medRxiv preprint doi: https://doi.org/10.1101/2021.03.11.21253356; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

Optimal vaccination strategies for COVID-19 based on dynamical social networks with real-time updating

Sibo Cheng[1], Rossella Arcucci[1], Christopher C. Pain[2], Yi-Ke Guo[1]

[1] *Data Science Instituite, Department of Computing, Imperial College London, UK*
[2] *Department of Earth Science & Engineering, Imperial College London, UK*

Abstract

Vaccination strategy is crucial in fighting against the COVID-19 pandemic. Since the supply is limited, contact network-based interventions can be most powerful to set an optimal strategy by identifying high-risk individuals or communities. However, due to the high dimension, only partial and noisy network information can be available in practice, especially for dynamical systems where the contact networks are highly time-variant. Furthermore, numerous mutations of SARS-CoV-2 impact considerably the current infectious probability, requiring real-time network updating algorithms. In this study, we propose a sequential network updating approach based on data assimilation techniques to combine different sources of temporal information. We then prioritise the individuals with high-degree or high-centrality, obtained from the assimilated networks, for vaccination. The assimilation-based approach is compared with the standard method (based on partially observed networks) and a random selection strategy in terms of vaccination effectiveness in a SIR model. The numerical comparison is first carried out using real-world face-to-face dynamical networks collected in a high school, following by sequential multi-layer networks, generated relying on the Barabasi-Albert model emulating the department of Computing at Imperial College London in the UK as an example.

Keywords: Network science, Data assimilation, COVID-19 vaccination, Centrality measure, Multi-layer networks

1. Introduction

The world is still in the middle of a pandemic involving COVID-19. The World Health Organization (WHO) and partners are working together on the response, tracking the pandemic, providing recommendations on critical steps, delivering necessary medical supplies to those in need and, finally, racing for the development and introduction of safe and reliable vaccines. Every year, vaccines save millions of lives. Vaccines work to identify and fend off the viruses and bacteria they attack by training and preparing the body's natural defences, the immune system. If the body is eventually exposed to such disease-causing germs, the body is ready to kill them instantly, avoiding illness. More than 50 vaccine candidates for COVID-19 are currently in trials[1]. And several have already been distributed in all countries to protect individuals. No other vaccine in the human history is so eagerly anticipated

[1] https://www.who.int/publications/m/item/draft-landscape-of-covid-19-candidate-vaccines

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.03.11.21253356; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

since until now no drugs are demonstrated to be available to treat COVID-19. Different types of vaccines are put into supply and recently started to be used [41], such as AstraZeneca, Pfizer, Moderna and Gamaleya. Until February 7th 2021, over ten millions vaccine doses have been already administered in the United Kingdom. Since the vaccination capacity is still limited until now, people who are most at risk, such as healthcare workers and aged population [39], are given priority. In fact, vaccination strategies play an essential role in preventing the rapid diffusion of COVID-19 pandemic, especially with the recent delay of vaccine delivery[2]. Clustering analysis is investigated following transmission cascades in local social communities. Among all connecting clusters, special attention has been given to educational settings, including high schools and universities [29]. Much effort has been devoted to maintaining the possibility of face-to-face teaching during the pandemic. On the other hand, thousands of clusters and outbreaks of COVID-19 have been reported in educational establishments. Hence, finding an optimal vaccination strategy for students and staff become vital to protect children and young people.

A continuous effort has been made for several decades to develop the simulation of infectious diseases based on observed social networks, including, for instance, H1N1 influenza (face-to-face contact network) [8] and HIV (sexual contact network) [30]. Social network-based analysis for disease spread modelling has been widely implemented since the outbreak of COVID-19 [36], [19], with the help of SIR (Susceptible-Infected-Recovered) or SEIR (Susceptible-Exposed-Infected-Recovered) model. When the network structure of contacts is (at least) partially observable, network-based interventions are most helpful to determining an optimal vaccination strategy under a limited capacity which has been proved in a variety of infectious diseases [38], [48]. These strategies are usually based on some individual-level measures, such as node degree or graph centrality, which requires knowledge of the full network. Furthermore, a significant variance of COVID infectious probability is observed [14] according to ages or activities. Meanwhile, many connecting clusters of COVID-19 have been identified at schools and workplaces[54], where individuals share similar characteristics. Thus the infectious probability of intra-connections inside those clusters could be considered homogeneous. This fact leads to the idea of multi-layer networks modelling where the infectious probability may vary from layer to layer.

Much effort is given to use network-based information for formulating optimal policy responses to COVID-19, including social distancing and countrywide lockdown [5]. However, the observation of social networks is often noisy (with either missing connections or mistaken edge weights), and for most of the time, incomplete [48]. Obtaining precise knowledge is most challenging since face-to-face contact networks are strongly time-variant. The noise-level could be up to 74% (missing edges) for observed connection networks, as mentioned by [32]. On the contrary, as pointed out by [2], contact tracing applications can significantly reduce the rate of infection in the studied population when the participation rate is above 60%. In other words, it is critical to maintaining an error level inferior to 40%. Therefore, a considerable gap can be found between the requiring precision and the available data of the tem-

[2] https://www.bbc.co.uk/news/world-europe-55771223

medRxiv preprint doi: https://doi.org/10.1101/2021.03.11.21253356; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

poral networks. A real-time updating of prior network knowledge is thus essential to improving vaccine efficiency.

In this paper, by investigating how the accuracy of network data could impact the vaccination effectiveness, we propose a real-time network updating approach based on sequential DA techniques. Originally developed in meteorological and environmental science, DA has been applied to a wide variety of industrial domains, including geophysical modelling [7], hydrology [11] and economics [42]. Recently, a sequential DA algorithm is also used for real-time parameter identification in the SIR model for COVID spread simulation [52]. An important advantage of using DA, compared to other statistical models for network reconstruction(e.g [46]) is that DA is widely used for large-dimensional problems with noisy and limited prior data. DA and dynamical network data have been combined in [12] which proposes a graph clustering approach to efficient localization of error covariances within an ensemble-variational DA framework. [28] presents a relation between statistical inference using graphical models and optimal sequential estimation algorithms such as Kalman filtering. In this work, DA is employed for a real-time update of the network, including novel information from dynamical observations. It contributes to leverage the information embedded in different noisy/incomplete observations by an optimisation process to reconstruct the current network. This is computationally feasible for large-scale problems thanks to the sparsity of the contact networks. Here, we propose two DA models for different parametrizations:

1. The first consists of reconstructing the complete contact network structures by observing edges in temporal sub-networks (as described in Section 4);
2. The second adjusts inhomogeneous infectious probabilities in a multi-layer network modelling (as described in Section 5).

These two models are respectively applied to

1. a real-world dynamical network data set describing the contacts of a French high school students in a week [22], collected using wearable sensors;
2. generated scale-free multi-layer networks, where each layer stands for a social community/cluster, determined by individual characteristics such as age or activity.

Preliminary analysis is performed to understand the data structure (clustering, classes, grades) of the high school contact networks and to demonstrate the time-variance. The same data set, collected in a high school in Lyon, has been used to simulate the COVID outbreak and to estimate the reproductive ratio R_0 in [36]. It is also shown in their work that the study of contact networks in schools or work-places could lead to more optimal contact-limiting strategies, such as self-isolation or countrywide lockdown. As for multi-layer systems, the dynamical networks are generated using the Barabasi-Albert model [1], with a power law degree distribution. The latter exists widely in real social networks. Since the mutations of SARS-CoV-2 have been continuously arisen, the infectious probability in each network layer is supposed to be time-variant, following an additive stochastic process. In both cases, the SIR simulation is carried out with realistic assumptions of COVID-19 to simulate the SARS-CoV-2 propagation while real-time observations are generated synthetically based on preliminary network analysis. The DA models proposed in

medRxiv preprint doi: https://doi.org/10.1101/2021.03.11.21253356; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

this paper are general, which could be applied to various scenarios with different types of real-world dynamical networks and observation data.

In summary, in this work we

- simulate the COVID-19 propagation and vaccination impact using real or generated multi-layer networks with the SIR model.
- propose a DA framework, with two different network parametrizations, to sequentially update the network structure based on noisy prior information and real-time observations.
- compare different graph measures, such as node degree, betweenness centrality for vaccination prioritization criteria on prior and assimilated networks.

The paper is organized as follows. Section 2 introduces the graph-based diffusion modelling and vaccination strategies. Data assimilation principle and adaptation of graph data are presented in Section 3. Section 4 shows numerical experiments in real-world social contact networks, and Section 5 shows experiments with multi-layer networks. Section 6 closes the paper with conclusions and future work.

