Network Reconstruction and Prediction of Epidemic Outbreaks for General Group-Based Compartmental Epidemic Models

Bastian Prasse and Piet Van Mieghem

***Abstract*—The underlying core of most epidemic models is the graph that specifies the contacts between healthy and infected individuals. However, in the majority of applications, the contact network is unknown. To understand and predict an epidemic outbreak nonetheless, network reconstruction methods aim to estimate the contact network from viral state observations. This work considers general compartmental epidemic models (GEMF) in discrete time, which describe the viral spread between groups of individuals. The reconstruction of the network translates into a set of linear equations that is severely ill-conditioned. Counterintuitively, we show that the contact network cannot be reconstructed from one epidemic outbreak with any finite machine precision, although an accurate prediction of the epidemic outbreak is possible.**

***Index Terms*—Epidemic models, network reconstruction, prediction of epidemic outbreaks**

I. Introduction

THE field of epidemics encompasses a plethora of phenomena and is rooted in the description of infectious diseases [1], with seminal works by Bernoulli [2] and Snow [3]. Beyond infectious diseases, the spread of opinions, trends and fake news on on-line social networks can be described as an epidemic of a viral infection, whereby individuals infect one another with the opinion, trend, etc. The vast majority of epidemic models assigns every individual to a *compartment* such as susceptible, infected, or recovered from the disease. Epidemic processes over networks assume that the spreading may occur from one to another individual only if the two individuals have contact [4], for instance by a friendship or sexual relation.

The contact graph between individuals has a great impact on the spread of the virus [4], [5]. However, in the study of real-world epidemics, there is often not much known about the contact graph other than high-level properties such as, for instance, the degree distribution [6]. To obtain a better understanding of the viral spread, network reconstruction methods aim to infer the unknown contact graph from observing the viral state evolution. If the contact graph can be reconstructed, then the epidemic outbreak can be predicted. However, as we will show in this work, the prediction of epidemic outbreaks is surprisingly less related to network reconstruction, despite the clear dependence of the dynamic equations of epidemic spread on the contact graph (see equation (4) below). *In particular, we show that, for the majority of applications, the network cannot be reconstructed although the epidemic outbreak can be predicted.*

The majority of network reconstruction methods focussed on inferring the contact network from viral state observations of every single individual [7]–[12]. Network reconstruction methods from viral state observations of single individuals are subject to two fundamental limitations. First, it is hardly practical to determine the viral state of every individual at every time in real-world epidemics. Second, an accurate network reconstruction requires a tremendous number n of viral state observation [8]. Thus, inferring the contact network between single individuals only seems possible long after the virus reached the endemic state or, if the virus dies out, by observing multiple epidemic outbreaks, both of which seems impractical. To overcome the challenges of reconstructing the contact network of individual-based models, we describe the evolution of the virus on a coarser level between groups, or communities, of similar individuals. The prevalence of a virus within a group is accessible by sampling representative individuals.

In this work, we focus on the viral spread over a network with N nodes, where each node corresponds to a group of individuals such as households or geographical regions. We consider that the viral spread between groups follows a discrete-time version of the Generalised Epidemic Mean-Field (GEMF) model [13] with heterogeneous spreading parameters on a directed contact network. The GEMF model considers M viral state compartments, which unifies a myriad of diverse epidemic models. For instance, in the Susceptible-Infected-Susceptible (SIS) process there are $M = 2$ compartments, and in the Susceptible-Infected-Recovered (SIR) process there are $M = 3$ compartments. The viral state of node i at continuous time $t \geq 0$ is denoted by $v_i(t) = (v_{i,1}(t), \ldots, v_{i,M}(t))^T \in [0,1]^M$, where $v_{i,p}(t)$ describes the fraction of individuals of group i in compartment p at time t.

Originally, Sahneh *et al.* [13] derived the GEMF model as a mean-field approximation of *individual-based* Markovian spreading processes, where every node i corresponds to a

Manuscript received November 12, 2019; revised January 20, 2020 and March 18, 2020; accepted April 10, 2020. Date of publication April 16, 2020; date of current version December 30, 2020. Recommended for acceptance by Dr. Caterina Scoglio. *(Corresponding author: Bastian Prasse.)*

The authors are with the Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, 2628, CD, Delft, The Netherlands (e-mail: B.Prasse@tudelft.nl; P.F.A.VanMieghem@tudelft.nl).

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier 10.1109/TNSE.2020.2987771

single individual, whose viral state equals either one of M compartments. Then, the probability that the viral state of individual i equals the p-th compartment at time t is approximated by the state $v_{i,p}(t)$ of the GEMF model. In contrast, our interpretation of the viral state $v_{i,p}(t)$ as the fraction of individuals of group i in compartment p is in line with [14]–[17]. Furthermore, Paré *et al.* [17] provided a validation of the NIMFA epidemic model on real-world epidemic data, where the nodes of the network correspond to groups of individuals, namely either households or counties. Ideally, individuals in the same group are interchangeable for describing the epidemic outbreak. The number of individuals in different groups i does not need to be the same.

The nomenclature is presented in Section II. We propose the GEMF epidemic model in discrete time in Section III. The network reconstruction for the GEMF epidemic model is equivalent to solving a set of linear equations, as we show in Section IV. Section V discusses fundamental limits of reconstructing the contact network from GEMF viral state observations. We propose a network reconstruction method in Section VI based on the *least absolute shrinkage and selection operator* (LASSO). The network reconstruction method is evaluated in Section VII for random graphs and real networks, also in the presence of model errors, and the results show that the prediction of an epidemic outbreak and the reconstruction of the contact network are fundamentally different tasks.

II. Nomenclature

The $N \times N$ identity matrix is denoted by I_N. The number of compartments in the GEMF model is denoted by $M \in \mathbb{N}$, and the $M \times 1$ all-one vector is denoted by u. For an $N \times 1$ vector x, $\mathrm{diag}(x)$ denotes the $N \times N$ diagonal matrix with the vector x on its diagonal. The spectral radius of a square matrix A is denoted by $\rho(A)$. For any $N \times N$ matrix A, we define the $N^2 \times 1$ vector that is obtained by stacking the columns of A as $\mathrm{vec}(A) = (a_{11}, \ldots, a_{N1}, a_{12}, \ldots, a_{N2}, \ldots)^T$. The Kronecker product of a $k \times l$ matrix A and a $p \times q$ matrix B is denoted by $A \otimes B \in \mathbb{R}^{kp \times lq}$.

III. The Discrete-Time GEMF Epidemic Model

In Section III-A, we define the general discrete-time GEMF epidemic model. We give important special cases of the GEMF model in Section III-B. In Section III-C, we introduce curing probability control for the GEMF model.

A. General GEMF Epidemic Model

We generalise the GEMF model [13] to heterogeneous spreading parameters and directed graphs. We state the GEMF model in discrete time and denote the viral state of group i at discrete time $k \in \mathbb{N}$ by $v_i[k] \in [0,1]^M$. Since $v_{i,p}[k]$ denotes the fraction of individuals of group i in compartment p and each individual is in exactly one compartment, it holds that $v_{i,1}[k] + \ldots + v_{i,M}[k] = 1$ at any time k. For every two compartments $p, q = 1, \ldots, M$, we denote the $N \times N$ zero-one adjacency matrix as A_{pq} with elements $a_{pq,ij}$. The adjacency matrices A_{pq} specify the contact network. If there is a directed link from compartment q of group j to compartment p of group i, then $a_{pq,ij} = 1$, and $a_{pq,ij} = 0$ otherwise. For instance, if compartment q denotes individuals that are in quarantine, then it holds that $A_{pq} = 0$ for all compartments $p \neq q$ since the quarantine-compartment q is isolated from all compartments $p \neq q$. In the GEMF model, there are two kinds of viral state transitions from time k to $k+1$. *Nodal transitions* occur at a node i independently of the viral state $v_j[k]$ of the other nodes $j \neq i$. The $M \times M$ nodal transition probability matrix S_i specifies the probabilities of nodal transitions at node i. The probability that, via a nodal transition, an individual in group i changes from compartment p to compartment q equals $(S_i)_{pq}$. In contrast, *edge-based transitions* do depend on the viral state $v_j[k]$ of the neighbours j of node i and, hence, on the contact network. The $M \times M$ edge-based transition probability matrix $B_{m,ij}$ specifies the probabilities of edge-based transitions at node i due to (for instance, an infection from) the individuals of group j in compartment m. More precisely, the probability that an individual in group i changes from compartment p to compartment q due to a fraction $v_{j,m}[k]$ of individuals of group j in compartment m equals $(B_{m,ij})_{pq} v_{j,m}[k]$. We emphasise that the edge-based transition probability matrix $B_{m,ii}$ from group i to group i is not necessarily zero, because the individuals in group i can possibly interact with each other.

The edge-based transition probability matrix $B_{m,ij}$ is related to the adjacency matrices as A_{pm}, for all compartments $p = 1, \ldots, M$, as follows. Since individuals of group j in compartment m have an impact on individuals of group i in compartment p only if there is a link from compartment m of group j to compartment p of group i, it holds

$$\left(B_{m,ij}\right)_{pq} = (\tilde{B}_{m,ij})_{pq} a_{pm,ij} \tag{1}$$

for some $M \times M$ matrix $\tilde{B}_{m,ij}$. Hence, it holds $a_{pm,ij} = 1$ only if[1] there is a compartment q such that $(B_{m,ij})_{pq} > 0$. More precisely, we can obtain the entries $a_{pm,ij}$ of the adjacency matrix A_{pm} by

$$a_{pm,ij} = \begin{cases} 1 & \text{if} \quad \exists q = 1, \ldots, M: \ \left(B_{m,ij}\right)_{pq} > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Any GEMF model can be visualised as in Figure 1 by the *transition graph*, which we define as follows. All compartments of two (arbitrary) groups i, j are represented by a node in the transition graph. Regarding the compartments of group i, two nodes in the transition graph are connected by a directed link if there is a transition between the respective compartments of group i. (The transitions between the compartments of the other group j are omitted, since the transitions between the compartments of one group i suffice to specify the GEMF model.) A node-based transition of group i from compartment p to compartment q is represented by a simple arrow "$\rightarrow$" that is labelled with the transition probability $(S_i)_{pq}$. An edge-based transition of group i from compartment p to compartment q is represented by an arrow with the multiplier "$\otimes$" in the middle.

[1] Here, we make the technical assumption: if there is a link from compartment m of group j to compartment p of group i, then the probability $(B_{m,ij})_{pq}$ is positive for at least one compartment q.

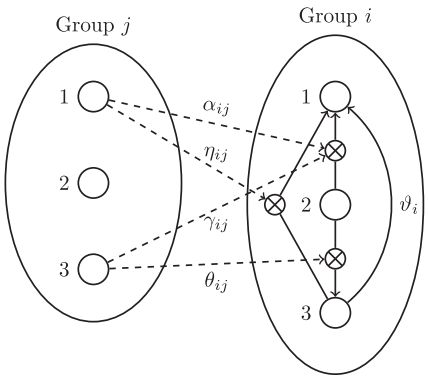

Fig. 1. The transition graph for an exemplary GEMF model with $M = 3$ compartments. The solid lines correspond to possible transitions between the three compartments of group i. The dashed lines illustrate which compartment of group j influences which edge-based transition between two compartments of group i.

If compartment m of group j has an influence on the edge-based transition of group i from compartments p to compartment q, then there is an arrow from compartment m of group j to the respective multiplier "$\otimes$", which is labelled with the transition probability $(B_{m,ij})_{pq}$. We emphasise that, by definition (1), $(B_{m,ij})_{pq} = 0$ if the respective link $a_{pm,ij} = 0$.

