How population clustering leads to over estimation of predicted pandemic curves

Mathias L. Heltberg[1,2,3,6*], Christian Michelsen[1,6], Emil S. Martiny[1], Lasse Engbo Christensen[4], Mogens H. Jensen[1] , Tariq Halasa[5], and Troels C. Petersen[1*]

1. Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17, 2100 Copenhagen E,

2. Laboratoire de Physique, Ecole Normale Superieure, Rue Lhomond 15, Paris 07505,

3. Infektionsberedskab, Statens Serum Institute, Artillerivej, 2300 Copenhagen S

4. DTU Compute, Section for Dynamical Systems, Department of Applied Mathematics and Computer Science, Technical University of Denmark, Anker Engelunds Vej 101A, 2800 Kongens Lyngby

5. Animal Welfare and Disease Control, University of Copenhagen, Gronnegårdsvej 8, 1870 Frederiksberg C

6. These authors contributed equally

(Dated: September 28, 2020)

The modeling of COVID-19 has become a critical aspect of present life. In this fundamental task, 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 (SEIR) individuals. However, these SEIR models are based on very simple assumptions which constitute obvious approximations. In this paper, we introduce an agent based model that allows us to include spatial clustering effects and based on Danish population data, we estimate how this impacts the long term development and the early prediction of a pandemic. Our results suggest that population clustering has a major impact on the long term development implying that our initial estimates on the herd immunity level, based on infection spreading in the early phases, might be over estimated by a factor of two.

I. INTRODUCTION

Since the emergence of reports from Wuhan, China, on the spread of a peculiar disease in late 2019, the authorities closely monitored a cluster of pneumonia cases in the city in December [1]. The disease was then reported to the WHO in late December 2019 [2] who identified the causative pathogen as SARS-CoV-2 [3]. Subsequently, the virus spread over the five continents affecting many countries, paralyzing societies, resulting in a large number of deaths and severe economic damage worldwide [8]. Following the wide spread of the disease now known as "COVID–19", 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 (e.g. [9][10][11][12][13]. Generally, the models used to represent the disease, have been variants

*Electronic address: heltberg@nbi.ku.dk

of the SIR model, and vary in complexity including simple deterministic compartmental models assuming homogeneous mixing within the population (e.g. [13] [14]), meta-population compartmental models that include heterogeneity by subdividing the population into groups where mixing within the groups is assumed homogenous (e.g. [17] [15] [18]), individual based models that include heterogeneity by including specific characteristics for each individual but without including spatial specifications (e.g. [4][10] [16]), and finally spatio-temporal agent-based models that include the highest level of heterogeneity by including individual characteristics together with the spatial clustering of individuals (e.g. [5]). An important aspect in the spread of Sars-CoV-2 is the presence of superspreaders. These can be defined as infectious individuals that are responsible for many new infections, possibly the majority of new infections [21]). This phenomenon has been identified both in animals and humans. For instance, [24] pointed out the presence of high-level shedders of Escherichia coli O157 in cattle herds that are responsible for many new infections. Following the emergence of SARS in late 2002 and investigations of its spread from China to other countries, one traveler was identified to be responsible for more than 100 SARS cases in Singapore ([20]). A similar observation was encountered in a large Hospital in Tianjin, China, where many cases where linked to one infectious individual ([22]). In the current Sars-CoV-2 epidemic, superspreading events have been observed and were responsible for different outbreaks (e.g. [19]; [6]). Although increased disease severity, high viral load and extensive social interactions have been suggested as potential explanatory factors ([26]), it is still unclear which of these processes, drive the superspreading phenomenon. To improve our understanding of this phenomenon, it is of fundamental importance that we study the implications of these factors on the spread of Sars-CoV-2, which will improve our ability to prevent outbreaks and control the disease spread rapidly. Models of Sars-CoV-2 provided valuable information that helped to understand the dynamics of its spread and guided the authorities to prevent and control the spread, saving many lives and resources. However, it remains unclear whether the differences in modeling heterogeneity between these models may result in different predictions and hence conclusions. Earlier work has generally pointed out the importance of including heterogeneity when modeling the spread of infectious disease, as contact patterns between individuals shape population-level disease dynamics ([25]). In addition, epidemic curve and size may differ depending on the mixing assumption ([7] [27]. Furthermore, [16] pointed out the importance of accounting for heterogeneities caused by superspreaders when modeling the spread of Sars-CoV-2. Thus, it is important to assess the impact of these assumptions in COVID–19 models on model predictions, in order to provide reliable predictions. Furthermore, assessment of the importance of including spatial clustering is necessary as spatial dependency of COVID–19 spread has been observed ([23] [28]). To our knowledge, no mathematical modeling study has been conducted combining these elements for Sars-CoV-2 spread. This paper investigates how spatial population clustering affect the progress of a pandemic. Using an event-driven Gillespie algorithm and a realistic spatially distributed population, we construct an interaction network and model the development in diseases with parameters similar to Sars-CoV-2. Based on our simulations, we use the initial part of the disease to predict the outcome of the disease and the herd immunity level, using a standard deterministic SEIR model. Doing this, we calculate the predicted bias on the forecast and thereby outline which heterogeneities are most critical to collect data on, if one wants to predict the worst-case outcome of a disease with greater precision. This paper is structured as follows: In section II we introduce the models and compare them for parameters maintaining homogeneity. We then move to section III where we include spatial dependence and introduce heterogeneity between agents. We use this in section IV to determine how much this affects the initial predictions of the size of the epidemic.