2. Graph-based diffusion modelling and vaccination strategies

2.1. SIR model

The analysis of the diffusion is conducted using a standard SIR model [3], which is a model without an explicit analytical solution of three interrelated non-linear ordinary differential equations:

$$\frac{dS}{dt} = -\beta \frac{IS}{N} \tag{1}$$

$$\frac{dI}{dt} = \beta \frac{IS}{N} - \gamma I \tag{2}$$

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

where S denotes the susceptible population size, I denotes the infected people that are not separated from the population. The recovered population R represents the patients that not infectious anymore, vaccinated or being quarantined from the population. The total number of the population is N. The parameters β and γ denote the transmission of the virus infection and the recovery rate, respectively. The SIR assumption has been widely adapted to simulate the COVID-19 propagation [13],[52] since reported COVID reinfection cases (e.g [51]) are still rare until now. The SIR model has also been broadly used in network-based disease simulations through random walk based simulations [30]. Each node symbolize an individual in the social network, whose status can alter from susceptible to infected (S-I) or infected to recovered (I-R), according to the random walk through temporal edges [17].

medRxiv preprint doi: https://doi.org/10.1101/2021.03.11.21253356; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

2.2. Graph-based vaccination strategy

Both disease spread simulation and optimal vaccination modelling based on social networks have received an increasing interest for different types of infectious disease [43]. We consider an directed or undirected graph $\mathcal{G}$ that is a pair of sets $\mathcal{G} = (V, E)$, where $V = \{v_1, v_2...v_n\}$ represents the set of individuals (graph nodes) and the set E contains the edges, each connecting a pair of individuals. Each graph edge $e \in E$ is represented by a triple $e = (v_i, v_j, w_{i,j})$ where v_i, v_j are the two endpoints and $w_{i,j} \in \mathbb{R}$ is the edge weight. For unweighted graphs $w_{i,j} \in \{0, 1\}$, while for weighted graphs $w_{i,j}$ could represent the frequency or the intimacy of the contact. In epidemic spread modelling, the infectious probability $p_{i,j}$ from the individual i to j (and vice versa) is often in function of $w_{i,j}$, $p_{i,j} = \mathcal{IP}(w_{i,j})$. We remind that $p_{i,j}$ may also depend on individual-level characteristics of v_i and v_j, such as age or activities. The connecting graph, also called as a social network, can be fully represented by the associated adjacency matrix $\mathbf{A} = \{A_{i,j}\}_{i,j=1,...,n}$. We use three Boolean vectors $\{\mathbf{I}_t, \mathbf{L}_t, \mathbf{R}_t\} \in \{\{0, 1\}^n\}^3$ to indicate the status of each individual, either infected, vaccinated or recovered in the SIR model at time t. The recovery period $T_\gamma \in \mathbb{N}$ is a random variable generated individually for each individual.

If we adopt the edge-wise function $\mathcal{IP}(.)$ in the whole network,

$$\mathcal{IP}(\mathcal{G})_{i,j} = \mathcal{IP}(A_{i,j}), \tag{4}$$

the infectious probability vector $\mathbf{I}_t^p \in (0, 1)^n$ at time t in this SIR model reads

$$\mathbf{I}_t^p = \Big(\mathcal{IP}(\mathbf{A}_{t-1})\, \mathbf{I}_{t-1}\Big) \odot (\mathbf{1}_n - \mathbf{R}_t) \odot (\mathbf{1}_n - \mathbf{L}_{t-1}) \odot (\mathbf{1}_n - \mathbf{I}_{t-1}), \tag{5}$$

where $\mathbf{1}_n = [1, 1...1]^T$ and $\odot$ stands for the vector-wise Hadamard product.

Following a uniform probability distribution, the vector of infections $\mathbf{I}_t$ is simulated using $\mathbf{I}_t^p$ and $\mathbf{I}_{t-1}$. The only controllable variable in Eq. 5 is the vaccination vector $\mathbf{L}_t$.

Different graph-based vaccination strategies can be employed to enhance the immunization impact with a limited vaccination capacity. The state of art approaches are usually determined by observed individual- or community- level social connections, often involving classical graph measures, for instance, graph degree, betweenness centrality [21] or community links[9]. Much effort has also been devoted to putting these strategies into practical settings where significant positive impacts have been observed [26]. Since available graph data often include non-negligible uncertainties (missing vertices or edges), statistical models are commonly employed to provide an optimal estimation of these graph measures. Practical approaches involve, for example, the fixed choice designs (FCD) [37] and the nomination strategy [18], both based on an estimation of the graph degree. Even with partially observed dynamical networks, the vaccination strategy could be significantly improved in terms of reducing the maximum infected number and delaying the disease propagation, compared to a random choice [53]. Nevertheless, precise knowledge of the network structure is crucial to determine an optimal vaccination strategy. It is

medRxiv preprint doi: https://doi.org/10.1101/2021.03.11.21253356; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

especially compulsory for communities-based approaches (e.g [25], [9]), since graph clustering algorithms can be sensitive to noises. However, the data collection of dynamical social networks remains cumbersome, especially for large dimensional problems. In this paper, we conducted our analysis based on three classical strategies, considered as less sensitive to data noise, compared to community-based approaches.

Random

The individuals to be vaccinated are randomly chosen, according to the number of doses limited, where no network knowledge is used.

Highest degree

For each temporal network, we choose to vaccinate people with the most contacts based on prior knowledge. Only observable individuals are taken into account. The degree $d(v)$ of node v in a network is simply defined as the sum of the column (or the row for undirected graphs) of the adjacency matrix,

$$d(v) = \sum_{k=1}^{n} |A_{k,v}| \tag{6}$$

Highest Centrality

The betweenness centrality [21] $g(v)$ of node v is defined as the number of shortest paths of all pairs of nodes in the graph that pass by the node v,

$$g(v) = \sum_{u \neq q \neq v} \frac{\sigma_{uq}^{\mathbf{A}}(v)}{\sigma_{uq}^{\mathbf{A}}} \quad u, q \in V \tag{7}$$

where $\sigma_{uq}^{\mathbf{A}}$ represents the total number of shortest paths from node u to node q and $\sigma_{uq}^{\mathbf{A}}(v)$ is the number of those paths that pass through v.

Other graph measures relying on detailed understandings of the network (e.g [9]) could also be used to establish a vaccine strategy. However, in real applications, the precise knowledge of the network is often out of reach. Here, our criteria for choosing graph-based vaccination strategies are two-folds: computationally efficient and non-sensitive to observation noises. The latter ensures the "validity" of the methodology even when working with incompleted networks. To enhance our estimation of dynamical contact networks, we make use of data assimilation algorithms.

3. Data assimilation principle and adaptation of graph data

In this section we introduce the variational data assimilation concept and the resolution using a linear estimator. We also introduce the novel approach which combines DA techniques with dynamical network data.

3.1. Variational assimilation and BLUE

DA algorithms aim to combine different sources of noisy information in order to provide a more reliable estimation of the current system. The state variables could be either a physical field or a sequence of parameters. The true state, denoted by $\mathbf{x}^{\text{true}}$, stands for the theoretical value of the state at some given coordinates/time, often out of reach in real-world applications. The objective of the assimilation is to

medRxiv preprint doi: https://doi.org/10.1101/2021.03.11.21253356; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

gain an optimal approximation $\mathbf{x}^a$ of the true state $\mathbf{x}^{\text{true}}$, based on the prior information which are two parts: an initial state estimation $\mathbf{x}^b$ (so-called the background state) and an observation vector $\mathbf{y}$. The former is often issued from prior numerical simulations/predictions while the latter can be obtained via physical measures of some control variables. Their tolerances, regarding theoretical values, are quantified by ϵ_b and ϵ_y,

$$\begin{aligned} \epsilon_b &= \mathbf{x}^b - \mathbf{x}^{\text{true}} \sim \mathcal{N}(0, \mathbf{B}) \\ \epsilon_y &= \mathbf{y} - \mathcal{H}(\mathbf{x}^{\text{true}}) \sim \mathcal{N}(0, \mathbf{O}), \end{aligned} \tag{8}$$

where the observation operator $\mathcal{H}$ from the state space to the observable space is supposed to be known. The probability distributions of the prior error are supposed to be centred Gaussian, characterized respectively by the covariance matrices $\mathbf{B}$ and $\mathbf{O}$.

The key idea in variational methods is to find a balance between the background and the observations using maximum a posteriori (MAP) method [10]. This leads to the loss function weighted by the inverse of $\mathbf{B}$ and $\mathbf{O}$,

$$J_{\text{3D-VAR}}(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \mathbf{x}^b)^T \mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}^b) + \frac{1}{2}(\mathbf{y} - \mathcal{H}(\mathbf{x}))^T \mathbf{O}^{-1}(\mathbf{y} - \mathcal{H}(\mathbf{x})) \tag{9}$$

$$= \frac{1}{2}||\mathbf{x} - \mathbf{x}^b||^2_{\mathbf{B}^{-1}} + \frac{1}{2}||\mathbf{y} - \mathcal{H}(\mathbf{x})||^2_{\mathbf{O}^{-1}}. \tag{10}$$

The optimisation problem defined by the objective function of Eq. (10) is called three-dimensional variational method (*3D-VAR*), which can also be considered as the general equation of variational methods without considering the transition model error. The output of Eq. 10 is denoted as $\mathbf{x}^a$, i.e.

$$\mathbf{x}^a = \underset{\mathbf{x}}{\text{argmin}}\Big(J(\mathbf{x})\Big). \tag{11}$$

If $\mathcal{H}$ can be approximated by some linear operator $\mathbf{H}$, Eq. 11 can be solved via BLUE (Best Linearized Unbiased Estimator) formulation,

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{K}(\mathbf{y} - \mathbf{H}\mathbf{x}^b) \tag{12}$$

$$\mathbf{P}_{\text{A}} = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{B} \tag{13}$$

where $\mathbf{P}_{\text{A}} = \text{Cov}(\mathbf{x}^a - \mathbf{x}_{\text{true}})$ is the analyzed error covariance and the $\mathbf{K}$ matrix, given by

$$\mathbf{K} = \mathbf{B}\mathbf{H}^T(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{O})^{-1} \tag{14}$$

is so called the Kalman gain matrix. In the rest of this paper, we denote $\mathbf{H}$ as the linearized transformation operator. The case when $\mathcal{H}$ is non-linear is more challenging for finding the minimum of Eq. 9, especially for high-dimensional problems. The resolution often involves gradient descent algorithms (such as "L-BFGS-B" or adjoint-based numerical techniques).

3.2. Online assimilation with graph data

The essential idea is to perform real-time updating of the partially observed dynamical networks based on other available information, such as sub-graph structures

medRxiv preprint doi: https://doi.org/10.1101/2021.03.11.21253356; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

or the current number of infected. To this end, the prior observed network $\mathbf{A}_t^b$ at time t is considered as the background state (i.e, $\mathbf{x}_t^b = \mathbf{A}_t^b$) while other information is embedded in the observation vector $\mathbf{y}_t$.

