

# Quantifying the over-estimation of predicted pandemic curves caused by population clustering

Mathias L. Heltberg,<sup>1,2,3,†,\*</sup> Christian Michelsen,<sup>1,†</sup>

Emil S. Martiny,<sup>1</sup> Lasse Engbo Christensen,<sup>4</sup>

Mogens H. Jensen,<sup>1</sup> Tariq Halasa,<sup>5</sup> and Troels C. Petersen<sup>1</sup>

<sup>1</sup>Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17, 2100 Copenhagen E

<sup>2</sup>Laboratoire de Physique, Ecole Normale Supérieure, Rue Lhomond 15, Paris 07505

<sup>3</sup>Infektionsberedskab, Statens Serum Institute, Artillerivej, 2300 Copenhagen S

<sup>4</sup> DTU Compute, Section for Dynamical Systems,

Department of Applied Mathematics and Computer Science,

Technical University of Denmark, Anker Engelunds Vej 101A, 2800 Kongens Lyngby

<sup>5</sup> Animal Welfare and Disease Control, University of Copenhagen,

Gronnegårdsvej 8, 1870 Frederiksberg C

†These authors contributed equally.

\*To whom correspondence should be addressed; [heltberg@nbi.ku.dk](mailto:heltberg@nbi.ku.dk).

**The modeling of Sars-CoV-2 has become a critical aspect of present life. A standard procedure has become the implementation of more or less detailed differential equations which estimate the time development in the number of Susceptible, (Exposed,) Infected, and Recovered (SIR/SEIR) individuals. However, these models are based on very simple assumptions which constitute obvious approximations. In this paper, we introduce an agent based model including spatial clustering. Based on Danish population data, we estimate how this impacts the early prediction of a pandemic and the long term developments. Our results show that early phase SEIR model predictions overestimate the peak number of infected and the equilibrium level by at least a factor two. These results are robust to variation of parameters influencing connection distances within a realistic range and independent of distribution of infection rates.**

Since the emergence of reports of a new disease from Wuhan, China, the pathogen now known as SARS-CoV-2 has spread dramatically, paralyzing societies, resulting in a large number of deaths and severe economic damage worldwide (1, 2). Mathematical models were developed to estimate the reproduction number and guide the authorities in an attempt to minimize the damage caused by this novel virus (3, 4, 5, 6). Generally, the models have been variants of the SIR/SEIR model and vary in complexity including simple deterministic compartmental models (6, 7), meta-population compartmental models (8, 9, 10), individual based models without including spatial specifications (4, 11, 12), and finally spatio-temporal agent based models (13).

One important aspect in the modelling is the ability to predict the infection peak height and the steady state level of individuals having been infected based on the early

rise in the number of infected before governmental interference.

Earlier work has pointed out the importance of including heterogeneity when modeling the spread of infectious disease such as contact patterns between individuals (14), population mixing assumptions (15), heterogeneities caused by super-spreaders (12), and the spatial dependency of COVID-19 (16, 17). However, these mathematical models have not combined heterogeneous elements nor quantified how much the early SIR/SEIR predictions for Sars-CoV-2 might be biased.

We address these deficiencies by constructing a stochastic agent based model (ABM) on a realistic, spatially distributed population where each individual (agent) could be in one of four states and switch according to the rates schematized in Fig. 1A. We used this ABM to investigate how heterogeneities can bias our prediction in the early phases of a pandemic. In order to reduce the space of parameters, we do not include political countermeasures during the pandemic, but rather test the levels of herd immunity and endemic steady state. Before including heterogeneity, we compared the ABM to the corresponding SEIR model, and found them to agree within 5% for all parameter configurations tested (see Supplementary Information).

We constructed a network of spatially distributed contacts using data based on:

- The geographic location of people in Denmark (from Boligsiden (18))
- The average number of contacts per individual per day of 11 (from HOPE (19)).

Given an average infectious period of 4 days, we approximate the average number of effective contacts to be  $\mu = 40$ .

- The average commuting distance  $\rho = 0.1 \text{ km}^{-1}$  and the fraction of long distance commutes  $\epsilon_\rho = 4\%$  (from Statistics Denmark (20))

This is schematized in Fig. 1B (see also Supplementary Information) and all ten parameters in this model are explained in Table 1. Having introduced heterogeneity, the

distribution of connections in this network are created automatically through the population clustering, see Fig. 2A. This naturally leads to individuals living in densely populated areas having higher number of connections.

In an example simulation with 100 initially infected individuals,  $N_{\text{init}} = 100$ , we observed a spatial difference in areas affected by the disease (Fig. 2B), as expected. One region reached local endemic steady state (green arrow, Fig. 2B) while other regions of similar density were highly infected (red arrow, Fig. 2B) and yet other districts were almost unaffected (grey arrow, Fig. 2B).

To quantify the effect of population clustering, we compared the ABM result to the reference SEIR model of similar parameters. Generally, we observed that the epidemic developed faster with a higher infection peak  $I_{\text{peak}}$ , but also subsided quicker, leading to a lower number of infected once reaching endemic steady state,  $R_\infty$  (Fig. 2C and Fig. 2D).

