Modeling the Impact of Public Traffic on COVID-19 Spread Rates in Germany

Tobias Krug
Department of Electrical and Computer Engineering, Chair of Automatic Control Engineering (LSR)
Technical University of Munich
Munich, Germany
tobias.krug@tum.de

***Abstract*—This report builds on the idea of the networked SEIR model to recover pandemic spread parameters and apply the identified model for simulation/prediction of pandemic activity based on COVID-19 case data, public transport schedule information and estimated mobility behavior of a population. A data-driven methodology to the problems of transient network structure recovery and time-varying strength of adjacency estimation is presented. A case study on German data is implemented and the resulting models' performances are evaluated numerically and graphically.**

***Index Terms*—Epidemic modeling, System identification, Networked control systems, Data engineering**

I. Introduction

A. Problem Statement

This report deals with the specific situation of COVID-19 spreading in Germany. It aims to contribute to the field of COVID-19 spread models. Hereby, it helps to tackle the current lack of concepts that consider the unique combination of extensive public transport infrastructure and regionally differing COVID-19 legislation as is faced in Germany.

The lack of such a model is in equal parts attributable to:

- the data collection avoidance due to the population's mindset valuing data privacy (GDPR et al.) and
- the niche role of digital solutions to everyday activities like commuting, which would allow data collection in the first place.

Hence, acquisition of actual usage data of public transport networks is complex and revolves around the combination of different (proxy) data sources; namely public transport schedules, behavior data and models of passengers and data sets to account for legislation and pandemic parameters on various geographical scopes.

B. Related Work

Pandemics in general and the COVID-19 pandemic caused by the spread of the SARS-CoV-2 virus specifically are of concern for a number of different academic entities and fields. The works that are of specific relevance for this report can be clustered into six categories:

1) varying approaches to model pandemics [1]
2) varying extensions of SIR models [2], [3] and the analysis of the model's accuracy and its sensitivity [4]
3) global Markov Model approach [5]
4) impact of quarantine/lockdown on the pandemic spread [6], [7] and other special groups at risk [8]
5) concepts to control the pandemic [9], [10], its economic impact [11] and how the right to demonstrate might conflict with pandemic containment [12]
6) approaches to track and predict spread in real time with mobility data [13] or online search terms [14]

The approach in this report builds on the paper and MATLAB implementation of Vrabac et al. [15]. The following section presents the prerequisites to apply it to data from Germany alongside the measures taken to improve the performance of the networked SEIR model.

C. Problem Setup

The problem setup consists of

- raw, unrelated data from various sources and
- a prototypic MATLAB implementation of the networked SEIR model from [15].

Based on this, a data processing pipeline to fuse the different data sets and transform them into a simulation ready format is presented. Subsequently, the parameters of a networked SEIR model are identified and tested using both factual initial conditions and behavior data and counterfactual behavior data to determine the sensitivity of the pandemic spread to changes of the population's mobility behavior.

II. Technical Approach

In this section, I present the concept to use mobility data, behavior data and pandemic data to calibrate and test a networked SEIR model. The theoretical concept of the networked SEIR model is introduced in subsection II-A, which is followed by implementation specifics in subsection II-B.

A. Networked SEIR model

The idea of a networked SEIR model builds on the widely employed concept of epidemic models, that divide populations into compartments, whose relations are described as a set of differential equations. The basic SIR model is extended to the SEIR as introduced by [16] to account for epidemic scenarios in which humans are not immediately infectious after contact but have a static incubation period in becoming so. Therefore it is applicable to analyze virus-induced epidemics like COVID-19. Additional to the standard SEIR model, two sets of spread

rates β are introduced in equations 2 and 3 of [15]. While $\beta_{I1}, \beta_{I2}, \beta_{I3}$ relate to the standard SEIR model spread from infectious to exposed compartments, the new set of spread rates $\beta_{E1}, \beta_{E2}, \beta_{E3}$ allow to model the spread from exposed compartments to other exposed compartments.

The networked SEIR model from [15] takes this idea even further, in that it aims to identify spread parameters for a population divided into many subpopulations, which itself are divided into the SEIR compartments. Each of these subpopulations is modeled to influence the compartments of geographically adjacent regions by parameter $A1$, their own compartments by means of community spreading with parameter $A2$ and due to the travel behavior of humans inscribed by $A3$. $A1$, $A2$ and $A3$ are all so-called adjacency matrices, which model the strength of relationship between network entities. Entities relate to political counties in this analysis.

1) Identifying the networked SEIR model parameters: The identification of the networked SEIR model can be expressed as an optimization problem, that results from discretization of the system of continuous partial differential equations. The following brief discussion is based on the paper and prototypic implementation in [15].

The optimization problem revolves around finding spread parameters that describe the pandemic spread across a number of counties (**??**), for a number of time steps (**??**) by using pandemic data with the sample rate h, which equals 1 in most scenarios.

$$\begin{aligned} m &= N_{counties} = const. \\ n &= N_{timesteps} = const. \\ j &\in \{1, 2, ..., n-1\} \end{aligned} \tag{1}$$

For each of these counties and days of analysis, a difference equation for the compartments exposed, infectious and removed describes the rate of change. The difference equations are expressed in vector form to simply the optimization problem, hence the input data matrices $E_{(n\times m)}$, $I_{(n\times m)}$ and $R_{(n\times m)}$ are reshaped[1] as vectors e, i and r. As we intend to identify one set of spread parameters common to all regions under analysis, the input case data matrices are expected to be normalized using each region's number of residents.