Figure 1 illustrates an exemplary transition graph for a GEMF model with $M = 3$ compartments. In the following, we show how the transition graph in Figure 1 fully specifies the GEMF model, i.e. the node-based and edge-based transitions. In Figure 1, there is exactly one simple arrow from compartment 3 to compartment 1, which is labelled with the transition probability ϑ_i. Hence, the nodal transition probability matrix S_i equals

$$S_i = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \vartheta_i & 0 & 0 \end{pmatrix}. \tag{3}$$

There are two arrows from compartment 1 of group j to the edge-based transitions of group i: from compartment 2 to compartment 1 (labelled with α_{ij}), and from compartment 3 to compartment 1 (labelled with ν_{ij}). Hence, the edge-based transition probability matrix $B_{1,ij}$ equals

$$B_{1,ij} = \begin{pmatrix} 0 & 0 & 0 \\ \alpha_{ij} & 0 & 0 \\ \nu_{ij} & 0 & 0 \end{pmatrix}.$$

There is no arrow from compartment 2 of group j to a transition of group i. Hence, compartment 2 of group j has no influence on the transitions of group i, and it holds that $B_{2,ij} = 0$. From the two arrows starting at compartment 3 of group j, we obtain the edge-based transition probability matrix $B_{3,ij}$ as

$$B_{3,ij} = \begin{pmatrix} 0 & 0 & 0 \\ \gamma_{ij} & 0 & \theta_{ij} \\ 0 & 0 & 0 \end{pmatrix}.$$

The matrices S_i and $B_{m,ij}$, where $m = 1, 2, 3$, for all groups i, j fully specify the transitions of the GEMF model. Furthermore, the links $a_{pm,ij}$ from compartment m of group j to compartment p of group i can be obtained from Figure 1 as follows. The link labelled with α_{ij} connects compartment 1 of group j to an edge-based transition *starting* at compartment 2 of group i (ending at compartment 1 of group i), which yields that $a_{21,ij} = 1$ if $\alpha_{ij} > 0$. Similarly, the link labelled with ν_{ij} yields that $a_{31,ij} = 1$ if $\nu_{ij} > 0$. Both of the links labelled with γ_{ij} and θ_{ij} connect compartment 3 of group j with edge-based transitions *starting* at compartment 2 of group i, which yields that $a_{23,ij} = 1$ if $\gamma_{ij} > 0$ or $\theta_{ij} > 0$ (or both). For the other compartments p, m, which have not been mentioned yet, it holds that $a_{pm,ij} = 0$.

Definition 1 (Discrete-Time GEMF Epidemic Model): The discrete-time GEMF epidemic model describes the evolution of the viral state $v_i[k] \in \mathbb{R}^M$ for every group $i = 1, \ldots, N$ as

$$v_i[k+1] = \left(I_M - Q_i^T\right) v_i[k] - \sum_{j=1}^{N} \sum_{m=1}^{M} v_{j,m}[k] Q_{m,ij}^T v_i[k], \tag{4}$$

where $k \in \mathbb{N}$ denotes the discrete time slot. Here, the $M \times M$ Laplacian matrices of the nodal transition probability matrix S_i and the edge-based transition probability matrix $B_{m,ij}$ are denoted by $Q_i = \mathrm{diag}(S_i u) - S_i$ and $Q_{m,ij} = \mathrm{diag}(B_{m,ij} u) - B_{m,ij}$.

In Appendix A, we derive the discrete-time GEMF model (4) from the continuous-time GEMF model [13] by applying Euler's method. If the initial viral state $v_i[1]$ of every node i satisfies $v_{i,1}[1] + \ldots + v_{i,M}[1] = 1$, then [13] it holds that $v_{i,1}[k] + \ldots + v_{i,M}[k] = 1$ at any time $k \geq 1$. Thus, the GEMF model (4) with MN compartments can be reduced to $(M-1)N$ non-linear difference equations.

Originally, the GEMF model was formulated for multi-layer networks [13]. The discrete-time GEMF model (4) does not explicitly model distinct network layers but directly *sums* the influences across all network layers. For instance, consider that infected individuals in group j infect susceptible individuals in group i via a link in the workplace network (network layer $l = 1$) with the transition probability $\beta_{ij}^{(1)}$ or via a link in the friendship contact network (network layer $l = 2$) with the transition probability $\beta_{ij}^{(2)}$. Then, an equivalent GEMF model is obtained by a total transition probability of $\beta_{ij} = \beta_{ij}^{(1)} + \beta_{ij}^{(2)}$ on one network layer. Since the value of the transition probability β_{ij} completely determines the viral state dynamics of the GEMF model (4), it is only possible to estimate the transition probability β_{ij} from viral state observations $v_i[1], v_i[2], \ldots$, but not the distinct addends $\beta_{ij}^{(1)}$ and $\beta_{ij}^{(2)}$ of the different layers.

B. Special Cases of the GEMF Epidemic Model

In this work, we consider four special cases of the GEMF model (4). First, we consider the SIS epidemic model with two compartments: the susceptible (or healthy) compartment $\mathcal{S}$ and the infected compartment $\mathcal{I}$. At any time k, an individual in group i changes from the infected compartment $\mathcal{I}$ to the susceptible compartment $\mathcal{S}$ with the curing,[2] probability

[2] For the models in Section III-B we refer to the transition probabilities δ_i and β_{ij} as curing probability and infection probability, respectively, to stress their physical meaning.

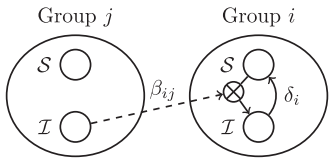

Fig. 2. The transition graph for the SIS epidemic model.

δ_i, which is the only nodal transition. Furthermore, there is exactly one edge-based transition, which is the infection of an individual. At any time k, an individual of group i changes from the susceptible compartment $\mathcal{S}$ to the infected compartment $\mathcal{I}$ with the probability $\sum_{j=1}^{N} \beta_{ij}\mathcal{I}_j[k]$, where $\mathcal{I}_j[k]$ denotes the fraction of infected individuals of group j at time k, and β_{ij} is the infection probability from group j to group i. Figure 2 shows the transition graph of the SIS epidemic model.

The derivations in this work hold for the general GEMF model (4). However, for the sake of explanation, we put a particular focus on the SIS epidemic model, whose systems equations are given as:

Definition 2 (SIS Epidemic Model[3]): For every group i, the viral state of the SIS epidemic model equals $v_i[k] = (\mathcal{S}_i[k], \mathcal{I}_i[k])^T$. Here, $\mathcal{S}_i[k]$ and $\mathcal{I}_i[k]$ denote the fraction of susceptible and infected individuals in group i at time $k \in \mathbb{N}$, respectively. For every group i at any time $k \in \mathbb{N}$, the viral state evolves according to

$$\mathcal{I}_i[k+1] = (1-\delta_i)\mathcal{I}_i[k] + (1-\mathcal{I}_i[k]) \sum_{j=1}^{N} \beta_{ij}\mathcal{I}_j[k] \qquad (5)$$

and the fraction of susceptible individuals follows as $\mathcal{S}_i[k] = 1 - \mathcal{I}_i[k]$. Here, β_{ij} denotes the *infection probability* from group j to group i, and δ_i denotes the *curing probability* of group i.

The second model that we consider has four compartments: the susceptible compartment $\mathcal{S}$, the exposed compartment $\mathcal{E}$, the infectious compartment $\mathcal{I}$, and the recovered (or removed) compartment $\mathcal{R}$. An individual transitions the compartments in the order $\mathcal{S} \to \mathcal{E} \to \mathcal{I} \to \mathcal{R}$. Individuals in the exposed compartment $\mathcal{E}$ have been infected by the disease but, in contrast to individuals in the infectious compartment $\mathcal{I}$, are not contagious yet. Individuals in the recovered compartment $\mathcal{R}$ have had the disease, but are not susceptible nor infectious any more (for instance, by immunisation or death). In the SEIR epidemic model, the only edge-based transition occurs from the susceptible compartment $\mathcal{S}$ to the exposed compartment $\mathcal{E}$, analogously to the $\mathcal{S} \to \mathcal{I}$ transition in the SIS epidemic model. Furthermore, there are two nodal transitions in the SEIR epidemic model. First, the transition from the exposed compartment $\mathcal{E}$ to the infectious compartment $\mathcal{I}$, which occurs with the incubation probability γ_i for an individual in group i. Second, the transition from the infectious compartment $\mathcal{I}$ to the recovered compartment $\mathcal{R}$, which occurs with the curing probability δ_i for an individual in group i. Both the transition graph and the systems equation of the SEIR epidemic model are stated in Appendix D.

The third model is the SIR epidemic model [20], which is closely related to the SEIR model. In the SIR epidemic model, the exposed compartment $\mathcal{E}$ is omitted and individuals transition the compartments in the order $\mathcal{S} \to \mathcal{I} \to \mathcal{R}$. Appendix C contains both the systems equations and the transition graph of the SIR epidemic model for completeness.

Lastly, we consider a two-staged infection process, with two different diseases and five compartments: the susceptible compartments $\mathcal{S}_l$, the infectious compartments $\mathcal{I}_l$, and the recovered (or removed) compartment $\mathcal{R}$, where $l = 1, 2$ denotes the disease. In the SISIR epidemic model, individuals transition the compartments in the order $\mathcal{S}_1 \to \mathcal{I}_1 \to \mathcal{S}_2 \to \mathcal{I}_2 \to \mathcal{R}$. There are two edge-based transitions in the SISIR model, the infectious transitions $\mathcal{S}_1 \to \mathcal{I}_1$ and $\mathcal{S}_2 \to \mathcal{I}_2$, which occur analogously to the $\mathcal{S} \to \mathcal{I}$ transition in the SIS model, but with infection rates $\beta_{l,ij}$ that depend on the respective disease $l = 1, 2$. The two nodal transitions $\mathcal{I}_1 \to \mathcal{S}_2$ and $\mathcal{I}_2 \to \mathcal{R}$ occur with the curing probabilities $\delta_{1,i}$ and $\delta_{2,i}$, respectively, for an individual in group i. The main motivation for studying the SISIR model is technical: the two contact networks corresponding to the two viruses are completely unrelated. Hence, the fact that node i can infect node j with virus 1 does not imply that node i can infect node j with virus 2. Thus, effectively a contact network with 2 N nodes has to be reconstructed from the viral state observations of N groups. Both the transition graph and the systems equation of the SISIR epidemic model are stated in Appendix E.

C. Curing Probability Control

So far, we assumed that the curing rates δ_i are constant or, equivalently, that the nodal transition probability matrices S_i do not change over time k. However, public health agencies react to an emerging epidemic outbreak by vaccinations and other disease control measures that do vary as time k evolves. In the SIS, SIR and SEIR epidemic models, we consider that the curing rates of every group i are time-dependent, i.e., the curing probability δ_i is replaced by

$$\tilde{\delta}_i[k] = \delta_i + \Delta\delta_i[k]. \qquad (6)$$

Here, the scalar $\Delta\delta_i[k] \geq 0$ is the known *curing probability control* at time k (for instance the fraction of vaccinations). Khanafer and Başar [21] studied a similar curing probability control approach to reduce the prevalence of virus. The constant curing probability term $\delta_i > 0$ in (6) corresponds to natural immunities and other influences which are unknown and have to be reconstructed from viral state observations to provide a full understanding of the viral spread. For the SISIR epidemic model, we consider that the curing rates of both diseases $l = 1, 2$ are time-dependent and equal to $\tilde{\delta}_{l,i}[k] = \Delta\delta_{l,i}[k] + \delta_{l,i}$ for every group i. For the general GEMF model (4), the concept of time-varying curing rates (6) is generalised by replacing the nodal transition probability matrix S_i by the time-dependent $M \times M$ matrix $\tilde{S}_i[k] = S_i + \Delta S_i[k]$. Here, the time-dependent $M \times M$ matrix $\Delta S_i[k]$ describes the known

[3] The equations (5) are also known as the discrete-time N-Intertwined Mean-Field Approximation (NIMFA) of the SIS process [18], [19]. 1.n this work, we refer to the system (5) as SIS model for consistency with the other special cases of the GEMF model.

controlled interventions to the viral state evolution, and the constant $M \times M$ matrix S_i is due to unknown terms of the nodal transitions. In Section V, we will show that a non-zero curing probability control $\Delta\delta_i[k] > 0$ is beneficial for the task of network reconstruction.

IV. Formulating the Network Reconstruction Problem as Linear Equations

The focus of this work is the inverse problem of estimating the parameters of the GEMF model (4) from viral state observations. More precisely:

Definition 3 (GEMF Network Reconstruction Problem): Assume that the controlled interventions $\Delta S_i[k]$ to the viral state evolution are either known or zero at every time k. Estimate the nodal transition probability matrix S_i and the edge-based transition probability matrix $B_{m,ij}$ for all nodes i, j and all compartments m from observations of the $M \times 1$ viral state vector $v_i[k]$ of every group i at every discrete time $k = 1, \ldots, n+1$, where $n \in \mathbb{N}$ denotes the *number of observed transitions*.

We emphasise that the adjacency matrices A_{pm} can be obtained from the matrices $B_{m,ij}$ by (2). For given viral state observations $v_i[1], \ldots, v_i[n+1]$ of every group i, the GEMF model (4) is linear with respect to the Laplacian matrices Q_i and $Q_{m,ij}$.

