sum K_i \geq \Gamma$ and (2) $K_i \in [0, L]$ can be found using the following algorithm.

1. Find diagonal $Q$ such that $Q^{-1}GQ$ is symmetric. We denote $Q^{-1}GQ$ as $\hat{G}$.

2. Set some $K_i$ to 0 and some $K_i$ to $L$ and rearrange the rows and columns of $D + KG$ in such a way that those $K_i$ fixed at 0 or $L$ are at the lower right corner. The resulting matrix can then be written as

$$
\left[ \begin{array}{cc}
D_A & 0 \\
0 & D_B
\end{array} \right] + \left[ \begin{array}{cc}
K_A & 0 \\
0 & K_B
\end{array} \right] = \hat{G}_{11} \hat{G}_{12} \hat{G}_{21} \hat{G}_{22}
$$

3. The eigenvalue $\lambda$ satisfies $(\lambda^T - D_A^T)(\hat{G}_{11} - \hat{G}_{12}(D_B + K_B \hat{G}_{22} - \lambda I)^{-1}K_B \hat{G}_{21})^{-1} = \Gamma - \text{trace}(L)$, and can be found through a simple numerical procedure. $K_A$ can be found as $K_A = \text{diagonalise}(\lambda^T - D_A^T)(\hat{G}_{11} - \hat{G}_{12}(D_B + K_B \hat{G}_{22} - \lambda I)^{-1}K_B \hat{G}_{21})^{-1})$. Check whether $K$ is feasible.

4. Repeat with different $K_i$ set to their boundary values until a global minimum $\lambda$ is found.

When $G$ is positive-definite in addition to diagonally symmetrisable, we can show that $\lambda_{\text{max}}(KG)$ is a convex function. In this case, the search algorithms given in Theorems 6 and 7 can be simplified. We can check whether a solution is a local optimum and stop or continue the search accordingly, since a local optimum is guaranteed to be a global optimum. Theorem 8 is for the $KG$ case, and Theorem 9 is for the $D + KG$ case.

**Theorem 8:** Consider a topology matrix $G$ that is non-negative, irreducible, positive-definite and diagonally symmetrisable. The $K = K^*$ that minimises the dominant eigenvalue of $KG$ subject to the constraints (1) $\sum K_i \geq \Gamma$ and (2) $K_i \in [0, L]$ can be found using the following algorithm.

1. Find diagonal $Q$ such that $Q^{-1}GQ$ is symmetric. We denote $Q^{-1}GQ$ as $\hat{G}$.

2. Find $K = \text{diagonalise}(\hat{G}^{-1}1/(1^T\hat{G}^{-1}1))$. If $0 \leq K_i \leq L$, this solution is optimal. Otherwise, go to step 3.

3. Set some $K_i$ to 0 and some $K_i$ to $L$, remove the rows and columns of $G$ corresponding to the $K_i$ chosen as 0 and rearrange the remaining rows and columns of $KG$ in such a way that those $K_i$ fixed at $L$ are at the lower right corner. The resulting matrix can then be written as

$$
\left[ \begin{array}{cc}
K_A & 0 \\
0 & L_I
\end{array} \right] \left[ \begin{array}{cc}
\hat{G}_{11} & \hat{G}_{12} \\
\hat{G}_{21} & \hat{G}_{22}
\end{array} \right]
$$

4. The eigenvalue $\lambda$ satisfies $\lambda^T(\hat{G}_{11} - \hat{G}_{12}(L\hat{G}_{22} - \lambda I)^{-1}L\hat{G}_{21})^{-1} = \Gamma - \text{trace}(L)$, and can be found through a simple numerical procedure.

5. $K_A$ can be found as $K_A = \text{diagonalise}(\lambda^T - D_A^T)(\hat{G}_{11} - \hat{G}_{12}(L\hat{G}_{22} - \lambda I)^{-1}L\hat{G}_{21})^{-1})$. If $0 \leq K_i \leq L$, and the left eigenvectors associated with the dominant eigenvalue of $D + KG$ have the pattern that the entries corresponding to $K_i = L$ are less than those corresponding to $K_i = 0$, this solution is optimal. Otherwise, go back to step 3.

**Theorem 9:** Consider a topology matrix $G$ that is non-negative, irreducible, positive-definite, and diagonally symmetrisable. The $K = K^*$ that minimises the dominant eigenvalue of $D + KG$ subject to the constraints (1) $\sum K_i \geq \Gamma$ and (2) $K_i \in [0, L]$ can be found using the following algorithm.