$$e = reshape(E^T_{(n\times m)}, n*m, 1) \tag{2a}$$
$$i = reshape(I^T_{(n\times m)}, n*m, 1) \tag{2b}$$
$$r = reshape(R^T_{(n\times m)}, n*m, 1) \tag{2c}$$
$$\Delta_e = e_{(n+1:end)} - e_{(1:(m-1)*n)} \tag{2d}$$
$$\Delta_i = i_{(n+1:end)} - i_{(1:(m-1)*n)} \tag{2e}$$
$$\Delta_r = r_{(n+1:end)} - r_{(1:(m-1)*n)} \tag{2f}$$
$$\Delta = [\Delta_e; \Delta_i; \Delta_r] \tag{3}$$

While the adjacency matrices $A1$ (models adjacent county spread) and $A2$ (models intra-county spread) are static by nature, $A3$ (models mobility-driven spread) introduces time-dependent strength of connectedness, e.g. number of vehicles moving–and hence changes with each iteration j. To further consider the time-dependent mobility behavior of the population and policy behavior of the government, the parameters ψ_{1_j}, ψ_{2_j} and ψ_{3_j} are introduced. Depending on the data set, these can be either scalar values representing similar behavior in all counties or diagonal matrices, wherein each element of the diagonal describes the mobility behavior and policy information of one region under analysis. Extending the works of Vrabac et al., this work determines the parameters ψ in a data-driven approach in subsection II-B.

$$A1_{s_j} = \psi_{1_j} A1 \tag{4a}$$
$$A2_{s_j} = \psi_{2_j} A2 \tag{4b}$$
$$A3_{s_j} = \psi_{3_j} A3_j \tag{4c}$$

The daily levels of each compartment (2a), (2b) and (2c) are then used to determine the discretized, time-dependent version of the networked SEIR model as described by (6). The networked SEIR model is then parametrized using the parameter vector $[\beta_{E1}, \beta_{E2}, \beta_{E3}, \beta_{I1}, \beta_{I2}, \beta_{I3}, \sigma, \gamma]$. These parameters are identified based on the optimization problem stated as (7), which is the minimum square normed difference of the parametrized networked SEIR model compartments and the actual daily difference of compartments given by the input data.

$$S_j = \mathbb{1} - diag(e_j + i_j + r_j) \tag{5a}$$
$$e_j = e_{(n*(j-1)+1:n*j)} \tag{5b}$$
$$i_j = i_{(n*(j-1)+1:n*j)} \tag{5c}$$
$$r_j = r_{(n*(j-1)+1:n*j)} \tag{5d}$$
$$\begin{aligned} \delta_{e_j} = h[&S_j A1_{s_j} e_j, S_j A2_{s_j} e_j, S_j A3_{s_j} e_j, \\ &S_j A1_{s_j} i_j, S_j A2_{s_j} i_j, S_j A3_{s_j} i_j, -e_j, 0] \end{aligned} \tag{5e}$$
$$\delta_{i_j} = h[0, 0, 0, 0, 0, 0, e_j, i_j] \tag{5f}$$
$$\delta_{r_j} = h[0, 0, 0, 0, 0, 0, 0, i_j] \tag{5g}$$
$$\delta = [\delta_e; \delta_i; \delta_r] \tag{6}$$
$$\begin{aligned} &[\beta^*_{E1}, \beta^*_{E2}, \beta^*_{E3}, \beta^*_{I1}, \beta^*_{I2}, \beta^*_{I3}, \sigma^*, \gamma^*] = \\ &\underset{\beta_{E1}, \beta_{E2}, \beta_{E3}, \beta_{I1}, \beta_{I2}, \beta_{I3}, \sigma, \gamma}{\arg\min} \\ &\{\|\delta[\beta_{E1}, \beta_{E2}, \beta_{E3}, \beta_{I1}, \beta_{I2}, \beta_{I3}, \sigma, \gamma]^T - \Delta\|\} \end{aligned} \tag{7}$$

To ensure plausible results, the optimization problem is constrained. Firstly, the spread parameters are required to be non-negative (8a). Secondly, the incubation rate σ is upper-bounded by a static time delay introduced to estimate the exposed compartment (8b). This delay shifts the data of the infectious compartment data - as derived from government agency data - by $\Delta_{t_{preExposed}}$ days into the past. Thirdly, the cure rate γ is upper-bounded (8c). Lastly, the combined spread rates described by the parameters β and the adjacency matrices A are upper-bounded by the inverse of the sample rate h for each time step j and region $k \in \{1, 2, ..., m\}$ (8d).

[1]Notation refers to the MATLAB implementation https://www.mathworks.com/help/matlab/ref/reshape.html

$$\beta_{E1}, \beta_{E2}, \beta_{E3}, \beta_{I1}, \beta_{I2}, \beta_{I3}, \sigma, \gamma >= 0 \tag{8a}$$

$$\sigma <= 1/\Delta_{t_{preExposed}} \tag{8b}$$

$$\gamma <= 1 \tag{8c}$$

$$|A1_{s_{j_{(k,1:m)}}}|(\beta_{E1} + \beta_{I1}) + |A2_{s_{j_{(k,1:m)}}}|(\beta_{E2} + \beta_{I2}) + |A3_{s_{j_{(k,1:m)}}}|(\beta_{E3} + \beta_{I3}) <= 1/h \tag{8d}$$

