IISE Transactions on Healthcare Systems Engineering

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uhse21

Inference of human contact networks based on epidemic data

O. Arda Vanli & Davison Elijah Tsekeni

To cite this article: O. Arda Vanli & Davison Elijah Tsekeni (16 Sep 2024): Inference of human contact networks based on epidemic data, IISE Transactions on Healthcare Systems Engineering, DOI: 10.1080/24725579.2024.2401466

To link to this article: https://doi.org/10.1080/24725579.2024.2401466

Published online: 16 Sep 2024.

IISE TRANSACTIONS ON HEALTHCARE SYSTEMS ENGINEERING
https://doi.org/10.1080/24725579.2024.2401466

Inference of human contact networks based on epidemic data

O. Arda Vanli and Davison Elijah Tsekeni

Department of Industrial and Manufacturing Engineering, Florida A&M University, Florida State University, Tallahassee, FL, USA

ABSTRACT

A key challenge in epidemic modeling is the lack of adequate data on population interactions (such as traffic flow or mobility patterns) which result in the spread of infectious diseases. Knowledge of social contact patterns is crucial for public health professionals to devise effective non-pharmaceutical interventions to control epidemics. This paper focuses on inferring social contact rates from reported infection counts during the spread of an infectious disease, addressing the increased difficulty that arises when dealing with "sparse" contact networks where only a small subset of the edges have non-zero weights. Specifically, a new geographically constrained lasso approach for network reconstruction for non-homogeneous mixing Susceptible-Infected-Removed (SIR) disease spread models is presented. The new network reconstruction method can explicitly account for the spatial proximity of network nodes in estimating the disease transmission rates and predicting the future evolution of the epidemic dynamics. Extensive numerical experiments are presented to show the proposed method outperforms existing approaches in terms of accuracy of contact identification under various graph topologies. A case study based on real data from the COVID-19 pandemic is presented to demonstrate the application of the approach for inferring contact structures and a counterfactual scenario analysis to assess effectiveness of containment strategies.

KEYWORDS

Compartmental models; epidemics on networks; infectious disease modeling; contact networks; stochastic SIR model

1. Introduction

Social network modeling has received significant attention in the literature due to its effectiveness in representing communications, contacts, or interactions between individuals, communities, or organizations. Efficient analysis of data from social networks is crucial for developing an understanding of the dynamics and structural changes of many network phenomena, in particular in health sciences where it is if interest to model and understand how human contacts impact the spread of infectious diseases. The overwhelming majority of the infectious disease models originate from the basic compartmental models introduced in the early twentieth century (Kermack & McKendrick, 1927) and categorize individuals according to their status with respect to an infectious disease to describe the way they transition from compartments to compartments as the disease spreads through contacts. These early scalar models, which assume homogeneous or "random" mixing, have later been enhanced by defining them on spatial patches of metapopulations to allow for movements of individuals (Arino, 2009) or by embedding them into networks (Pastor-Satorras et al., 2015) to more realistically represent the influence of contacts between individuals on the spread of the disease. In addition to infectious disease analysis, these models were also shown to be applicable for the study of the "spread" or adoption of new behaviors on social networks, including the spread of obesity (Hill et al., 2010), mass shootings (Towers et al., 2015), and smoking, alcohol consumption, and depression (Christakis & Fowler, 2013), among others. The focus of this paper is on making inferences about contact rates among individuals or sub-populations during the spread of an infectious disease based on observed infection count data using network models. Making accurate inferences on contact rates will allow the public health practitioners to test hypotheses about transmission of certain diseases, thereby enabling them to devise better intervention and containment strategies (e.g., by identifying subregions in which the largest transmissions take place).

The basic Susceptible-Infected-Removed (SIR) compartmentalization-based epidemiological models assume a homogeneous (or random) mixing of individuals, which implies that all members of the population have the same contact pattern and the disease transmission rate through the population is uniform. This assumption is often violated in real-world epidemics, however. For example, the per-capita contact rate parameter, assumed to be a constant in the basic SIR model, is shown to vary significantly between young/old persons and urban/rural settings (Vynnycky & White, 2010). Recent studies have shown that neglecting the age group or spatial heterogeneity in population mixing and contact patterns can introduce large errors in predictions of local pandemic timing and magnitudes even though the

CONTACT O. Arda Vanli oavanli@eng.fsu.edu Department of Industrial and Manufacturing Engineering, Florida A&M University, Florida State University, Tallahassee, FL, USA.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/uhse.

aggregate behavior at larger scales mirrors a basic SIR-like pattern. A possible solution to this problem is to incorporate a social contact matrix in the disease spread models (Fumanelli et al., 2012) and account for non-homogeneous mixing patterns (Vynnycky & White, 2010). These approaches allow estimating the non-uniform transmission rates between members of different age groups (Mossong et al., 2008; Wallinga et al., 2006), however, require conducting large social contact surveys, including diary-based surveys (Read et al., 2008), web-based surveys (Smieszek et al., 2014) and wireless sensors (Salathé et al., 2010). For implementing non-pharmaceutical interventions and disease control strategies, SIR compartmental models have been particularly effective. To quantify associations between timings of stay-at-home orders and population mobility, epidemic models have been integrated with mobility data (Audirac et al., 2022). Interventions in the form of vaccination, reduction of local contact rates, and restriction on long-range movement were effective in controlling the disease spread (Wan et al., 2008). A common approach in modeling the impact of interventions in SIR models is to assume that the disease transmission parameter is constant up to the time point when the control measure is introduced and after that, it decays to some specified value according to a parametric function (Chowell et al., 2004).

In understanding the dynamics and control of infectious diseases with SIR models the most commonly applied method involves imposing a general structure of modeling person to person transmission, in the form of the Who Acquires Infection From Whom (WAIFW) matrix (Vynnycky & White, 2010), which represents prior assumptions about the mixing patterns in the population. Accurate quantification of the social contact and transmission rates is highly critical in the calculation of important epidemic parameters and the resulting estimates of intervention, such as the basic reproductive number and the minimum immunization coverages (Van Effelterre et al., 2009). In the absence of direct observations on relevant mixing patterns, such a simplifying structure is required to estimate the contact rates between different groups of populations (Hens et al., 2009; Wallinga et al., 2006), however, limitations remain for estimating the contact rates from limited epidemiological data. To address some of these limitations, this paper investigates a new network reconstruction method to infer contact rates and their impacts on disease spread.

Network-based approaches (Kiss et al., 2017; Kolaczyk & Csárdi, 2014) provide a convenient framework to incorporate the heterogeneity in contact rates in mathematical modeling of the SIR epidemics without the need for large social contact surveys. The lack of accurate data on contact patterns is usually handled in these approaches by using random graphs and regular lattices on which the epidemic processes are represented (Colizza et al., 2007). Epidemic processes on networks assume that spreading occurs from one individual to another if they come into contact. Therefore, the network-based epidemic models often need to be constructed at the individual level, and the contact graph between individuals is one of the most critical parameters of the model for making inferences about the virus's spread (Pastor-Satorras et al., 2015). Network-based epidemic analysis allows the quantification of high-level network properties such as the degree distribution (Bansal et al., 2007). In addition, "approximate" contact network modeling utilizing surveys of individuals (e.g., Eubank et al., 2004), census data (e.g., Meyers et al., 2003), or other collected data (e.g., Meyers et al., 2003) was shown to be helpful, for the cases where knowledge on every disease-causing contact between individuals is not available. To overcome the challenges of incomplete information on the contact network of individual-based models, several authors (Prasse & Van Mieghem, 2020) describe the evolution of the virus on a coarser level between groups, or communities, of similar individuals.

In studying the effect of mixing or mobility on the spread of an infectious disease, an extensively studied class of epidemic models are the metapopulation models (Arino, 2009; Keeling et al., 2004). A metapopulation model divides the region or population under consideration into distinct patches and uses a different epidemic model to describe the spread of an infectious disease among the members of each patch. The patches are then coupled with connections representing the possibility for individuals in the various epidemiological compartments to travel between locations. Metapopulation models themselves can be considered as networks with the groups represented by nodes and the interactions among groups represented as the edges. Recently it has been demonstrated by Colizza et al. (2007) that metapopulation models can be viewed as a special case of reaction-diffusion processes (particles representing people moving between different locations) which allows for an efficient solution of the coupled infection dynamics.

Significant advances in the control theory for epidemics on networks (see e.g., Fu et al., 2013; Small & Cavanagh, 2020) has provided methodology for stability analysis of control strategies. For example, Hota et al. (2021), presented a discrete-time SIR epidemic model defined on a network, a framework similar to ours, to develop closed-loop control. This line of research mainly focuses on the stability and optimality of controllers, however, and does not address the main challenges in estimation of contact rate parameters among large number of regions. Diffusion of awareness is another critical aspect that can be studied on networks (Fu et al., 2013, p. 184). When a disease breaks out in a human population, changes in behavior in response to the outbreak can alter the progression of the infectious agent: people aware of a disease in their proximity can take measures to reduce their susceptibility. As an example, Wang et al. (2019) showed how the propagation of disease awareness through a network can significantly alter the epidemic threshold and disease prevalence. The increasingly powerful artificial intelligence (AI) and deep learning algorithms have also been successfully adopted in infectious disease modeling. AI methods build on the application of physics informed neural networks (PINNs) to efficiently solve nonlinear ordinary and partial differential equation systems that incorporate transportation between populations and their impact on the dynamics of infectious diseases (Raissi et al.,

2019). While these significant advances provide enhanced ability to mathematical model and forecast infectious diseases, they do not readily allow making statistical inferences of social contact networks from limited epidemic data.

Network reconstruction (Gomez-Rodriguez et al., 2012) methods have recently received increased attention in the fields of graph theory and statistical learning due to its effectiveness in estimating social contact networks from epidemic data. Network reconstruction is the problem in which one aims to extract the topology of a contact network of a set of nodes, by observing data taken from the system that evolves on it (Timme & Casadiego, 2014) The network is described by a (possibly weighted) adjacency matrix, and the reconstruction problem is typically formulated as an optimization. In the context of epidemics, network reconstruction entails estimating the transmission rates (among regions or individuals) from the observed infection data. Recently, Prasse and Van Mieghem (2020) have formulated the reconstruction problem for a wide class of epidemic models using the least absolute shrinkage and selection operator (lasso) which consists of a least-square optimization and a L_1 regularization to enforce some degree of sparsity. The nodes of the graph represent sub-populations, for example, cities in the transportation network of a country. Beaufort et al. (2022) utilized a similar network formulation, however, instead of assuming that the interactions between populations are described by a static contact structure, they used a reaction-diffusion dynamics model to allow for dynamic interactions. The network reconstruction problem is closely related to the graph learning problem for graphical models to understand causal relationships among process variables. In graphical models, the graph captures how the joint distribution over all the random variables is decomposed into a product of factors, representing the conditional independence between variables. For estimating the true network structure, regularization techniques that incorporate an extra penalty for model complexity have been used to obtain sparse solutions (Friedman et al., 2008; Meinshausen & Bühlmann, 2006). Popular choices of the penalty term include L_0 regularization (van de Geer & Bühlmann, 2013), L_1 regularization (Han et al., 2016; Yuan & Lin, 2007), and concave penalty (Aragam & Zhou, 2015).