Once the current contact network is updated thanks to Eq. 10, vaccination strategies could be implemented on the analyzed network $\mathbf{x}_t^a = \mathbf{A}_t^a$ (i.e step 1 $\rightarrow$ step 2 in Fig. 1) which is a more accurate approximation of the true state. The degree and the betweenness centrality of the assimilated network is given by

$$d_t^a(v) = \sum_{k=1}^{n} |(\mathbf{A}_t^a)_{k,v}|, \quad g^a(v) = \sum_{u \neq q \neq v} \frac{\sigma_{uq}^{\mathbf{A}^a}(v)}{\sigma_{uq}^{\mathbf{A}^a}}. \tag{15}$$

where $(\mathbf{A}_t^a)_{k,v}$ denotes the k, vst element of the adjacency matrix $\mathbf{A}_t^a$. Similar expressions of $d_t^b(v)$ and $g^b(v)$ taken on the background state can be given using $\mathbf{A}^b$ and $\sigma^{\mathbf{A}^b}$. The principle of real-time assimilation with graph data is illustrated in Fig. 1 where the virus propagation is simulated using the SIR model as described in Section 2.2 between two vaccination steps. Compared to the overlapped graph, the advantage of working with temporal networks is that the temporal correlation could be considered. In fact, an individual can be active solely during a relatively small period of time as shown later in Section 4.1. Therefore, instead of using overlapped graph (even available), analysing temporal networks can result in a more optimal real-time vaccination strategy.

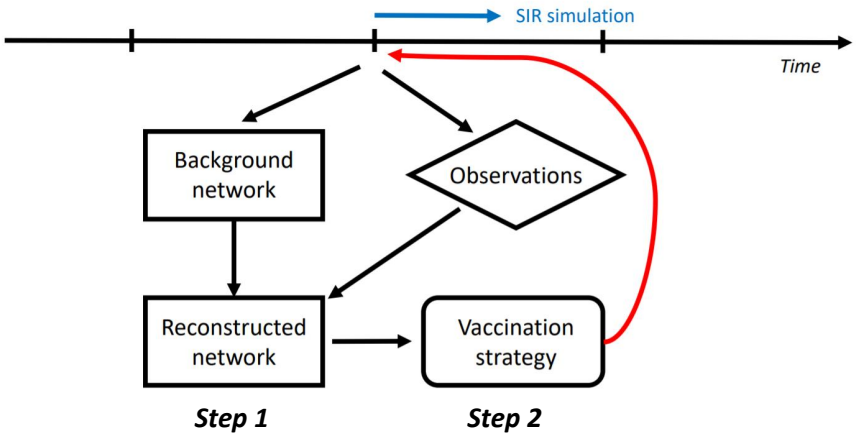


Figure 1: Illustration of real-time DA updating for partially observed contact networks

A major challenge for implementing DA algorithms with graph data is the computational cost since the adjacency matrix $\mathbf{A}_t$, considered as the state variable is a two-dimensional vector. We can rely on the assumption of graph sparsity and appropriate parameterization to reduce the computational burden. In this work, we propose two DA frameworks for dynamical networks updating, respectively introduced in Section 4 and 5. The former aims to reconstruct the full network with observations of sub-graphs while the latter attempts to adjust the parameterized community-wise infectious probability, relying on a multi-layer modelling. These two modellings, relatively at local and global scale, show also the flexibility of this data assimilation framework.

medRxiv preprint doi: https://doi.org/10.1101/2021.03.11.21253356; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

4. Numerical experiments in real-world social contact networks

4.1. Assumptions and preliminary analysis

This study is based on the recently (before the COVID outbreak) collected face-to-face contact data in a French high school [22], which has been used to simulate COVID outbreak [36]. The connection networks of 329 students (coverage of 86% of the students) in a high school of Lyon are available for 7374 time steps in a week. For the sake of simplicity, we condense the dynamical graph to 78 time steps by overlapping every 100 consecutive networks. Each time condensed time step symbolizes $30 \sim 60$ minutes. The temporal networks remain sparse since the average graph density (i.e. number of non-zero edges divided by the number of node pairs) equals to 0.76%. All contact networks are assumed undirected, which means the associated adjacency matrices are all symmetric (i.e., $\mathbf{A}_t = \mathbf{A}_t^T$) and the virus could spread in both directions of an edge. According to [36], the infectious probability (of a 20-second contact) in this network can be estimated as $p \approx 0.1\% \sim 1\%$. However, this estimated probability might be contested regarding newly discovered SARS-CoV-2 variants [27]. In this paper, in order to adequately investigate the optimality of different vaccination strategies, we fix the infectious probability to $p = 2\%$. The average recovery period in the SIR model is set to 60 time steps (around 4 to 5 days), following a uniform probability distribution, i.e. $T_\gamma \sim \text{unif}(55, 65)$.

We start by performing some preliminary analysis of the network data to better understand the underlying graph structures. The overlapped network (i.e. $\sum_{t=1}^{78} \mathbf{A}_t$) of all time steps is shown in Fig. 2(a) where a clear community structure can be observed. Identifying these communities is crucial for simulating the disease spread [31], especially for highly infectious virus like SARS-CoV-2 [47], as well as determining optimal vaccination strategies. To this end, we make use of the Fluid community detection algorithm proposed by [45] which is advantageous for sparse graphs since the algorithm complexity is *linear* to the number of non-zero edges in the network, i.e. $\mathcal{O}(|E|)$.

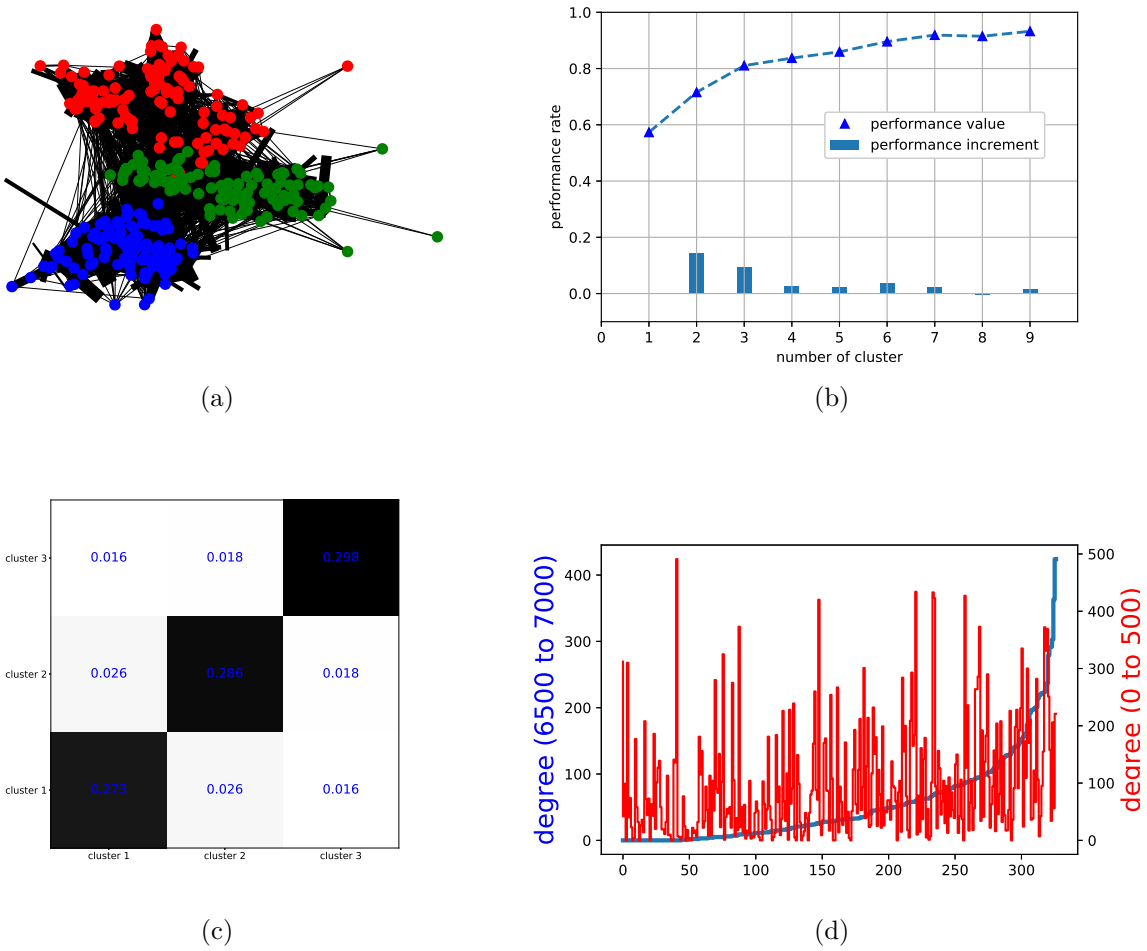
In real applications, specifying the number of communities remains usually cumbersome. Here, we apply several times the community detection algorithms against different assumed community numbers k_c, before evaluating the performance rate $p^r(\mathcal{C})$ [20] of the obtained partition $\mathcal{C}$. The latter, defined as

$$p^r(\mathcal{C}) = \frac{|E_c| + \left(n(n-1) - |E_{\bar{c}}|\right)}{\frac{1}{2}n(n-1)}. \tag{16}$$

where $|E_c|, |E_{\bar{c}}|$ indicate respectively the number of edges intra- and inter-clusters. The performance rate is commonly used as an indicator for finding the optimal community number, which is a standard approach for graph clustering problems. According to the result presented in Fig. 2(b) where we observe clearly a stationary of performance rate starting from $k_c = 4$, we choose to proceed with the optimal number of clusters $k_c^o = 3$. The final clustering result is displayed in Fig. 2(a) where clusters/communities are shown in red, green and blue color. The three detected communities are equivalently distributed, as shown by the reordered adjacency matrix (Fig. 2(c)), respectively with 106, 110 and 111 nodes. From a practical point of view, these communities could be considered as different grades or classes in the

medRxiv preprint doi: https://doi.org/10.1101/2021.03.11.21253356; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

high school, with a similar structure to the graph data presented in [23].