1. Find diagonal $Q$ such that $Q^{-1}GQ$ is symmetric. We denote $Q^{-1}GQ$ as $\hat{G}$.

2. Find $\lambda$ that satisfies $\hat{G}^{-1}(\lambda^T - 1^TD)1 = \Gamma$ through a simple iteration. And then find $K$ from $K = \text{diagonalise}(\hat{G}^{-1}(\lambda^T - 1^TD))$. If $0 \leq K_i \leq L$, this solution is optimal. Otherwise, go to step 3.

3. Set some $K_i$ to 0 and some $K_i$ to $L$, and rearrange the rows and columns of $D + KG$ in such a way that those $K_i$ fixed at 0 or $L$ are at the lower right corner. The resulting matrix can then be written as

$$
\left[ \begin{array}{cc}
D_A & 0 \\
0 & D_B
\end{array} \right] + \left[ \begin{array}{cc}
K_A & 0 \\
0 & K_B
\end{array} \right] = \hat{G}_{11} \hat{G}_{12} \hat{G}_{21} \hat{G}_{22}
$$

4. The eigenvalue $\lambda$ satisfies $(\lambda^T - D_A^T)(\hat{G}_{11} - \hat{G}_{12}(D_B + K_B \hat{G}_{22} - \lambda I)^{-1}K_B \hat{G}_{21})^{-1} = \Gamma - \text{trace}(K_B)$, and can be found through a simple numerical procedure.

5. $K_A$ can be found as $K_A = \text{diagonalise}(\lambda^T - D_A^T)(\hat{G}_{11} - \hat{G}_{12}(D_B + K_B \hat{G}_{22} - \lambda I)^{-1}K_B \hat{G}_{21})^{-1})$. If $0 \leq K_i \leq L$, and the left eigenvectors associated with the dominant eigenvalue of $D + KG$ have the pattern that the entries corresponding to $K_i = L$ are less than those corresponding to...
0 < K_i < L, and the latter are less then those corresponding to \( K_i = 0 \), this solution is optimal. Otherwise, go back to step 3.

2.4 Computation and implementation issues

Let us make a couple of notes on our methods. First, the reader may wonder why we have developed algorithms that are specialised to the stated decentralised design task rather than using standard optimisation algorithms, so we will briefly comment on the computational and conceptual advantages provided by this approach. Second, we will briefly discuss the robustness of the design strategy.

2.4.1 Advantages of our optimisation approach:

Computationally, the design (optimisation) problem studied here is hard, because it is a decentralised design/control problem with a cost based on an associated dynamics (e.g. a settling-rate cost). Decentralised problems of this sort are known in general to be NP-hard \([59]\). The particular decentralised problem of designing a diagonal matrix (say \( K \)) so as to place the eigenvalues of \( KG \) or \( K + G \) in desirable locations subject to constraints falls within this class of challenging problems: in fact, there is no known general algorithm for deciding whether there is \( K \) that places the eigenvalues in the unit circle (or in the closed left-half plane), let alone doing so with minimal resource \([8, 60, 61]\).

Essentially, the difficulty in the problem stems from the fact that the cost depends in a complicated (and implicit) way on the design variables, and yet there are not sufficient degrees of freedom to construct the dynamics (modes) at will. Because of these limitations, standard optimisation packages cannot be used to find the optimal design. Specialised semi-definite programming and in turn linear matrix inequality techniques can be applied for some classes of symmetric topology matrices \( G \) \([62, 63]\), but these techniques are not applicable for the broad class of positive matrices considered here, nor can they be specialised to guarantee solution using only a small number of iterations as we have done for the symmetrisable case. More specifically, we have studied the \( KG \) and \( D + G \) design problems for networks of up to 1000 regions (individuals) and find that typically at most three iterations of our algorithm are needed to exactly find the optimal. Also, we note that some clever manipulation is needed to pose the design problem as an SDP problem even in the symmetric case, so that implementation of the algorithm through this approach is also complicated. The essential difficulties in decentralised design/control problems have led to a burgeoning interest in recent years on design for modern networks \([4, 5, 8]\) using graph-theoretic and matrix-theoretic methods; broadly, our techniques are aligned with this methodology and permit low-computation design for relevant network topologies.