Many realistic disease transmission networks characterized by sparse adjacency matrices (having only a small number of nonzero weights) present a main computational challenge for the existing network construction methods. In addition, the lack of adequate data on how populations interact (such as data on traffic flow or mobility patterns) is a barrier against making inferences for disease transmission-causing contact patterns. Our paper aims to fill this research gap, by proposing a new regularized statistical estimation technique that enables using geographic information of population units, and not requiring any mobility or traffic data, in understanding and inferring the impacts of social contact on pandemic spread from limited epidemic data. This paper will present a new geographically constrained lasso algorithm for contact network reconstruction in disease spread models. Specifically, a new penalty modifier-based lasso formulation is proposed to account for both geodesic distances and adjacencies among the population sub-groups for making inferences of contact networks. The paper makes original contributions in several areas. First, by contrast to the existing network reconstruction approaches, the proposed method provides an ability to explicitly account for the spatial proximity of the nodes in estimating the disease transmission rates and predicting the future evolution of the epidemic dynamics. Second, a chain-binomial-based stochastic SIR model is formulated to enable simulating epidemic scenarios obeying various graph topologies and assess the benefits of the proposed approach over existing methods. Finally, how to develop epidemic intervention strategies are discussed based on the solutions of the proposed method. The method is illustrated and compared to the existing lasso methods with numerical simulation experiments and on a real infection data set of the COVID-19 pandemic in Florida.

2. Models of epidemics

The most basic model in epidemiology is the homogeneous mixing, continuous time Susceptible-Infected-Removed (SIR) compartmental model without demographic processes (i.e., no births or deaths). This model describes the dynamics of an epidemic in which an individual is in one of three disjoint compartments at any given time: not yet infected and susceptible to disease (S), infected and infectious (I); recovered and unable to spread the disease or reinfected (R). The model is governed by the following system of (coupled) ordinary differential equations (Vynnycky & White, 2010), where rates of flow between compartments are determined by parameters that depend on the history of the disease:

$$\frac{dS(t)}{dt} = -\beta \frac{S(t)I(t)}{N} \tag{1}$$

$$\frac{dI(t)}{dt} = \beta \frac{S(t)I(t)}{N} - \gamma I(t) \tag{2}$$

$$\frac{dR(t)}{dt} = \gamma I(t) \tag{3}$$

where S, I, and R, are the numbers of susceptible, infected, and recovered individuals, respectively, in a given period and N is the number of individuals in the population.

The parameter γ denotes the rate at which infectious individuals recover (become immune) per unit time, also implying that infectious individuals undergo an average recovery period of $1/\gamma$ time units before progressing to the recovered class. Susceptible individuals contact with the virus at the per-capita rate $\beta I/N$ where β is the transmission rate per person per unit time that measures the number of times two specific individuals come into effective contact per unit time (Vynnycky & White, 2010). The parameter β, therefore, captures both the rate at which epidemiologically relevant contacts are made and the probability that the contact between infectious and susceptible individuals leads to the transmission of infection (Rock et al., 2014). The

homogeneous mixing assumption means that at any point in time, every susceptible individual has the same probability of contacting every other individual in the population. The quantity $\beta I/N$, sometimes referred to as the "force of infection," the product of the transmission rate β and the probability I/N that the contact is infectious, therefore, gives the rate at which susceptible individuals become infected at time t.

The continuous time SIR model is typically solved numerically, for example, using the forward Euler time-stepping approach, resulting in a set of difference equations (Allen, 2008; Rock et al., 2014) for the numbers of infected and recovered individuals at discrete time periods, denoted t, as

$$I_{t+\Delta t} = I_t - \Delta t(\gamma I_t - S_t \beta I_t/N) \tag{4}$$

$$= {}_{It} - \Delta t\gamma I_t + \Delta t(N - I_t - R_t)\beta I_t/N \tag{5}$$

$$R_{t+\Delta t} \quad = R_t + \Delta t\gamma I_t \tag{6}$$

where Δt is the discretization time step and the number of susceptible individuals is $S_t = N - I_t - R_t$. As discussed in our case study later, a typical frequency in which public health data on infection counts is reported daily. Therefore, in what follows, the integration step size in the discrete-time formulation is assumed to be a single day, i.e., $\Delta t = 1$.

The interpretation of the parameters in the discrete-time model differs somewhat from the continuous-time model. While β is defined as the rate of contact resulting in transmission in the continuous-time model, it can be interpreted as the probability of contacts resulting in transmission in the discrete-time model. It is known that when the time step used in the discrete-time formulation is sufficiently small then rates can be interpreted as probabilities (Vynnycky & White, 2010, p. 152).

Public health authorities typically report the number of confirmed new cases arising in a given period, referred to as "incidence." Incidence is different from "prevalence," or the number of infectious individuals I_t, that is modeled in the SIR model. Prevalence, I_t and number of recovereds, R_t, can be obtained from the reported incidence data. Let C_t be the incidence data, which satisfies $C_t = S_{t-1} - S_t$. Then for a SIR model, the number of infecteds is $I_t = (1-\gamma)I_{t-1} + C_t$ and the number of recovereds is $R_t = R_{t-1} + \gamma I_{t-1}$.

The solution of the SIR model allows for the prediction of the important parameters of particular interest in public health studies, such as final epidemic size, reproductive number, epidemic thresholds, and peak height and peak time (House & Keeling, 2011). Among those, final epidemic size is how many individuals ultimately become infected during the entire time an epidemic lasts and is computed as the number of recovereds as t approaches infinity. Reproductive number, or the average number of secondary infections, is defined as $R_0 = \beta/\gamma$ and determines how fast the infections spread. If R_0 is greater than 1, the epidemic prevails, while, by contrast, if R_0 is less than 1, the initially infected individuals recover without infecting other susceptible individuals and the epidemic dies out. Epidemic control measures may be designed for the objective of delaying the peak time, reducing the final epidemic size, or reducing the reproductive number.

3. Proposed method

To be able to make inferences about the contact rates in populations with non-homogeneous viral spread between groups, the proposed methodology utilizes a discrete-time version of the SIR model and a contact network on which it is defined. Recall that (as discussed in Section 2) the contact rate parameter of interest β captures both the rate of contacts and the probability that the contact leads to transmission. The network representation partitions the population or region into n non-overlapping sub-regions indexed by $i = 1, 2, ..., n$ with which to model non-homogeneous contact patterns of individuals among and within the sub-regions. We define the human contact network as a directed, weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = \{v_1, ..., v_n\}$ being a set of n nodes representing the sub-regions and $\mathcal{E} = \{e_{12}, ..., e_{ij}, ..., e_{n-1n}\}$ being a set of $\binom{n}{2}$ weighted edges.

Nodes $v_i \in \mathcal{V}$ may represent groups of individuals such as households or geographical sub-regions. Edges $e_{ij} \in \mathcal{E}$ between nodes v_i and v_j (for $i, j = 1, 2, ..., n$ where $i \neq j$) represent the frequency of contacts between the nodes at a specific time point. The contacts between the nodes in the network are characterized by its adjacency matrix, an $n \times n$ matrix $\boldsymbol{A} = (A_{ij})$ where A_{ij} is equal to the weight of the edge between node i and node j. Individuals in the same sub-region may interact with each other, that is, self-edges may exist, $A_{ii} \neq 0$.

The number of neighbors of node i (its number of contacts, or the number of edges attached to it), found as the sum of the ith row (or ith column) in the adjacency matrix, is called the degree of node i and the degree distribution of a network is the frequency distribution of the degrees throughout the entire set of nodes. For weighted networks the degree of a node is the sum of weights of all edges incident to the node, and the degree of a node is often named the "strength" of the node. The distribution of strength, also called the weighted degree distribution, is then defined similarly to the degree distribution. The degree is one of the most fundamental measures of node centrality, which addresses the question of the most important nodes in the network. For modeling the spread of infectious diseases on networks, it is reasonable to assume that individuals or sub-regions with connections to many others have a strong impact on the scale of the disease spread (Newman, 2010, p.168) and node degree and node betweenness centrality measures have been used to identify target nodes for epidemic control (e.g., vaccination) strategies (Salathé et al., 2010).

Let (S_{it}, I_{it}, R_{it}) be the viral state of sub-region i in discrete-time periods $t = 1, 2, ..., T$. Similarly let N_i denote

the population of sub-region i, such that the total population is $N = \sum_{i=1}^{n} N_i$, and let γ_i be the curing rate for individuals in sub-region i. Let β_{ij} be the probability of transmission that specifies the proportion of contacts of individuals in sub-region j with sub-region i that result in transmission. Note that $\beta_{ii} \neq 0$ since the individuals within a city do interact with each other. For every sub-region i, the viral state evolves according to the discrete-time SIR model

$$S_{it} = N_i - I_{it} - R_{it} \tag{7}$$

$$I_{i,t+1} = (1-\gamma_i)I_{it} + S_{it}\sum_{j=1}^{n}\beta_{ij}I_{jt}/N_j \tag{8}$$

$$R_{i,t+1} = R_{it} + \gamma_i I_{it} \tag{9}$$

Note that in this paper, we focus on static networks, specifically networks in which contacts are assumed to be fixed during the infectious period of an individual. This means that both the vertex set and the edge set remain fixed throughout the disease-spreading period. The assumption of fixed contacts is reasonable for diseases that spread slowly compared to the rate at which individuals change the numbers and identities of their contacts (Bansal et al., 2007). The time-varying behavior of networks in the detection of anomalies in connections between nodes has been extensively studied in the statistical analysis of network data elsewhere (Park et al., 2013; Yu et al., 2018). An overview of methods was provided in recent review papers by Savage et al. (2014) and Woodall et al. (2017).

3.1. Geographically constrained lasso inference of contact networks

We propose a least absolute shrinkage and selection operator (lasso) method to estimate the contact probabilities β_{ij} (probabilities of contact resulting in transmission) between individuals of sub-regions i and j and the curing rates γ_i within sub-region i, as well as predict the future evolution of the epidemic dynamics from an observed set of epidemic data. For given curing rates, the discrete-time SIR equations (Eqs. (7–9)) can be represented as a linear regression problem, as β_{ij} appear linearly in the equations. Suppose prevalence data I_{it} is observed for time periods $t = 1, 2, \ldots, T$ for sub-regions $i = 1, 2, \ldots, n$. The data according to the model (8) can be represented using the linear regression model.