(a) (b)

(c) (d)

Figure 2: Preliminary analysis of the dynamical high-school connection network: (a): Overlapped contact network, (b): Performance rate $p^r(\mathcal{C})$ against assumed community number (c): Reordered adjacency matrix after clustering (d): Node degree distribution of the first and the last 50 time steps.

4.2. DA modelling and numerical results

Since it is infeasible to collect contact networks via wireless equipment in all educational settings post lockdown, the objective of this study is to enhance the vaccination strategy when only partial/noisy information is available, for instance, via tracing applications. For this reason, the full contact networks $\mathbf{A}_t^{\text{true}}$ are supposed to be out of reach. In terms of background states and observations, we suppose the temporal network is only partially observable *a priori* where 50% to 70% of nodes are missing in the background estimation of the network $\mathbf{A}_t^b \in \mathbb{R}^{329\times329}$. The missing nodes are selected randomly and kept invariant at all time steps. In reality, the missing nodes could refer to, for example, people who haven't installed the tracing application on their smartphones. On the other hand, we dispose an observation vector $\mathbf{y}_t$ which contains the sub-networks for each of these three detected clusters. In other words, we suppose the intra-community contacts of students in each class/-grade are fully observable by $\mathbf{y}_t$. The objective is to performing DA algorithms

medRxiv preprint doi: https://doi.org/10.1101/2021.03.11.21253356; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

sequentially to correct the knowledge of the background network relying on the observed sub-networks. The transformation operator $\mathbf{H}$ is thus linear (sub-Identity matrix) and the DA problem is solved via BLUE as shown in Eq. 10. $\mathbf{x}_b^t = \text{vect}(\mathbf{A}_t^b)$ and $\mathbf{y}_t$ are vectorized with Identity error covariances $\mathbf{B}$ and $\mathbf{O}$ as demonstrated in Fig. 3.

After each vaccination, the SIR model is applied to simulate the virus propagation until the next time step, as summarized in Eq. 5. An essential advantage of BLUE-type formulation with invariant prior covariances is that the Kalman gain matrix could be computed offline *a priori* since it is invariant to current $\mathbf{x}_b$ and $\mathbf{y}$. Thus the computational cost of DA can be reduced considerably. The vaccination capacity is fixed $2\%(= 6$ individuals of all students for all strategies (random, highest degree, highest centrality) presented in Section 2.2, based on prior or assimilated graphs.

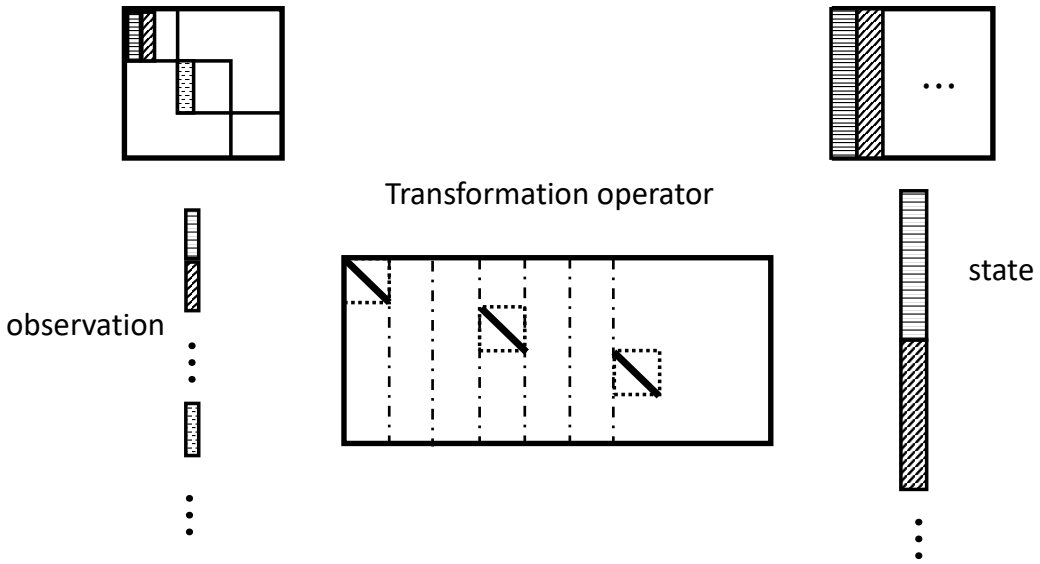


Figure 3: DA process at the current time where intra-connections are observable

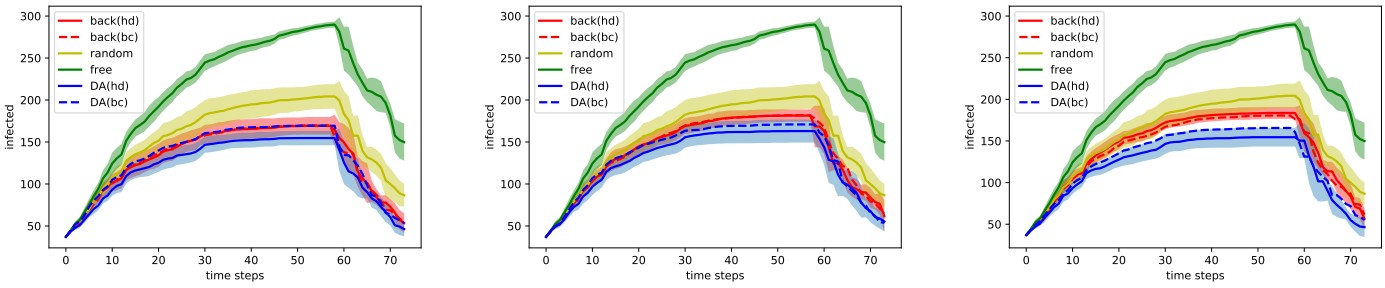
The evolution of the number of infected $|\mathbf{I}_t|$, according to different vaccination strategies, is displayed in Fig. 4 where the percentage of missing nodes in the background state is fixed respectively as 50%, 60% and 70%. To acquire robust numerical results, each type of simulation with or without vaccinations is repeated 10 times and their average values are drawn in solid or dashed curves in Fig. 4. Standard deviations of the simulations (except dashed lines) are also displayed with transparent shades to ensure the robustness of the comparison. The averaged maximum number of infected of each strategy is shown in Table 1. We remind that vaccinations take place at every time step for 6 selected students ($\approx 2\%$ of the population) after the simulation of virus propagation with a contamination probability of 2% for each temporal edge. The initial infected $\mathbf{I}_{t=0}$, commonly used for all simulations, is randomly simulated with a probability of $P\Big((\mathbf{I}_{t=0})_k\Big) = 10\%$ for $k = 1, \ldots, 329$.

From Fig. 4, we observe that almost all averaged curves rise to a high point and

medRxiv preprint doi: https://doi.org/10.1101/2021.03.11.21253356; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

peaked around $t = 50 - 60$ when all individuals are either infected or vaccinated. Since the vaccination process takes place in a relatively short period (a week), we suppose the infected individuals are not detected in real-time. As a consequence, a student can be vaccinated after being infected by the virus, leading to vaccine failure. This fact emphasizes the importance of the vaccination strategy chosen. What can be clearly observed in Fig. 4 is the decreasing of the infected number, according to the vaccination strategy in the order of free (no vaccination) $\rightarrow$ random $\rightarrow$ background $\rightarrow$ assimilated (DA). This order is globally consistent, regardless of time. In the first place, all vaccination strategies manage to significantly reduce the number of infected and delay virus propagation, compared to the free simulation (green curve). In terms of maximum infected numbers, for all three cases, the peak value reduced in average $26\%, 34\%, 34\%, 40\%$ and 37%, respectively for random, background with highest degree, background with highest centrality, assimilated with highest degree and assimilated with highest centrality. All other strategies are dominated by the assimilated curves, especially when proceeding with the highest degree strategy. The difference, in particular between background and assimilated curves, can be more significant if working with large-scale networks. On the other hand, for background-network-based strategies, a growth of maximum infected number against prior error level is noticed in Table. 1 while the results based on assimilated networks remain robust. This fact promotes the use of data assimilation on network data when prior error level can not be precisely specified. We remind that the missing nodes at each time step are generated independently with no temporal correlation, explaining why reasonably good results can be obtained with 70% missing nodes. In summary, numerical results show that the DA-based real-time updating of networks improve the impact of vaccination considerably, resulting in reducing virus spread.

In these experiments, the use of node degree (solid curves) and betweenness centrality, for both background (red) and assimilated (blue) cases, exhibits similar performances. Such fact indicates a high-level (non-negligible) inter-clusters connections where a contrary case can be found in section 5.