2) Simulating the networked SEIR Model: Based on the parametrized networked SEIR model, we can calculate the pandemic spread per compartment, region and day. The entry point for the simulation is the set of initial conditions $\{s_{t0}, i_{t0}, e_{t0}, r_{t0}\}$, which describe the initial pandemic levels of each region, and the parameter vector $[\beta_{E1}, \beta_{E2}, \beta_{E3}, \beta_{I1}, \beta_{I2}, \beta_{I3}, \sigma, \gamma]$. Using (4a), (4b) and (4c) an iterative problem to determine the pandemic levels is stated by the following set of equations:

$$S_j = \mathbb{1} - diag(e_j + i_j + r_j) \tag{9a}$$

$$e_{j+1} = e_j + h(S_j((\beta_{E1} A1_{s_j} + \beta_{E2} A2_{s_j} + \beta_{E3} A3_{s_j}) e_j^T + (\beta_{I1} A1_{s_j} + \beta_{I2} A2_{s_j} + \beta_{I3} A3_{s_j}) i_j^T) - \sigma e_j^T)^T \tag{9b}$$

$$i_{j+1} = i_j + h(\sigma e_j - \gamma i_j) \tag{9c}$$

$$r_{j+1} = r_j + h(\gamma i_j) \tag{9d}$$

B. Implementation

To evaluate the aforementioned approach against German pandemic and mobility data, the networked SEIR model extending the works of Vrabac et al. and related data processing was implemented in MATLAB. This work covers the whole sequence from said data processing to model identification and simulation and is introduced in the following paragraphs. As the RKI data set [17] provides data with a county resolution, this work implemented the networked SEIR model approach using German counties as regions under analysis.

1) Data fusion pipeline: To identify the networked SEIR model as described in subsubsection II-A1, a number of different input data is required.

First and foremost, the COVID-19 pandemic information consisting of daily confirmed cases (confirmed as proxy for infectious, recovered and death as proxy for removed) is needed. This work relies on the official data set for Germany [17]. The data is processed to yield the data matrices $E_{(n\times m)}$, $I_{(n\times m)}$ and $R_{(n\times m)}$, which are then normalized by the number of residents of each county, available from the geopolitical information provided by the government agency BKG [18].

Secondly, the three adjacency matrices have to be calculated. $A1$ can be derived by calculating all adjacent counties for each county using county shape information from [18]. $A2$ is defined as a 2d eye matrix with both dimensions equaling the number of counties under analysis. The time-dependent $A3$ finally models mobility-driven spread and hence depends on mobility information. This work proposes a new method to calculate this matrix based on public transport schedule data [19] and geopolitical data [18]. The detailed approach is introduced in subsubsection II-B2.

Lastly, the adjacency matrices have to be adjusted to actual, time-dependent behavior and policy of the population under analysis. This work implements the scaling based on behavior data, which is recovered from aggregated, anonymized mobile phone data. The matrices $A1, A2$ are hereby scaled using county resolution, total mobility behavior [20]. As $A3$ models the public transport network, another data set that specifically recovered railway usage figures with nationwide resolution is applied for scaling [21]. Though not part of this work, it is possible to introduce other auxiliary information for adjacency scaling. This notably includes information on COVID-19 policies [22], economic figures [23] and fine-grained mobility behavior [24].

The last data processing stage then determines the joint time range, for which pandemic information, transient adjacency and behavior information is available for subsequent SEIR model identification and simulation.

The relationship of data sets and processing stages is depicted in Figure 1.

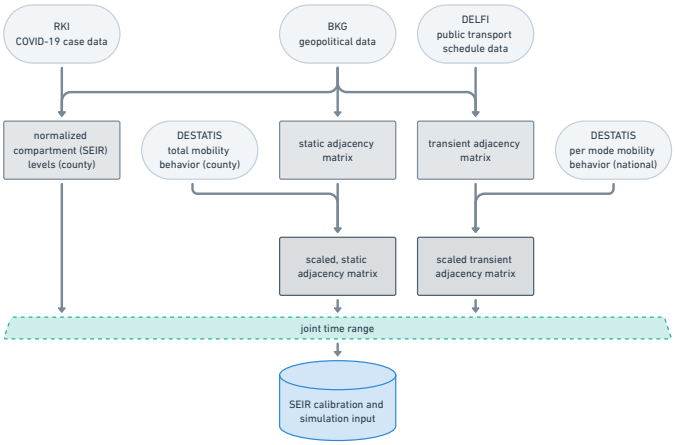


Fig. 1: Data processing pipeline

2) Transient adjacency matrix calculation: The main body of work of this report is the new method to derive the transient adjacency matrix $A3$ from publicly available data in a programmatic way. The entry point for this calculation is the public transport schedule data collected by a project from DELFI e.V. [19]. The data sets contain data on schedules of regional and national agencies operating across all of Germany. As the schedule data's intended use is traveling and routing services, the locations are not related to the geopolitical entity county used in this work. Hence, the coordinate information of each stop is tested against the shape information of all counties, which is retrieved from [18]. This allows to map stops to counties and subsequently calculate the adjacency matrix as the aggregated number of services between two counties.

To prepare the data set–which is provided in the standard GTFS format–for the calculation of the transient adjacency

matrix, the following steps on the respective files are required in the mentioned order:

1) agency.txt: reduce to agencies of interest
2) routes.txt: reduce to routes matching the filtered agency list
3) trips.txt: reduce to trips matching the route_id occurrences of the filtered routes list
4) stop_times.txt: reduce to stop times matching the filtered trips list
5) stops.txt: reduce to stop_id values occurring in the filtered stop times list
6) calendar: reduce to service_id values matching the filtered trips list
7) calendar_dates.txt: reduce to service_id values matching the filtered trips list
8) stops.txt: match latitude and longitude with county polygon from BKG county data and augment stops table with ARS code per stop[2]
9) stop_times.txt: augment stop_times table with ARS code per stop event with the stop_id $\rightarrow$ ARS mapping of the augmented stops table

Based on the processed schedule data and the successful mapping of stop locations to county codes, the adjacency matrix can finally be calculated as follows:

1) combine calendar (cyclic schedule) and calendar_dates (exceptions) tables to calculate the operation status of each service_id for the complete time period
2) iterate over all trips to yield a vector of ARS codes indicating the location of the stops along a trip and use the accompanying service_id to increment the mobility level for all ARS codes (counties) for every day, where this service_id is operated
3) normalize the data[3]
4) assemble a joint adjacency matrix of all individual GTFS data sets, considering the most recent data as most useful and output a joint adjacency matrix with daily mobility within counties and in between counties