Lemma 4: Consider the GEMF model (4) with a nodal transition matrix $\tilde{S}_i[k] = S_i + \Delta S_i[k]$, where the time-varying matrix $\Delta S_i[k]$ is known (or equals zero). Denote the $M \times M$ Laplacian matrix of the known control matrix $\Delta S_i[k]$ by $\Delta Q_i^T[k] = \mathrm{diag}(\Delta S_i[k]u) - \Delta S_i[k]$. For any group i, define the $Mn \times 1$ vector V_i as

$$V_i = \begin{pmatrix} v_i[2] - v_i[1] + \Delta Q_i^T[1] v_i[1] \\ \vdots \\ v_i[n+1] - v_i[n] + \Delta Q_i^T[n] v_i[n] \end{pmatrix},$$

and define the $Mn \times M^2$ matrices $W_i, R_{m,ij}$ as

$$W_i = -(I_M \otimes v_i[1], \ldots, I_M \otimes v_i[n])^T$$

$$R_{m,ij} = (v_{j,m}[1](I_M \otimes v_i[1]), \ldots, v_{j,m}[n](I_M \otimes v_i[n]))^T.$$

Furthermore, define the $Mn \times M^2(1+NM)$ matrix F_i as

$$F_i = \left(W_i, R_{1,i1}, \ldots, R_{1,iN}, R_{2,i1}, \ldots, R_{M,iN}\right)$$

and the $M^2(1+NM) \times 1$ *GEMF parameter vector* x_i as

$$x_i = \Big(\mathrm{vec}(Q_i)^T, \mathrm{vec}(Q_{1,i1})^T, \ldots, \mathrm{vec}(Q_{1,iN})^T, \times \mathrm{vec}(Q_{2,i1})^T, \ldots, \mathrm{vec}(Q_{M,iN})^T\Big)^T.$$

Then, the GEMF parameter vector x_i satisfies the linear system

$$V_i = F_i x_i. \tag{7}$$

Proof: Appendix B. ∎

The entries of the $M^2 \times 1$ vectors $\mathrm{vec}(Q_i)$ and $\mathrm{vec}(Q_{m,ij})$ are linear combinations of the entries of the nodal transition probability matrix S_i and the edge-based transition probability matrix $B_{m,ij}$. Thus, the GEMF network reconstruction problem results in a set of equations (7) that is linear with respect to the matrices S_i and $B_{m,ij}$. For every node i, the maximum number of unknowns is bounded by the number $M^2(1+NM)$ of entries of the GEMF parameter vector x_i. However, in most cases, many entries of the matrices S_i and $B_{m,ij}$ are a-priori known to be zero, since some nodal or edge-based transitions cannot occur. For instance, at most one entry of the matrix S_i in (3) is non-zero. Furthermore, since the viral state $v_i[k]$ of every group i obeys $v_{i,1}[k] + \cdots + v_{i,M}[k] = 1$, there are n redundant equations in (7), and every N-th row of (7) can be omitted. Hence, the set of linear equations (7) can be often be expressed more compactly for particular GEMF models. To give an example, for the group-based SIS epidemic model (5) the linear system (7) can be expressed compactly as follows.

Lemma 5: For any node i, the curing probability δ_i and the infection probabilities $\beta_{1i}, \ldots, \beta_{iN}$ of the SIS epidemic model (5) with time-varying curing rates $\tilde{\delta}_i[k] = \Delta\delta_i[k] + \delta_i$ satisfy

$$V_{\mathrm{SIS},i} = F_{\mathrm{SIS}}, i(\delta_i, \beta_{i1}, \ldots, \beta_{iN})^T. \tag{8}$$

Here, the $n \times 1$ vector V_{SIS}, i equals

$$V_{\mathrm{SIS}}, i = \begin{pmatrix} \mathcal{I}_i[2] - (1 - \Delta\delta_i[1])\mathcal{I}_i[1] \\ \vdots \\ \mathcal{I}_i[n+1] - (1 - \Delta\delta_i[n])\mathcal{I}_i[n] \end{pmatrix}$$

and the $n \times (N+1)$ matrix F_{SIS}, i is given by

$$F_{\mathrm{SIS}}, i = \begin{pmatrix} -\mathcal{I}_i[1] & \mathcal{S}_i[1]\mathcal{I}_1[1] & \ldots & \mathcal{S}_i[1]\mathcal{I}_N[1] \\ \vdots & \vdots & \ddots & \vdots \\ -\mathcal{I}_i[n] & \mathcal{S}_i[n]\mathcal{I}_1[n] & \ldots & \mathcal{S}_i[n]\mathcal{I}_N[n] \end{pmatrix}.$$

Analogously to Lemma 5, we state the linear system (7) more compactly for the group-based SIR, SEIR, and SISIR epidemic models in Appendix C, Appendix D, and Appendix E, respectively.

V. The Limits of Network Reconstruction

On the first sight, it seems straightforward to infer the network from GEMF viral state observations, since the network reconstruction is equivalent to solving the linear system (7). However, as we show in the following, the linear system (7) is extremely ill-conditioned, which is a severe limitation to the GEMF network reconstruction problem itself – *regardless of the specific network reconstruction method*. In Section V-A, we discuss the limits of reconstructing large networks from GEMF viral state observations. In Section V-B, we show the dramatic impact of model errors on the network reconstruction problem.

A. Reconstruction of Large Networks

The set of linear equations (7) can, in theory, be solved exactly if the rank of the matrix F_i equals the number of unknowns. However, any computer works with finite precision arithmetic, which causes small, but non-zero, round-off errors. In the worst case, even small round-off errors can accumulate and greatly affect the accuracy of the solution of the linear system (7). To solve the linear system (7) in practice, the *numerical* rank of the matrix F_i is decisive. The numerical rank of the matrix F_i equals the number of singular values of the matrix F_i that are greater than a small threshold ϵ_{rank}, which is set in accordance to the machine precision.

We perform numerical simulations to obtain the average numerical rank of the matrix F_i for the group-based SIS, SIR, SEIR and SISIR epidemic models. For the SIS, SIR, SEIR, and SISIR epidemic models, the adjacency matrices A_{12}, A_{12}, A_{13}, and both A_{12} and A_{34}, respectively, that correspond to the contact network between infected and susceptible nodes, are generated according to the Barabási-Albert random graph model [22], where the initial number of nodes is set to $m_0 = 3$ and the number of links per addition of a new node is set to $m = 3$. Furthermore, we set $a_{ii} = 1$ for every group i of the respective adjacency matrices, since we consider that individuals in group i can infect one another. On the one hand, we consider that there is no curing probability control, i.e. $\Delta\delta_i[k] = 0$ for every group i at every time k. On the other hand, we set the curing probability control term $\Delta\delta_i[k]$ to a uniformly distributed random number in $[0, \Delta\delta_{\max}, i]$ for every group i at every time k, where the maximum control value equals $\Delta\delta_{\max}, i = 0.01\delta_{\max}, i$. Further details on the simulation parameters are given in Appendix F.

Without curing probability control, i.e. $\Delta\delta_i[k] = 0$ for every group i at every time k, Figure 3 shows that the numerical rank of the matrix F_i, computed by the Matlab command `rank`, quickly stagnates as the number of groups N grows. Thus, the linear system (7) is very ill-conditioned. For instance, for the group-based SIS model (5), the numerical rank stagnates at approximately $\text{numrank}(F_i) \approx 15$, and the linear system (8) has practically not more than 15 independent equations. *Hence, large networks cannot be reconstructed from GEMF viral state observations of a single epidemic outbreak without curing probability control*[4], which is in agreement with other works [23], [24] that consider network reconstruction for *individual-based* epidemic models. For the SISIR model, the numerical rank of the matrix F_i is approximately twice as high as for the other epidemic models, which is intuitive since the contact network for the SISIR model is effectively of size $2\,N$. With curing probability control on the other hand, the numerical rank of the matrix F_i behaves very differently. In particular, the numerical rank of the matrix F_i equals the number of unknown parameters for the SIS, SIR, SEIR and SISIR epidemic model, also for large networks. *Hence, a time-varying control of the curing rates $\delta_i[k]$ is necessary for the reconstruction of large networks.*

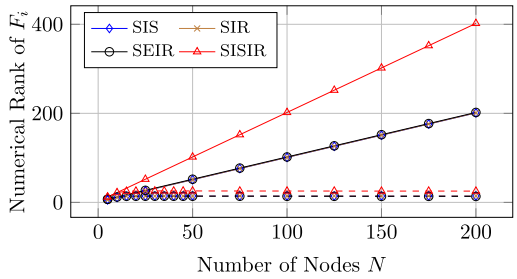

Fig. 3. The numerical rank of the matrix F_i versus the number of nodes N. The dashed and the solid lines depict the results without and with curing probability control, respectively. The results are averaged over 100 Barabási-Albert random graphs.

In theory, we see two alternatives to controlling the curing rates for the reconstruction of large networks. However, we argue that neither of these two alternatives is applicable to real-world epidemics. First, a greater number of linearly independent equations (7) can be achieved by observing multiple epidemic outbreaks [25] with different initial viral states $v[1]$. Each epidemic outbreak results in a different matrix F_i, which can be stacked such that the linear system (7) has sufficiently many independent equations. However, the numerical rank of the matrix F_i stagnates when the number of nodes N increases. Thus, the greater the network size N the more epidemic outbreaks need to be observed to reconstruct the network. We believe that it is far from practical to observe multiple outbreaks for real-world epidemics, in particular for novel viruses that demand rapid intervention.

Second, if some properties of the contact network are known a-priori, then less equations are possibly needed to solve the GEMF network reconstruction problem. For instance, if the maximum degree of a node i is upper-bounded by $d_{\max}$ and the infection rates are upper-bounded by $\beta_{ij} \leq \beta_{\max}$, then the constraint $\sum_{j=1}^{N} \beta_{ij} \leq \beta_{\max} d_{\max}$ can be included in the linear system (8) of the group-based SIS network reconstruction problem. However, the rank of the matrix F_i stagnates when the number of nodes N increases. Hence, the greater the network, the more constraints must be included in the linear system (7), which does not seem viable for a large network size N.

B. The Impact of Model Errors

A real-world virus does not exactly follow the difference equations of the GEMF model (4). Instead, the viral state $v_i[k]$ of any group i evolves according to $v_i[k+1] = f_{\text{GEMF}}, i(v_1[k], \ldots, v_N[k]) + w_i[k]$, where $f_{\text{GEMF}}, i(v_1[k], \ldots, v_N[k])$ denotes the right-hand side of (4), and the $M \times 1$ vector $w_i[k]$ denotes the *model error* at group i and time k. To ensure that $v_{i,1}[k] + \cdots + v_{i,M}[k] = 1$ at every time k, we set $w_{i,l}[k] = 0$ for exactly one compartment l. For the SIS, SIR, SEIR, and SISIR models, we choose the remaining compartment l as: $\mathcal{S}_i$, $\mathcal{S}_i$, $\mathcal{R}_i$, and $\mathcal{R}_i$, respectively.

[4] If the contact network is sparse, then compressed sensing methods [10] could be applied to reconstruct the network from the underdetermined system (7). For compressed sensing methods, the number of linearly independent equations that are required for reconstructing a network with s non-zero elements grows at least proportionally to $s \log(N/s)$. However, since the rank of the matrix F_i *stagnates* for a growing number of nodes N, also compressed sensing methods fail for large networks.

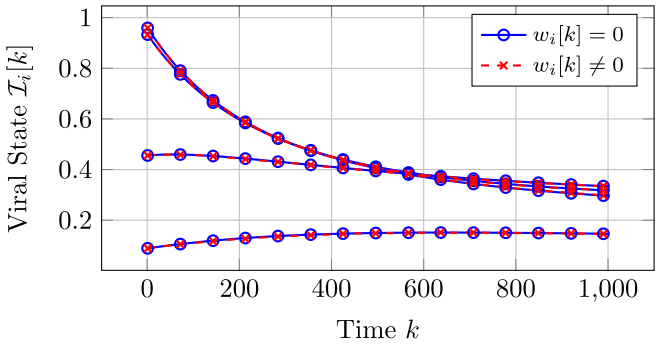

Fig. 4. The viral state $\mathcal{I}[k]$ of the group-based SIS epidemic model (5) for an Erdős-Rényi random graph with $N = 20$ groups without curing probability control ($\Delta\delta_{\max,i} = 0$), with and without model errors $w_i[k]$. The viral state $\mathcal{I}_i[k]$ of four of the twenty groups i is depicted.

To demonstrate the impact of model errors $w_i[k]$ on the network reconstruction problem, we perform numerical simulations of the SIS epidemic model (5) on a small Erdős-Rényi random graph with $N = 20$ nodes and link probability $p = 0.1$. We set all parameters to the same values as in Section V-A. We consider three cases for the maximum control value: $\Delta\delta_{\max}, i = 0$ (no curing probability control), $\Delta\delta_{\max}, i = 0.05\delta_i$, and $\Delta\delta_{\max}, i = \delta_i$. On the one hand, we consider a viral state evolution without model errors, i.e. $w_i[k] = 0$ for all nodes i and all times k. On the other hand, we consider that the SIS epidemic model (5) is subject to independently and identically distributed Gaussian model errors $w_{i,m}[k] \sim \mathcal{N}(0, \varsigma_i^2)$ with variance $\varsigma_i^2 = (0.05\Delta t)^2$.