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i \tag{10}$$

where $\boldsymbol{\beta}_i = (\beta_{i1}, \beta_{i2}, \ldots, \beta_{in})^T$ is the vector of transmission rates between sub-region i and the other sub-regions, and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iT-1})$ is the vector of model errors that are assumed to be normally and independently distributed. The T- dimensional observation vector $\boldsymbol{y}_i$ and the $(T-1)\times n$ regressor matrix $\boldsymbol{X}_i$ are defined as:

$$\boldsymbol{y}_i = \begin{bmatrix} I_{i2} - (1-\gamma_i)I_{i1} \\ I_{i3} - (1-\gamma_i)I_{i2} \\ \vdots \\ I_{iT} - (1-\gamma_i)I_{i,T-1} \end{bmatrix}, \tag{11}$$

$$\boldsymbol{X}_i = \begin{bmatrix} S_{i1}I_{11}/N_1 & \ldots & S_{i1}I_{n1}/N_n \\ \vdots & \ddots & \vdots \\ S_{i,T-1}I_{1,T-1}/N_1 & \ldots & S_{i,T-1}I_{n,T-1}/N_n \end{bmatrix}.$$

We remark here that the infection counts in a sub-region i depend on the transmission rates of all other sub-regions, which are coupled together and defined by the regressor matrix $\boldsymbol{X}_i$. To estimate the contact probabilities β_{ij} from the observed data, the following lasso problem is solved (Prasse & Van Mieghem, 2020) for sub-regions $i = 1, 2, \ldots, n$

$$\min_{\boldsymbol{\beta}_i} \ \|\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_i\|_2^2 + \alpha_i \ \|\boldsymbol{\beta}_i\|_1 \tag{12}$$

where the tuning parameter $\alpha_i \geq 0$ is determined by cross validation. This is the "ordinary" lasso approach to the network reconstruction problem, previously studied by Prasse and Van Mieghem (2020) and Youssef and Scoglio (2011) which we use as a benchmark in our comparisons.

It is clear that to solve this problem, $\boldsymbol{y}_i$ and $\boldsymbol{X}_i$ should be evaluated for which the number of susceptible individuals S_{it} and the curing rates γ_i needs to be known. One approach is to solve the lasso problem is by specifying the curing rate and assuming that at the beginning of the pandemic $R_{i1} = 0$, meaning that the sub-populations are completely susceptible. Accordingly, for a given curing rate γ_i, the vector $\boldsymbol{y}_i$ is specified. The recovereds R_{it} are determined for all $t \geq 2$ using Eq. (9) and then the number of susceptibles S_{it} are determined using Eq. (7) from which $\boldsymbol{X}_i$ is specified.

As it will be later shown in our illustrations, depending on the graph topology, the ordinary lasso may result in a fairly large number of non-zero contact probabilities in applications where the true relation corresponds to no contact (i.e., high false positive classification errors). This restricts the method's ability in obtaining sparse solutions, which limits its practical applicability in public health intervention because too many interactions are estimated as significant. As an efficient network reconstruction solution, this paper will formulate the reconstruction problem by assuming that edges between nodes that are geographically closer to each other tend to have larger contact rates compared to those between nodes that are farther apart. To this end, the proposed lasso method introduces a new multiplier ν_{ij} for the L_1 norm of the contact probability between sub-regions i and j. The multiplier is defined based on not only the spatial proximity of the sub-regions but also on the network adjacency structure, which could account for additional characteristics such as movement or contacts among sub-regions. Specifically, the proposed geographically constrained lasso problem determines the contact rates of the ith node by solving:

$$\min_{\boldsymbol{\beta}_i} \|\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}\|_{i2}^2 + \alpha_i \sum_{j=1}^{n} \nu_{ij} \left|\beta_{ij}\right| \tag{13}$$

where $\nu_{ij} \geq 0$ is the new multiplier for the jth contact probability β_{ij}. Therefore, the multiplier is a number that modifies the usual lasso tuning parameter α_i to allow differential shrinkage of the transmission rates. It is obvious that $\nu_{ij} = 1$ for all coefficients corresponds to the ordinary lasso where coefficients are penalized equally. By contrast, $\nu_{ij} = 0$ for some variables implies the variable is always included in the model (no shrinkage) and the case of $\nu_{ij} = \infty$ for some variables implies the variable is always excluded from the model.

This also has been termed the "penalty modifier" approach to lasso (Hastie et al., 2015, p. 51) in the literature. However, no clear method has been offered to specify the modifiers in a parsimonious way. In this study, to specify the proposed penalty modifier, a power law is adopted as some function of $(D_{ij})^{\omega}$ where $\boldsymbol{D} = (D_{ij})$ is the matrix of Euclidean distances between the centroids of sub-regions i and j and ω is an exponent such that $\omega \geq 0$. Specifically, a graph-weighted penalty modifier is proposed that combines the power law of Euclidean distances and the entries A_{ij} of the adjacency matrix $\boldsymbol{A}$ of the contact network as follows

$$\nu_{ij}(\boldsymbol{D}, \boldsymbol{A}) = \begin{cases} D_{ij}^{\omega}/A_{ij} & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases} \tag{14}$$

We note that this is analogous to the cost function used in modeling flows in transportation networks with gravity models (Kolaczyk & Csárdi, 2014, p. 163). A choice of $\omega = 2$ is typical in gravity models and will be used in this paper. The unique contribution of the proposed method therefore is to formulate the penalty modifiers not simply based on the geographical distances between the nodes but also based on the network structure available in the form of the graph adjacency matrix. While the ordinary lasso assumes that the tuning parameter is constant for all transmission rates, the proposed approach introduces a separate tuning parameter for different transmission rates. This helps improve the estimation and obtain sparse solutions, particularly when there is prior knowledge or preference over some of the coefficients.

The binary adjacency matrix $\boldsymbol{A}$ is constructed by assuming that nodes correspond to the centroids of the sub-regions and that two nodes are connected with an edge if the corresponding sub-regions share a common border. The definition (14) ensures that the sub-regions that are located nearby (small D_{ij}) and are adjacent ($A_{ij} = 1$) receive small penalty factors. If the sub-regions do not share a border ($A_{ij} = 0$) then the penalty factor is $\nu_{ij} = \infty$, regardless of how small the distance is, and the variable is completely excluded from the model. It also ensures that $\nu_{ij} = 1$ for contacts within the same sub-region, i.e., $i = , j$, and the coefficients will be treated as ordinary lasso. For contacts between different sub-regions $i \neq j$ that share a border (that is, $A_{ij} = 1$) the penalty factor is increased in proportion to the distance (D_{ij}) between the centroids of the sub-regions and the coefficients will receive additional shrinkage. The tuning parameters α_i are determined, as before, by cross-validation.

The optimization equation (Eq. (13)) is solved for a set of candidate values $\boldsymbol{\Omega} = \{\gamma_L, ..., \gamma_H\}$ equally spaced between $\gamma_L = 0.01$ and $\gamma_H = 1$. Using Eq. (9), we determine the potential sequences of recovered individuals $R_{i1}, ..., R_{iT}$ for each candidate value $\gamma_i \in \boldsymbol{\Omega}$. The curing probability γ and the resulting sequence $R_{i1}, ..., R_{iT}$ are selected as the elements of $\boldsymbol{\Omega}$ that result in the best fit of the SIR model to the reported number of infections.

Remark on directed networks: When the population is divided into n distinct groups, the ijth entry of the contact matrix β_{ij} is the transmission coefficient from an infective in group j to a susceptible in group i. While it is typical and more common to assume a symmetric contact structure (modeled by an undirected network), some applications require non-symmetric contact patterns (modeled by a directed network). For example, transmission of certain childhood infections among older and younger age groups (Van Effelterre et al., 2009) or transmission through donated blood products (Keeling & Eames, 2005) are instances when infection can only travel one way along an edge, and can be modeled directed graphs. To ensure generality and broader applicability we adopt directed contact networks in our method.

3.2. Simulation of epidemics on networks

To assess the accuracy of the network inference on realistic contact networks, this paper presents a new network-based stochastic SIR model to simulate non-homogeneous mixing epidemic processes with known transmission and curing rate parameters. The method is an extension of the stochastic SIR model (Bansal et al., 2007; Bjørnstad, 2018, p. 234) which adopts a definition of contact networks to enable simulations of non-homogeneous mixing SIR epidemics. The simulated epidemics originating from known transmission rates will serve as the ground truth, allowing us to assess the efficacy of the proposed and existing network construction methods.

The stochastic SIR model, also referred to as the "chain-binomial" model, was originally proposed by Bailey (1975, and further studied by Ferrari et al. (2005), Bjørnstad (2018, p. 36), Lekone and Finkenstädt (2006). On a contact network, an infection is transmitted between a susceptible node and an infectious node with a probability, β_{ij}, per time step. Let B_{it} denote the number of susceptible individuals who become infected and D_{it} the number of cases who are removed from the infectious class in i-th subpopulation during the time interval starting at time step t. The probability of any individual in sub-population i becoming infected (i.e., force of infection) at time t is $\lambda_{it} = 1 - \prod_{j=1}^{n} (1 - \beta_{ij})^{Z_{ijt}}$ where Z_{ijt} is the number of infected neighbors of sub-region i who preside in sub-region j at time step t (Vynnycky & White 2010, p. 152). For a network represented by an adjacency matrix $\boldsymbol{A}$, this is found as $Z_{ijt} = A_{ij}I_{jt}$. For the special case of a constant transmission rate $\beta_{ij} = \beta$, the probability of infection is simplified as $\lambda_{it} = 1 - (1 - \beta)^{Z_{it}}$ in which

$Z_{it} = \sum_{j=1}^{n} Z_{ijt}$, where Z_{it} is the total number of infected neighbors of node i. The total number of infected neighbors for all nodes is obtained as $\boldsymbol{Z}_t = \boldsymbol{A}\boldsymbol{I}_t$ where $\boldsymbol{Z}_t = (Z_{it})$ and $\boldsymbol{I}_t = (I_{it})$.

Therefore, B_{it} follows a binomial distribution with a sample size equal to the number of susceptibles and probability of infection equal to λ_{it}. Similarly, D_{it} follows a binomial distribution with probability of recovery γ_i and sample size equal to the number of infecteds. Accordingly, the SIR epidemic on a network is simulated using

$$B_{it} \sim \text{Binomial}(S_{it}, \lambda_{it}) \tag{15}$$

$$D_{it} \sim \text{Binomial}(I_{it}, \gamma_i) \tag{16}$$

$$\lambda_{it} = 1 - \prod_{j=1}^{n} (1 - \beta_{ij})^{Z_{ijt}} \tag{17}$$

$$S_{i,t+1} = S_{it} - B_{it} \tag{18}$$

$$I_{i,t+1} = I_{it} + B_{it} - D_{it} \tag{19}$$

$$R_{i,t+1} = R_{it} + D_{it}. \tag{20}$$

For studying the properties of epidemics on networks, this paper will consider several structured network types, commonly considered in the epidemics literature. This includes small-world networks, characterized by high levels of both local clustering and global connectivity (Watts & Strogatz, 1998), scale-free networks, characterized by degree distributions that follow a power law distribution with a small fraction of very highly connected hubs (Barabasi & Albert, 1999) and lattices in which all nodes have the same degree, and any given node is connected to physically proximate nodes. In the simulation study of this paper, simulated realizations of epidemics with contact networks following these basic network structures will be considered. As an example, Fig. 1 shows simulated lattice, small-world, and scale-free networks with 64 nodes.

4. Simulation study

In this section, we demonstrate the application of the proposed network construction and inference approach by simulation. The R package `igraph` is used to generate realizations from various random contact networks and the proposed network-based stochastic SIR simulation approach is used to generate epidemic trajectories due to the assumed contact rates and graph topologies. We limit our discussion to networks with arbitrary degree distributions, including lattice networks, and random networks with small-world and scale-free degree distributions. The proposed approach will be compared to the ordinary lasso network reconstruction method (Prasse & Van Mieghem, 2020) as the benchmark approach to make inference on the contact networks based on the simulated epidemic trajectories.