3) Parameters to tailor the SEIR model identification and simulation: The calculation of the transient adjacency matrix $A3$ can be tailored using the following set of filters to inspect a subset of public transport agencies:

- includeAgencyList: a positive list of strings to include in the analysis, e.g. "DB" to include "Deutsche Bahn" operated agencies
- excludeAgencyList: a negative list of strings to remove from the analysis, e.g. "Bus" to exclude any bus operating agency

The identification of the networked SEIR model and the subsequent simulation and counterfactual simulation can be configured with the following set of parameters:

- time delays:
  - tLagRemoved: time delay in days to be used for removed compartment estimation based on recovered case data
  - tLagDeath: time delay in days to be used for removed compartment estimation based on death case data; fatalities are considered in this work contrary to the standard SEIR model approach because COVID-19 lead to significant excess mortality [25] that the author assumes to exceed any birth rate changes
  - tLeadExposed: time delay in days by which positive test cases are shifted to estimate the exposed compartment
- tSmoothCases: time duration of smoothing for simulation input compartment data
- h: sampling time of COVID-19 input data
- dateStart: start date of simulation interval (if data is available)
- dateEnd: stop date of simulation interval (if data is available)
- countiesList: specify a list of ARS codes to simulate with
- mobilityLevel: a switch to determine the mobility behavior data to use for adjacency matrix scaling; available options are "nation" using [21], "county" using [20] and "combined" using both
- rPopulation: a threshold of population density to omit any lower-density counties from the analysis
- counterfactuals:
  - type: a switch to select the type of counterfactual to simulate for, available options are "fixed" (a static level of mobility compared to 2019 levels), "scaled" (a multiple of the recovered mobility behavior) and "cappedTop" (an upper bounded version of the recovered mobility behavior)
  - value: the value to apply according to the counterfactual method selected with type

III. Evaluation

A. Experimental Results

Based on the MATLAB implementation, the data sets are processed and the networked SEIR model approach is applied. The results are evaluated from two different perspectives: 1) presentation and interpretation of results (numerically and graphically) for identification and simulation of a networked SEIR model for the complete time period, using combined mobility behavior and comparing the results with and without the use of the transient adjacency matrix and 2) comparison to the original results from [15].

[2]The matching of latitude/longitude tuples with county polygons applies the MATLAB interior point check function iteratively. See https://www.mathworks.com/help/matlab/ref/delaunaytriangulation.isinterior.html

[3]Unfortunately, the schedule data does not provide information on the capacity of a given vehicle or service event. Similarly, the author was not able to acquire data sets with actual usage figures of public transport on a vehicle level. The adjacency matrix is therefore normalized to the maximum sum of services per departing county. The actual usage figures are then applied via the scaling factor ψ_3. Additionally, two other methods of normalization are implemented but produce inferior results: normalization to maximum service count per day for all counties as done by Vrabac et al. and normalization to the sum of services count per starting county per day, which relates to row-wise normalization and results in row-wise stochastic matrices.

1) Networked SEIR model identification and simulation results: The networked SEIR model using German public transport data and mobility behavior data proves to be a very effective approach to model the COVID-19 outbreak even for the multiple waves that occurred until the time of writing. Figure 2 shows a comparison of the results for the networked SEIR model with and without use of the transient adjacency matrix $A3$ as introduced in subsubsection II-B2. The plots show the denormalized data per compartment aggregated over the entire nation. As can be seen, the use of the transient adjacency matrix (subplots Figure 2a, Figure 2c and Figure 2e) drastically improves simulation performance and successfully recovers the second and third outbreak.

To yield these results, the settings according to Table I were applied. The identified parameters and error figures are compared in Table II.

As Table II indicates, including the $A3$ adjacency matrix results in higher error of the normalized compartment, but reduces the error on denormalized levels per compartment. This relates well to the orange curves for the simulated pandemic levels in Figure 2. The deviation on the normalized data presumably results from the incomplete model of the pandemic based on the mobility data alone, as other factors like policy, weather, virus variants, population structure are neglected in this analysis.

The deviations of the model from the real outbreak data starting in April 2021 are likely attributable to the onset of vaccinations, which are not accounted for in the SEIR model and reduce the susceptible compartment, therefore reducing the pandemic activity.

Additionally, Figure 2 presents the possibility to simulate counterfactuals and check their impact with the identified model. As the plots suggest, the seemingly simple policy of limiting the railway usage to 50% of 2019 levels might have reduced the pandemic activity by a large margin[4].

TABLE I: Settings for the identification and simulation for 01-Feb-2020 to 23-Jun-2021

Option	Value
tLagRemoved	7
tLagDeath	7
tLeadExposed	7[5]
tSmoothCases	7
h	1
mobilityLevel	combined
counterfactual type	cappedTop
counterfactual value	0.5

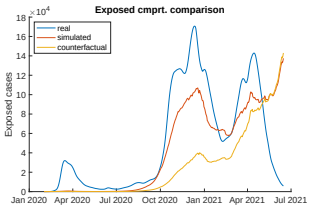
[4]Final levels of removed compartment: real: 3.7182×10^6, simulated: 3.0049×10^6, counterfactual: 1.8377×10^6

[5]The RKI notes a median of 5-6 days and a 95th percentile of 10-14 days incubation period based on a survey of studies. [26]

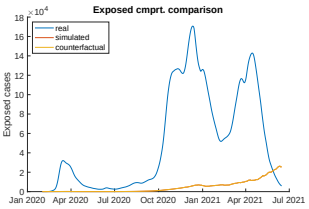
[6]Two error metrics are presented per compartment. The first one is calculated on normalized compartment levels, while the second one is calculated on aggregated, denormalized levels.