In order to explore how population clustering affects the epidemic, we chose a reference value of infection rates,  $\beta = 0.01$ , and an alternative value of  $\beta = 0.007$ . In the absence of spatial dependence ( $\rho = 0 \text{ km}^{-1}$ ) these correspond to initial reproduction numbers  $\mathcal{R}_0 \approx 1.7$  and 1.1, respectively. Increasing the spatial dependence (i.e. increasing  $\rho$ ) lead to a significant rise in the infection peak for the ABM,  $I_{\text{peak}}^{\text{ABM}}$ , compared to the (unaffected) SEIR model,  $I_{\text{peak}}^{\text{SEIR}}$  for both the reference value and the alternative lower value of  $\beta$  (black and blue points, Fig. 2E). We introduced heterogeneity in infection strengths ( $\sigma_\beta = 1$ , see Fig. 1B), thus making some individuals much more infectious than others (i.e. including ‘*super shedders*’). We found no significant impact from this effect (red points in Fig. 2E). Similarly, we introduced heterogeneity in connection weights ( $\sigma_\mu = 1$ , see Fig. 1B), thus making some individuals much more likely to form contacts than others (i.e. including ‘*super connecters*’). This leads to a significant effect for  $\rho = 0 \text{ km}^{-1}$ , which converges towards the other curves for  $\rho > 0.1 \text{ km}^{-1}$  (orange (only super connecters) and green

(super connecters and super shedders) points in Fig. 2E). The total number of infected individuals when the epidemic is over,  $R_\infty$ , converged towards half of the SEIR model prediction as a function of  $\rho$  except for  $\beta = 0.007$  where the endemic steady state level is larger than the one obtained by the SEIR model (Fig. 2F). Fixing  $\rho = 0.1 \text{ km}^{-1}$  and increasing the fraction of distance-independent contacts,  $\epsilon_\rho$ , we found that  $I_{\text{peak}}^{\text{ABM}}$  is almost unaffected for  $\epsilon_\rho < 0.5$  (Fig. 2G), while  $R_\infty^{\text{ABM}}$  increases linearly towards the SIER model  $R_\infty^{\text{SEIR}}$ , as expected (Fig. 2H).

Next we consider how these heterogeneities bias the traditional SEIR model predictions, especially the predictions based on fits to the number of infected (i.e. the curve to be flattened) in the beginning of the epidemic (see Supplementary Information). Without spatial dependence, the predicted curves fitted the number of infected individuals very well (Fig. 3A). Introducing spatial dependence ( $\rho = 0.1 \text{ km}^{-1}$ ) leads to a severe overestimation of the epidemic based on the number of early infection cases (Fig. 3B). This result can be interpreted by the fact that in societies where population density and thus individual contact number varies significantly, the early phase will be driven by people with many contacts (super connecters). This typically happens in cities where the population density is high. Increasing the spatial dependence  $\rho$  we found that the SEIR model predictions overestimated the infection peak height  $I_{\text{peak}}$  and the total number of infected  $R_\infty$  significantly even for very small spatial heterogeneities (Fig. 3C and Fig. 3D). We observed this general trend for all tested combinations of parameters and heterogeneities. In particular we found that if long-distance connections  $\epsilon_\rho$  are below 10 %, the bias in the estimated infection peak height,  $I_{\text{peak}}$ , was constant within statistical uncertainty (Fig. 3E). For the total number of infected,  $R_\infty$ , we observed an almost linear regression to the SEIR model as  $\epsilon_\rho$  approaches one. However, even when  $\epsilon_\rho = 0.25$ , the prediction bias was still a factor of two (Fig. 3F). We concluded from these curves a general trend;

if one fits a SEIR model to infection numbers during the beginning of an epidemic, and use these estimates to predict the characteristics of the epidemic at a national level, one overestimates the number of infected by at least a factor of two.

In summary, our research reveals that the degree of population clustering in Denmark creates a discrepancy between the early predictions made by the SEIR models and the underlying agent based interactions. It results in a significant overestimation of the impact of the disease, both in terms of maximal number of simultaneously infected (by a factor of 3) and the endemic steady state level (by a factor of 2.5). Such discrepancies have been observed for earlier pandemics, for instance the 1918 Spanish flu, where the predicted herd immunity level was severely overestimated (21). These results can be an important element in explaining these mismatches, even though other elements, as for instance social distancing and mutations to the viral strain, also play a part. During 2020, numerous countries have been faced with the task of laying out strategies to minimize the consequences of SARS-CoV-2, including the importance of '*flattening the curve*'. While this is truly crucial to avoid overpopulated hospitals, it should be taken seriously enough that we might specify to a higher degree of certainty which curve to be flattened. The mean reproduction number,  $\mathcal{R}_0$ , was around  $2.5 - 3$  across countries in the early phase (22). This led to predictions estimating that 60% of the populations would be infected at the end of the pandemic. Our results estimate this number to be closer to 30%. This has two important implications which are largely of a positive character. First, we do not have to fear as many infected as predicted by SEIR models and other models not including spatial clustering. This could already have significant impact on the outbreaks that will occur during the rest of the pandemic, which are likely to be smaller than expected, especially in countries that have already measured a high fraction of infected in the population. Secondly, our study emphasize the great benefits by making lock-downs early in the pan-

demic, when it is driven by '*super connectors*'. Since people in living in city-clusters are more likely to catch the infection, they are, on average, more likely to be affected in the beginning and by removing contacts from these individuals, one can avoid the worst peak while affecting the fewest number of people.

During this pandemic, mathematical predictions have been heavily criticized (23, 24) and it is now important to improve the accuracy of them, in order to increase the confidence in the predictions. Our work seriously questions the validity of predictions based on SEIR models, quantifies their biases from not including spatial clustering, and suggests that the precise effect of population clustering should be addressed more seriously. While our work has absolutely no political agenda, it should serve as an input in the current debate of how to handle the severe consequences of a crisis like SARS-CoV-2.