We consider networks with 16, 25, 36, and 64 nodes. For the lattice networks, these correspond to $4 \times 4, 5 \times 5, 6 \times 6$ and 8×8 grids, respectively. It is assumed that the nodes are spatially located on a two-dimensional uniform grid, where two adjacent nodes along either dimension are separated by a unit distance. Regardless of the distances between nodes, the node adjacency is determined based on the network topology (that is, human contacts occur only along neighboring nodes). As an example, for the 64-node networks shown in Fig. 1. The distance matrix $\boldsymbol{D}$ is the same regardless of graph topology (since the pairwise distances are determined according to the grid structure given by the lattice) and, the distance between, say, nodes 1 and 2, or between nodes 1 and 9 are equal to 1, i.e., $d_{12} = d_{19} = 1$. The adjacency matrix $\boldsymbol{A}$, however, is different depending on the simulated edges.

The parameters for the degree distributions are chosen so that each network has an identical average degree. For example, for a 64-node network, the average degree for all topologies is set to 3.5, which is the average degree of a two-dimensional lattice. Figure 1 shows a single simulated realization of the lattice, small-world, and scale-free networks with 64 nodes. The node population is set to $N_i = 20$ for all $i = 1, 2, \ldots, n$ where n is the number of nodes, and the curing rate is set to $\gamma_i = 0.3$ for all nodes. The contact rate per edge is set to $\beta_{ij} = 0.01$ if $A_{ij} \neq 0$ and $\beta_{ij} = 0$ if $A_{ij} = 0$, where $\boldsymbol{A} = (A_{ij})$ is the binary adjacency matrix that is determined based on the topology of the simulated network.

The proposed penalty modifier approach and the existing ordinary lasso techniques are applied to reconstruct the contact network and estimate the contact and curing rates from the simulated epidemic data. The proposed method is implemented using the penalty function (14), the geographical locations of the nodes $\boldsymbol{D}$, and the adjacency structure $\boldsymbol{A}$ of

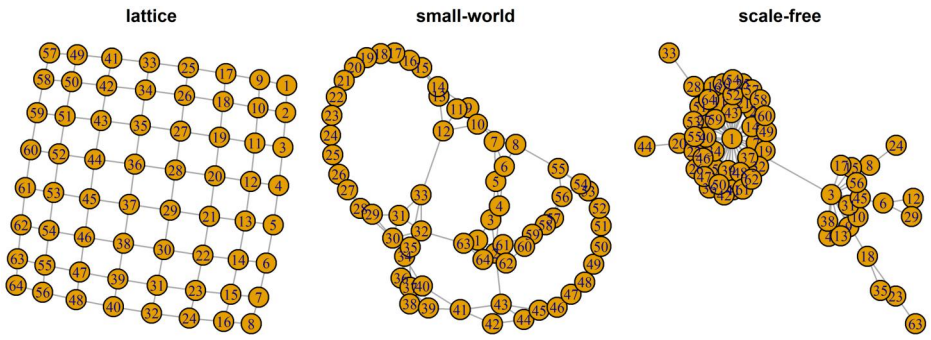

Figure 1. Simulated lattice, small world and scale-free contact networks with 64 nodes.

the networks. The existing lasso approach is implemented by solving the problem (12).

The benefit of the proposed approach over the benchmark method is studied based on classification error rates, computed with the estimated contact rates and the known true contact rates. Specifically, for a simulated network with true contact rates β_{ij} and estimated contact rates $\hat{\beta}_{ij}$ obtained with the network inference method for nodes $i, j \in \{1, 2, \ldots, n\}$ the False Negative Rate (FNR) and the False Positive Rate (FPR) are computed as:

$$\text{FNR} = \frac{1}{n_0} \sum_{i,j} J\left\{\hat{\beta}_{ij} = 0, \beta_{ij} \neq 0\right\} \tag{21}$$

$$\text{FPR} = \frac{1}{n_1} \sum_{i,j} J\left\{\hat{\beta}_{ij} \neq 0, \beta_{ij} = 0\right\} \tag{22}$$

where $J\{a, b\}$ is an indicator taking on the value of 1 when both conditions a and b in the curly braces are met, and 0 otherwise; n_0 and n_1 are the number of entries of the adjacency matrix $\boldsymbol{A}$ that are zero and non-zero, respectively. Improvement in an error rate is computed as $(M_P - M_L)/M_L$ where M_P and M_L are the (false positive or false negative) error rates calculated based on the inferred network structures obtained using the proposed approach and the ordinary lasso approach, respectively, in each simulation.

Figure 2 shows the proportion of susceptibles, infected, and recovereds (to the population size), for one simulated epidemic of the entire network. While it is difficult to make generalizations from only one realization of the epidemics, certain unique features of the network topologies can be noted. The proportion of infecteds has a similar peak value for all networks. However, the timing of the peak infecteds and the final value reached by recovereds (the size of the epidemic) show some variations. Notably, in small-world networks, the peak infecteds occur soonest (at time period 10), the final value of recovereds is reached quickly (around time 20), and the size of the epidemic is highest (about 0.95) compared to lattice and scale-free networks. This is a pattern consistent with results reported in the literature for epidemics on small-world networks (Bansal et al., 2007).

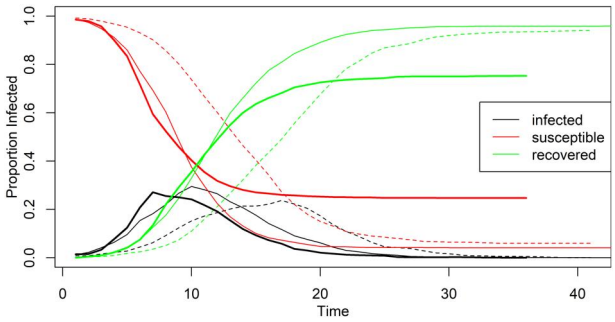

Figure 2. Simulated epidemics on different network topologies with 64 nodes: lattice (dashed line), small-world (solid thin line) and scale-free (solid thick line) networks.

The simulations are replicated 1000 times. Table 1 shows the number of nodes (n), maximum possible number of edges in a directed graph (n^2), the number of edges in the simulated graph, and the density of the network (the fraction of the maximum number of edges that are present, obtained by dividing the latter by the former), for different graphs. It is important to note that although a new network is generated in each simulation, the number of edges does not change for a given number of nodes because the mean degree is fixed. As can be seen, for a given network size (number of nodes), the number of edges (and, therefore, the network density) is largest in small-world networks, followed by lattices, which are followed by scale-free networks.

The false negative rate (FNR) and false positive rate (FPR) in recovering the contact rates by each method and the improvement attained by the proposed method obtained from simulations are shown in Fig. 3. Figure 3(a) shows the error rates. It is evident that the proposed method can greatly reduce both the FPR and FNR of the ordinary lasso in all network types and sizes. However, the reduction in FPR is noticeably greater than that in FNR regardless of the network type. Figure 3(b) shows the improvements (reduction) in FNR and FPR with proposed approach over the ordinary lasso. The reduction in both error rates increases with network size in all network type, however, the it is most pronounced in lattice networks. As indicated by Table 1, the network densities for all simulated networks are less than 0.5 (0.25 or smaller), suggesting that the false negative rates are expected to be much higher than false positive rates.

Further statistical analysis of the improvements is warranted to more accurately quantify the effects of various network properties on the improvements attained in classification error rates. A two-way Analysis of Variance (ANOVA) was conducted on the improvements, considering network topology and size (number of nodes) as factors. The models showed that both the topology and size have

Table 1. Edge topologies of the simulated networks.

Network	Number of nodes	Maximum number of edges	Number of edges	Density of network
Small world	16	256	64	0.25
	25	625	100	0.16
	36	1296	144	0.11
	64	4096	256	0.06
Lattice	16	256	48	0.19
	25	625	80	0.13
	36	1296	120	0.09
	64	4096	224	0.05
Scale free	16	256	30	0.12
	25	625	48	0.08
	36	1296	70	0.05
	64	4096	126	0.03

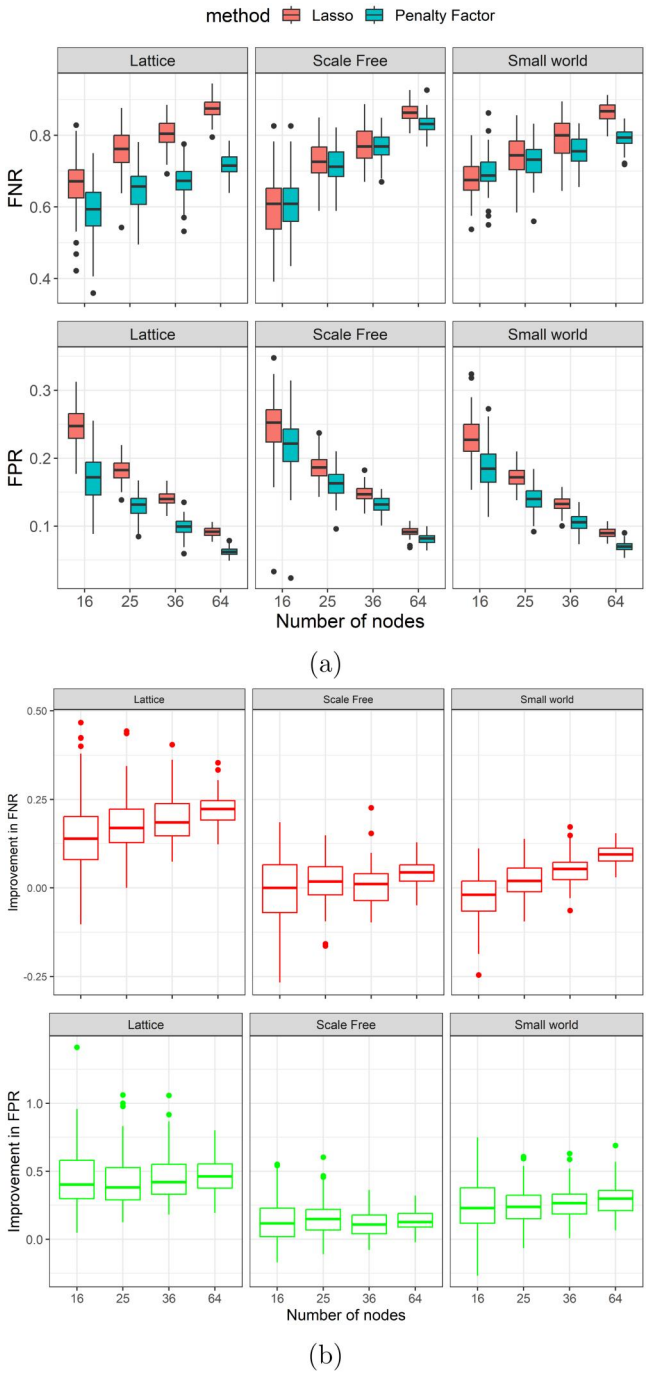

Figure 3. (a) Classification error rates FNR and FPR for SIR epidemic contact network reconstruction using the proposed approach and ordinary lasso. (b) Improvements in FNR and FPR by using the proposed method compared to ordinary lasso.

significant effects (at 0.05 level) on improvements in FPR and FNR. Figure 4 shows the estimates and the 95% confidence intervals for the effects of network type and size on the improvements. The largest improvements in FNR and FPR are observed in lattice networks, followed by small-world networks, and then scale-free networks. This outcome (also observed in the error rate plots previously) is expected since the nodes in a lattice are located in a spatial configuration and the proposed approach explicitly accounts for the geographic distance of the nodes in formulating the network reconstruction formulation. In general, the improvements in FNR and FPR are larger with larger network sizes for lattice and small-world topologies. For scale-free networks, while the improvements are significant, it does not scale up with network size as strongly as lattice or small-world networks.