TABLE II: Identification results and error metrics[6]of the networked SEIR model for 01-Feb-2020 to 23-Jun-2021

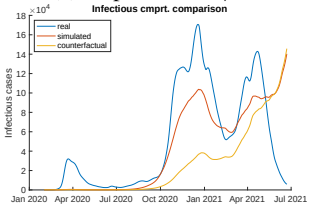
Metric	$A3 \neq 0$	$A3 = 0$
β_E	$[6.8\times10^{-5}, 0.1354, 0.0574]$	$[6.3\times10^{-8}, 0.1497, 0]$
β_I	$[1.8\times10^{-5}, 7.8\times10^{-7}, 3.3\times10^{-4}]$	$[9\times10^{-9}, 2.8\times10^{-8}, 0]$
σ	0.1395	0.1390
γ	0.1394	0.1392
$\frac{\|e_{real}-e_{sim}\|}{\|e_{real}\|}$	$1.955, 0.4588$	$1.132, 0.9379$
$\frac{\|i_{real}-i_{sim}\|}{\|i_{real}\|}$	$2.023, 0.4641$	$1.130, 0.9379$
$\frac{\|r_{real}-r_{sim}\|}{\|r_{real}\|}$	$2.039, 0.4588$	$1.076, 0.9457$



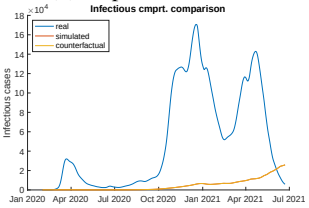
(a) Exposed, $A3 \neq 0$



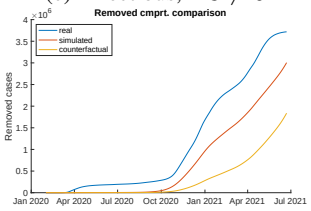
(b) Exposed, $A3 = 0$



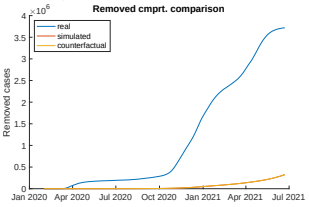
(c) Infectious, $A3 \neq 0$



(d) Infectious, $A3 = 0$



(e) Removed, $A3 \neq 0$



(f) Removed, $A3 = 0$

Fig. 2: Comparison of aggregated, denormalized compartment levels of networked SEIR model with and without $A3$[7]

2) Comparison to the results of Vrabac et al.: This work differs to the approach and results presented in [15] in the following categories:

- The scope of analysis is the entire set of counties of Germany instead of a subset of counties from the U.S.
- The time range includes not just the first COVID-19 outbreak but considers all three waves that occurred until the time of writing.
- The scaling of the adjacency matrix is performed based on measured mobility behavior–which correlates with

[7]The counterfactual line represents the outcome of a theoretical change in the usage behavior of public transport. The plots show the results of applying upper bounded (50%) behavior data as ψ_{3_j}. The simulated and counterfactual lines intersect for subplots (b), (d) and (f) as $A3 = 0$.

policy information–instead of guessed parameters.
- The analysis is validated against denormalized data.

IV. Discussion

The new approach to pandemic modeling based on networked entities represented by adjacency matrices that are programmatically calculated based on official data from government agencies shows good performance for recovery and prediction of the pandemic activity on the nation level.

As was shown, the small number of parameters depicted in Table I is sufficient to tailor the framework. At the same time, different data can be easily included for identification and simulation with the networked SEIR model by extending the behavior/policy factors ψ. Both aspects allow to consider other data sources for COVID-19 simulation or modify the framework to apply it to other epidemics.

Still, Figure 2 and the plots in section A show that the approach presented herein does not model the COVID-19 outbreak either completely or reliably. This is partly caused by neglected effects[8], but also likely influenced by underreporting of cases causing a wrong estimation of the actual pandemic levels. It is further affected by time-varying test policy and number of tests executed–and subsequently positive test ratios–as well as official COVID-19 legislation and varying spread characteristics of the different SARS-CoV-2 virus strains–though the spread supporting factors are still under investigation.

A. Sensitivity Analysis

An extensive sensitivity analysis was not performed due to the large size of the data sets and the extensive set of parameters and settings. Nevertheless, a number of different scopes of analysis and their impact is presented in section A.

V. Conclusion

To conclude, the networked SEIR model from [15] was applied to German geopolitical, transport and behavior data. The experimental results suggest that mobility induced spreading is an important factor, though community spreading within counties still accounts for the majority of transmission (see Table II and [10]). Counterfactual simulation with a modified behavior vector suggests that reduced long-distance rail-bound mobility played a role in reducing the pandemic spread and could be a measure to contain future outbreaks. Future works building on the approach should consider a) addition of a new compartment for vaccinated cases alongside declined immunity over time [27] and b) introduction of novel data sources like policy information, weather data or virus strain characteristics to improve the scaling factors ψ of the adjacency matrices for further improvement of the model's accuracy.