Figure 4 illustrates that the evolution of the viral state $v_i[k]$ is virtually unaffected by the model error $w_i[k]$. If a real-world epidemic evolved with an equally small model error $w_i[k]$ as in Figure 4, then the SIS epidemic model (5) would be considered an outstanding fit to the epidemic data. On the first sight, Figure 4 suggests that it is possible to reconstruct the network from GEMF viral state observations $v_i[k]$ with a negligibly small model error $w_i[k]$. However, the GEMF network reconstruction problem is dramatically sensitive to small perturbations by model errors $w_i[k]$. The upper sub-plot in Figure 5 shows that, without curing probability control, only around five singular values $\sigma_j(F_{\text{SIS}}, i)$ of the matrix F_{SIS}, i remain largely unaffected by model errors $w_i[k]$. Hence, without curing probability control, even small networks cannot be reconstructed from GEMF viral state observations, also when the model errors $w_i[k]$ seem negligibly small[5]. The lower sub-plot in Figure 5 shows that, for a sufficiently great curing probability control $\Delta\delta_i[k]$, the model error $w_i[k]$ only slightly perturbs the singular values $\sigma_j(F_{\text{SIS}}, i)$. *Hence, curing probability control is necessary to reconstruct the network in the presence of model errors* $w_i[k]$ – but controlling the curing rates is possibly not *sufficient*, since we only studied the perturbation of the singular values $\sigma_j(F_{\text{SIS}}, i)$ but not the perturbation of the *whole* matrix F_{SIS}, i.

[5] Furthermore, the sensitivity to model errors renders model-free network inference methods [26] not suitable for the GEMF network reconstruction problem, since model-free methods per definition induce model errors.

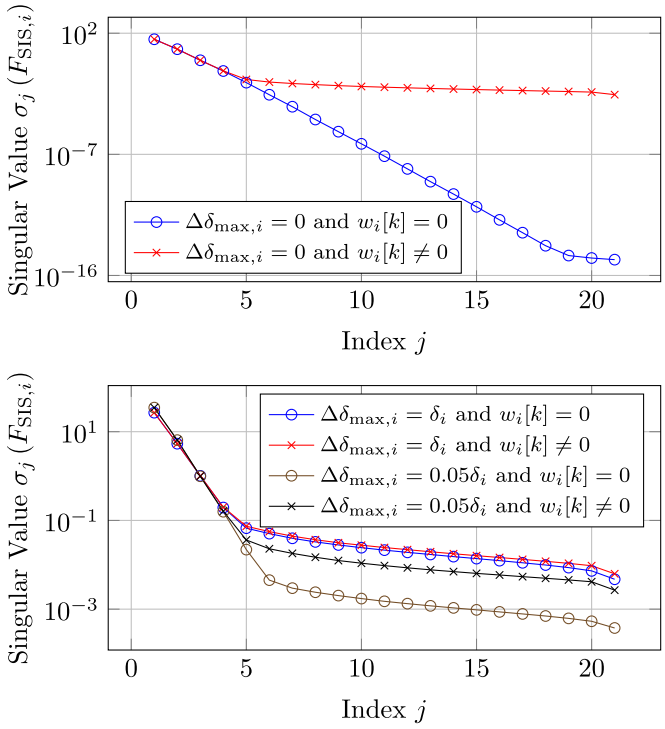

Fig. 5. The singular values $\sigma_j(F_{\text{SIS},i})$ of the matrix $F_{\text{SIS},i}$ of the linear system (8), with and without model errors $w_i[k]$. The upper sub-plot refers to no curing probability control ($\Delta\delta_{\max,i} = 0$), and the lower sub-plot considers a small and great value for the maximum value $\Delta\delta_{\max,i}$ of the curing probability control. The results are averaged over 100 Erdős-Rényi random graphs with $N = 20$ nodes.

VI. Network Reconstruction Algorithm

When the GEMF model (4) is subject to model errors $w_i[k]$, then the GEMF parameter vector x_i does not satisfy the linear system (7) with equality. Thus, we resort to finding the vector x_i as the minimiser of the Euclidean norm $\|V_i - F_i x_i\|_2^2$. More precisely, our network reconstruction method is based on the constrained LASSO [27]:

$$\begin{aligned} \hat{x}_i = \ & \arg\min_{x_i} \quad \|V_i - F_i x_i\|_2^2 + \rho_i \|x_i\|_1 \\ & \text{s.t.} \quad x_i \geq 0 \\ & (x_i)_j = 0 \quad \forall j \in \Omega_i \end{aligned} \tag{9}$$

Including the ℓ_1-regularisation term $\|x_i\|_1$ in the objective favours the estimation of a sparse GEMF parameter vector x_i, which is motivated by two reasons. First, the majority of real-world networks are indeed sparse [28]. Second, we follow the *bet on sparsity* principle: "Use a procedure does well in sparse problems, since no procedure does well in dense problems" [27]. Tuning the *regularisation parameter* $\rho_i > 0$ in the objective of (9) controls the trade-off between a good fit to the model (first addend) and the sparsity of the GEMF parameter vector x_i (second addend). We set the value of the scalar $\rho_i > 0$ by *cross-validation* [27]. In (9), the inequality $x_i \geq 0$ for the GEMF parameter vector x_i holds element-wise. The indices j in the set $\Omega_i \subset \{1, \ldots, M^2(1 + NM)\}$ refer to entries $(x_i)_j$ that must be zero for the particular GEMF model. For instance, the 3×3 nodal transition matrix S_i in (3) has

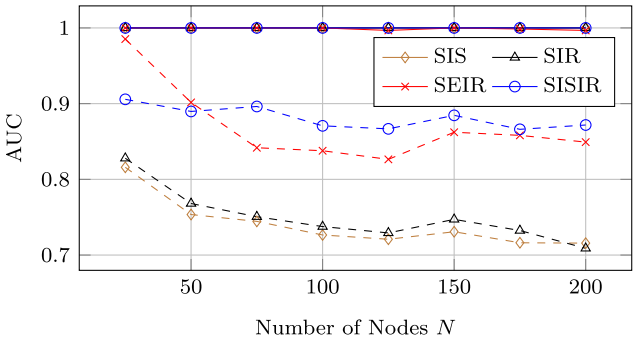

Fig. 6. The accuracy of the network reconstruction versus the network size N for four different epidemic models. The solid lines show the accuracy with curing probability control, $\Delta\delta_i[k] \neq 0$, and the dashed lines show the accuracy without curing probability control, $\Delta\delta_i[k] = 0$. The results are averaged over 100 Barabási-Albert random graphs.

eight zero entries, which results in the inclusion of eight indices in the set Ω_i. To solve (9) numerically, we apply the interior point algorithm provided by the Matlab command `quadprog`. If there are no model errors, i.e., $w_i[k] = 0$ for every group i at every time k, then we do not estimate the GEMF parameter vector x_i by the LASSO formulation (9). Instead, we apply the QR-solver provided by the Matlab command `mldivide` if the matrix F_i is of full rank, and we apply a basis pursuit approach [29] if the matrix F_i is not of full rank. For further details on the network reconstruction algorithm, we refer the reader to Appendix G.

VII. Numerical Evaluation

To evaluate the quality of the network reconstruction, we compute the area under the receiver-operating-characteristic curve (AUC) [30]. The AUC ranges from 0 to 1, where an AUC of $1/2$ is equivalent to flipping a coin to determine whether a link is presence or absence. If the estimated network equals the true network, then the AUC equals 1. We compute the AUC with respect to the estimates of the respective adjacency matrices A_{12}, A_{21}, and A_{13} of the SIS, SIR and SEIR model. For the SISIR model, we consider the mean of the two AUCs with respect to the adjacency matrices A_{12} and A_{34}. Furthermore, we define the *prediction error* $\epsilon_{\mathcal{I}}$ until the *prediction time* n_{pred} as

$$\epsilon_{\mathcal{I}} = \frac{1}{N} \frac{1}{n_{\text{pred}} - n} \sum_{k=n+1}^{n_{\text{pred}}} \sum_{i=1}^{N} \left| \mathcal{I}_i[k] - \hat{\mathcal{I}}_i[k] \right|.$$

Here, $\hat{\mathcal{I}}_i[k]$ denotes the predicted fraction of infectious individuals in group i at time k, which is obtained by iterating GEMF (4) without model errors from time $k = n$ to $k = n_{\text{pred}}$ with the parameter vector $\hat{x}_i$ that was estimated from the viral state observations $v_i[1], \ldots, v_i[n]$ for every node i. For the SISIR model, we define the prediction error $\epsilon_{\mathcal{I}}$ as the sum of the two prediction errors with respect to the two compartments $\mathcal{I}_1$ and $\mathcal{I}_2$. Unless stated otherwise, all parameters are set to the same values as in Section V-A.

A. Absence of Model Errors

For every group i, we set the maximum control value to $\Delta\delta_{\max}, i = 0.05\delta_i$ and the observation length to $n = 10\ N$.

TABLE I
The Prediction Error $\epsilon_{\mathcal{I}}$ and the AUC for Different Epidemic Models Without Curing Probability Control, Averaged Over 100 Barabási-Albert Random Graphs With $N = 200$ Nodes

	SIS	SIR	SEIR	SISIR
AUC	0.52	0.52	0.54	0.52
$\epsilon_{\mathcal{I}}$	$3.72 \cdot 10^{-4}$	$3.49 \cdot 10^{-5}$	$4.28 \cdot 10^{-5}$	$6.25 \cdot 10^{-5}$

Figure 6 shows that, without model errors $w_i[k]$, the network reconstruction is almost always exact – provided that the curing rates are controlled ($\Delta\delta_{\max}, i = 0.05\delta_i$). Without curing probability control ($\Delta\delta_{\max}, i = 0$), the reconstructed network differs considerably from the true network when the number of nodes N is large, in agreement with Figure 3.

To evaluate the prediction error $\epsilon_{\mathcal{I}}$ in the absence of curing probability control, we reduce the observation length to $n = 100$ and set the prediction time to $n_{\text{pred}} = 1000$. Table I shows that the prediction error $\epsilon_{\mathcal{I}}$ is practically zero, even though the AUC is very low. *Thus, without curing probability control, fundamentally different contact networks result in virtually the same viral state sequence*.

Figure 7 shows the absolute value of the Pearson correlation coefficient $|\text{corr}(x_i, \hat{x}_i)|$ of the entries of the i-th eigenvector x_i and the estimate $\hat{x}_i$ of the edge-based transition probability matrices between the infectious and the susceptible compartment[6]. Only the principal eigenvectors $x_1, \hat{x}_1$ are similar and the correlation between the eigenvectors $x_i, \hat{x}_i$ is very small for $i \geq 2$.

B. Presence of Model Errors

As illustrated by Figure 5, we cannot expect that an accurate network reconstruction is possible in the presence of model errors $w_i[k]$. However, Table I shows that, at least in the absence of model errors $w_i[k]$, the prediction of the epidemic outbreak is surprisingly less related to an accurate network reconstruction. We consider Barabási-Albert random graphs with $N = 100$ nodes. For every group i at every time k, we generate the model error $w_i[k]$ as a Gaussian random variable with standard deviation $\varsigma_i = 0.1\Delta t$, and we set the sampling time to $\Delta t = \Delta t_{\max}/5$, where the maximum sampling time $\Delta t_{\max}$ is given in Appendix F.

Figure 8 gives an impression on the prediction accuracy for the SIS process (5), when the network is reconstructed from the viral state sequence $v[1], \ldots, v[n]$ until the observation lengths $n = 50$ and $n = 100$, respectively. For an observation length $n = 50$, the AUC equals approximately 0.53 and the viral state prediction diverges from the true viral state $v[k]$ as time k evolves. However, the viral state prediction is accurate until discrete time $k \approx 125$, which is valuable for medium-term disease control measures. For an observation length $n = 100$, the AUC equals approximately 0.54 and the viral state prediction is relatively accurate at all times $k \geq n$ – taking the random model errors $w_i[k]$ into account. Hence, also in

[6] For the SISIR model, only the correlation of the eigenvectors corresponding to the contact graph from the infectious compartment $\mathcal{I}_1$ to the susceptible compartment $\mathcal{S}_1$ is depicted. The correlation corresponding to the compartments $\mathcal{I}_2$ and $\mathcal{S}_2$ behaves similarly.