# Figures and Tables

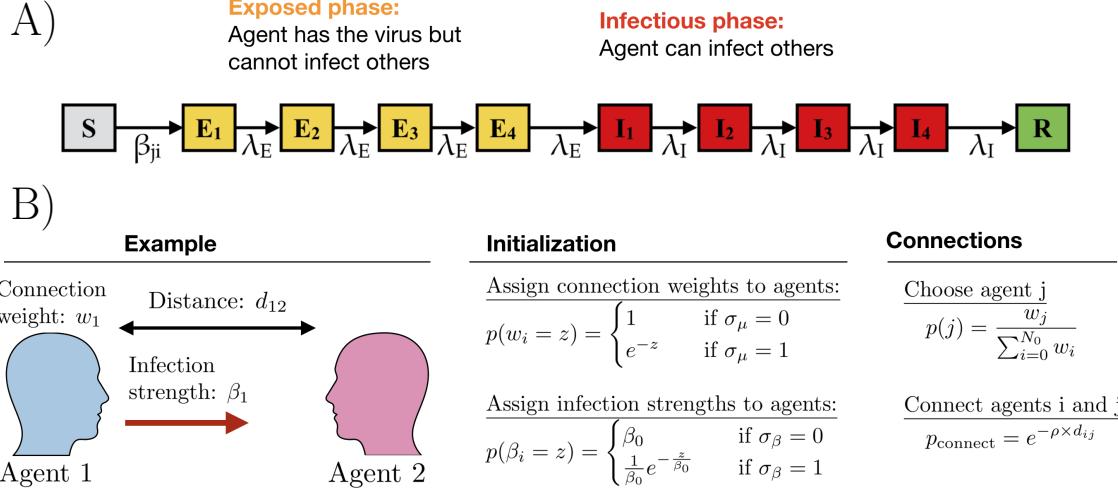


Figure 1: **A)** Illustration of the modified susceptible-exposed-infected-removed (SEIR) model used. It consists of ten consecutive states ( $S$ ,  $E_{1-4}$ ,  $I_{1-4}$ , and  $R$ ), with transition rates governed by  $\beta$ ,  $\lambda_E$ , and  $\lambda_I$ , respectively. **B)** Illustration of how the spatial network is generated and heterogeneities in individuals included.

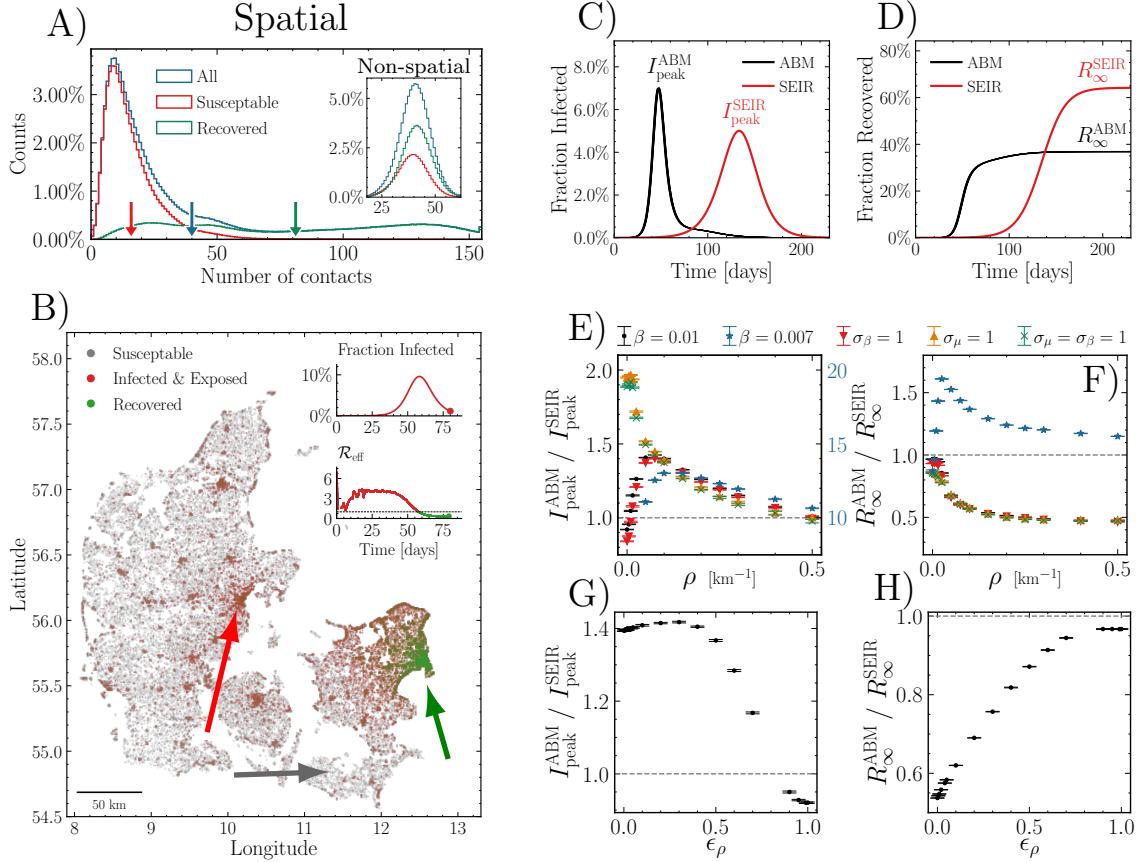


Figure 2: **A)** Histograms showing the number of susceptible (red) and recovered (green) individuals at the end of an epidemic with  $\rho = 0.1 \text{ km}^{-1}$ . The distribution before the epidemic is shown in blue. The arrows show the mean of each distribution. The inset shows the same for  $\rho = 0 \text{ km}^{-1}$ . **B)** Visualisation of the spatial position of individuals during the infection and which state they are in. Green arrow: Largest city in Denmark (Copenhagen): mostly recovered. Red arrow: Second largest city in Denmark (Aarhus): mostly infected. Grey arrow: low-population area: mostly susceptible (i.e. have not been infected). **C)** Number of infected individuals as a function of time. Data shown for the spatially distributed network ( $\rho = 0.1 \text{ km}^{-1}$ ). Simulation was repeated 10 times. **D)** cumulative sum of individuals who have had the disease as a function of time (with  $\rho = 0.1 \text{ km}^{-1}$ ). **E)** Relative difference in maximal number of infected,  $I_{\text{peak}}$ , between deterministic (SEIR) and ABM as a function of  $\rho$ , and shown for different parameters. Note the data for  $\beta = 0.007$  are shown in blue with a factor 10 scaling (right y-axis). **F)** Relative difference in total number of infected at the end of the epidemic,  $R_\infty$ , between deterministic (SEIR) and ABM as a function of  $\rho$ . Colors similar to E. **G)** Same as E, but as a function of  $\epsilon_\rho$ . **H)** Same as F, but as a function of  $\epsilon_\rho$ .