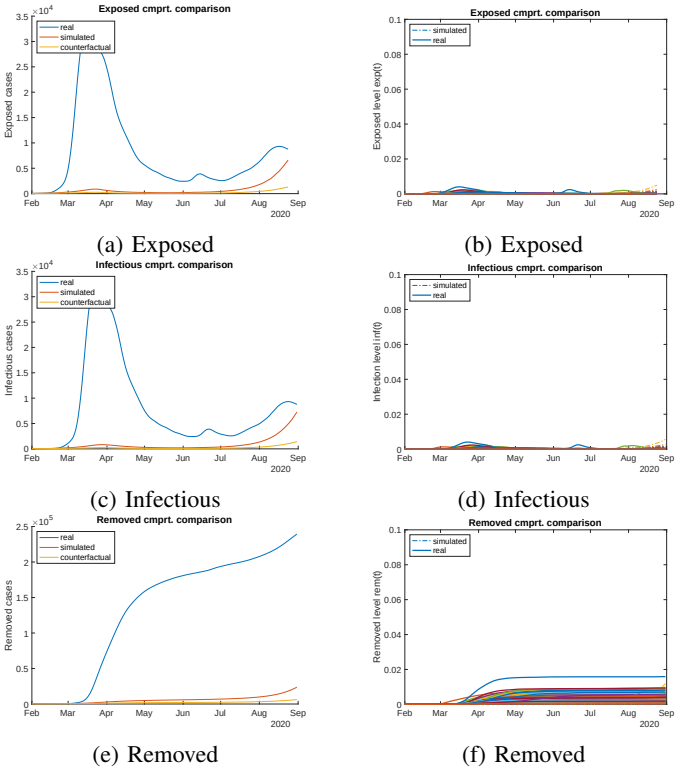
[8]Vaccinations are not considered and mandate the introduction of a new compartment with transitions from susceptible to vaccinated. Additionally, new studies show that neither recovered nor fully vaccinated people exhibit permanent protection from infections. Hence, a SEIRS model as proposed [27] with a transition from removed to susceptible should be investigated.

Appendix A
Variation of Settings of the Networked SEIR Model Identification and Simulation

This section presents the impact of different time ranges on the results of the parameter identification and subsequent simulation. To provide a better understanding of the effects of the options on the behavior of the networked SEIR model, both the aggregated, denormalized compartment levels and the normalized compartment levels per county are presented. As can be seen, aggregated levels are more consistent with the real pandemic activity, while the spread levels of the individual counties show large deviations resulting from this work's focus on the mobility behavior of the population and suggesting to consider additional data for the scaling factors ψ.

A. Wave 1 - from 01-FEB-2020 to 01-SEP-2020

Figure 3 shows how the networked SEIR model fails to recover the spread activity for wave 1. The root cause is unknown, but indicates, that mobility is only one of the many drivers of a pandemic.



(a) Exposed (b) Exposed

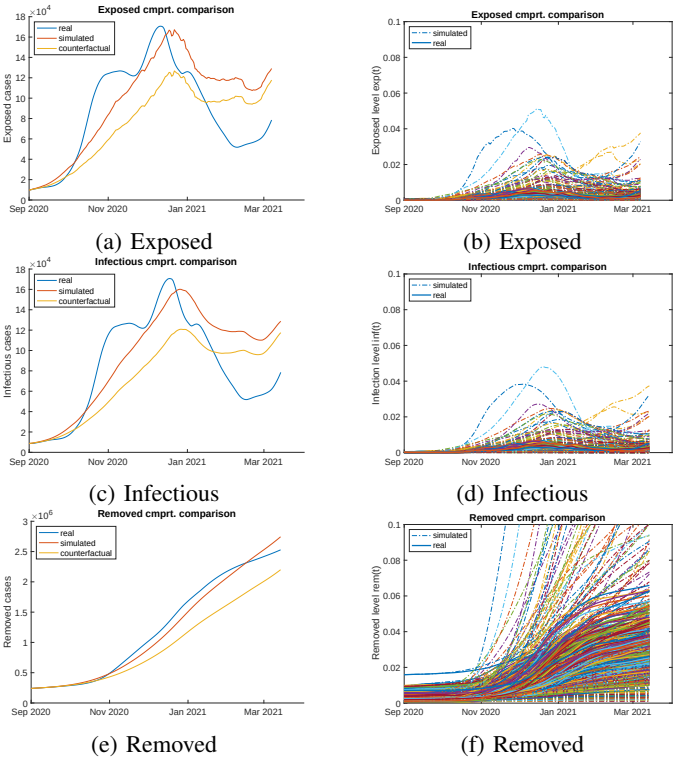
(c) Infectious (d) Infectious

(e) Removed (f) Removed

Fig. 3: Compartment levels, wave 1 with combined mobility; left: normalized per county, right: aggregated denormalized

B. Wave 2 - from 01-SEP-2020 to 15-MAR-2021

Figure 4 is a good example of the performance of the networked SEIR model. As can be seen, the simulated behavior is matching the real behavior both in time and magnitude of the outbreak. The counterfactual shows similar reduction in pandemic activity as was shown for the complete time period in section III.

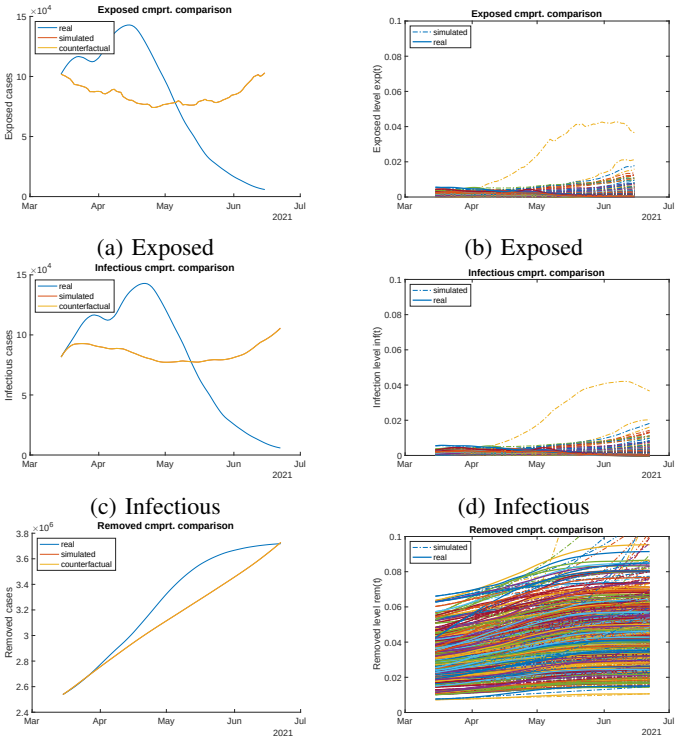


(a) Exposed (b) Exposed

(c) Infectious (d) Infectious

(e) Removed (f) Removed

Fig. 4: Compartment levels, wave 2 with combined mobility; left: normalized per county, right: aggregated denormalized