II. DETERMINISTIC AND AGENT-BASED MODELS

We initiated this investigation, by assuming that every individual could be in 4 different states:

- Susceptible state (S), where one could be infected by others
- Exposured (E), where one has the infection but cannot yet pass it on to others
- Infectious (I), where one has the infection and can pass it on to others
- Recovered (R), where one has reached the immunity level

Since the time spent in both the exposured and infectious state in reality does not follow an exponential distribution, we introduced 4 substates in each of these states to follow the more realistic gamma distribution. The choice of 4 substates in both exposed and infected states are not critical to the results, and 2-5 substates can be used without great difference to the results. Thereby every agent (i) can change states according to the rates schematized in Figure 1.

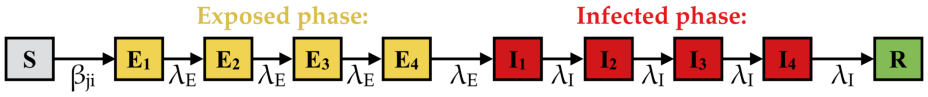

FIG. 1: Illustration of the modified susceptible-exposed-infected-removed (SEIR) model used. It consists of ten consecutive states (S, E_{1-4}, I_{1-4}, and R), with transition rates governed by β, λ_E, and λ_I, respectively.

We initialized N_0 agents on a network and chose an average degree of connectivity, μ. Then we generated a total of $\mu \times N_0$ links between the agents so each agent, i, had a specific number of connections, where all connections from this agent has an assigned interaction strength β_{ij}. In the process of connecting agents, we used a hit and miss method where a connection between two random agents are first suggested and a connection is created depending on an acceptance probability, p_0. This probability can in general depend on relations between the two suggested individuals for instance their mutual distance in space. Using this algorithm, we constructed a network with which we could simulate an epidemic starting from N_{init} initially exposed agents. This is the fundamental setup for the agent based model (ABM).
If the constructed network is homogenous, 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:

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

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.

III. 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 5.8×10^6 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, https://www.boligsiden.dk). The distribution has been approximated with a 2D kernel density estimate from which we randomly select agent locations. This inhomogenious 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 2A).

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_{max}, between the models and the sum of all agents having been infected at the end of the epidemic, R_∞. As such, we looked at the ratios $I_{\text{max}}^{\text{ABM}}/I_{\text{max}}^{\text{SEIR}}$ 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 (Figure 2B and 2C).