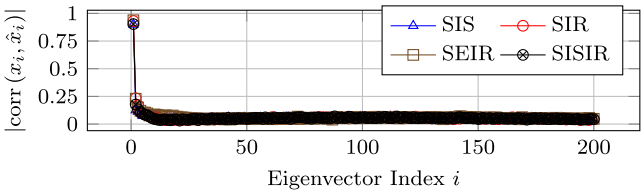

Fig. 7. The absolute value of the Pearson correlation coefficient of the entries of the eigenvectors x_i and the estimates $\hat{x}_i$, averaged over 100 Barabási-Albert random graphs with $N = 200$ nodes.

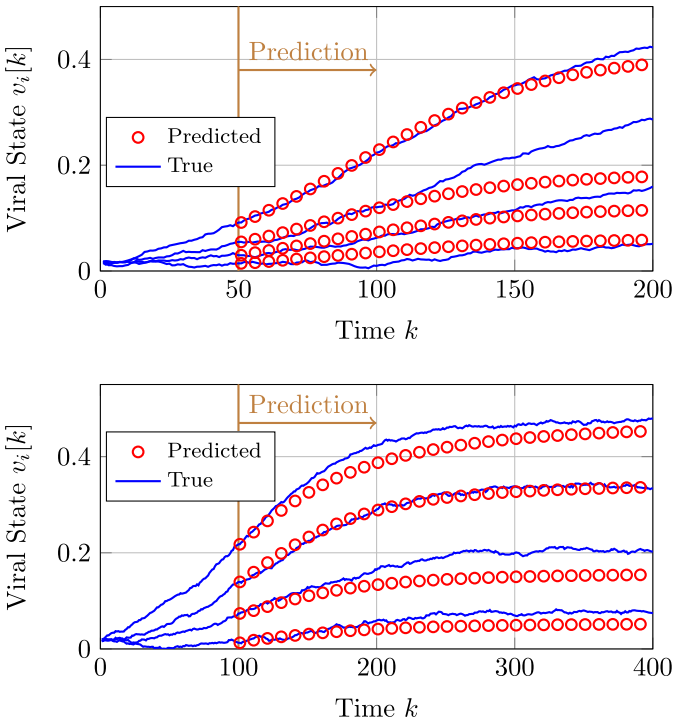

Fig. 8. The true and predicted viral state $v_i[k]$ of the SIS model (5) of four nodes of a Barabási-Albert random graph with $N = 100$ nodes subject to model errors $w_i[k]$. The upper and lower sub-plot refer to an observation length of $n = 50$ and $n = 100$, respectively.

the presence of model errors $w_i[k]$, a prediction of the viral state $v[k]$ is generally possible, and the greater the number of observations n the more accurate the long-term viral state prediction.

To evaluate the prediction accuracy versus the observation length n, we consider the contact network of the *Infectious: Stay Away* exhibition [31] with $N = 410$ nodes, accessed via the *Konect* network collection [32]. Every node i corresponds to an individual, and there is a link between two nodes if the corresponding two individuals had at least one face-to-face contact for more than 20 seconds. We set the infection rates β_{ij} proportional to the number of contacts between individual i and j, such that the infection probability β_{ij} of the two individuals i, j that had the most face-to-face contacts is three times as great as the infection probability of two individuals that only had a single face-to-face contact. The self-infection probabilities β_{ii} are set to zero for every group i. The curing probabilities δ_i are set as in Section V-A, such that the basic reproduction number equals $R_0 = 1.5$. Figure 9 shows the AUC and the prediction error ϵ_{pred} versus the observation

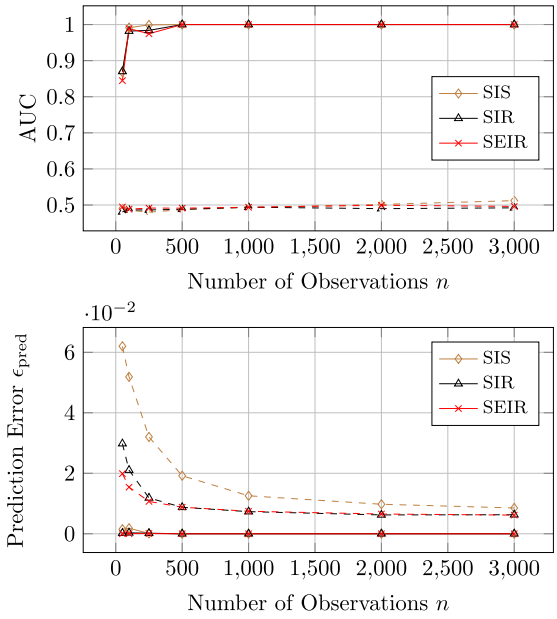

Fig. 9. The AUC and the prediction error ϵ_{pred} versus the observation length n of the contact network of the *Infectious: Stay Away* exhibition [31] with $N = 410$ nodes. The solid and dashed lines correspond to the absence and presence of model errors $w_i[k]$, respectively. The results are averaged over 10 realisations of the respective epidemic model with different initial viral states $v[1]$.

length n with and without model errors $w_i[k]$. In the presence of model errors $w_i[k]$, the prediction error ϵ_{pred} converges quickly to a small value, even though the AUC remains at around 0.5 for all observation lengths n.

VIII. Conclusions

In this work, we considered the reconstruction of the contact network and the prediction of epidemic outbreaks for general group-based compartmental epidemic models. Our contribution is composed of two parts.

In the first part, we proposed the GEMF model in discrete time, which generalises a plethora of diverse compartmental group-based epidemic models. We suggested the transition graph as an equivalent and compact visual representation of any particular GEMF model. Furthermore, the GEMF model can take multi-layer contact networks into consideration. Thus, the GEMF model is a powerful framework to study general spreading processes.

In the second part, we studied the network reconstruction problem for the GEMF model. Reconstructing the network gives rise to a set of linear equations that is severely ill-conditioned. The ill-condition of the linear system has three crucial implications for the network reconstruction from one epidemic outbreak. First, the network can only be reconstructed if there are no model errors and the curing rates are controlled. Second, fundamentally different networks result in virtually the same viral state sequences. Third, even though the contact network cannot be reconstructed without curing probability control or in the presence of model errors, the prediction of the epidemic outbreak is possible, provided that it is known by which GEMF compartmental model the viral state sequence was generated.

In summary, a given viral state sequence, and in particular small perturbations thereof, corresponds to a diverse set of potentially underlying contact networks. Furthermore, the task of predicting an epidemic outbreak is significantly easier than reconstructing the contact network. Specifying the set of contact networks that result in virtually the same viral state sequence stands on the agenda of future research.

REFERENCES

[1] R. M. Anderson and R. M. May, *Infectious Diseases of Humans: Dynamics and Control*. London, U.K.: Oxford Univ. Press, 1992.

[2] D. Bernoulli, "Essai d'une nouvelle analyse de la mortalité causée par la petite vérole et des avantages de l'inoculation pour la prévenir," *Histoire de l'Acad. Roy. Sci.(Paris) avec Mém. des Math. et Phys. and Mém*, vol. 1, pp. 1–45, 1760.

[3] J. Snow, *On the Mode of Communication of Cholera*. London, U.K.: John Churchill, 1855.

[4] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, "Epidemic processes in complex networks," *Rev. Modern Phys.*, vol. 87, no. 3, pp. 925–979, 2015.

[5] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks," *Phys. Rev. Lett.*, vol. 86, no. 14, 2001, Art. no. 3200.

[6] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, and Y. Åberg, "The web of human sexual contacts," *Nature*, vol. 411, no. 6840, pp. 907–908, 2001.

[7] T. P. Peixoto, "Network reconstruction and community detection from dynamics," *Phys. Rev. Lett.*, vol. 123, Sep. 2019, Art. no. 128301.

[8] B. Prasse and P. Van Mieghem, "Exact network reconstruction from complete SIS nodal state infection information seems infeasible," *IEEE Trans. Netw. Sci. Eng.*, vol. 6, no. 4, pp. 748–759, Oct.-Dec. 2019.

[9] A. Vajdi and C. Scoglio, "Identification of missing links using susceptible-infected-susceptible spreading traces," *IEEE Trans. Netw. Sci. Eng.*, vol. 6, no. 4, pp. 917–927, Oct.–Dec. 2019.

[10] Z. Shen, W.-X. Wang, Y. Fan, Z. Di, and Y.-C. Lai, "Reconstructing propagation networks with natural diversity and identifying hidden sources," *Nature Commun.*, vol. 5, pp. 1–10, 2014.

[11] M. Gomez Rodriguez, J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 1019–1028.

[12] S. Myers and J. Leskovec, "On the convexity of latent social network inference," in *Proc. Advances Neural Inf. Process. Syst.*, 2010, pp. 1741–1749.

[13] F. D. Sahneh, C. Scoglio, and P. Van Mieghem, "Generalized epidemic mean-field model for spreading processes over multilayer complex networks," *IEEE/ACM Trans. Netw.*, vol. 21, no. 5, pp. 1609–1620, Oct. 2013.

[14] A. Lajmanovich and J. A. Yorke, "A deterministic model for gonorrhea in a nonhomogeneous population," *Math. Biosci.*, vol. 28, no. 3-4, pp. 221–236, 1976.

[15] A. Fall, A. Iggidr, G. Sallet, and J.-J. Tewa, "Epidemiological models and Lyapunov functions," *Math. Modelling Natural Phenomena*, vol. 2, no. 1, pp. 62–83, 2007.

[16] Y. Wan, S. Roy, and A. Saberi, "Designing spatially heterogeneous strategies for control of virus spread," *IET Syst. Biol.*, vol. 2, no. 4, pp. 184–201, 2008.

[17] P. E. Paré, J. Liu, C. L. Beck, B. E. Kirwan, and T. Başar, "Analysis, estimation, and validation of discrete-time epidemic processes," *IEEE Trans. Control Syst. Technol.*, vol. 28, no. 1, pp. 79–93, Jan. 2020.

[18] P. Van Mieghem, "The N-Intertwined SIS epidemic network model," *Computing*, vol. 93, no. 2-4, pp. 147–169, 2011.

[19] B. Prasse and P. Van Mieghem, "The viral state dynamics of the discrete-time NIMFA epidemic model," *IEEE Trans. Netw. Sci. Eng.*, 2019. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8864012

[20] M. Youssef and C. Scoglio, "An individual-based approach to SIR epidemics in contact networks," *J. Theoret. Biol.*, vol. 283, no. 1, pp. 136–144, 2011.

[21] A. Khanafer and T. Başar, "On the optimal control of virus spread in networks," in *Proc. IEEE 7th Int. Conf. Netw. Games, Control Optim. (NetGCoop)*, 2014, pp. 166–172.

[22] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[23] M. Gomez-Rodriguez, L. Song, H. Daneshmand, and B. Schölkopf, "Estimating diffusion networks: Recovery conditions, sample complexity & soft-thresholding algorithm," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 3092–3120, 2016.

[24] P. Netrapalli and S. Sanghavi, "Learning the graph of epidemic cascades," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 40, no. 1, pp. 211–222, 2012.

[25] M. Timme and J. Casadiego, "Revealing networks from dynamics: An introduction," *J. Phys. A: Math. Theoret.*, vol. 47, no. 34, 2014, Art. no. 343001.

[26] J. Casadiego, M. Nitzan, S. Hallerberg, and M. Timme, "Model-free inference of direct network interactions from nonlinear collective dynamics," *Nature Commun.*, vol. 8, no. 1, 2017, Art. no. 2192.

[27] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning With Sparsity: The Lasso and Generalizations*. Boca Raton, FL, USA: CRC Press, 2015.

[28] A.-L. Barabási, *Network Science*. Cambridge, U.K.: Cambridge Univ. Press, 2016.

[29] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Rev.*, vol. 43, no. 1, pp. 129–159, 2001.

[30] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

[31] L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, and W. Van den Broeck, "What's in a crowd? Analysis of face-to-face behavioral networks," *J. Theoret. Biol.*, vol. 271, no. 1, pp. 166–180, 2011.

[32] J. Kunegis, "Konect: The Koblenz network collection," in *Proc. 22nd Int. Conf. World Wide Web*, 2013, pp. 1343–1350.

Bastian Prasse received the BSc degree in computer engineering (Dean's List) from RWTH Aachen University, in 2012, the MSc degree in systems and control theory from the Royal Institute of Technology (KTH), Stockholm, Sweden, and the MSc degree in computer engineering (Dean's List) from RWTH Aachen University, Aachen, Germany, both in 2015. He is currently working toward the Ph.D. degree since April 2017 at the Delft University of Technology, the Netherlands. His main research interests include network reconstruction and network control.