(a) (b) (c)

Figure 4: Evolution of infected against different prior error level (percentage of unobserved vertices): (a) 50%, (b) 60%, (c) 70%. Standard deviations are also displayed by transparent shades.

medRxiv preprint doi: https://doi.org/10.1101/2021.03.11.21253356; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

strategy \ prior error level	50 %	60 %	70 %
Free	88%	88%	88%
Random	62%	62%	62%
Background (hd)	51%	55%	56%
Background (bc)	52%	55%	55%
Assimilated (hd)	47%	50%	47%
Assimilated (bc)	51%	52%	50%

Table 1: Maximum number of infected (in percentage) against different vaccination strategies

5. Experiments with multi-layer networks

5.1. multi-layer modelling of scale-free networks

As stated in recent researches [35], the infectious probability of COVID-19 can differ significantly from different communities of population, according to their age, gender, activity and so on. For example, both the transmissibility and the mortality rate is reported to be higher for aged people, soliciting appropriate strategies to protect this fraction of the population. Furthermore, SARS-CoV-2 variants may also vary geographically [4], leading to inhomogenous transition probabilities. Since the outbreak of the COVID-19 pandemic, a continuous effort has been given to understanding the behaviour of virus infection with respect to individual-level (e.g. aged people [40]) and community-level (e.g. healthcare workers [50]) characteristics. These phenomena lead to the idea of using multi-layer networks, where different types of connections exist between graph nodes (see Fig. 5(a)), for simulating the virus spread in social networks. In general, multi-layer (also known as "multiplex") networks [15] are widely used to study graph diffusion problems [24] and to define generalized versions of Pagerank[16]. Recently, multi-layer modelling has also been applied to COVID-19 spread simulation [49] where each layer refers to a potential contamination community, such as school, workplace or transport. Appropriate use of these layer information can optimise vaccination strategies as mentioned in [6], by prioritising the population with high risk and high transmissibility.

Since the collection of large-scale face-to-face contact multi-layer dynamical networks is extremely complicated, we rely on conceptual modelling in this work to further examine the performance of the novel approach. Dynamical contact networks of 1000 individuals and 5 layers (each of 200 nodes) are synthetically generated, where each layer suggests a specific group of the population, according to their age or activities (e.g. students, healthcare workers). Assuming all the edges in the temporal networks are fully observable, our objective here is to calibrate the time-variant infectious probabilities $\{p_{i,t}\}_{i=1,\dots,5}$ based on the observation of infected number in each of those layers $\{\mathbf{I}_{i,t}\}_{i=1,\dots,5}$. The temporal variance of $\{p_{i,t}\}_{i=1,\dots,5}$ can be a consequence of SARS-CoV-2 mutations. More precisely, the values of $\{p_{i,t}\}_{i=1,\dots,5}$ update every 5 time steps, following a stochastic process,

$$p_{i,5t_m+1} = \max(p_{i,5t_m} + \delta_{p,m}, 0) \quad \text{for} \quad t_m \in \mathbb{N} \tag{17}$$

medRxiv preprint doi: https://doi.org/10.1101/2021.03.11.21253356; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

where $\delta_{p,m} \sim unif(-0.04\%, 0.04\%)$ and the observation vector consists of incremental infected numbers $\Delta I_{i,t} = I_{i,t} - I_{i,t-1}$. For inter-layer connections, the infectious probability is determined by the layer of the receiving nodes, i.e.

$$\mathcal{IP}(\mathcal{G}_t)_{i,j} = \mathcal{IP}\Big((\mathbf{A}_t)_{i,j} \times p_{i,t}\Big), \tag{18}$$

as shown in Fig. 5(a). It is worth mentioning that the associated adjacency matrix $\mathbf{A}_t$ is no longer symmetric under this assumption. Nevertheless, the network virus spread modelling in section2.2 remains valid.

As for the generation of temporal networks, we depend on the concept of scale-free networks [44] where the degree distribution follows a power law,

$$P_{\text{sf}}(k) \sim k^{-\gamma} \tag{19}$$

where $P_{\text{sf}}(k)$ stands for the probability of a node to have k connections while $2 \leq \gamma \leq 3$ is a chosen parameter. To simulate intra-connections in each layer, we use the Barabasi-Albert (BA) model [1], which is scale-free with $\gamma = 3$, incorporating two important concepts in graph theory: growth and preferential attachment[33], widely existed in social networks. Therefore, BA model is a reference tool to generate real-world-like networks, including web connections or citation networks. For generating a BA network, nodes are added to the network consecutively where the probability of the new node to be connected with the existing node v writes

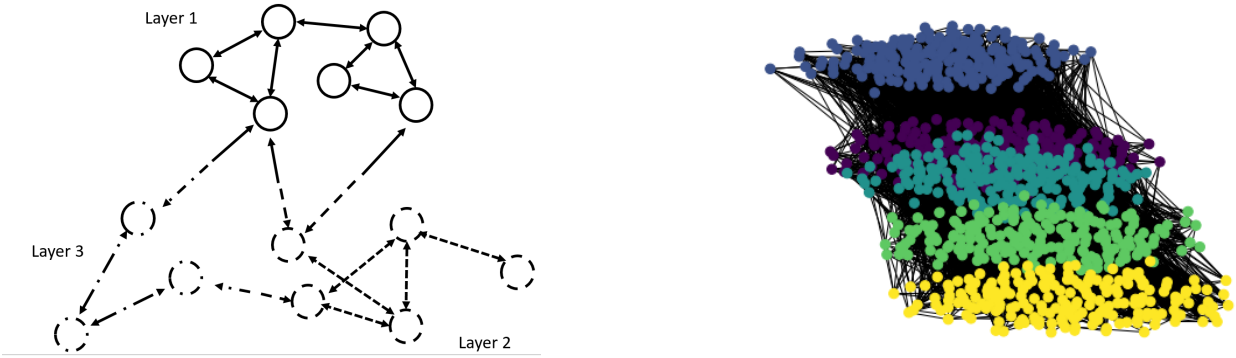
$$P_{BA}(v) = \frac{d(v)}{\sum_j d(j)}. \tag{20}$$

The denominator in Eq. 20 represents twice the current number of edges in the network. Individuals with a higher degree own a stronger ability to grab links added to the BA network, which is an adequate assumption for social networks. Moreover, the inter-layers connections are generated randomly with a density of 0.5%, much sparser than intra-layer edges. Eventually, an example of a complete temporal network is drawn in Fig. 5(b) where the five layers are shown in different colors.

Since temporal edges are supposed to be known in this modelling, we aim to estimate $\{p_{i,t}\}_{i=1..5}$ based on the evolution of the infected numbers in all five layers. In fact, we can predict $\Delta\{I_{i,t}\}_{i=1..5}$ via a prior estimation of $\{p_{i,t}\}$, establishing a state-observation mapping $\mathbf{H} \in \mathbb{R}^{5\times5}$ for DA algorithms. More precisely, the DA problem could be addressed as

$$\mathbf{x}^b = \begin{pmatrix} p^b_{1,t} \\ p^b_{2,t} \\ p^b_{3,t} \\ p^b_{4,t} \\ p^b_{5,t} \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} \Delta I_{1,t} \\ \Delta I_{2,t} \\ \Delta I_{3,t} \\ \Delta I_{4,t} \\ \Delta I_{5,t} \end{pmatrix}, \quad \mathbf{H} = 200 \times \mathbf{N}\left(\mathbf{A}_t\mathbf{I}_t\right) \odot (\mathbf{1}_n - \mathbf{L}_t) \tag{21}$$

medRxiv preprint doi: https://doi.org/10.1101/2021.03.11.21253356; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .



(a) (b)

Figure 5: (a): Illustration of multi-layers network modelling where the infectious probability depends on the layer of the reception node. (b): Different layers in one temporal contact network where, for example, the yellow layer could represent the community of academic staff in the department of Computing at Imperial College London and the other layers stand for students of different grades ($\mathbf{CI}_a$).

where

$$\mathbf{N} = \begin{pmatrix} 1_{1\times 200}, 0_{1\times 200}, 0_{1\times 200}, 0_{1\times 200}, 0_{1\times 200} \\ 0_{1\times 200}, 1_{1\times 200}, 0_{1\times 200}, 0_{1\times 200}, 0_{1\times 200} \\ 0_{1\times 200}, 0_{1\times 200}, 1_{1\times 200}, 0_{1\times 200}, 0_{1\times 200} \\ 0_{1\times 200}, 0_{1\times 200}, 0_{1\times 200}, 1_{1\times 200}, 0_{1\times 200} \\ 0_{1\times 200}, 0_{1\times 200}, 0_{1\times 200}, 0_{1\times 200}, 1_{1\times 200} \end{pmatrix}. \tag{22}$$

The simulation/vaccination framework is similar to the one in Section 4 with a vaccination rate of $\approx 2\%$ of the population at each time step. This means all people will be vaccinated before $t = 50$. For all assimilations, the error covriances are set to be identity matrices as in Section 4. Our goal is to determine an optimal vaccination order based on available noisy information. In order to cover more possible scenarios, we set various initial probabilities $\{p_{i,0}\}$, as shown in table 2, denoted as $\mathbf{CI}_a, \ldots, \mathbf{CI}_f$. For the sake of simplicity, $\{p_{i,0}\}$ always follow a decreasing order from layer 1 to layer 5. Typically, the initial probabilities in $\mathbf{CI}_f$ are more homogeneous, compared to $\mathbf{CI}_a$ or $\mathbf{CI}_e$. To give an example, $\mathbf{CI}_a$ could be used to simulate, for instance, the scenario in the department of computing at Imperial College where nearly 800 students plus faculty members can be found. The layer with high infectious probability may consist of professors, (senior) researchers and HR officers while the other four layers can stand for graduate or undergraduate students of different grades. The former community possesses definitively a much higher average age, in contrast to the latter. Furthermore, each community holds a dense intra-connections, coherent with our model assumption. The diversity of initial conditions ($\mathbf{CI}_a, \ldots, \mathbf{CI}_f$) ensures the robustness of the proposed approach.