(a) Exposed (b) Exposed

(c) Infectious (d) Infectious

(e) Removed (f) Removed

Fig. 5: Compartment levels, wave 3 with combined mobility; left: normalized per county, right: aggregated denormalized

C. *Wave 3 - from 15-MAR-2021 to 23-JUN-2021*

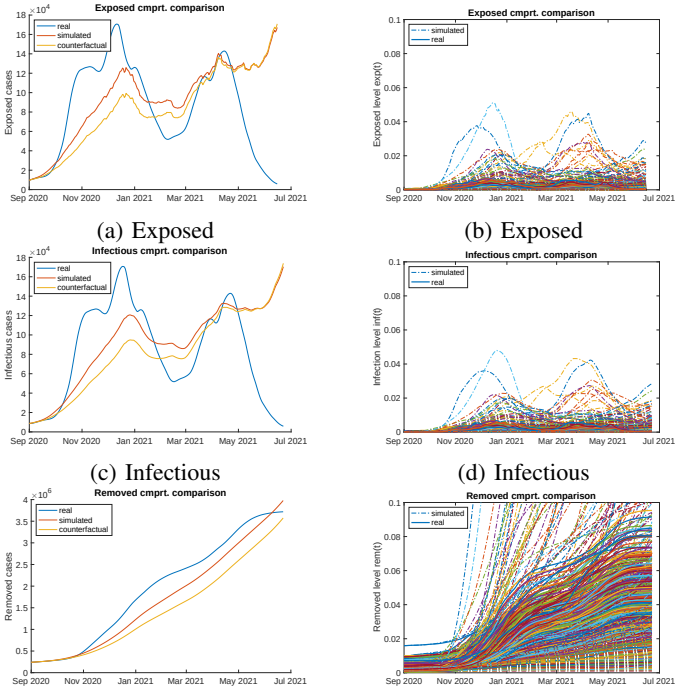
The analysis of the third wave in Figure 5 is a powerful example for the shortcomings of the approach. The plots clearly show that the simulation is not able to match the actual, measured spread activity and instead predicts an almost constant level of infections–though yielding similar removed compartment levels. This suggests to consider a new compartment for vaccinated cases, as the observed decline of real cases likely results from a decrease in the susceptible compartment due to the success of the vaccination program.

D. *Waves 2 and 3 - from 01-SEP-2021 to 23-JUN-2021*

The combined analysis of the second and third wave in Figure 6 shows a similar behavior as presented in section III except for the huge deviations fore the omitted first wave. The plots support the networked SEIR model's ability to recover the two distinct waves purely from the measured mobility behavior of the population under analysis.

APPENDIX B
DATA SETS AND FORMATS

This section introduces specialties and caveats about the data sets used in the report and how their specific formats work. The interested reader is guided to other resources for further reading. The data sets are aggregated and made available in their raw form in the following public git repository free of charge using the git LFS technology: https://github.com/hashkode/covid-data



(a) Exposed (b) Exposed

(c) Infectious (d) Infectious

(e) Removed (f) Removed

Fig. 6: Compartment levels, waves 2&3 with combined mobility; left: normalized per county, right: aggregated denormalized

A. Mobility data - GTFS

The mobility data used herein is provided by the DELFI e.V. and licensed under CC BY 4.0.[19] The data set is provided in the format of GTFS, which is a standardized format originating from Google. The reference documentation for the format describing all involved files, keys and possible uses can be found here: GTFS Reference.

B. COVID-19 data - RKI

The COVID-19 data used in this project is taken from the Robert Koch-Institut (RKI). It consists of:

- Daily caseload per age group, gender and county for Germany in the form of a timetable [17] licensed under DL-DE/BY-2.0 and
- daily vaccination information on a national and state level [28] also licensed under DL-DE/BY-2.0.

C. Mobility behavior data - DESTATIS

The mobility data that is used to prepare the behavior vector mentioned in (4a), (4b) and (4c) is provided by the German Statistics Office named DESTATIS. Two different data sets are used herein:

- Relative change in total mobility behavior compared to 2019 levels per county as a timetable [20] licensed for free use with reference DESTATIS - Copyright allgemein and
- Relative change mobility behavior for different modes of travel compared to 2019 levels aggregated on the national level as a timetable [21] licensed for free use with reference DESTATIS - Copyright allgemein.

REFERENCES

[1] L. Zino and M. Cao, "Analysis, Prediction, and Control of Epidemics: A Survey from Scalar to Dynamic Network Models," *arXiv:2103.00181 [cs, eess, math]*, Feb. 2021.

[2] X.-X. Liu, S. J. Fong, N. Dey, R. G. Crespo, and E. Herrera-Viedma, "A new SEAIRD pandemic prediction model with clinical and epidemiological data analysis on COVID-19 outbreak," *Applied Intelligence*, Jan. 2021.

[3] A. M. Ramos, M. R. Ferrández, M. Vela-Pérez, A. B. Kubik, and B. Ivorra, "A simple but complex enough ϑ-SIR type model to be used with COVID-19 real data. Application to the case of Italy," *Physica D: Nonlinear Phenomena*, vol. 421, p. 132839, July 2021.