Piet Van Mieghem received the Masters (magna cum laude, 1987) and Ph.D. (summa cum laude, 1991) degrees in electrical engineering from the K.U. Leuven, Leuven, Belgium. He is a Professor at the Delft University of Technology and Chairman of the section Network Architectures and Services (NAS) since 1998. His main research interests lie in modeling and analysis of complex networks and in new Internet-like architectures and algorithms for future communications networks. Before joining Delft, he worked at the Interuniversity Micro Electronic Center (IMEC) from 1987 to 1991. During 1993-1998, he was a Member of the Alcatel Corporate Research Center in Antwerp, Belgium. He was a Visiting Scientist at MIT (1992–1993), Visiting Professor at UCLA (2005), Visiting Professor at Cornell University (2009), and at Stanford University (2015). He is the Author of four books: *Performance Analysis of Communications Networks and Systems* (Cambridge Univ. Press, 2006), *Data Communications Networking* (Techne, 2011), *Graph Spectra for Complex Networks* (Cambridge Univ. Press, 2011), and *Performance Analysis of Complex Networks and Systems* (Cambridge Univ. Press, 2014). He was Member of the Editorial Board of Computer Networks (2005–2006), the IEEE/ACM TRANSACTIONS ON NETWORKING (2008–2012), the Journal of Discrete Mathematics (2012–2014) and Computer Communications (2012–2015). Currently, he serves on the Editorial Board of the *OUP Journal of Complex Networks*.

APPENDIX A
DERIVATION OF THE DISCRETE-TIME GEMF MODEL

In Subsection A.1, we give a brief description the continuous-time GEMF model [13] for completeness. In Subsection A.2, we extend the continuous-time GEMF model to *heterogeneous* spreading parameters. In Subsection A.3, we show that applying Euler's method to the continuous-time GEMF model of Sahneh *et al.* [13] results in the discrete-time model (4).

A.1 Continuous-Time GEMF Model with Homogeneous Spreading Parameters

There are two kinds of transition in the GEMF model. First, there are *nodal transitions*. Node i changes from compartment p to compartment q with the transition rate δ_{pq}. The $M \times M$ *nodal transition rate matrix* A_δ is defined as

$$(A_\delta)_{pq} = \delta_{pq}, \quad 1 \leq p, q \leq M.$$

The second kind of transitions in the GEMF model are *edge-based* transitions. The GEMF model is formulated for multilayer networks. The layers are denoted by $l = 1, ..., L$, where L denotes the number of layers. For every layer l, there is an $N \times N$ adjacency matrix A_l with elements $a_{l,ij}$ for every pair of nodes i, j. If there is a directed link on layer l from node j to node i, then it holds $a_{l,ij} = 1$. If there is no link on layer l from node j to node i, then it holds $a_{l,ij} = 0$. To every network layer l, there is exactly one influencer compartment $c_l \in \{1, ..., M\}$. If a node i has neighbours j on graph layer l, i.e. $a_{l,ij} = 1$, which are in compartment c_l, then node i changes from compartment p to compartment q with the transition rate $\beta_{l,pq}$. For every layer l, the $M \times M$ *edge-based transition rate matrix* A_{β_l} is defined as

$$(A_{\beta_l})_{pq} = \beta_{l,pq}, \quad 1 \leq p, q \leq M.$$

The matrices A_δ and A_{β_l} are adjacency matrices and define the nodal transition rate graph and, for every layer l, an edge-based transition rate graph. The Laplacian matrix of the nodal transition rate graph and the edge-based transition rate graphs, respectively, are denoted by

$$Q_\delta = \text{diag}\left(A_\delta u\right) - A_\delta$$

and

$$Q_{\beta,l} = \text{diag}\left(A_{\beta_l} u\right) - A_{\beta_l}.$$

Finally, the GEMF model in continuous time describes the evolution of the $M \times 1$ viral state vector $v_i(t)$ as

$$\frac{dv_i(t)}{dt} = -Q_\delta^T v_i(t) - \sum_{l=1}^{L} \left(\sum_{j=1}^{N} a_{l,ij} v_{j,c_l}(t) \right) Q_{\beta,l}^T v_i(t) \tag{10}$$

for every node i. We refer the reader to [13] for further details of the GEMF model.

A.2 Continuous-Time GEMF Model with Heterogeneous Spreading Parameters

In real-world epidemics, heterogeneous spreading parameters are more likely than homogeneous spreading parameters. For instance, in an SIS epidemic process, an elderly individual is more susceptible to getting infected than younger individuals. Hence, if β_{1j} and β_{2j} denote the infection rates from an individual j to an elderly individual 1 and a younger individual 2, respectively, then it holds that $\beta_{1j} > \beta_{2j}$. Similarly, the curing rate δ_1 of an elderly individual 1 is lower than the curing rate δ_2 of a younger individual 2.

To consider heterogeneous spreading parameters, we replace the nodal transition rates δ_{pq} from compartment p to compartment q by the rates $\delta_{pq,i}$, which depend on the node i. Hence, the $M \times M$ nodal transition rate matrix A_δ is replaced by the $M \times M$ matrix $A_{\delta,i}$ whose elements are given by

$$(A_{\delta,i})_{pq} = \delta_{pq,i}, \quad 1 \leq p, q \leq M,$$

for every node i. Analogously, we replace the edge-based transition rates $\beta_{l,pq}$ from compartment p to compartment q on layer l by the rates $\beta_{l,pq,ij}$, which depend on the nodes i, j. Hence, the $M \times M$ adjacency matrix A_{β_l} of the edge-based transition rates on layer l is replaced by the $M \times M$ adjacency matrix $A_{\beta_l,ij}$ whose elements are defined by

$$(A_{\beta_l,ij})_{pq} = \beta_{l,pq,ij}, \quad 1 \leq p, q \leq M.$$

With heterogeneous spreading parameters, the GEMF model (10) becomes

$$\frac{dv_i(t)}{dt} = -Q_{\delta,i}^T v_i(t) - \sum_{l=1}^{L} \sum_{j=1}^{N} v_{j,c_l}(t) a_{l,ij} Q_{\beta,l,ij}^T v_i(t), \tag{11}$$

Here, the Laplacian matrix of the nodal transition rate graph and the edge-based transition rate graphs, respectively, with heterogeneous spreading parameters are denoted by

$$Q_{\delta,i} = \text{diag}\left(A_{\delta,i} u\right) - A_{\delta,i}$$

and

$$Q_{\beta,l,ij} = \text{diag}\left(A_{\beta,l,ij} u\right) - A_{\beta,l,ij}.$$

A.3 Discrete-Time GEMF Model with Heterogeneous Spreading Parameters

Before formulating the GEMF model in discrete time, we rewrite the differential equation (11). We define the set of layers l whose influence compartment c_l equals m as

$$\mathcal{L}_m = \{l = 1, ..., L | c_l = m\}.$$

Then, we can rewrite (11) as

$$\begin{aligned} \frac{dv_i(t)}{dt} = & - Q_{\delta,i}^T v_i(t) \\ & - \sum_{j=1}^{N} \sum_{m=1}^{M} \sum_{l \in \mathcal{L}_m} v_{j,m}(t) a_{l,ij} Q_{\beta,l,ij}^T v_i(t), \end{aligned}$$

which is equivalent to

$$\frac{dv_i(t)}{dt} = -Q_{\delta,i}^T v_i(t) - \sum_{j=1}^{N}\sum_{m=1}^{M} v_{j,m}(t)\left(\sum_{l\in\mathcal{L}_m} a_{l,ij} Q_{\beta,l,ij}^T\right) v_i(t). \quad (12)$$

Euler's method approximates the derivative as

$$\left.\frac{dv_i(t)}{dt}\right|_{t=k\Delta t} \approx \frac{v_i\left((k+1)\Delta t\right) - v_i\left(k\Delta t\right)\right)}{\Delta t} \quad (13)$$

for a small[1] sampling time Δt and a discrete time slot $k \in \mathbb{N}$. We denote $v_i[k] = v_i(k\Delta t)$ and, using Euler's method (13) with equality, obtain from (12) that

$$v_i[k+1] = v_i[k] - \Delta t Q_{\delta,i}^T v_i[k] - \sum_{j=1}^{N}\sum_{m=1}^{M} v_{j,m}[k]\left(\Delta t \sum_{l\in\mathcal{L}_m} a_{l,ij} Q_{\beta,l,ij}^T\right) v_i[k].$$

Finally, we identify the Laplacian matrices of the discrete-time GEMF model (4) as

$$Q_i = \Delta t Q_{\delta,i}$$

and

$$Q_{m,ij} = \Delta t \sum_{l\in\mathcal{L}_m} a_{l,ij} Q_{\beta,l,ij}.$$

Thus, the nodal transition probability matrix S_i and the edge-based transition probability matrix $B_{m,ij}$ of the discrete-time GEMF model (4) are related to the matrices $A_{\delta,i}, A_{\beta,l,ij}$ of the continuous-time GEMF model (11) via

$$S_i = \Delta t A_{\delta,i}$$

and

$$B_{m,ij} = \Delta t \sum_{l\in\mathcal{L}_m} a_{l,ij} A_{\beta,l,ij}. \quad (14)$$

From (14) follows that the edge-based transition probability matrix $B_{m,ij}$ describes the influence of individuals of group j in compartment m on node i, *summed* over all layers l that are in the set $\mathcal{L}_m$.

APPENDIX B
PROOF OF LEMMA 4

The GEMF model (4) model with a time-varying nodal transition matrix $\tilde{S}_i[k] = S_i + \Delta S_i[k]$ is given by

$$v_i[k+1] = \left(I_M - Q_i^T - \Delta Q_i^T[k]\right) v_i[k] - \sum_{j=1}^{N}\sum_{m=1}^{M} v_{j,m}[k] Q_{m,ij}^T v_i[k]$$

for every group $i = 1, ..., N$. Here, the $M \times M$ Laplacian matrix of the known control matrix $\Delta S_i[k]$ equals

$$\Delta Q_i^T[k] = \mathrm{diag}\left(\Delta S_i[k] u\right) - \Delta S_i[k].$$

1. For the SIS epidemic model, we derived an upper bound on the sampling time Δt that ensures the stability of the steady-state [19].

For any $M \times M$ matrix A and any $M \times 1$ vector x, it holds

$$Ax = \left(I_M \otimes x^T\right) \mathrm{vec}(A^T),$$

which follows from the definition of the matrix vectorisation and the Kronecker product. Hence, we can rewrite the GEMF equations (4) as

$$v_i[k+1] - v_i[k] + \Delta Q_i^T[k] v_i[k] = -\left(I_M \otimes v_i^T[k]\right) \mathrm{vec}(Q_i) - \sum_{j=1}^{N}\sum_{m=1}^{M} v_{j,m}[k]\left(I_M \otimes v_i^T[k]\right) \mathrm{vec}(Q_{m,ij}). \quad (15)$$

To complete the proof, we stack (15) for the observation times $k = 1, ..., n$ and obtain

$$\begin{pmatrix} v_i[2] - v_i[1] + \Delta Q_i^T[1] v_i[1] \\ \vdots \\ v_i[n+1] - v_i[n] + \Delta Q_i^T[n] v_i[n] \end{pmatrix} = - \begin{pmatrix} I_M \otimes v_i^T[1] \\ \vdots \\ I_M \otimes v_i^T[n] \end{pmatrix} \mathrm{vec}(Q_i) - \sum_{j=1}^{N}\sum_{m=1}^{M} \begin{pmatrix} v_{j,m}[1]\left(I_M \otimes v_i^T[1]\right) \\ \vdots \\ v_{j,m}[n]\left(I_M \otimes v_i^T[n]\right) \end{pmatrix} \mathrm{vec}(Q_{m,ij}).$$

APPENDIX C
SIR EPIDEMIC MODEL

Youssef and Scoglio [20] derived a mean-field approximation of the individual-based SIR model in continuous time. Applying Euler's method to the continuous-time SIR model in [20] yields the group-based SIR epidemic model, whose transition graph is given by Figure 10.

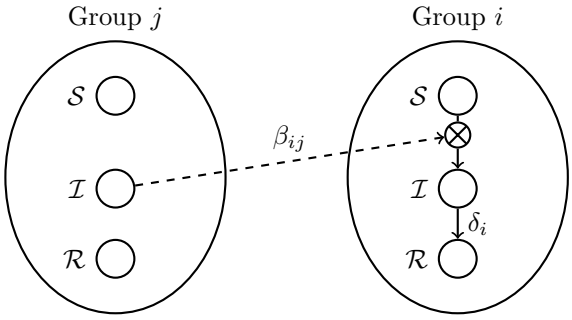

Fig. 10. The transition graph for the SIR epidemic model.