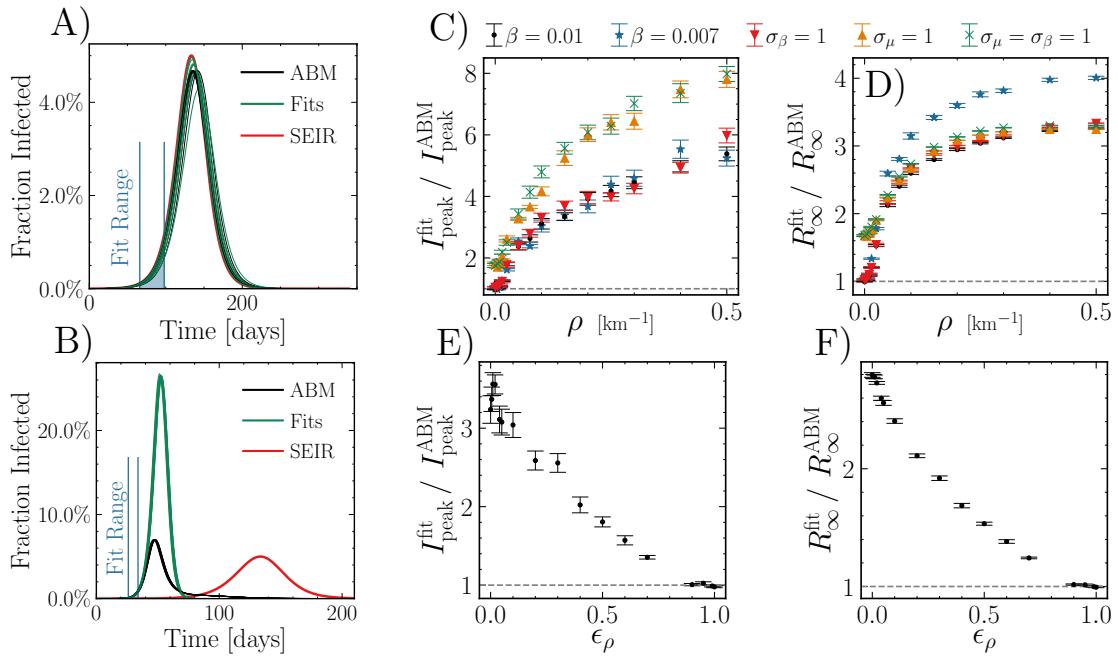


Figure 3: **A)** Number of infected individuals for the ABM in black, the SEIR model in red and the SEIR fits to the ABM data in green. Blue lines show the interval where parameters are fitted (also shown below the curves). Here  $\rho = 0 \text{ km}^{-1}$ . **B)** Same as A but with population clustering ( $\rho = 0.1 \text{ km}^{-1}$ ). **C)** Relative difference in maximal number of infected,  $I_{\text{peak}}$ , between the fit and the ABM for different values of  $\rho$ . Simulations repeated 10 times for each data-point. **D)** Relative difference in total number of infected at the end of the epidemic,  $R_\infty$ , between the fit and the ABM for different values of  $\rho$ . **E)** Same as C, but as a function of  $\epsilon_\rho$ . **F)** Same as D, but as a function of  $\epsilon_\rho$ .

Variable	Description	Value	Range	Units
$N_0$ :	Population size	$5.8 \cdot 10^6$	$10^5 - 10^7$	–
$N_{\text{init}}$ :	Number of individuals initially infected	100	$1 - 10^4$	–
$\mu$ :	Average number of network contacts	40	$10 - 100$	–
$\beta$ :	Typical infection strength	0.01	$0.001 - 0.1$	$\text{day}^{-1}$
$\lambda_E$ :	Rate to move through $\frac{1}{4}$ of latency period	1	$0.5 - 4$	$\text{day}^{-1}$
$\lambda_I$ :	Rate to move through $\frac{1}{4}$ of infectious period	1	$0.5 - 4$	$\text{day}^{-1}$
$\sigma_\mu$ :	Population clustering spread	0	$0 - 1$	–
$\sigma_\beta$ :	Interaction strength spread	0	$0 - 1$	–
$\rho$ :	Typical acceptance distance	0.1	$0 - 0.5$	$\text{km}^{-1}$
$\epsilon_\rho$ :	Fraction of distance-independent contacts	0.04	$0 - 1$	–

Table 1: Overview of the ten parameters applied in this study, their typical value, and the ranges we have considered. The first six parameters are standard SEIR parameters, whereas the last four parameters define the heterogeneity in the model. These four parameters do not affect the SEIR model.

## References

1. M. Chinazzi, *et al.*, *Science* **368**, 395 (2020).
2. W.H.O., Listings of WHO's response to COVID-19, [www.who.int/news/item/29-06-2020-covidtimeline](http://www.who.int/news/item/29-06-2020-covidtimeline).
3. R. M. Anderson, H. Heesterbeek, D. Klinkenberg, T. D. Hollingsworth, *The Lancet* **395**, 931 (2020).
4. J. Hellewell, *et al.*, *The Lancet Global Health* **8**, e488 (2020).
5. M. J. Keeling, T. D. Hollingsworth, J. M. Read, *medRxiv* p. 2020.02.14.20023036 (2020).
6. T. Kuniya, *Journal of Clinical Medicine* **9**, 789 (2020).
7. R. Li, *et al.*, *Science* **368**, 489 (2020).
8. K. Prem, *et al.*, *The Lancet Public Health* **5**, e261 (2020).
9. B. A. D. van Bunnik, *et al.*, *medRxiv* p. 2020.05.04.20090597 (2020).
10. L. Danon, E. Brooks-Pollock, M. Bailey, M. J. Keeling, *medRxiv* p. 2020.02.12.20022566 (2020).
11. S. L. Chang, N. Harding, C. Zachreson, O. M. Cliff, M. Prokopenko, *arXiv:2003.10218 [cs, q-bio]* (2020).
12. K. Sneppen, R. J. Taylor, L. Simonsen, *medRxiv* p. 2020.05.17.20104745 (2020).
13. G. J. Milne, S. Xie, *medRxiv* p. 2020.03.20.20040055 (2020).