5. Case study

The proposed contact network reconstruction approach is illustrated using infection count data from the first wave of the COVID-19 pandemic. The study period spans 155 days, from October 10, 2020 to March 8, 2021. The existing ordinary lasso approach was also implemented based on the same data and the results are compared to those of our approach. We focus on the state of Florida as the study region, and the daily infection case counts reported in the state's 67 counties are used for contact network inference.

COVID-19 was declared by the World Health Organization as a pandemic on March 11, 2020 (Cucinotta & Vanelli, 2020). The study period coincides with the first phase of the COVID-19 pandemic, during which no vaccines were yet available (the first vaccines only became publicly available starting in December 2020). The daily count of confirmed new COVID-19 infections (incidence) for Florida counties is sourced from the New York Times COVID-19 data repository (New York Times, 2023). The county populations required to construct the SIR models are obtained from the 2020 U.S. Census (U.S. Census Bureau, 2020).

Florida's counties are modeled as nodes in the network, and the contacts between individuals across counties during the study period are modeled as the edge weights of the network. The daily infection counts data exhibit large variations due to possible underreporting and weekend/holiday effects and to mitigate the impact of variability we apply a seven-day moving average smoothing to the count data before conducting the network inference. The moving averaged count data are used to estimate the contact rates between counties and the recovery rates of counties of the SIR model using both the proposed and existing methods. The prevalences are obtained from the incidenced using the approach outlined in Section 2.

In the proposed approach, it is assumed that contact of people between all pairs of counties occurs through neighboring counties. Figure 5 shows the adjacency graph of the counties, from which the adjacency matrix is derived. In this graph, the nodes correspond to the centroids of the counties, and two nodes are connected by an edge if the corresponding counties share a common border. The geographic analysis package `spdep` in R is used to construct the maps and the adjacency matrices. The penalty modifier ν_{ij} is determined according to Eq. (14), where $\boldsymbol{D} = (D_{ij})$ represents the Euclidean distances between the centroids of county i and county j for all $i, j = 1, 2, \ldots, 67$, and $\boldsymbol{A} = (A_{ij})$ denotes the entries of the adjacency matrix of the counties.

As a point of reference, the homogeneous-mixing SIR model, given by Eqs. (4) to (6), is estimated using the COVID-19 counts aggregated to entire Florida. The prevalence I_t predicted with the fitted model and the observed

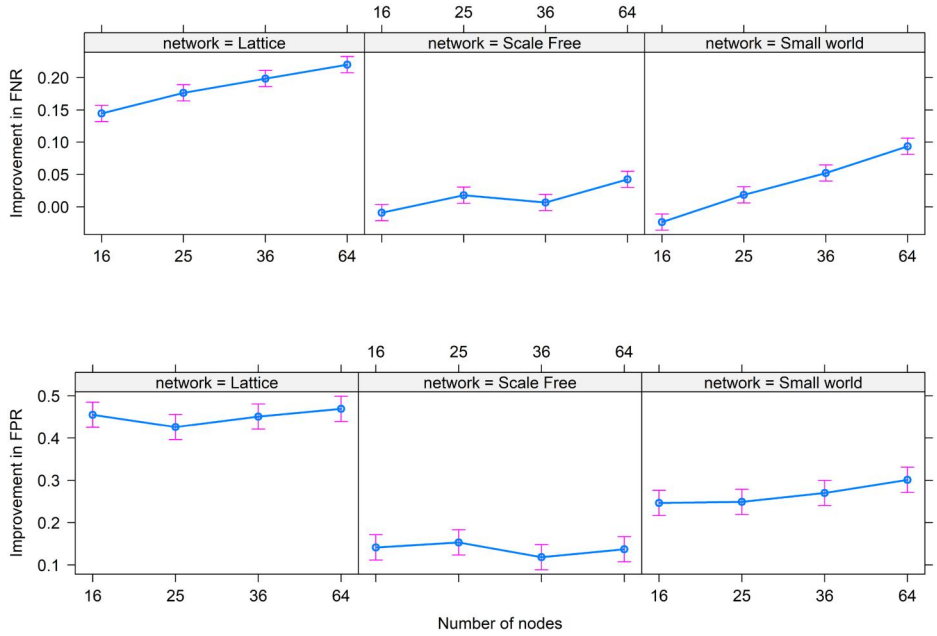

Figure 4. Effects of network size and topology on improvements in classification error rates.

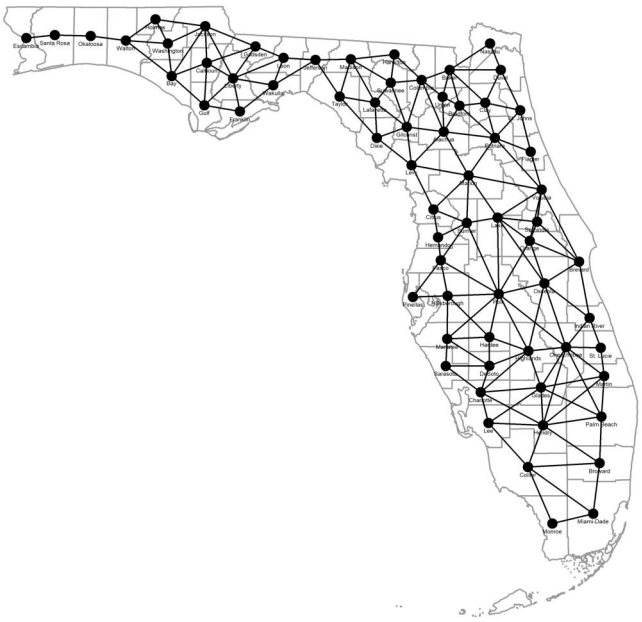

Figure 5. Florida county adjacency graph.

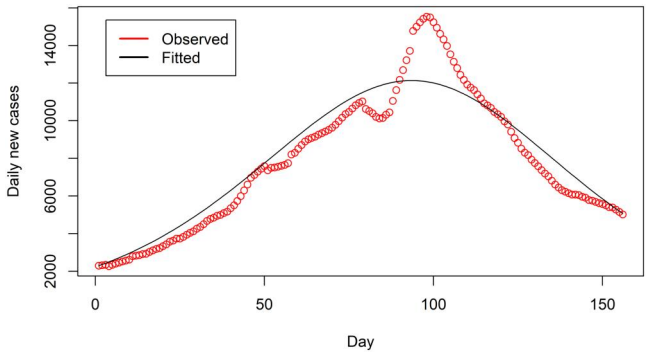

Figure 6. Aggregated infection case counts and predictions obtained from the homogeneous mixing SIR model for Florida. The study period of 155 days covers dates between October 4, 2020 and March 8, 2021.

prevalence are shown in Fig. 6. From the analysis of the aggregated data the estimated transmission rate, curing rate, and reproduction numbers are obtained as $\hat{\beta} = 0.71, \hat{\gamma} = 0.69$ and $\hat{R}_0 = 1.03$, respectively. We note, however, this model assumes the contact rate is constant throughout the state. The proposed method by contrast explicitly models non-homogeneity of contact rates.

Using the proposed network reconstruction method (and a non-homogeneous mixing SIR model) the infection count trajectories for each county are predicted as shown in Fig. 7. The estimated contact rates between the counties obtained with both the proposed and existing methods are shown in Fig. 8 in graph and heatmap formats. From the diagonals of the contact rate matrices (Fig. 8(c,d)), it is evident that almost all nodes have self-edges. To maintain clarity and reduce clutter, however, the self-edges are omitted in the graph format (Fig. 8(a,b)).

Despite the L_1 regularization applied to achieve a sparse solution in the ordinary lasso approach, the resulting contact network, shown in Fig. 8(b), has a large number of significant contact rates, some of which being the contacts between counties that are a large distance apart. By contrast, the proposed geographically constrained lasso approach results in substantially sparser, and probably more realistic, contact networks, as shown in Fig. 8(a), with fewer edges that typically connect neighboring counties. This shows the importance of explicitly accounting for the geodesic distances in the network reconstruction solution. Out of the 4489 ($= 67^2$) possible number of edges (including self edges) of a complete graph, the ordinary lasso identifies 536 edges to have (non-zero) contacts while the proposed approach identifies only 146 edges to have (non-zero) contacts. With the proposed method, it was observed that the most frequent contacts occurred between counties such as Hillsborough & Union, Hamilton & Pasco, and Franklin & Palm Beach, all having a contact rate $\beta > 1$).

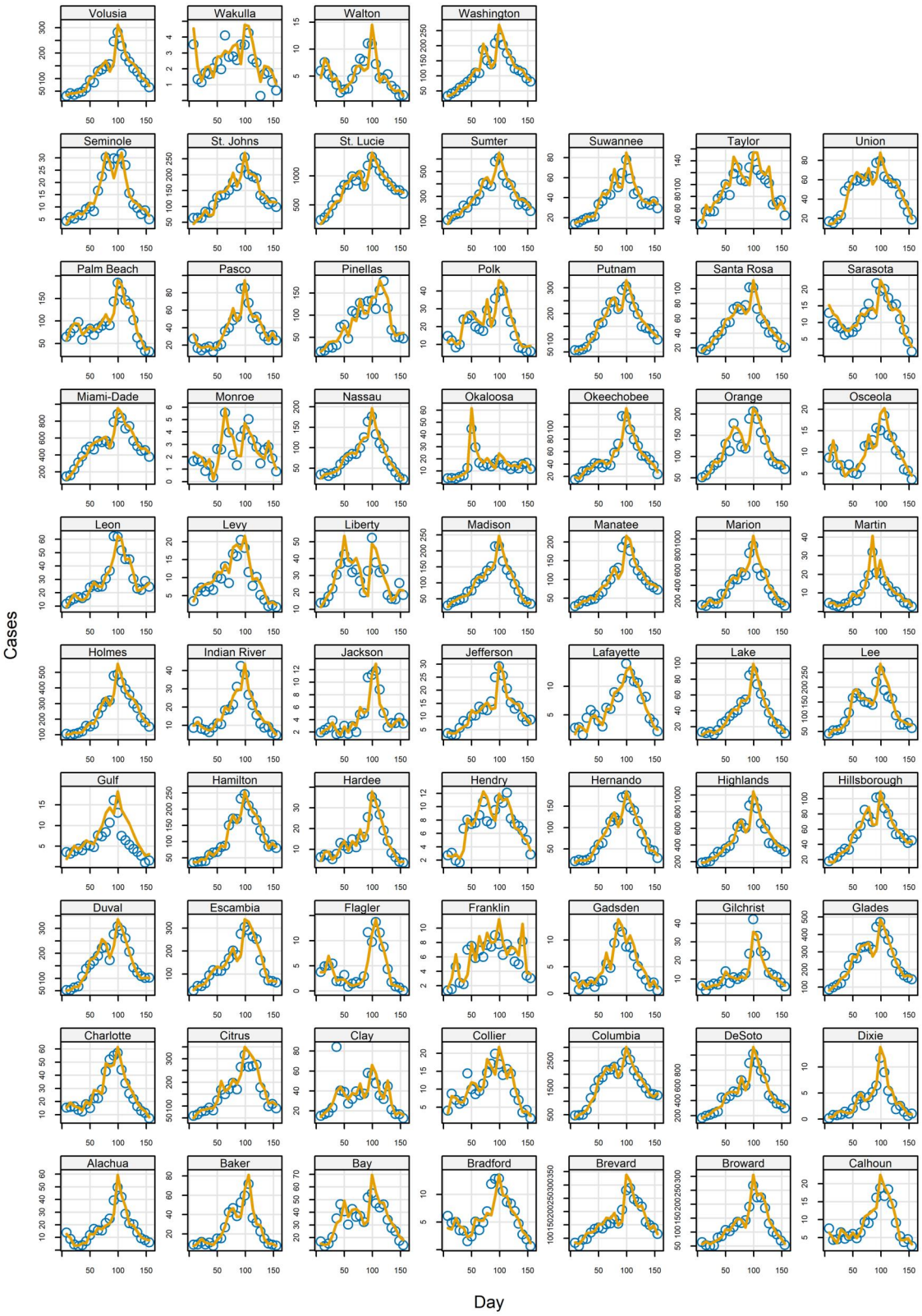

Figure 7. Predicted and observed infection case counts for Florida counties obtained with the proposed network reconstruction approach (study period covers 155 days between October 4, 2020 and March 8, 2021).