Definition 6 (SIR Epidemic Model [20])**.** *For every group i, the viral state of the discrete-time SIR epidemic model equals $v_i[k] = (\mathcal{S}_i[k], \mathcal{I}_i[k], \mathcal{R}_i[k])^T$. Here, $\mathcal{S}_i[k]$, $\mathcal{I}_i[k]$ and $\mathcal{R}_i[k]$ denote the fraction of susceptible, infected, and recovered individuals in group i at time $k \in \mathbb{N}$, respectively. For every group i, the viral state evolves in discrete time k according to*

$$\mathcal{I}_i[k+1] = (1-\delta_i)\mathcal{I}_i[k] + (1 - \mathcal{I}_i[k] - \mathcal{R}_i[k]) \sum_{j=1}^{N} \beta_{ij} \mathcal{I}_j[k]$$
$$\mathcal{R}_i[k+1] = \mathcal{R}_i[k] + \delta_i \mathcal{I}_i[k] \quad (16)$$

and the fraction of susceptible individuals follows as

$$\mathcal{S}_i[k] = 1 - \mathcal{I}_i[k] - \mathcal{R}_i[k] \tag{17}$$

at any time $k \in \mathbb{N}$. Here, β_{ij} denotes the infection probability *from group j to group i, and δ_i denotes the* curing probability *of group i.*

Stacking the SIR equations (16) for $\mathcal{I}_i[k+1]$ and $\mathcal{R}_i[k+1]$ for the observation times $k = 1, ..., n$ yields with (17) Lemma 7.

Lemma 7. *For any node i, the curing probability constant δ_i and the infection probabilities $\beta_{1i}, ..., \beta_{iN}$ of the group-based SIR epidemic model (16) with time-varying curing rates $\tilde{\delta}_i[k] = \Delta\delta_i[k] + \delta_i$ satisfy*

$$\tilde{V}_{\text{SIR},i} = F_{\text{SIR},i}\left(\delta_i, \beta_{i1}, ..., \beta_{iN}\right)^T.$$

Here, the $2n \times 1$ vector $\tilde{V}_{\text{SIR},i}$ equals

$$\tilde{V}_{\text{SIR},i} = \left(\tilde{V}^T_{\text{SIR},i}[1], ..., \tilde{V}^T_{\text{SIR},i}[n]\right)^T,$$

with the 2×1 vectors

$$\tilde{V}_{\text{SIR},i}[k] = \begin{pmatrix} \mathcal{I}_i[k+1] - (1 - \Delta\delta_i[k])\mathcal{I}_i[k] \\ \mathcal{R}_i[k+1] - (1 + \Delta\delta_i[k])\mathcal{R}_i[k] \end{pmatrix}.$$

Furthermore, the $2n \times (N+1)$ matrix $F_{\text{SIR},i}$ equals

$$F_{\text{SIR},i} = \begin{pmatrix} F^T_{\text{SIR},i}[1] & ... & F^T_{\text{SIR},i}[n] \end{pmatrix}^T$$

with the $2 \times (N+1)$ matrices

$$F_{\text{SIR},i}[k] = \begin{pmatrix} -\mathcal{I}_i[k] & \mathcal{S}_i[k]\mathcal{I}_1[k] & ... & \mathcal{S}_i[k]\mathcal{I}_N[k] \\ \mathcal{I}_i[k] & 0 & ... & 0 \end{pmatrix}.$$

APPENDIX D
SEIR EPIDEMIC MODEL

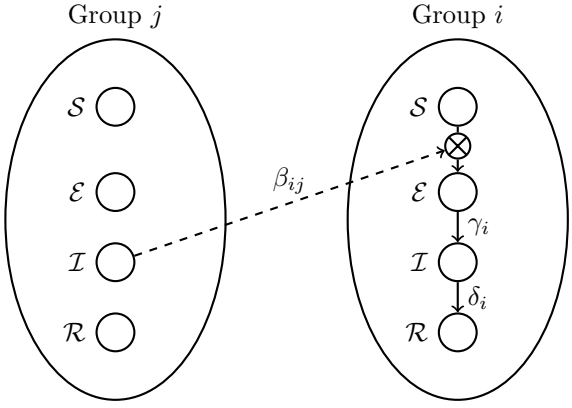

Fig. 11. The transition graph for the SEIR epidemic model.

From Figure 11, we obtain the nodal-based transition matrix S_i and the edge-based transition matrices $B_{3,ij}$ of the SEIR model as

$$S_i = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & \gamma_i & 0 \\ 0 & 0 & 0 & \delta_i \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

and

$$B_{3,ij} = \begin{pmatrix} 0 & \beta_{ij} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

For the compartments $m = 1, 2, 4$, the edge-based transition matrices equal $B_{1,ij} = B_{2,ij} = B_{4,ij} = 0$. Thus, the Laplacian matrices equal

$$Q_i = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \gamma_i & -\gamma_i & 0 \\ 0 & 0 & \delta_i & -\delta_i \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{18}$$

and

$$Q_{3,ij} = \begin{pmatrix} \beta_{ij} & -\beta_{ij} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{19}$$

For the compartments $m = 1, 2, 4$, the Laplacian matrices equal $Q_{1,ij} = Q_{2,ij} = Q_{4,ij} = 0$. With (18) and (19), the SEIR model specified by Figure 11 follows with (4) as:

Definition 8 (SEIR Epidemic Model)**.** *For every group i, the viral state of the SEIR epidemic model equals $v_i[k] = (\mathcal{S}_i[k], \mathcal{E}_i[k], \mathcal{I}_i[k], \mathcal{R}_i[k])^T$. Here, $\mathcal{S}_i[k]$, $\mathcal{E}_i[k]$, $\mathcal{I}_i[k]$ and $\mathcal{R}_i[k]$ denote the fraction of susceptible, exposed, infectious, and recovered individuals in group i at time $k \in \mathbb{N}$, respectively. For every group i, the viral state evolves in discrete time k according to*

$$\begin{aligned} \mathcal{S}_i[k+1] &= \mathcal{S}_i[k] - \mathcal{S}_i[k] \sum_{j=1}^{N} \beta_{ij}\mathcal{I}_j[k] \\ \mathcal{E}_i[k+1] &= (1 - \gamma_i)\mathcal{E}_i[k] + \mathcal{S}_i[k] \sum_{j=1}^{N} \beta_{ij}\mathcal{I}_i[k] \\ \mathcal{I}_i[k+1] &= (1 - \delta_i)\mathcal{I}_i[k] + \gamma_i\mathcal{E}_i[k] \end{aligned} \tag{20}$$

and the fraction of recovered individuals follows as

$$\mathcal{R}_i[k] = 1 - \mathcal{S}_i[k] - \mathcal{E}_i[k] - \mathcal{I}_i[k]$$

at any time $k \in \mathbb{N}$. Here, β_{ij} denotes the infection probability *from group j to group i, γ_i denotes the* incubation probability *of group i, and δ_i denotes the* curing probability *of group i.*

Stacking the SEIR equations (20) for $\mathcal{S}_i[k+1]$, $\mathcal{E}_i[k+1]$ and $\mathcal{I}_i[k+1]$ for the observation times $k = 1, ..., n$ yields Lemma 9.

Lemma 9. *For any node i, the incubation probability γ_i, the curing probability constant δ_i and the infection probabilities $\beta_{1i}, ..., \beta_{iN}$ of the group-based SIR epidemic model (20) with time-varying curing rates $\tilde{\delta}_i[k] = \Delta\delta_i[k] + \delta_i$ satisfy*

$$\tilde{V}_{\text{SEIR},i} = F_{\text{SEIR},i}\left(\gamma_i, \delta_i, \beta_{i1}, ..., \beta_{iN}\right)^T.$$

Here, the $3n \times 1$ vector $\tilde{V}_{\text{SEIR},i}$ equals

$$\tilde{V}_{\text{SEIR},i} = \begin{pmatrix} \tilde{V}^T_{\text{SEIR},i}[1] & ... & \tilde{V}^T_{\text{SEIR},i}[n] \end{pmatrix}^T,$$

with the 3×1 *vectors*

$$\tilde{V}_{\text{SEIR},i}[k] = \begin{pmatrix} \mathcal{S}_i[k+1] - \mathcal{S}_i[k] \\ \mathcal{E}_i[k+1] - \mathcal{E}_i[k] \\ \mathcal{I}_i[k+1] - (1 - \Delta\delta_i[k])\mathcal{I}_i[k] \end{pmatrix}.$$

Furthermore, the $3n \times (N+2)$ *matrix* $F_{\text{SEIR},i}$ *equals*

$$F_{\text{SEIR},i} = \begin{pmatrix} F_{\text{SEIR},i}^T[1] & \dots & F_{\text{SEIR},i}^T[n] \end{pmatrix}^T$$

with the $3 \times (N+2)$ *matrices*

$$F_{\text{SEIR},i}[k] = \begin{pmatrix} 0 & 0 & -\mathcal{S}_i[k]\mathcal{I}_1[k] & \dots & -\mathcal{S}_i[k]\mathcal{I}_N[k] \\ -\mathcal{E}_i[k] & 0 & \mathcal{S}_i[k]\mathcal{I}_1[k] & \dots & \mathcal{S}_i[k]\mathcal{I}_N[k] \\ \mathcal{E}_i[k] & -\mathcal{I}_i[k] & 0 & \dots & 0 \end{pmatrix}.$$

APPENDIX E
SISIR EPIDEMIC MODEL

An exemplary application of the SISIR model is the description of two viruses, which spread outside and inside a quarantine. The state $\mathcal{S}_1$ corresponds to healthy individuals. Individuals that are infected by the first virus are in the state $\mathcal{I}_1$ and are moved upon detection of the infection (with the curing probability δ_1) to the state $\mathcal{S}_2$, which corresponds to the quarantine. In the quarantine, the spread of another virus takes place, which is modelled by an SIR process (the states $\mathcal{S}_2$,$\mathcal{I}_2$,$\mathcal{R}$).

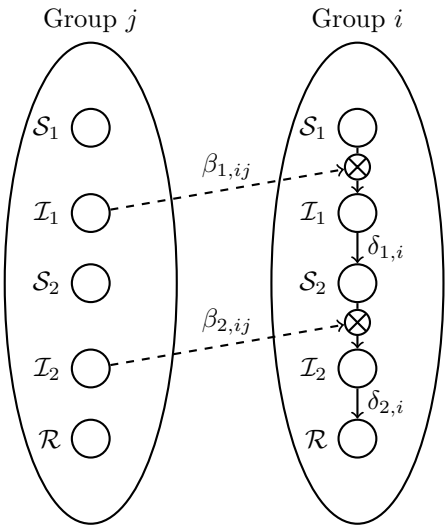

Fig. 12. The transition graph for the SISIR epidemic model.

From Figure 12, we obtain the nodal-based transition matrix S_i and the edge-based transition matrices $B_{2,ij}$ and $B_{4,ij}$ of the SISIR model as

$$S_i = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \delta_{1,i} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \delta_{2,i} \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

and

$$B_{2,ij} = \begin{pmatrix} 0 & \beta_{1,ij} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

and

$$B_{4,ij} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \beta_{2,ij} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

For the compartments $m = 1, 3, 5$, the edge-based transition matrices and their Laplacians equal $B_{1,ij} = B_{3,ij} = B_{5,ij} = 0$ and $Q_{1,ij} = Q_{3,ij} = Q_{5,ij} = 0$. The other Laplacian matrices equal

$$Q_i = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & \delta_{1,i} & -\delta_{1,i} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \delta_{2,i} & -\delta_{2,i} \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

and

$$Q_{2,ij} = \begin{pmatrix} \beta_{1,ij} & -\beta_{1,ij} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

and

$$Q_{4,ij} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \beta_{2,ij} & -\beta_{2,ij} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Thus, the SISIR specified by Figure 12 follows with (4) as:

Definition 10 (SISIR Epidemic Model)**.** *For every group* i*, the viral state of the SISIR epidemic model equals* $v_i[k] = (\mathcal{S}_{1,i}[k], \mathcal{I}_{1,i}[k], \mathcal{S}_{2,i}[k], \mathcal{I}_{2,i}[k], \mathcal{R}_i[k])^T$*. Here,* $\mathcal{S}_{l,i}[k]$ *and* $\mathcal{I}_{l,i}[k]$ *denote the fraction of individuals in group* i *at time* $k \in \mathbb{N}$ *that are susceptible to and infected by disease* $l = 1, 2$*, respectively. At time* $k \in \mathbb{N}$*,* $\mathcal{R}_i[k]$ *denotes the fraction of recovered individuals in group* i*. For every group* i*, the viral state evolves in discrete time* k *according to*

$$\mathcal{S}_{1,i}[k+1] = \mathcal{S}_{1,i}[k] - \mathcal{S}_{1,i}[k] \sum_{j=1}^{N} \beta_{1,ij} \mathcal{I}_{1,j}[k] \tag{21}$$

$$\mathcal{I}_{1,i}[k+1] = (1 - \delta_{1,i})\mathcal{I}_{1,i}[k] + \mathcal{S}_{1,i}[k] \sum_{j=1}^{N} \beta_{1,ij} \mathcal{I}_{1,j}[k]$$

$$\mathcal{S}_{2,i}[k+1] = \mathcal{S}_{2,i}[k] + \delta_{1,i}\mathcal{I}_{1,i}[k] - \mathcal{S}_{2,i}[k] \sum_{j=1}^{N} \beta_{2,ij} \mathcal{I}_{2,j}[k]$$

$$\mathcal{I}_{2,i}[k+1] = (1 - \delta_{2,i})\mathcal{I}_{2,i}[k] + \mathcal{S}_{2,i}[k] \sum_{j=1}^{N} \beta_{2,ij} \mathcal{I}_{2,j}[k]$$

and the fraction of recovered individuals follows as

$$\mathcal{R}_i[k] = 1 - \mathcal{S}_{1,i}[k] - \mathcal{I}_{1,i}[k] - \mathcal{S}_{2,i}[k] - \mathcal{I}_{2,i}[k]$$

at any time $k \in \mathbb{N}$. Here, for the disease $l = 1, 2$, $\beta_{l,ij}$ denotes the infection probability *from group j to group i, and $\delta_{l,i}$ denotes the* curing probability *of group i.*

Stacking the SISIR equations (21) for $\mathcal{S}_{l,i}[k+1]$ and $\mathcal{I}_{l,i}[k+1]$, where $l = 1, 2$, for the observation times $k = 1, ..., n$ yields Lemma 11.