14. S. Bansal, B. T. Grenfell, L. A. Meyers, *Journal of The Royal Society Interface* **4**, 879 (2007).
15. L. Kong, J. Wang, W. Han, Z. Cao, *International Journal of Environmental Research and Public Health* **13** (2016).
16. D. Kang, H. Choi, J.-H. Kim, J. Choi, *International Journal of Infectious Diseases* **94**, 96 (2020).
17. D. Giuliani, M. M. Dickson, G. Espa, F. Santi, *arXiv:2003.06664 [stat]* (2020).
18. Boligsiden.dk, [www.boligsiden.dk](http://www.boligsiden.dk).
19. HOPE Project, [www.hope-project.dk](http://www.hope-project.dk).
20. Statistics Denmark, Statistikbanken, [www.statistikbanken.dk](http://www.statistikbanken.dk).
21. V. Andreasen, C. Viboud, L. Simonsen, *The Journal of infectious diseases* **197**, 270 (2008).
22. P. Boldog, *et al.*, *Journal of Clinical Medicine* **9**, 571 (2020).
23. I. Holmdahl, C. Buckee, *New England Journal of Medicine* (2020).
24. L. Wynants, *et al.*, *BMJ* **369** (2020).

# Supplementary Information

## Deterministic and agent-based models

In this section we describe the deterministic SEIR model and compare it to the agent based model (ABM). We initiated this investigation by assuming that every individual could be in 4 different states:

**S** Susceptible, where one can be infected by others.

**E** Exposed, where one has the infection but cannot yet pass it on to others.

**I** Infectious, where one has the infection and can pass it on to others.

**R** Recovered, where one has been infected but is now immune.

Since the time spent in both the exposed and infectious state in reality does not follow an exponential distribution, we introduced 4 sub-states in each of these states to follow the more realistic gamma distribution. The choice of 4 sub-states in both exposed and infected states are not critical to the results, and two to five sub-states can be used without great difference to the results. Thereby every agent (*i*) can change states according to the associated rates.

If the constructed network is homogeneous, we can formulate and describe the development of the disease, using a so-called SEIR model. This model assumes that all agents have approximately the same number of contacts and the same infection rates. Based on this, the dynamics of the infections are described with the differential equations shown below:

$$\begin{aligned}
\dot{S} &= -\tilde{\beta}SI \quad \text{with} \quad \tilde{\beta} = \beta \frac{\mu}{2N_0} \\
\dot{E}_1 &= \beta_D SI - \lambda_E E_1 \\
\dot{E}_i &= \lambda_E E_{i-1} - \lambda_E E_i \quad (i = 2, 3, 4) \\
\dot{I}_1 &= \lambda_E E_4 - \lambda_I I_1 \\
\dot{I}_i &= \lambda_I I_{i-1} - \lambda_I I_i \quad (i = 2, 3, 4) \\
\dot{R} &= \lambda_I I_4.
\end{aligned} \tag{1}$$

Models like this have been used to a great amount in 2020, and the purpose of the SEIR model in this work is to compare it with the ABM. To simulate the ABM model we use an event-driven Gillespie algorithm. To optimize speed, no agent can be infected (or attempted to be infected) twice. Thus, at each event we decide whether an agent should change state or an infected agent should pass on the infection to one of his neighbours. All this is selected relative to all the rates of every event and thus the time step until the next event is based on the sum of all possible rates.

### **Construction of spatial network**

We initialized  $N_0$  agents on a network generated a total of  $\mu \times N_0$  links between two agents, with an assigned interaction strength  $\beta_{ij}$  for each link. The average contact number,  $\mu$ , was fixed to 20, based on results from the danish HOPE project, gathering data on population behaviour since April 2020 (25). In order to include a realistic, spatial distribution of the population, we randomly selected agent locations from a 2D kernel density estimate, we had generated based on housing sales in Denmark 2007-2019 (Data given with permission from Boligsiden, (26)). To connect the agents, we used a hit and miss method, where two random agents is first suggested and then connected with probability,  $p(d) = e^{-\rho \cdot d_{ij}}$ . Here

$d_{ij}$  is the distance between agents and  $\rho$  is a parameter with units of inverse distance. We choose  $\rho = 0.1 \text{ km}^{-1}$  (i.e. 10 km) which is the average distance travelled by labour force (Statistics Denmark (27)). To allow some long distance interactions, we introduced a parameter  $\epsilon_\rho = 4\%$  representing the fraction of distance independent connections. This value, we based on the fraction of workers travelling longer than 50 km to work (Statistics Denmark (27)).

### Event driven Gillespie Algorithm

In the Gillespie algorithm we consider a number,  $M$ , of different reactions,  $R_1 \dots R_M$ . At  $t = 0$ , we consider the initial number of possible reactions. We now calculate the PDF for the time until the next reaction occur

We consider the probability that the next reaction is of type  $\epsilon$ , and it occurs in the time-interval  $[t + \tau, t + \tau + dt]$ . We therefore consider:

$$P(\tau, \epsilon)d\tau = \frac{\text{No reaction in } [t, t + \tau]}{P_{\text{not}}(\tau)} \cdot \frac{\text{Reaction } \epsilon \text{ occurs}}{R_\epsilon d\tau}$$