5.1. Design of interventions

An important use of contact network inference for public health and epidemic control is to develop intervention measures. The contact rate estimates obtained with the proposed approach will be used for devising targeted interventions in controlling disease spread for high contact rate areas. In particular we consider two main control measures (also discussed in Wan et al., 2008): (*i*) reducing local contact rates *within* counties, which can be implemented in the form of

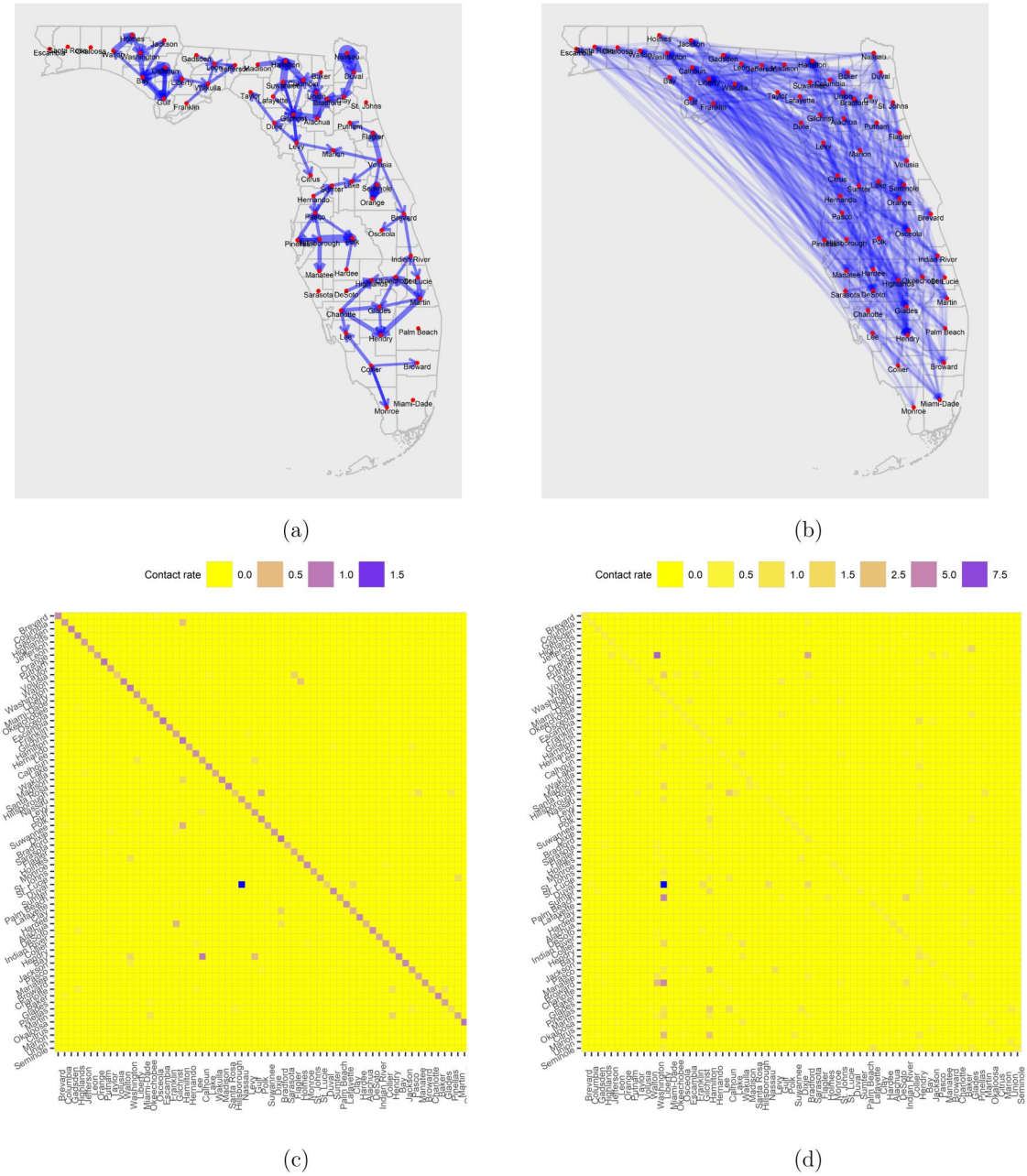

Figure 8. Contact rate estimates of Florida counties obtained from the infection count data; top rows: graph format, bottom rows: heat map format; (a) & (c) proposed approach (b) & (d) ordinary lasso approach.

social distancing and lockdowns and (*ii*) reducing contact rates *across* counties, which can be implemented in the form of restriction on long-range movement. In this study we adjust the contact rate matrices to reflect two counterfactual scenarios of movement restrictions and draw counterfactual predictions based on these, to retrospectively assess the usefulness of the restrictions.

The transmission rates obtained from the full (155-day) infection data based on the proposed method are used and the control measure is introduced counterfactually on day 80. The impact of control intervention on the future trajectory of the epidemic is accounted for by following the approach of Chowell et al. (2004). That is, the contract rates are kept equal to their estimated value $\hat{\beta}_{ij}$, the entries of the matrix shown in Fig. 8(c), up to the time point at which intervention is introduced and the contact rates that are larger than a threshold value β^* are reduced to the threshold β^* after this point according to the control measure. Threshold value β^* is selected as a percentile of the contact rates obtained under the no intervention configuration. Movement restrictions for both within-county travel (diagonal terms) and across-county travel (off-diagonal terms) are considered.

Figure 9 shows the Pareto plots of the transmission rates due to within-county and across-county contacts (obtained from Fig. 8(c)). Note that there are 67 within-county transmission rates and $4,422 (= 67^2 - 67)$ cross-county transmission rates. For visualization purposes, only the top 40 within-county transmission rates and the top 30 cross-county rates are shown in the figures. The transmission rates

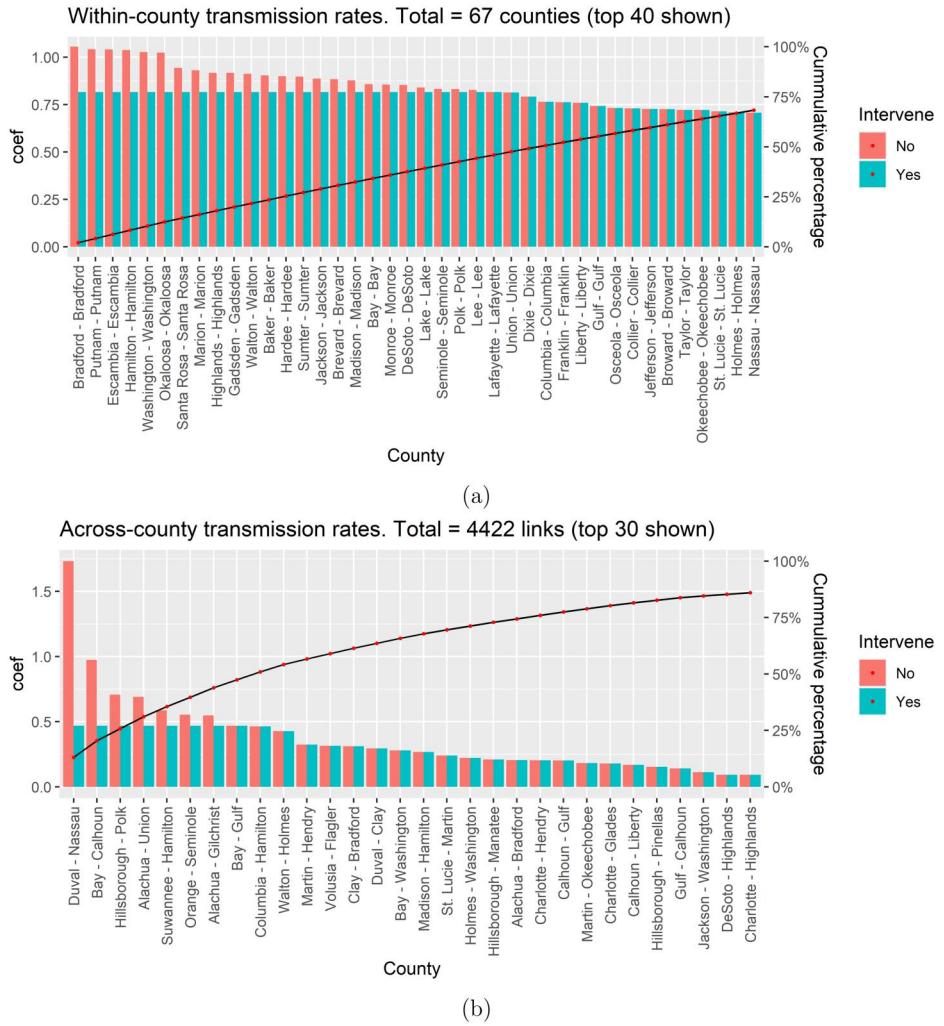

Figure 9. Pareto charts of transmission rates (a) due to within-county contacts and (b) due to cross-county contacts. The rates after thresholding to the 45th percentile are also shown.