medRxiv preprint doi: https://doi.org/10.1101/2021.03.11.21253356; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5
$\mathbf{CI}_a$	2.5%	1%	1%	1%	1%
$\mathbf{CI}_b$	3.5%	1.5%	1%	0.5%	0.5%
$\mathbf{CI}_c$	2.5%	2.5%	2.5%	0.5%	0.5%
$\mathbf{CI}_d$	4.5%	1.5%	1%	0.5%	0.5%
$\mathbf{CI}_e$	3.5%	2.5%	1%	1%	0%
$\mathbf{CI}_f$	2%	2%	1.5%	1%	1%

Table 2: Initial infectious probability $\{p_{i,0}\}$ in different layers

The experiments set-up is similar to the one in Section 4. While computing the node degree and the betweenness centrality, the graph edges are weighted by either the background ($\{p^b_{i,t}\}$) or the analyzed ($\{p^a_{i,t}\}$) layer probabilities. Since the layer information is unattainable *a priori*, background networks are set to be homogeneous (i.e $\{p^b_{1,t} \equiv p^b_{2,t} \equiv p^b_{3,t} \equiv p^b_{4,t} \equiv p^b_{5,t}\}$). The evolution of the infected number, issued from a Monte Carlo test of 10 simulations, is illustrated in Fig. 6. The stand deviation is represented by colored transparent zones. We also display the result of using exact $\{p_{i,t}\}$ (instead of $\{p^b_{i,t}\}$ (red) or $\{p^a_{i,t}\}$(green)) for vaccination in yellow. This curve is thus considered as the optimal target for the assimilation-based strategy. When vaccinating the nodes with the highest degree, a substantial advantage of the DA approach (solid green line) compared to the background (solid red line) one, can be noticed in all six sub-figures of Fig. 6. In fact, both the maximum infected number and the averaged standard deviation have been significantly reduced, as confirmed in table. 3. On the contrary, DA gives much less impact when selecting the individuals with highest centrality as shown by dashed lines in Fig. 6. A reasonable explanation could be the phenomena of brokerage [34]. The endpoints of the few inter-layer edges play an essential role in virus spread. These nodes, also known as "broker", do not necessarily own a high degree in the graph. However, since many shortest paths pass by them from one layer to another, the betweenness centrality may peak at these nodes with or without adjusting $\{p_{i,t}\}$. Such fact shows that when the precise knowledge of inhomogeneous infectious probability is out of reach, proceeding with highest centrality might be a robust choice. Nevertheless, both the dashed green line and the dashed red line are dominated by the solid green line (assimilated networks with highest degree) in all six sub-figures.

We also notice that for Fig. 6(a,b,d) where the five layers exhibit more variance for the initial probabilities, the assimilated curve is much closer to the optimal one. In fact, optimally vaccinating a very inhomogeneous network requires less accurate knowledge of layer probabilities as long as the most infectious layers can be identified. For example, proceeding with $(5\%, 1\%, 1\%, 1\%, 1\%)$ and $(7\%, 0.5\%, 0.5\%, 0.5\%, 0.5\%)$ for vaccine priorities may lead to very similar results.

medRxiv preprint doi: https://doi.org/10.1101/2021.03.11.21253356; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

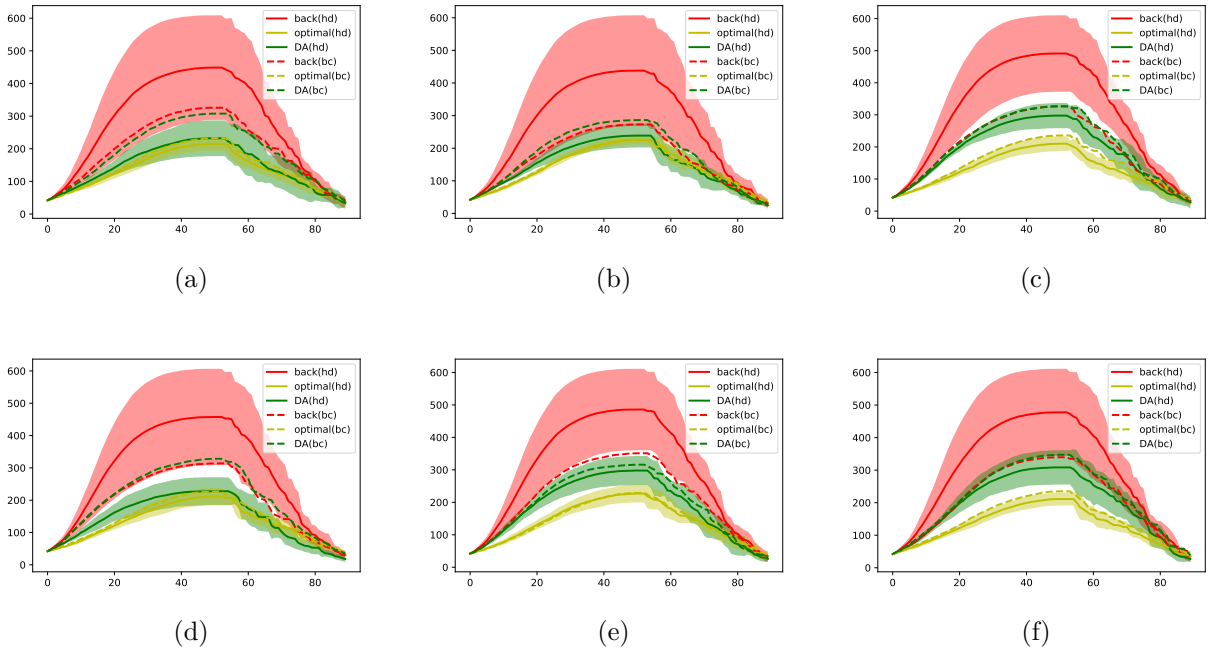


Figure 6: Evolution of infected number (average of 10 simulations) following initial conditions $\mathbf{CI}_a$ .. $\mathbf{CI}_f$

	highest degree					
	max			std		
	prior	DA	true	prior	DA	true
$\mathbf{CI}_a$	44.9%	23.2%	21.5%	10.5%	3.7%	1.7%
$\mathbf{CI}_b$	43.8%	23.9%	22.6%	10.9%	2.4%	1.4%
$\mathbf{CI}_c$	49.1%	29.8%	21.0%	8.1%	2.8%	1.7%
$\mathbf{CI}_d$	45.8%	22.8%	21.1%	9.7%	2.9%	1.8%
$\mathbf{CI}_e$	48.6%	29.8%	22.7%	8.4%	3.2%	2.0%
$\mathbf{CI}_f$	47.8%	30.9%	21.1%	9.1%	3.6%	1.6%

Table 3: Averaged maximum infected number and averaged standard deviation when using node degree as order of vaccination priority

	highest centrality					
	max			std		
	prior	DA	true	prior	DA	true
$\mathbf{CI}_a$	32.6%	30.8%	23.3%	6.0%	5.1%	1.6%
$\mathbf{CI}_b$	27.3%	28.6%	22.5%	2.8%	2.6%	2.2%
$\mathbf{CI}_c$	32.7%	32.6%	23.6%	4.4%	3.4%	2.0%
$\mathbf{CI}_d$	31.3%	32.8%	22.7%	30.0%	30.0%	18.0%
$\mathbf{CI}_e$	35.1%	31.6%	22.9%	36.3%	24.3%	17.8%
$\mathbf{CI}_f$	34.0%	34.7%	23.5%	3.8%	3.9%	2.2%

Table 4: Averaged maximum infected number and averaged standard deviation when using betweenness centrality as order of vaccination priority

medRxiv preprint doi: https://doi.org/10.1101/2021.03.11.21253356; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

The evolution of the normalized true layer probabilities $\frac{p_{i,t}}{\sum_k p_{k,t}}$, as well as their posterior (analyzed) estimation $\frac{p^a_{i,t}}{\sum_k p^a_{k,t}}$ is drawn in Fig. 7. Each row represents one simulation (not the average) from $\mathbf{CI}_a$ to $\mathbf{CI}_f$. For the most infectious layer (first column in Fig. 7), $p_{1,0}$ (in red) is always under-estimated since at the beginning of assimilation we proceed with homogeneous networks using $p^b_{i,0}$. This gap is then rapidly reduced with the increasing of $p^a_{1,t}$, result in a more optimal vaccination strategy. Since vaccinating infected individuals is ineffective, the early phase (around first 20 time steps) of the outbreak is crucial to delay the COVID spread because the most active individuals (either in terms of degree or centrality) can be infected very quickly. Therefore, the DA correction at the beginning of the vaccination process plays an essential role in reducing the propagation speed. On another note, the strong oscillation of red curves implies a high instability of the observation vector $\mathbf{y}_t = [\Delta I_{i,t}]_{i=1..5}$ due to sampling uncertainties. For real applications, the balance between the background state $\mathbf{x}^b_t$ and the observation $\mathbf{y}_t$, adjusted by error covariances $\mathbf{B}$ and $\mathbf{O}$, is worth further investigation. In addition, the correlation between different layers (for example, healthcare workers and hospitalised patients) can also be taken into account by the extra-diagonal elements in $\mathbf{B}$ and $\mathbf{O}$.

In summary, the assimilation-based vaccination strategy shows competitive performance in this multi-layer modelling even though the assimilated layer probabilities are not always accurate. Using the assimilated temporal networks with "highest degree" dominate other approaches, with a smaller average infected number and lower standard deviation.