Therefore we want to describe  $P_{\text{not}}(\tau)$  in terms of the rates. Since at each timestep  $\epsilon$ , the probability for no reaction to appear is:

$$P_{\text{not}}(dt) = 1 - \sum_{i=1}^N R_i dt$$

We can thus define  $\tau \equiv n \cdot dt$  and then:

$$P_{\text{not}}(\tau) = P_{\text{not}}(dt)^n = \left(1 - \sum_{i=1}^N R_i \frac{\tau}{n}\right)^n = e^{-r\tau}$$

where  $r \equiv \sum_{i=1}^N R_i$

This means that we should generate a random number according to the exponential

distribution, and a random number according to a uniform distribution. Here we can use the transformation method, and we can then create the update process, where at each step we jump a step in time  $\tau$  to next reaction, and picks the reaction according to r. Schematically the Gillespie algorithm can be described as:

- Pick two random numbers,  $\nu_1$  and  $\nu_2$ . Calculate time until next reaction:

$$\tau = -\frac{\ln \nu_1}{r}$$

- Pick the next reaction:

$$\epsilon = \frac{\sum_{i=1}^{k-1} r_i}{\sum_{j=1}^n r_j} < \nu_2 \leq \frac{\sum_{i=1}^k r_i}{\sum_{j=1}^n r_j}$$

- Update the system according to the chosen reaction.

In this way the system can be updated, and adjusting the reactions to each time step. All rates in this project are either by one agent moving to a new state, or an agent passing on an infection. Thus we compare all rates of all possible events, and choose the next event according to their relative weights.

### Comparison of SEIR and homogeneous ABM

We started out testing that the ABM model reproduces the results of the deterministic SEIR model when based on the same assumptions. Therefore, we consider a population where all agents have the same probability to be selected and two suggested agents are connected with  $p_0 = 1$ . We set  $N_{\text{init}} = 100$  to represent an initial outbreak and choose a population size of  $5.8 \cdot 10^6$  individuals, reflecting the population of a small country (Denmark).

For the ABM model, we visualized the spread of the disease by assigning a position to

every agent. In order to include more realistic assumptions about the spatial distribution of the population, we distribute the network nodes, i.e. the agents, according to housing sales in Denmark 2007-2019 (Data given with permission from Boligsiden, (26)). The distribution has been approximated with a 2D kernel density estimate from which we randomly select agent locations. This inhomogeneous distribution represents a realistic mix of urban and rural population for Denmark, but we believe that many of these conclusions can be transferred to countries of similar population structure. By doing this, we could observe, as expected, that the infected agents were well mixed across the country. By this we mean that even though people are spatially distributed in city clusters, the pandemic is observable in all places at the same time (Figure S1A).

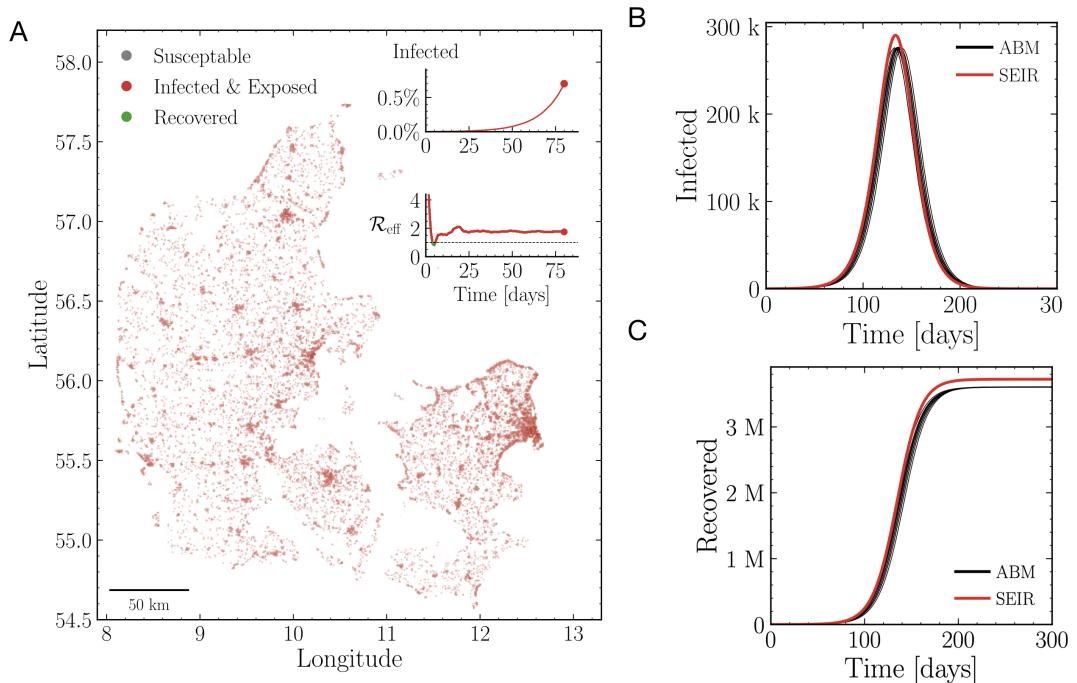


Figure S1: A) Illustration susceptible, infected and recovered agents without spatial connection dependence. B) Comparison between ABM simulations (black) and SEIR (red) simulations in the number of infected individuals. C) Comparison between ABM simulations (black) and SEIR (red) simulations in the number of recovered agents.

Comparing the ABM and SEIR, we used the current number of infected agents,  $I(t)$ , and the cumulative sum of all agents who have been infected at that time,  $R(t)$ . Specifically, we compared the maximum peak height of the number of infections,  $I_{\text{peak}}$ , between the models and the sum of all agents having been infected at the end of the epidemic,  $R_\infty$ . As such, we looked at the ratios  $I_{\text{peak}}^{\text{ABM}}/I_{\text{peak}}^{\text{SEIR}}$  and  $R_\infty^{\text{ABM}}/R_\infty^{\text{SEIR}}$ . We find that the two models give similar results, but we note that the ABM model rises a little later (Fig. S1B and Fig. S1C).