[4] S. H. A. Khoshnaw, M. Shahzad, M. Ali, and F. Sultan, "A quantitative and qualitative analysis of the COVID–19 pandemic model," *Chaos, Solitons & Fractals*, vol. 138, p. 109932, Sept. 2020.

[5] Z. E. O. Frihi, J. Barreiro-Gomez, S. E. Choutri, and H. Tembine, "Toolbox to simulate and mitigate COVID-19 propagation," *SoftwareX*, vol. 14, p. 100673, June 2021.

[6] C. R. Wells, J. P. Townsend, A. Pandey, S. M. Moghadas, G. Krieger, B. Singer, R. H. McDonald, M. C. Fitzpatrick, and A. P. Galvani, "Optimal COVID-19 quarantine and testing strategies," *Nature Communications*, vol. 12, p. 356, Jan. 2021.

[8] A. Kouidere, L. E. Youssoufi, H. Ferjouchia, O. Balatif, and M. Rachik, "Optimal Control of Mathematical modeling of the spread of the COVID-19 pandemic with highlighting the negative impact of quarantine on diabetics people with Cost-effectiveness," *Chaos, Solitons & Fractals*, vol. 145, p. 110777, Apr. 2021.

[7] Z. Memon, S. Qureshi, and B. R. Memon, "Assessing the role of quarantine and isolation as control strategies for COVID-19 outbreak: A case study," *Chaos, Solitons & Fractals*, vol. 144, p. 110655, Mar. 2021.

[9] O. Pinto Neto, D. M. Kennedy, J. C. Reis, Y. Wang, A. C. B. Brizzi, G. J. Zambrano, J. M. de Souza, W. Pedroso, R. C. de Mello Pedreiro, B. de Matos Brizzi, E. O. Abinader, and R. A. Zângaro, "Mathematical model of COVID-19 intervention scenarios for São Paulo—Brazil," *Nature Communications*, vol. 12, p. 418, Jan. 2021.

[10] A. Aravindakshan, J. Boehnke, E. Gholami, and A. Nayak, "Preparing for a future COVID-19 wave: Insights and limitations from a data-driven evaluation of non-pharmaceutical interventions in Germany," *Scientific Reports*, vol. 10, p. 20084, Nov. 2020.

[11] J. P. Caulkins, D. Grass, G. Feichtinger, R. F. Hartl, P. M. Kort, A. Prskawetz, A. Seidl, and S. Wrzaczek, "The optimal lockdown intensity for COVID-19," *Journal of Mathematical Economics*, vol. 93, p. 102489, Mar. 2021.

[12] M. Lange and O. Monscheuer, "Spreading the Disease: Protest in Times of Pandemics," *SSRN Electronic Journal*, 2021.

[13] K. Leung, J. T. Wu, and G. M. Leung, "Real-time tracking and prediction of COVID-19 infection using digital proxies of population mobility and mixing," *Nature Communications*, vol. 12, p. 1501, Mar. 2021.

[14] V. Lampos, M. S. Majumder, E. Yom-Tov, M. Edelstein, S. Moura, Y. Hamada, M. X. Rangaka, R. A. McKendry, and I. J. Cox, "Tracking COVID-19 using online search," *npj Digital Medicine*, vol. 4, pp. 1–11, Feb. 2021.

[15] D. Vrabac, M. Shang, B. Butler, J. Pham, R. Stern, and P. E. Pare, "Capturing the Effects of Transportation on the Spread of COVID-19 with a Networked SEIR Model," p. 6, 2020.

[16] J. L. Aron and I. B. Schwartz, "Seasonality and period-doubling bifurcations in an epidemic model," *Journal of Theoretical Biology*, vol. 110, pp. 665–679, Oct. 1984.

[17] Robert Koch-Institut (RKI), "RKI COVID19." https://t1p.de/egmz, 2021.

[18] "Verwaltungsgebiete 1:250 000 mit Einwohnerzahlen (Ebenen), Stand 31.12.." https://t1p.de/0akc, Dec. 2020.

[19] DELFI e.V., "OpenData ÖPNV." https://t1p.de/sy3lq, 2021.

[20] S. B. (Destatis), "Veränderungsrate der Mobilität ggü. 2019," July 2021.

[21] Statistisches Bundesamt (Destatis), "Verkehrsmittel im Fernverkehr." https://t1p.de/cpyf, June 2021.

[22] T. Hale, N. Angrist, R. Goldszmidt, B. Kira, A. Petherick, T. Phillips, S. Webster, E. Cameron-Blake, L. Hallas, S. Majumdar, and H. Tatlow, "A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker)," *Nature Human Behaviour*, vol. 5, pp. 529–538, Apr. 2021.

[23] Statistisches Bundesamt (Destatis), "Konjunkturindikatoren." https://t1p.de/e0hh.

[24] "COVID-19 Community Mobility Report." https://t1p.de/vctj.

[25] World Health Organization, "The true death toll of COVID-19: Estimating global excess mortality." https://www.who.int/data/stories/the-true-death-toll-of-covid-19-estimating-global-excess-mortality.

[26] Robert Koch-Institut (RKI), "Epidemiologischer Steckbrief zu SARS-CoV-2 und COVID-19." https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Steckbrief.html, July 2021.

[27] O. N. Bjørnstad, K. Shea, M. Krzywinski, and N. Altman, "The SEIRS model for infectious disease dynamics," *Nature Methods*, vol. 17, pp. 557–558, June 2020.

[28] Robert Koch-Institut (RKI), "RKI - Coronavirus SARS-CoV-2 - Tabelle mit den gemeldeten Impfungen nach Bundesländern und Impfquoten nach Altersgruppen." https://t1p.de/em9w, 2021.