We also contend that the analytical design method presented here provides a conceptual advantage over numerical techniques. In fact, the presented algorithms admit a simple conceptual interpretation, for the symmetrisable case: resources should be provided to regions (individuals) to equalise the impact of the infectives from each region to the other ones. In the case where the resource limits do not permit such allocation to a particular region (individual), neighbouring regions must be provided with additional resources to compensate for the resource deficiency and, equivalently, if a region has more resource allocation than needed due to the lower resource bound, resources in neighbouring regions can be curtailed. We believe that this simple conceptualisation of the optimal solution is valuable, especially, for achieving robustness: even when the model parameters and/or control models are inaccurate, one can allocate resources with the aim of equalising spread impact and expect to obtain a good if not optimal solution.

2.4.2 Robustness of the design: Second, an important practical concern in applying these algorithms to the epidemic-control problems introduced in Section 2 is whether or not the parameters in the models can be approximated with sufficient fidelity to achieve a high-performance design. In discussing these concerns, we first note that the parametric data needed for the optimisations are only those needed for analysing the basic reproduction ratios of the two studied models (the multi-group model and the contact model). We thus refer the reader to existing work \([3, 24]\) for discussions on how these parameters can be identified. Fundamentally, we notice that the parameters in both models are concerned with either the spread of interaction or with the infectiousness of the virus, and so counting of flows/interactions as well as infection relative to total interaction can be used to find the parameters. Also, the interaction parameters in particular can also be identified from historical epidemics in the network of interest. However, we are proposing control actions and recognising that the model parameters may indeed be uncertain (especially those that may need to be obtained in real time), in that we recognise that the robustness of the developed algorithm requires further study. Our preliminary efforts in this direction are very promising: for the Hong Kong SARS virus example, we have found that the optimal heterogeneous optimal design far outperforms a homogeneous design even when the parameters are uncertain by as much as 40% (see Section 3 for details). This high tolerance of the design to parameter errors is not surprising since the optimum is deeply tied to the network structure, and more specifically to the total impact of a region/node. These structural features are well characterised even when individual parameters are uncertain. We note that this conceptual justification of robustness is yet one more advantage of addressing the decentralised design problem explicitly rather than trying to apply standard optimisation tools.

3 Control of the Hong Kong SARS epidemics

Recall that Riley et al. \([3]\) developed a spatial model for the spread of SARS in Hong Kong’s 18 districts (Fig. 2) and
proposed a homogeneous control for reducing the basic reproduction ratio $R_0$ to 1. In that model, the nominal corrective factor $f_{ji}$ takes three different values contingent on the structural relationship of $i$ and $j$: (1) $i = j$, (2) districts $i$ and $j$ are contiguous and (3) districts $i$ and $j$ are not contiguous.

Here, we find an optimal heterogeneous control that uses the same total resource amount as the controller in [3]. This controller reduces the basic reproduction ratio to 0.64. Thus, we see that an epidemic can be stopped more quickly with the same control resources, by allocating more resources to some districts than others. Equivalently, it is easily shown that $R_0 = 1$ can be achieved even when the total control resource is reduced to 79% of that with equal allocation (Table 1).

This intelligent allocation takes advantage of the spatial structure of the population, by placing more control resources in the districts that are important for the spread of an epidemic (Fig. 2). In this way, the limited control resources are best able to reduce the rate at which the epidemic diminishes. Such a control would reduce the impact on people’s daily lives in some districts (which have less control resources allocated) and overall, while still stopping the virus spread quickly.

For illustration, we also consider how the heterogeneous resource allocation changes when the transmission coefficient within a district is increased (compared with the mixing rate between districts). As expected, increasing the local mixing rate makes the use of spatial information less important, and also makes the allocation more homogeneous (Table 1).

In all of these experiments, we see that the most resources are placed in districts 5 and 7, and the smallest resources are placed in district 1. This is expected since districts 5 and 7 are the ones with pivotal locations (e.g. with many neighbours) and hence the successful control of these districts is important in the control of disease spread. In contrast, district 1 is almost isolated and hence has the least contribution to the spread of diseases in Hong Kong.