### **Robustness test of homogeneous parameters**

We tested how the standard parameters affected the difference between the deterministic and the ABM models. By comparing the fundamental  $I_{\text{peak}}$  and  $R_\infty$  as a function of the average number of connections  $\mu$  we found strong resemblance if  $\mu$  is large, but at very small values of  $\mu$  there is a fundamental difference and the ABM shows a significantly lower result, because the epidemic never starts at all (Fig. S2A and Fig. S2B). Then we tested the value of  $\beta$ , i.e. the interaction strength between two connected agents, and found that this does not significantly affect the results (Fig. S2C and Fig. S2D). We tested the importance of the number of initially infected on the peak height and the number of infected. Here we found for all tested values a very small dependence on the number of initially infected (Fig. S2E and Fig. S2F). Next, we investigated the effect of the size of the network in the ABM and found that the  $I_{\text{peak}}$  and  $R_\infty$  in the ABM is smaller by only 3 % and 4 %, respectively (Fig. S2G and Fig. S2H). We checked the stochastic noise of the ABM and found that it followed a normal distribution with a standard deviation of 0.08 % for  $I_{\text{peak}}$  and 0.02 % for  $R_\infty$ . We also tested the importance of  $\lambda_E$  and  $\lambda_I$ , and here we found a smaller than 5 % dependence on the choice of parameters. From this analysis of the ABM, we conclude that it follows the SEIR for the parameters introduced, and

our next goal is to include the effects of geometry and clustering that are present in real epidemics but which are not included in SEIR.

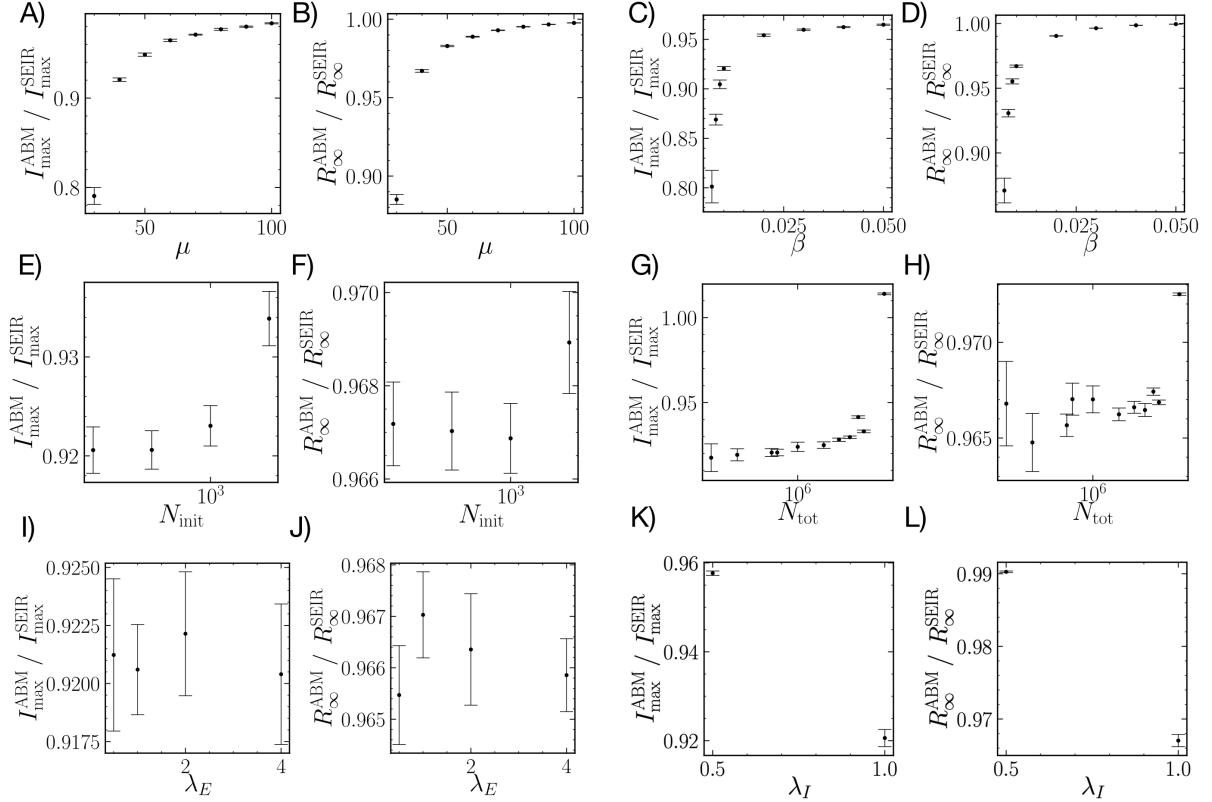


Figure S2: A) Ratio between maximal value of number of infected for ABM and SEIR model for similar parameters as a function of  $\mu$ . B) Same as A) but for the number of recovered at the end of the simulation. C) Same as A, but as a function of  $\beta$ . D) Same as B, but as a function of  $\beta$ . E) Same as A, but as a function of  $N_{\text{init}}$ . F) Same as B, but as a function of  $N_{\text{init}}$ . G) Same as A, but as a function of  $N_{\text{tot}}$ . H) Same as B, but as a function of  $N_{\text{tot}}$ . I) Same as A, but as a function of  $\lambda_E$ . J) Same as B, but as a function of  $\lambda_E$ . K) Same as A, but as a function of  $\lambda_I$ . L) Same as B, but as a function of  $\lambda_I$ .

## Fits and predictions

We defined an early phase to be the period of time between 0.1 % and 1 % of the population were infected (blue lines Fig. 3A). We then fitted  $\beta$  and a time delay,  $\tau$ , to the SEIR model with a  $\chi^2$ -fit (assuming Poissonian statistics) and kept  $\lambda_E$  and  $\lambda_I$  fixed to the true numbers

(used in the simulation). The initial number of infected,  $N_{\text{init}}$ , was also fixed to the true numbers. The fit parameters were then inserted into the SEIR model, and  $I_{\text{peak}}^{\text{fit}}$  and  $R_{\infty}^{\text{fit}}$  were extracted from the fitted model and compared to the  $I_{\text{peak}}^{\text{ABM}}$  and  $R_{\infty}^{\text{ABM}}$  from the ABM simulation.