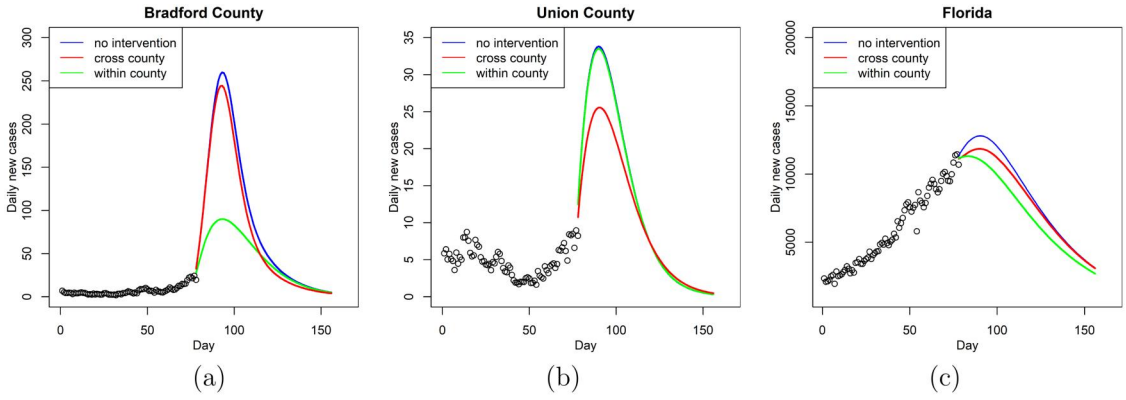

Figure 10. Estimated case counts based on counterfactual intervention measures and no intervention taken on day 80 based on data available up to day 155. Transmission rates are thresholded to the 45-th percentiles for all counties (a) Bradford County, (b) Union County and (c) resulting changes in Florida.

after using the 45th percentile as the threshold value are also shown. In the comparisons, we considered the effectiveness of movement restrictions that threshold to 45th percentiles of within and across county contacts. To assess the effectiveness of these intervention measures, the network SIR model is solved to determine the infection trajectories of the counties (after the intervention implemented at day 80) with the new transmission parameters.

Figure 10 shows the predictions of infection case counts $I_{i,t}$ of several counties as examples and the aggregated counts for entire Florida under the intervention scenarios. The effectiveness of interventions on Bradford and Union Counties are shown in Fig. 10(a,b), and that for Florida is shown in Fig. 10(c). From Fig. 10(a) for Bradford, the within-county travel restriction is significantly more effective than the cross-county travel restriction. By contrast, for

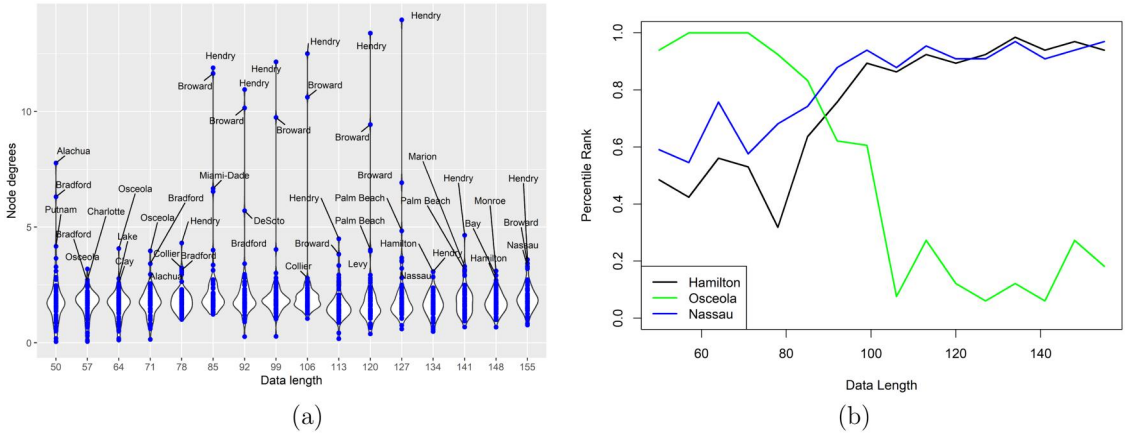

Figure 11. (a) Degree distributions of the estimated contact networks over the study periods. (b) Percentile ranks of the degrees of some of the counties over time.

Union, Fig. 10(b) shows that cross-county travel restriction is significantly more effective than the within-county restriction. These results make intuitive sense when the transmission rate estimates shown in Fig. 9 are considered: from Fig. 9(a): Bradford is the top county for the within-county transmission rates and from Fig. 9(b): Union appears as one of the top counties involved in high cross-county transmission rates. Based on Fig. 10(c), the aggregate impact of the within-county contact interventions in infection counts in Florida are comparatively more effective than those for across-county contacts.

The effectiveness of the intervention measures can also be summarized by reporting the final epidemic size (for entire Florida). For this, the number of recovereds $R_{i,t}$ corresponding to the 3 scenarios (shown as blue, red and green curves in Fig. 10(a)) were computed for all counties and the final epidemic size was found as total number of recovereds at day 155, that is, $\sum_{i=1}^{67} R_{i,155}$. Under the no-intervention, restricting across-county contacts and restricting within-county movements scenarios (all implemented as 45th percentile thresholding), the final epidemic sizes are estimated as 1 115 516 people, 1 082 171 people and 1 012 687 people, respectively. Therefore, with the two intervention strategies, 3.0% and 9.2% reduction in the overall size of the epidemic is achieved. For comparison, the effectiveness of a higher degree of restrictions, corresponding to thresholding to 70th percentile of the contact rates, was also evaluated. It can be shown that under across-county and within-county movement restrictions, the final epidemic sizes can be obtained as 1 041 057 people and 836 810 people, respectively, which represents reductions of 7.0% and 25%, respectively, over the no intervention scenario.

5.2. Dynamic evolution of the contact network

The reported solutions were obtained using the entire infection data available at a specific time point (from the first day until day 155). The network reconstruction problem can also be solved sequentially with gradually increasing amounts of infection data as they become available during the pandemic. Such a dynamic network analysis would be of practical value to healthcare practitioners in devising timely containment and intervention strategies by studying how the network evolves and whether new contact patterns across communities emerge over time.

To this end, the proposed network reconstruction approach was applied sequentially using infection count data sets with progressively increasing sizes. We considered study periods that span periods between the first day until day 50, until day 57, until day 64, and so on until day 155. That is, the length of the data set was increased seven days at a time which resulted in a total of 16 different solutions to be generated. The final solution, which consists of all available data, corresponds to the results presented earlier. To assess how the network structure changes over time we displayed the (weighted) degree distributions of the contact networks obtained from solving the problem with the data set available at each period.

Figure 11(a) shows the degree distributions of the networks obtained from increasing-size data sets. The solution obtained with data length 155 correspond to the estimated contact rates in Fig. 8. Throughout, the majority of the counties exhibit low degrees, with a degree of 3 or less. However, there is a significant tail in the distributions, particularly evident in the periods between days 85 and 127, where two or more counties have substantially higher degrees (larger than 9) than the rest of the counties. While some counties (e.g., Hendry and Broward) consistently appear in the top most connected counties throughout all time steps, some counties exhibit time trends in their connectivity.

Figure 11(b) shows how the percentile ranks change over time for some of counties (Osceola, Hamilton, and Nassau) exhibiting interesting time trends. Osceola exhibit high connectivity in the periods leading up to day 71 but loses influence afterwards, while Hamilton and Nassau counties start with low connectivity but increase influence in the later periods. Such identification of trends in the node degrees would be helpful to practitioners in predicting regions of future outbreaks and timely planning of the sites of potential intervention measures for higher effectiveness.

6. Conclusions

A methodology is presented for making inferences on contact networks from observed infection data. Given that geographically near populations are more likely to make contact with each other, the method starts from the traditional lasso framework and consists of formulating a penalty modifier explicitly based on the spatial proximity of the regions.

A simulation study with known transmission rates for several random graph topologies is presented to study the efficacy of correctly identifying the links on which contact takes place and the links on which no contact takes place. It is shown that the proposed geographically constrained lasso approach outperforms the ordinary lasso that does not account for spatial information in uncovering the transmission rates.

A case study based on actual data from the COVID-19 pandemic is presented to estimate the contact rates based on a non-homogeneous mixing SIR model based on the daily infection data reported among the counties of Florida. It is shown how the approach can be used to identify links among counties with higher transmission rates than others and such information can be used to devise effective movement restriction measures that result in significant reduction in predicted epidemic sizes and effective control of disease spread.

The proposed method solely uses the spatial proximity of the regions in making inferences about the human contacts that lead to disease transmission. As discussed, relatively little is typically known about contact networks in real-life epidemics. In our approach, we assume that contacts between all pairs of sub-regions occur through neighboring sub-regions. No explicit use was made from mobility or human movement in the form of GPS or cell-phone location data as some previous authors have considered (Citron et al., 2021; Tizzoni et al., 2014). It is expected that movement or mobility data will enhance information about human contact and improve prediction accuracy of the future trajectory of epidemics. The definition of the penalty modifier can be further enhanced if there is knowledge available about people or vehicle movement on roadways between respective sub-regions. As future work, the method can be further enhanced by incorporating such movement data when available, into the network reconstruction approach to not only improve prediction accuracy but also uncover the causes of the high infection counts in such regions.

Acknowledgements

This research was supported by National Science Foundation (NSF) grant CMMI-2101091. The authors would like to acknowledge the funding support from NSF.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by National Science Foundation.

References

Allen, L. J. (2008). An introduction to stochastic epidemic models. In F. Brauer, P. van den Driessche, & J. Wu (Eds.), *Mathematical epidemiology* (pp. 81–130). Springer.

Aragam, B., & Zhou, Q. (2015). Concave penalized estimation of sparse Gaussian Bayesian networks. *The Journal of Machine Learning Research*, *16*(1), 2273–2328.

Arino, J. (2009). Diseases in metapopulations. In Z. Ma, Y. Zhou, & J. Wu (Eds.), *Modeling and dynamics of infectious diseases* (pp. 64–122). World Scientific.

Audirac, M., Tec, M., Meyers, L. A., Fox, S., & Zigler, C. (2022). Impact of the timing of stay-at-home orders and mobility reductions on first-wave COVID-19 deaths in US counties. *American Journal of Epidemiology*, *191*(5), 900–907. https://doi.org/10.1093/aje/kwac027

Bailey, N. T. J. (1975). *The mathematical theory of infectious diseases and its applications*. Griffin.

Bansal, S., Grenfell, B. T., & Meyers, L. A. (2007). When individual behaviour matters: Homogeneous and network models in epidemiology. *Journal of the Royal Society, Interface*, *4*(16), 879–891. https://doi.org/10.1098/rsif.2007.1100

Barabasi, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512. https://doi.org/10.1126/science.286.5439.509

Beaufort, L. B., Massé, P. Y., Reboulet, A., & Oudre, L. (2022). Network reconstruction problem for an epidemic reaction–diffusion system. *Journal of Complex Networks*, *10*(6), cnac047. https://doi.org/10.1093/comnet/cnac047

Bjørnstad, O. N. (2018). *Epidemics: Models and data using R*. Springer.