6. Conclusion and Future Work

Despite that continuous effort, suchlike vaccination or countrywide lockdown, has been given, it is still unclear these days how the COVID-19 pandemic could play out. Determining an optimal vaccination strategy is essential for long-term combating against the COVID, especially with arising SARS-CoV-2 mutations. When the ability to vaccinate the entire population is restricted, using temporal contact network information can improve significantly the vaccination impact on slowing down the disease propagation. This may allow the loosening of some restrictions, which is crucial to rescue the economy from the current pandemic. In this paper, we propose a data assimilation framework to monitor the evolution of social contact networks based on different information sources. The assimilated networks are used to govern vaccination strategies by prioritising high-risk individuals. An important strength of this framework compared to other network reconstruction methods, stands for the flexibility of dealing with available data and the efficiency for large-scale networks. We have applied the proposed approach to real high school contact networks with synthetic observations and to real-world-like dynamical multi-layer networks generated using Barbasi-Albert model. The latter is used to simulate virus propagation with inhomogeneous community-level infectious probabilities. In both applications, the proposed method exhibits a significant advantage in terms of effectiveness (smaller infected number) and robustness (lower deviation). The choice of graph measures for identifying high-risk individuals, such as node degree or betweenness centrality has also been discussed through numerical results in this study.

medRxiv preprint doi: https://doi.org/10.1101/2021.03.11.21253356; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

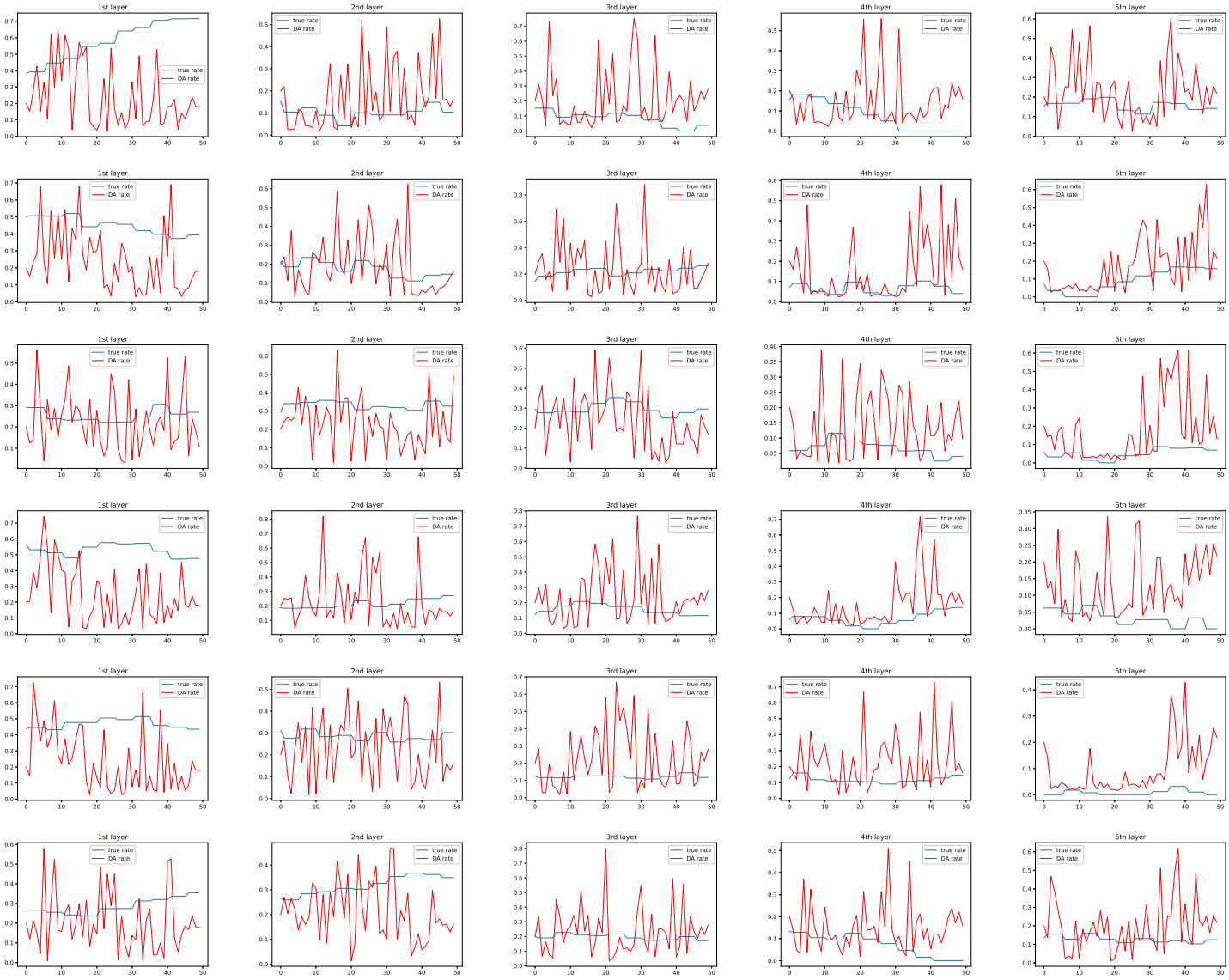


Figure 7: The true and estimated infectious probability (normalized) for one simulation with different initial conditions $\mathbf{CI}_a ... \mathbf{CI}_f$ in each row

We notice that some recent work focuses on establishing data-driven models to predict individual- or community-level infectious probabilities by learning on personal data, involving height, weight and health records. Computational fluid dynamics (CFD) simulations are also being developed to simulate SARS-CoV-2 transmission in schools and offices. Future work can be considered to improve individual-level modelling by incorporating these features in the contact networks. Our work opens promising perspectives to governing optimal vaccination strategies when more network information (e.g. from tracing applications) can be available.

Acknowledgements

This work is supported by the EP/V036777/1 Risk EvaLuatIon fAst iNtelligent Tool (RELIANT) for COVID19 and the EP/T000414/1 PREdictive Modelling with QuantIfication of UncERtainty for MultiphasE Systems (PREMIERE). This

medRxiv preprint doi: https://doi.org/10.1101/2021.03.11.21253356; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

research was partially funded by the Leverhulme Centre for Wildfires, Environment and Society through the Leverhulme Trust, grant number RC-2018-023.

Bibliography

[1] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, 2002.

[2] H. Alsdurf, E. Belliveau, Y. Bengio, T. Deleu, P. Gupta, D. Ippolito, R. Janda, M. Jarvie, T. Kolody, S. Krastev, T. Maharaj, R. Obryk, D. Pilat, V. Pisano, B. Prud'homme, M. Qu, N. Rahaman, I. Rish, J.-F. Rousseau, A. Sharma, B. Struck, J. Tang, M. Weiss, and Y. W. Yu. Covid white paper. *preprint arXiv:2005.08502*, 2020.

[3] R. M. Anderson. Discussion: the kermack-mckendrick epidemic threshold theorem. *Bulletin of mathematical biology*, 53(1-2):1, 1991.

[4] R. S. Baric. Emergence of a highly fit sars-cov-2 variant. *New England Journal of Medicine*, 383(27):2684–2686, 2020. PMID: 33326716.

[5] P. Block, M. Hoffman, I. J. Raabe, J. B. Dowd, C. Rahal, R. Kashyap, and M. C. Mills. Social network-based distancing strategies to flatten the covid-19 curve in a post-lockdown world. *Nature Human Behaviour*, 4(6):588–596, Jun 2020.

[6] J. H. Buckner, G. Chowell, and M. R. Springborn. Optimal dynamic prioritization of scarce covid-19 vaccines. *medRxiv : the preprint server for health sciences*, 2020.

[7] A. Carrassi, M. Bocquet, L. Bertino, and G. Evensen. Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *Wiley Interdisciplinary Reviews: Climate Change*, 9(5):e535, 2018.

[8] S. Cauchemez, A. Bhattarai, T. L. Marchbanks, R. P. Fagan, S. Ostroff, N. M. Ferguson, and D. Swerdlow. Role of social networks in shaping disease transmission during a community outbreak of 2009 h1n1 pandemic influenza. *Proceedings of the National Academy of Sciences*, 108(7):2825–2830, 2011.

[9] Y. Chen, G. Paul, S. Havlin, F. Liljeros, and H. E. Stanley. Finding a better immunization strategy. *Physical Review Letters*, 101:058701, 2008.

[10] S. Cheng, J.-P. Argaud, B. Iooss, D. Lucor, and A. Ponçot. Background error covariance iterative updating with invariant observation measures for data assimilation. *Stochastic Environmental Research and Risk Assessment*, 33(11):2033–2051, 2019.

[11] S. Cheng, J.-P. Argaud, B. Iooss, D. Lucor, and A. Ponçot. Error covariance tuning in variational data assimilation: application to an operating hydrological model, accepted for publication in *Stochastic Environmental Research and Risk Assessment*, 2020.

medRxiv preprint doi: https://doi.org/10.1101/2021.03.11.21253356; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

[12] S. Cheng, J.-P. Argaud, B. Iooss, A. Ponçot, and D. Lucor. A graph clustering approach to localization for adaptive covariance tuning in data assimilation based on state-observation mapping. *arXiv preprint arXiv:2001.11860*, 2020.

[13] I. Cooper, A. Mondal, and C. G. Antonopoulos. A SIR model assumption for the spread of covid-19 in different communities. *Chaos, Solitons & Fractals*, 139:110057, 2020.

[14] N. G. Davies, P. Klepac, Y. Liu, K. Prem, M. Jit, C. A. B. Pearson, B. J. Quilty, A. J. Kucharski, H. Gibbs, S. Clifford, A. Gimma, K. van Zandvoort, J. D. Munday, C. Diamond, W. J. Edmunds, R. M. G. J. Houben, J. Hellewell, T. W. Russell, S. Abbott, S. Funk, N. I. Bosse, Y. F. Sun, S. Flasche, A. Rosello, C. I. Jarvis, R. M. Eggo, and C. C.-. w. group. Age-dependent effects in the transmission and control of covid-19 epidemics. *Nature Medicine*, 26(8):1205–1211, 2020.