### Robustness test for fits with homogeneous parameters

We tested how the predicted value from the fit, was affected by specific values in the standard parameters that are defined for the homogeneous ABM. We found that there was no bias for small values of  $\mu$ , and as  $\mu$  increased the ratio was very close to 1 (Fig. S3A and Fig. S3B). Testing for values of  $\beta$  this converged rapidly to 1 as  $\beta$  was increasing, but the discrepancy was smaller than 5% for all values (Fig. S3C and Fig. S3D). As a function of the number of initially infected, we did not find a large discrepancy for the parameters tested (Fig. S3E and Fig. S3F). We also tested as a function of the network size, and here we found the largest discrepancy for the largest network, which was surprising, but since this was still on the order of 5% we concluded that it did not affect the overall conclusion (Figure S3G and S3H). Finally we tested the dependency on  $\lambda_E$  and  $\lambda_I$  but the predictions generated from the SEIR fit, matched the outcome form the ABM to a high level of precision for all tested values (Fig. S3I–F).

### Author contributions

MLH and CM developed the models and performed the simulations. All authors contributed to the writing of the paper. MLH, CM, MHJ, TH and TCP conceived and designed the project.

### Data availability

All scripts and figures can be found on [www.github.com/ChristianMichelsen/NetworkSIR](https://www.github.com/ChristianMichelsen/NetworkSIR).

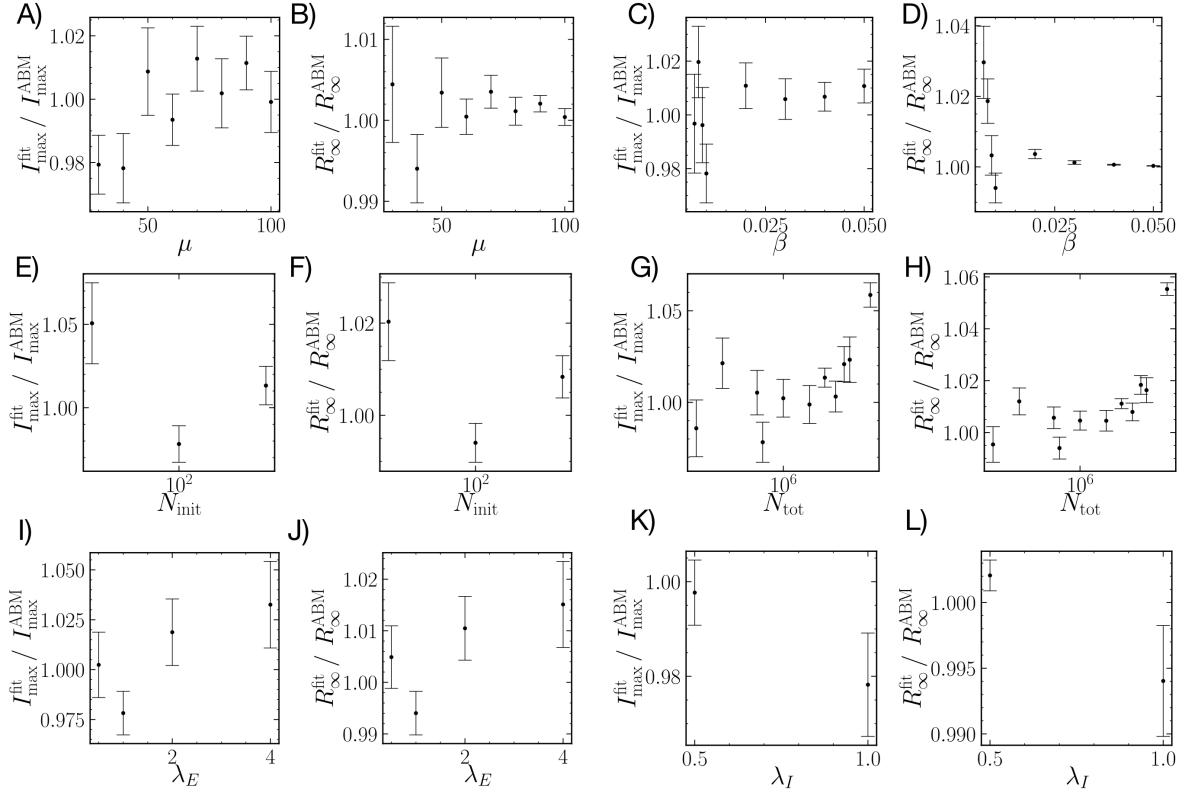


Figure S3: A) Ratio between maximal value of predicted number of infected and real maximal value for ABM as a function of  $\mu$ . B) Same as A) but for the number of recovered at the end of the simulation. C) Same as A, but as a function of  $\beta$ . D) Same as B, but as a function of  $\beta$ . E) Same as A, but as a function of  $N_{init}$ . F) Same as B, but as a function of  $N_{init}$ . G) Same as A, but as a function of  $N_{tot}$ . H) Same as B, but as a function of  $N_{tot}$ . I) Same as A, but as a function of  $\lambda_E$ . J) Same as B, but as a function of  $\lambda_E$ . K) Same as A, but as a function of  $\lambda_I$ . L) Same as B, but as a function of  $\lambda_I$ .

### Acknowledgements

The authors are grateful to the Danish expert group of SARS-CoV-2 modelling, especially Robert L. Skov, Kåre Mølbak, Viggo Andreasen, Kaare Græsbøl, Theis Lange, Carsten Kirkeby, Frederik P. Lyngse, Matt Denwood, Uffe Thygesen and Laust Hvas Mortensen. Furthermore we thank Kim Sneppen for valuable discussions.

## **References**

25. HOPE Project, [www.hope-project.dk](http://www.hope-project.dk).
26. Boligsiden.dk, [www.boligsiden.dk](http://www.boligsiden.dk).
27. Statistics Denmark, Statistikbanken, [www.statistikbanken.dk](http://www.statistikbanken.dk).