Chowell, G., Hengartner, N. W., Castillo-Chavez, C., Fenimore, P. W., & Hyman, J. M. (2004). The basic reproductive number of Ebola and the effects of public health measures: The cases of Congo and Uganda. *Journal of Theoretical Biology*, *229*(1), 119–126. https://doi.org/10.1016/j.jtbi.2004.03.006

Christakis, N. A., & Fowler, J. H. (2013). Social contagion theory: Examining dynamic social networks and human behavior. *Statistics in Medicine*, *32*(4), 556–577. https://doi.org/10.1002/sim.5408

Citron, D. T., Guerra, C. A., Dolgert, A. J., Wu, S. L., Henry, J. M., Sánchez C, H. M., & Smith, D. L. (2021). Comparing metapopulation dynamics of infectious diseases under different models of human movement. *Proceedings of the National Academy of Sciences*, *118*(18), e2007488118. https://doi.org/10.1073/pnas.2007488118

Colizza, V., Pastor-Satorras, R., & Vespignani, A. (2007). Reaction-diffusion processes and metapopulation models in heterogeneous networks. *Nature Physics*, *3*(4), 276–282. https://doi.org/10.1038/nphys560

Cucinotta, D., & Vanelli, M. (2020). WHO declares COVID-19 a pandemic. *Acta Bio-Medica: Atenei Parmensis*, *91*(1), 157–160. https://doi.org/10.23750/abm.v91i1.9397

Eubank, S., Guclu, H., Anil Kumar, V. S., Marathe, M. V., Srinivasan, A., Toroczkai, Z., & Wang, N. (2004). Modelling disease outbreaks in realistic urban social networks. *Nature*, *429*(6988), 180–184. https://doi.org/10.1038/nature02541

Ferrari, M. J., Bjørnstad, O. N., & Dobson, A. P. (2005). Estimation and inference of R0 of an infectious pathogen by a removal method. *Mathematical Biosciences*, *198*(1), 14–26. https://doi.org/10.1016/j.mbs.2005.08.002

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441. https://doi.org/10.1093/biostatistics/kxm045

Fu, X., Small, M., & Chen, G. (2013). *Propagation dynamics on complex networks: Models, methods and stability analysis*. John Wiley.

Fumanelli, L., Ajelli, M., Manfredi, P., Vespignani, A., & Merler, S. (2012). Inferring the structure of social contacts from demographic data in the analysis of infectious diseases spread. *PLoS Computational Biology*, *8*(9), e1002673. https://doi.org/10.1371/journal.pcbi.1002673

Gomez-Rodriguez, M., Leskovec, J., & Krause, A. (2012). Inferring networks of diffusion and influence. *ACM Transactions on Knowledge*

Discovery from Data, *5*(4), 1–37. https://doi.org/10.1145/2086737.2086741

Han, S. W., Chen, G., Cheon, M. S., & Zhong, H. (2016). Estimation of directed acyclic graphs through two-stage adaptive lasso for gene network inference. *Journal of the American Statistical Association*, *111*(515), 1004–1019. https://doi.org/10.1080/01621459.2016.1142880

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). Statistical learning with sparsity. *Monographs on Statistics and Applied Probability*, *143*, 143.

Hens, N., Goeyvaerts, N., Aerts, M., Shkedy, Z., Van Damme, P., & Beutels, P. (2009). Mining social mixing patterns for infectious disease models based on a two-day population survey in Belgium. *BMC Infectious Diseases*, *9*(1), 5. https://doi.org/10.1186/1471-2334-9-5

Hill, A. L., Rand, D. G., Nowak, M. A., & Christakis, N. A. (2010). Infectious disease modeling of social contagion in networks. *PLoS Computational Biology*, *6*(11), e1000968. https://doi.org/10.1371/journal.pcbi.1000968

Hota, A. R., Godbole, J., & Pare, P. E. (2021). A closed-loop framework for inference, prediction, and control of SIR epidemics on networks. *IEEE Transactions on Network Science and Engineering*, *8*(3), 2262–2278. https://doi.org/10.1109/TNSE.2021.3085866

House, T., & Keeling, M. J. (2011). Epidemic prediction and control in clustered populations. *Journal of Theoretical Biology*, *272*(1), 1–7. https://doi.org/10.1016/j.jtbi.2010.12.009

Keeling, M. J., Bjørnstad, O. N., & Grenfell, B. T. (2004). Metapopulation dynamics of infectious diseases. In I. Hanski, & O. E. Gaggiotti (Eds.), *Ecology, genetics and evolution of metapopulations* (pp. 415–445). Academic Press.

Keeling, M. J., & Eames, K. T. (2005). Networks and epidemic models. *Journal of the Royal Society, Interface*, *2*(4), 295–307. https://doi.org/10.1098/rsif.2005.0051

Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A. Containing Papers of a Mathematical and Physical Character*, *115*(772), 700–721.

Kiss, I. Z., Miller, J. C., & Simon, P. L. (2017). *Mathematics of epidemics on networks*. Springer.

Kolaczyk, E. D., & Csárdi, G. (2014). *Statistical analysis of network data with R* (Vol. 65). Springer.

Lekone, P. E., & Finkenstädt, B. F. (2006). Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study. *Biometrics*, *62*(4), 1170–1177. https://doi.org/10.1111/j.1541-0420.2006.00609.x

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, *34*(3), 1436–1462. https://doi.org/10.1214/009053606000000281

Meyers, L. A., Newman, M. E. J., Martin, M., & Schrag, S. (2003). Applying network theory to epidemics: Control measures for Mycoplasma pneumoniae outbreaks. *Emerging Infectious Diseases*, *9*(2), 204–210. https://doi.org/10.3201/eid0902.020188

Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Tomba, G. S., Wallinga, J., Heijne, J., Sadkowska-Todys, M., Rosinska, M., & Edmunds, W. J. (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Medicine*, *5*(3), e74. https://doi.org/10.1371/journal.pmed.0050074

Newman, M. (2010). *Networks*. Oxford University Press.

New York Times. (2023). *Coronavirus (Covid-19) data in the United States*. https://github.com/nytimes/covid-19-data

Park, Y., Priebe, C. E., & Youssef, A. (2013). Anomaly detection in time series of graphs using fusion of graph invariants. *IEEE Journal of Selected Topics in Signal Processing*, *7*(1), 67–75. https://doi.org/10.1109/JSTSP.2012.2233712

Pastor-Satorras, R., Castellano, C., Van Mieghem, P., & Vespignani, A. (2015). Epidemic processes in complex networks. *Reviews of Modern Physics*, *87*(3), 925–979. https://doi.org/10.1103/RevModPhys.87.925

Prasse, B., & Van Mieghem, P. (2020). Network reconstruction and prediction of epidemic outbreaks for general group-based compartmental epidemic models. *IEEE Transactions on Network Science and Engineering*, *7*(4), 2755–2764. https://doi.org/10.1109/TNSE.2020.2987771

Raissi, M., Ramezani, N., & Seshaiyer, P. (2019). On parameter estimation approaches for predicting disease transmission through optimization, deep learning and statistical inference methods. *Letters in Biomathematics*, *6*(2), 1–26. https://doi.org/10.30707/LiB6.2Raissi

Read, J. M., Eames, K. T., & Edmunds, W. J. (2008). Dynamic social networks and the implications for the spread of infectious disease. *Journal of the Royal Society, Interface*, *5*(26), 1001–1007. https://doi.org/10.1098/rsif.2008.0013

Rock, K., Brand, S., Moir, J., & Keeling, M. J. (2014). Dynamics of infectious diseases. *Reports on Progress in Physics. Physical Society*, *77*(2), 026602. https://doi.org/10.1088/0034-4885/77/2/026602

Salathé, M., Kazandjieva, M., Lee, J. W., Levis, P., Feldman, M. W., & Jones, J. H. (2010). A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(51), 22020–22025. https://doi.org/10.1073/pnas.1009094108

Savage, D., Zhang, X., Yu, X., Chou, P., & Wang, Q. (2014). Anomaly detection in online social networks. *Social Networks*, *39*, 62–70. https://doi.org/10.1016/j.socnet.2014.05.002

Small, M., & Cavanagh, D. (2020). Modelling strong control measures for epidemic propagation with networks—A COVID-19 case study. *IEEE Access: Practical Innovations, Open Solutions*, *8*, 109719–109731. https://doi.org/10.1109/ACCESS.2020.3001298

Smieszek, T., Barclay, V. C., Seeni, I., Rainey, J. J., Gao, H., Uzicanin, A., & Salathé, M. (2014). How should social mixing be measured: Comparing web-based survey and sensor-based methods. *BMC Infectious Diseases*, *14*(1), 136. https://doi.org/10.1186/1471-2334-14-136

Timme, M., & Casadiego, J. (2014). Revealing networks from dynamics: An introduction. *Journal of Physics A: Mathematical and Theoretical*, *47*(34), 343001. https://doi.org/10.1088/1751-8113/47/34/343001

Tizzoni, M., Bajardi, P., Decuyper, A., Kon Kam King, G., Schneider, C. M., Blondel, V., Smoreda, Z., González, M. C., & Colizza, V. (2014). On the use of human mobility proxies for modeling epidemics. *PLoS Computational Biology*, *10*(7), e1003716. https://doi.org/10.1371/journal.pcbi.1003716

Towers, S., Gomez-Lievano, A., Khan, M., Mubayi, A., & Castillo-Chavez, C. (2015). Contagion in mass killings and school shootings. *PLoS One*, *10*(7), e0117259. https://doi.org/10.1371/journal.pone.0117259

U.S. Census Bureau. (2020). *County population by characteristics: 2020-2022*. https://www.census.gov/data/tables/time-series/demo/popest/2020s-counties-detail.html

van de Geer, S., & Bühlmann, P. (2013). L_0-penalized maximum likelihood for sparse directed acyclic graphs. *The Annals of Statistics*, *41*(2), 536–567. https://doi.org/10.1214/13-AOS1085

Van Effelterre, T., Shkedy, Z., Aerts, M., Molenberghs, G., Van Damme, P., & Beutels, P. (2009). Contact patterns and their implied basic reproductive numbers: An illustration for varicella-zoster virus. *Epidemiology and Infection*, *137*(1), 48–57. https://doi.org/10.1017/S0950268808000563

Vynnycky, E., & White, R. (2010). *An introduction to infectious disease modelling*. OUP Oxford.

Wallinga, J., Teunis, P., & Kretzschmar, M. (2006). Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. *American Journal of Epidemiology*, *164*(10), 936–944. https://doi.org/10.1093/aje/kwj317

Wan, Y., Roy, S., & Saberi, A. (2008). Designing spatially heterogeneous strategies for control of virus spread. *IET Systems Biology*, *2*(4), 184–201. https://doi.org/10.1049/iet-syb:20070040

Wang, Z., Guo, Q., Sun, S., & Xia, C. (2019). The impact of awareness diffusion on SIR-like epidemics in multiplex networks. *Applied Mathematics and Computation*, *349*, 134–147. https://doi.org/10.1016/j.amc.2018.12.045

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440–442. https://doi.org/10.1038/30918

Woodall, W. H., Zhao, M. J., Paynabar, K., Sparks, R., & Wilson, J. D. (2017). An overview and perspective on social network monitoring. *IISE Transactions*, *49*(3), 354–365. https://doi.org/10.1080/0740817X.2016.1213468

Youssef, M., & Scoglio, C. (2011). An individual-based approach to SIR epidemics in contact networks. *Journal of Theoretical Biology*, *283*(1), 136–144. https://doi.org/10.1016/j.jtbi.2011.05.029

Yu, L., Woodall, W. H., & Tsui, K. L. (2018). Detecting node propensity changes in the dynamic degree corrected stochastic block model. *Social Networks*, *54*, 209–227. https://doi.org/10.1016/j.socnet.2018.03.004

Yuan, M., & Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, *94*(1), 19–35. https://doi.org/10.1093/biomet/asm018