We have also pursued a preliminary robustness analysis for our design. In particular, we have analysed the performance

![Figure 2 Map of Hong Kong’s 18 districts (obtained from http://en.wikipedia.org/wiki/Hong_Kong#Administrative_divisions)](http://en.wikipedia.org/wiki/Hong_Kong#Administrative_divisions)

We use the model parameters in [3]: $f_{ii} = 1$, $f_{ij} = 0.57$ when districts $i$ and $j$ are neighbours, $f_{ji} = 0.02$ when $i$ and $j$ are not adjacent, $\bar{\beta} = 0.062$, and $T = 10.6$. 
of our design in the case where the actual parameters are different from the ones used for the design. For illustration, let us consider one of the designs shown in Table 1, a heterogeneous one corresponding to $f_{ii} = 3$. In this case, we see that a heterogeneous design reduces $R_0$ to 0.87 if the total amount of resource is that for which a homogeneous design reduces $R_0$ to 1. Now say that each interaction parameter $f_{ii} = 3$ is in fact in error by at most 40%, or more specifically that the parameter is in fact $a_{ii}/f_{ii}$, where $a_{ii}$ is uniformly distributed between 0.6 and 1.4. Over 100 trials, our design achieves average $R_0 = 0.89$, when compared with average $R_0 = 0.88$ if the design were re-optimised for the new parameters. We thus see that the design still significantly outperforms a homogeneous design with 40% error in the parameters.

4 Acknowledgments

The authors are grateful for the support provided by the National Science Foundation (under Grant ECS0528882) and the National Aeronautics and Space Administration (under Grant NNA06CN26A).

5 References


Table 1 Resource allocation needed to reduce $R_0$ to 1

<table>
<thead>
<tr>
<th>district</th>
<th>$f_{ii} = 1$</th>
<th>$f_{ii} = 1.5$</th>
<th>$f_{ii} = 3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>district 1</td>
<td>0.9096</td>
<td>0.6222</td>
<td>0.3031</td>
</tr>
<tr>
<td>district 2</td>
<td>0.9096</td>
<td>0.5368</td>
<td>0.2525</td>
</tr>
<tr>
<td>district 3</td>
<td>0.5816</td>
<td>0.4338</td>
<td>0.2546</td>
</tr>
<tr>
<td>district 4</td>
<td>0</td>
<td>0</td>
<td>0.1919</td>
</tr>
<tr>
<td>district 5</td>
<td>0</td>
<td>0.0160</td>
<td>0.0915</td>
</tr>
<tr>
<td>district 6</td>
<td>0.5816</td>
<td>0.4242</td>
<td>0.1705</td>
</tr>
<tr>
<td>district 7</td>
<td>0</td>
<td>0</td>
<td>0.0834</td>
</tr>
<tr>
<td>district 8</td>
<td>0.9096</td>
<td>0.5926</td>
<td>0.2708</td>
</tr>
<tr>
<td>district 9</td>
<td>0</td>
<td>0.0789</td>
<td>0.1734</td>
</tr>
<tr>
<td>district 10</td>
<td>0</td>
<td>0</td>
<td>0.1419</td>
</tr>
<tr>
<td>district 11</td>
<td>0.5816</td>
<td>0.4553</td>
<td>0.2194</td>
</tr>
<tr>
<td>district 12</td>
<td>0</td>
<td>0.2118</td>
<td>0.1798</td>
</tr>
<tr>
<td>district 13</td>
<td>0.5816</td>
<td>0.4457</td>
<td>0.1985</td>
</tr>
<tr>
<td>district 14</td>
<td>0.9096</td>
<td>0.5428</td>
<td>0.2587</td>
</tr>
<tr>
<td>district 15</td>
<td>0.9096</td>
<td>0.4784</td>
<td>0.2475</td>
</tr>
<tr>
<td>district 16</td>
<td>0.9096</td>
<td>0.4784</td>
<td>0.2475</td>
</tr>
<tr>
<td>district 17</td>
<td>0</td>
<td>0.1919</td>
<td>0.1908</td>
</tr>
<tr>
<td>district 18</td>
<td>0</td>
<td>0.1919</td>
<td>0.1908</td>
</tr>
</tbody>
</table>

The $K_i$ for each district are shown in the table. Notice that smaller $K_i$ correspond to more resources. $\sum K_i$ equal the subtraction of total utilised resources from $N$ (the number of districts, e.g. 18 in this case). Note that the reduction in resource used for $f_{ii} = 1$ is $(18 - 7.7842)/(18 - 5.0138) \times 100\% = 79\%$. 

P $K_i$ using heterogeneous control | 7.7842 | 5.7007 | 3.6667 |

P $K_i$ using homogeneous control | 5.0138 | 4.4009 | 3.22 |