[15] M. De Domenico, C. Granell, M. A. Porter, and A. Arenas. The physics of spreading processes in multilayer networks. *Nature Physics*, 12(10), 2016.

[16] M. De Domenico, A. Solé-Ribalta, E. Omodei, S. Gómez, and A. Arenas. Ranking in interconnected multilayer networks reveals versatile nodes. *Nature Communications*, 6(1):6868, 2015.

[17] R. Durrett. Some features of the spread of epidemics and information on a random graph. *Proceedings of the National Academy of Sciences*, 107(10):4491–4498, 2010.

[18] J. Fernández-Gracia, J.-P. Onnela, M. L. Barnett, V. M. Eguíluz, and N. A. Christakis. Influence of a patient transfer network of us inpatient facilities on the incidence of nosocomial infections. *Scientific Reports*, 7(1):2930, 2017.

[19] J. Firth, J. Hellewell, P. Klepac, S. Kissler, A. Kucharski, and L. Spurgin. Using a real-world network to model localized covid-19 control strategies. *Nature Medicine*, 26, 2020.

[20] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75 – 174, 2010.

[21] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977.

[22] M. G'enois and A. Barrat. Can co-location be used as a proxy for face-to-face contacts? *EPJ Data Science*, 7(1):11, 2018.

[23] H. Guclu, J. Read, C. J. Vukotich, Jr, D. D. Galloway, H. Gao, J. J. Rainey, A. Uzicanin, S. M. Zimmer, and D. A. T. Cummings. Social contact networks and mixing among students in k-12 schools in pittsburgh, PA. *PLOS ONE*, 11:1–19, 2016.

[24] M. Gueuning, S. Cheng, R. Lambiotte, and J.-C. Delvenne. Rock–paper–scissors dynamics from random walks on temporal multiplex networks. *Journal of Complex Networks*, 8(2), 2019.

medRxiv preprint doi: https://doi.org/10.1101/2021.03.11.21253356; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

[25] M. Génois, C. Vestergaard, J. Fournet, A. Panisson, I. Bonmarin, and A. Barrat. Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers. *Network Science*, 3:326 – 347, 2014.

[26] G. Harling and J.-P. Onnela. Impact of degree truncation on the spread of a contagious process on networks. *Network Science*, 6(1):34–53, 2018.

[27] Y. J. Hou, S. Chiba, P. Halfmann, C. Ehre, M. Kuroda, K. H. Dinnon, S. R. Leist, A. Schäfer, N. Nakajima, K. Takahashi, R. E. Lee, T. M. Mascenik, R. Graham, C. E. Edwards, L. V. Tse, K. Okuda, A. J. Markmann, L. Bartelt, A. de Silva, D. M. Margolis, R. C. Boucher, S. H. Randell, T. Suzuki, L. E. Gralinski, Y. Kawaoka, and R. S. Baric. Sars-cov-2 d614g variant exhibits efficient replication ex vivo and transmission in vivo. *Science*, 370(6523):1464–1468, 2020.

[28] A. T. Ihler, S. Kirshner, M. Ghil, A. W. Robertson, and P. Smyth. Graphical models for statistical inference and data assimilation. *Physica D: Nonlinear Phenomena*, 230(1-2):72–87, 2007.

[29] S. Ismail, V. Saliba, J. Bernal, M. Ramsay, and S. Ladhani. Sars-cov-2 infection and transmission in educational settings: a prospective, cross-sectional analysis of infection clusters and outbreaks in england. *The Lancet Infectious Diseases*, 2020.

[30] M. J. Keeling and K. T. Eames. Networks and epidemic models. *Journal of the Royal Society Interface*, 2(4):295–307, 2005.

[31] I. Z. Kiss, D. M. Green, and R. R. Kao. Disease contact tracing in random and clustered networks. *Proceedings of the Royal Society B: Biological Sciences*, 272(1570):1407–1414, 2005.

[32] J. H. Koskinen, G. L. Robins, P. Wang, and P. E. Pattison. Bayesian analysis for partially observed network data, missing ties, attributes and actors. *Social Networks*, 35(4):514 – 527, 2013.

[33] P. Krapivsky and D. Krioukov. Scale-free networks as preasymptotic regimes of superlinear preferential attachment. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 78:026114, 2008.

[34] S.-W. Kwon, E. Rondi, D. Z. Levin, A. D. Massis, and D. J. Brass. Network brokerage: An integrative review and future research agenda. *Journal of Management*, 46(6):1092–1120, 2020.

[35] A. T. Levin, W. P. Hanage, N. Owusu-Boaitey, K. B. Cochran, S. P. Walsh, and G. Meyerowitz-Katz. Assessing the age specificity of infection fatality rates for covid-19: systematic review, meta-analysis, and public policy implications. *European Journal of Epidemiology*, 35(12):1123–1138, 2020.

[36] S. Mauras, V. Cohen-Addad, G. Duboc, M. D. la Tour, P. Frasca, C. Mathieu, L. Opatowski, and L. Viennot. Analysis of mitigation of covid-19 outbreaks in workplaces and schools by hybrid telecommuting. *medRxiv*, 2020.

medRxiv preprint doi: https://doi.org/10.1101/2021.03.11.21253356; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

[37] C. McCarty, P. D. Killworth, and J. Rennell. Impact of methods for reducing respondent burden on personal network structural measures. *Social Networks*, 29(2):300 – 315, 2007.

[38] L. Meyers. Contact network epidemiology: Bond percolation applied to infectious disease prediction and control. *Bulletin of the American Mathematical Society*, 44:63–86, 2006.

[39] M. C. Mills and D. Salisbury. The challenges of distributing COVID-19 vaccinations. *EClinicalMedicine*, 31:100674, 2021.

[40] A. L. Mueller, M. S. McNamara, and D. A. Sinclair. Why does COVID-19 disproportionately affect older people? *Aging (Albany NY)*, 12(10):9959–9981, 2020.

[41] A. Mullard. How COVID vaccines are being divied up around the world. *Nature*, 2020.

[42] P. Nadler, R. Arcucci, and Y. Guo. Data assimilation for parameter estimation in economic modelling. In *15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS) 2019*, pages 649–656, 2019.

[43] M. Newman. Spread of epidemic disease on networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 66:016128, 2002.

[44] M. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[45] F. Parés, D. G. Gasulla, A. Vilalta, J. Moreno, E. Ayguadé, J. Labarta, U. Cortés, and T. Suzumura. Fluid communities: A competitive, scalable and diverse community detection algorithm. In *Complex Networks & Their Applications VI*, pages 229–240, Cham, 2018. Springer International Publishing.

[46] T. P. Peixoto. Network reconstruction and community detection from dynamics. *Physical Review Letters*, 123:128301, 2019.

[47] K. Prem, Y. Liu, T. W. Russell, A. J. Kucharski, R. M. Eggo, N. Davies, M. Jit, P. Klepac, S. Flasche, S. Clifford, C. A. B. Pearson, J. D. Munday, S. Abbott, H. Gibbs, A. Rosello, B. J. Quilty, T. Jombart, F. Sun, C. Diamond, A. Gimma, K. van Zandvoort, S. Funk, C. I. Jarvis, W. J. Edmunds, N. I. Bosse, and J. Hellewell. The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. *Lancet Public Health*, 5(5):e261–e270, 2020.

[48] J. Rushmore, D. Caillaud, R. J. Hall, R. M. Stumpf, L. A. Meyers, and S. Altizer. Network-based vaccination improves prospects for disease control in wild chimpanzees. *Journal of the Royal Society Interface*, 11(97):20140349, 2014.

[49] L. F. Scabini, L. C. Ribas, M. B. Neiva, A. G. Junior, A. J. Farfán, and O. M. Bruno. Social interaction layers in complex networks for the dynamical epidemic modeling of covid-19 in brazil. *Physica A: Statistical Mechanics and its Applications*, 564:125498, 2021.

medRxiv preprint doi: https://doi.org/10.1101/2021.03.11.21253356; this version posted March 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

[50] N. Shaukat, D. M. Ali, and J. Razzak. Physical and mental health impacts of covid-19 on healthcare workers: a scoping review. *International Journal of Emergency Medicine*, 13(1):40, 2020.

[51] R. L. Tillett, J. R. Sevinsky, P. D. Hartley, H. Kerwin, N. Crawford, A. Gorzalski, C. Laverdure, S. C. Verma, C. C. Rossetto, D. Jackson, M. J. Farrell, S. Van Hooser, and M. Pandori. Genomic evidence for reinfection with SARS-CoV-2: a case study. *Lancet Infect Dis*, 21(1):52–58, 2021.

[52] S. Wang, X. Yang, L. Li, P. Nadler, R. Arcucci, Y. Huang, Z. Teng, and Y. Guo. A bayesian updating scheme for pandemics: Estimating the infection dynamics of covid-19. *IEEE Computational Intelligence Magazine*, 15(4):23–33, 2020.

[53] Y. Yang, A. McKhann, S. Chen, G. Harling, and J.-P. Onnela. Efficient vaccination strategies for epidemic control using network information. *Epidemics*, 27:115 – 122, 2019.

[54] S. E. F. Yong, D. E. Anderson, W. E. Wei, J. Pang, W. N. Chia, C. W. Tan, Y. L. Teoh, P. Rajendram, M. P. H. S. Toh, C. Poh, V. T. J. Koh, J. Lum, N. M. Suhaimi, P. Y. Chia, M. I. Chen, S. Vasoo, B. Ong, Y. S. Leo, L. Wang, and V. J. M. Lee. Connecting clusters of COVID-19: an epidemiological and serological investigation. *Lancet Infect Dis*, 20(7):809–815, 2020.