Lemma 11. *For any node i, the curing probability constant $c_{l,i}$ and the infection probabilities $\beta_{l,1i}, ..., \beta_{l,iN}$, for both diseases $l = 1, 2$, of the group-based SISIR epidemic model (21) with time-varying curing rates $\tilde{\delta}_{l,i}[k] = \Delta\delta_{l,i}[k] + \delta_{l,i}$ satisfy*

$$\tilde{V}_{\text{SISIR},i} = F_{\text{SISIR},i} \left(\delta_{1,i}, \delta_{2,i}, \beta_{1,i1}, ..., \beta_{1,iN}, \beta_{2,i1}, ..., \beta_{2,iN}\right)^T.$$

Here, the $4n \times 1$ vector $\tilde{V}_{\text{SISIR},i}$ equals

$$\tilde{V}_{\text{SISIR},i} = \left(\tilde{V}^T_{\text{SISIR},i}[1] \quad ... \quad \tilde{V}^T_{\text{SISIR},i}[n]\right)^T,$$

with the 4×1 vectors

$$\tilde{V}_{\text{SISIR},i}[k] = \begin{pmatrix} \mathcal{S}_{1,i}[k+1] - \mathcal{S}_{1,i}[k] \\ \mathcal{I}_{1,i}[k+1] - (1 - \Delta\delta_{1,i}[k])\mathcal{I}_{1,i}[k] \\ \mathcal{S}_{2,i}[k+1] - \mathcal{S}_{2,i}[k] - \Delta\delta_{1,i}[k]\mathcal{I}_{1,i}[k] \\ \mathcal{I}_{2,i}[k+1] - (1 - \Delta\delta_{2,i}[k])\mathcal{I}_{2,i}[k] \end{pmatrix}$$

for any time $k = 1, ..., n$. Furthermore, the $4n \times (2N+2)$ matrix $F_{\text{SISIR},i}$ equals

$$F_{\text{SISIR},i} = \left(F^T_{\text{SISIR},i}[1] \quad ... \quad F^T_{\text{SISIR},i}[n]\right)^T$$

where the $4 \times (2N+2)$ matrices $F_{\text{SISIR},i}[k]$ are given by

$$F_{\text{SISIR},i}[k] = \begin{pmatrix} 0 & 0 & -F^{(1)}_{\text{SISIR},i}[k] & 0 \\ -\mathcal{I}_{1,i}[k] & 0 & F^{(1)}_{\text{SISIR},i}[k] & 0 \\ \mathcal{I}_{1,i}[k] & 0 & 0 & -F^{(2)}_{\text{SISIR},i}[k] \\ 0 & -\mathcal{I}_{2,i}[k] & 0 & F^{(2)}_{\text{SISIR},i}[k] \end{pmatrix}$$

for any time $k = 1, ..., n$, and the $1 \times N$ vectors $F^{(l)}_{\text{SISIR},i}[k]$ equal

$$F^{(l)}_{\text{SISIR},i}[k] = \left(\mathcal{S}_{l,i}[k]\mathcal{I}_{l,1}[k] \quad ... \quad \mathcal{S}_{l,i}[k]\mathcal{I}_{l,N}[k]\right).$$

APPENDIX F
SIMULATION PARAMETERS

We describe the precise setting of the simulation parameters for Subsection 5.1. In [19], an upper bound $\Delta t_{\max}$ on the sampling time Δt of Euler's method was derived that ensures the stability of the steady-state $\mathcal{I}_\infty$ of the discrete-time SIS[2] epidemic model (5). We set the sampling time to $\Delta t = \Delta t_{\max}/100$. For the Barabási-Albert of Figure 3, the resulting sampling time Δt ranges from $5.4 \cdot 10^{-4}$ to $1.2 \cdot 10^{-3}$. If there is a link between node i and j, then we set the infection rates β_{ij} and β_{ji} (respectively, $\beta_{l,ij}$ and $\beta_{l,ji}$ for the SISIR model) to a uniformly distributed random number in $[0.5\Delta t, 0.6\Delta t]$. Hence, $\beta_{ij} \neq \beta_{ji}$ holds in general, and $\beta_{ii} > 0$ due to infections between individuals in the same group i. If there is no link between node i and j, then we set the infection rates to $\beta_{ij} = 0$ and $\beta_{ji} = 0$. We set the "initial curing rates" $\delta_i^{(0)}$ to a uniformly distributed random number in $[0.5\Delta t, 0.6\Delta t]$. Then, we set the curing rates δ_i to a multiple of the initial curing rates $\delta_i^{(0)}$, i.e. $\delta_i = c\delta_i^{(0)}$ for every node i and some scalar c such that the basic reproduction number equals $R_0 = 1.5$. For every group i of the SIS, SIR and SEIR epidemic models, the initial fraction of infected individuals $\mathcal{I}_i[1]$ is set to a uniformly distributed random number in $[0, 1]$, and the initial fraction of susceptible individuals $\mathcal{S}_i[1]$ is set to $\mathcal{S}_i[1] = 1 - \mathcal{I}_i[1]$. For the SEIR epidemic model, the incubation probability γ_i is set to a uniformly distributed number in $[0.5\Delta t, 0.6\Delta t]$ For every group i and both diseases $l = 1, 2$ in the SISIR epidemic model, the initial fractions of infected individuals $\mathcal{I}_{l,i}[1]$ is set to a uniformly distributed random number in $[0, 0.5]$, and the initial fraction of susceptible individuals is set to $\mathcal{S}_{l,i}[1] = (1 - \mathcal{I}_{1,i}[1] - \mathcal{I}_{2,i}[1])/2$ for $l = 1, 2$. Hence, the initial fraction of recovered individuals in the SIR, SEIR and SISIR model and the initial fraction of exposed individuals in the SEIR model are $\mathcal{R}_i[1] = 0$ and $\mathcal{E}_i[1] = 0$, respectively. The observation length is set to $n = 1000$.

2. The stability of the equilibria of the general discrete-time GEMF model (4) is an open question.

APPENDIX G
DETAILS OF THE NETWORK RECONSTRUCTION ALGORITHM

In Subsection G.1, we provide the details of the network reconstruction algorithm in the presence of model errors $w_i[k]$. Subsection G.2 gives the network reconstruction algorithm in case of no model errors $w_i[k]$.

G.1 Network Reconstruction in the Presence of Model Errors

The GEMF parameter vector x_i in the LASSO problem (9) implicitly depends on the sampling time Δt. For instance, the SIS model (5) reads

$$\begin{aligned}\mathcal{I}_i[k+1] =& (1 - \Delta t\delta_{\text{cont},i})\mathcal{I}_i[k] \\ &+ (1 - \mathcal{I}_i[k]) \sum_{j=1}^{N} \Delta t\beta_{\text{cont},ij}\mathcal{I}_j[k],\end{aligned}$$

where the *continuous-time* spreading parameters $\delta_{\text{cont},i}, \beta_{\text{cont},ij}$ and the discrete-time spreading parameters δ_i, β_{ij} are related via $\delta_{\text{cont},i} = \delta_i/\Delta t$ and $\beta_{\text{cont},ij} = \beta_{ij}/\Delta t$. More generally, the GEMF parameter vector x_i equals $x_i = x_{\text{cont},i}\Delta t$ for some vector $x_{\text{cont},i}$ that is independent of the sampling time Δt. The sampling time Δt is not related to the contact network and, hence, should not have an influence on the estimation of the GEMF parameter vector x_i. By multiplying the objective in the LASSO problem (9) by $(\Delta t)^{-2}$, we obtain an equivalent optimisation problem as

$$\begin{aligned} \min_{x_{\text{cont},i}} \quad & \left\| \frac{1}{\Delta t} V_i - F_i x_{\text{cont},i} \right\|_2^2 + \frac{\rho_i}{\Delta t} \left\| x_{\text{cont},i} \right\|_1 \\ \text{s.t.} \quad & x_{\text{cont},i} \geq 0 \\ & \left(x_{\text{cont},i}\right)_j = 0 \quad \forall j \in \Omega_i \end{aligned} \tag{22}$$

In contrast to (9), the LASSO problem (22) is independent of the sampling time Δt. The estimate $\hat{x}_i(\rho_i)$ for the GEMF parameter vector x_i follows from multiplying the solution of (22) by the sampling time Δt.

To set the regularisation parameter $\rho_i > 0$ in the LASSO problem (22) by cross-validation, we generate candidate values for ρ_i that are logarithmically equidistantly

spaced from $\rho_{\min,i}$ to $\rho_{\max,i}$. We set the maximum value to $\rho_{\max,i} = 2\frac{1}{\Delta t}\|F_i^T V_i\|_\infty \cdot 10^{-4}$ and the minimum value to $\rho_{\min,i} = 10^{-4}\rho_{\max,i}$. We denote the set of all candidate values for the scalar ρ_i by $\Theta_i = \{\rho_{\min,i}, ..., \rho_{\max,i}\}$. We apply cross validation [27] to obtain the value of the scalar ρ_i in the set Θ_i that results in the minimum mean squared error $\|V_i - F_i x_i\|_2^2$. Our network reconstruction method is given in pseudo-code by Algorithm 1.

Algorithm 1 GEMF Network Reconstruction

Input: viral states $v_i[1]$, ..., $v_i[n+1]$ and $\Delta S_i[1]$, ..., $\Delta S_i[n]$ for all nodes i
Output: estimate for the GEMF parameter vector $\hat{x}_i$ for all nodes i
for $i = 1, ..., N$ **do**
  $\rho_{\max,i} \leftarrow 2\|F_i^T V_i\|_\infty \cdot 10^{-8}$
  $\rho_{\min,i} \leftarrow 10^{-4}\rho_{\max,i}$
  $\Theta_i \leftarrow$ 20 logarithmically equidistant values from $\rho_{\min,i}$ to $\rho_{\max,i}$
  for $\rho_i \in \Theta_i$ **do**
    estimate $\mathrm{MSE}\,(\hat{x}_i(\rho_i))$ by 5-fold cross validation on F_i, V_i and solving (22) on the respective training set
  end for
  $\rho_{\text{opt},i} \leftarrow$ minimiser of the estimates of $\mathrm{MSE}\,(\hat{x}_i(\rho_i))$
  $\hat{x}_i \leftarrow$ the solution $\hat{x}_i(\rho_{\text{opt},i})$ to (22) on the whole data set F_i, V_i
end for

G.2 Network Reconstruction in the Absence of Model Errors

If there are no model errors, i.e. $w_i[k] = 0$ at every time k for every group i, then the linear system (7) is satisfied with equality. Depending on the (numerical) rank of the matrix F_i, we employ two methods to estimate the parameter vector x_i. First, if the rank of the matrix F_i equals the number of unknown components of the parameter vector x_i, then we solve the linear system $F_i x_i = V_i$ with the QR-solver provided by the Matlab command `mldivide`. Second, if the rank of the matrix F_i is lower than the number of unknown components of the parameter vector x_i, then we estimate the parameter vector x_i by the basis pursuit [29] approach:

$$\begin{aligned} \hat{x}_i = \arg\min_{x_i} \quad & \|x_i\|_1 \\ \text{s.t.} \quad & F_i x_i = V_i \\ & x_i \geq 0 \\ & (x_i)_j = 0 \quad \forall j \in \Omega_i \end{aligned} \tag{23}$$

To solve the linear programme (23) numerically, we apply the dual simplex algorithm provided by the Matlab command `linprog`.