We tested how the standard parameters (see Table 1, row 1-5) affected the difference between the deterministic and the ABM models. By comparing the fundamental I_{max} and R_∞ as a function of the average number of connections μ we found strong resemblance if μ is large, but at very small values of μ there is a fundamental difference and the ABM shows a significantly lower result, because the epidemic never starts at all (Figure 2D). Next, we investigated the effect of the size of the network in the ABM and found that the I_{max} and R_∞ in the ABM is smaller by only 3 % and 4 %, respectively (Data not shown). Finally, we tested the value of β, i.e. the interaction strength between two connected agents, and found that this does not significantly affect the results (Figure 2E). 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_{max} and 0.02 % for R_∞. 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.

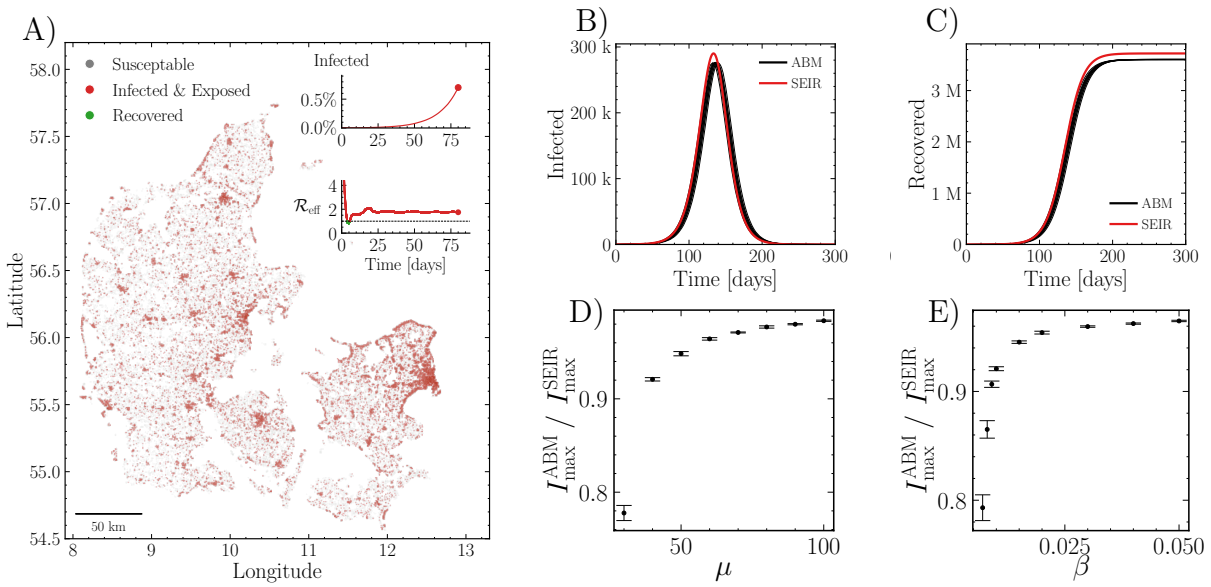

FIG. 2: **A)** Visualisation of the spatial position of infected agents during the infection. The top inset shows the number of infected agents as a function of time and the lower inset shows the effective reproduction number $\mathcal{R}_{\text{eff}}$. **B)** Number of infected, I, as a function of time. Data shown for standard parameters (see Table 1, row 1-5), and the simulation was repeated 10 times. **C)** The cumulative sum of agents who have had the disease, R, as a function of time. Data shown for standard parameters (See Table 1, row 1-5). **D)** Relative difference in maximal number of infected, I_{max}, between deterministic (SEIR) and ABM for different values of μ, the average degree of connectivity. **E)** Same as **D**, but as a function of infection strength, β.

IV. SPATIALLY WEIGHTED CONNECTIVITY LEADS TO CLUSTERING AND ENHANCED INFECTION PRESSURE

In this section we introduce the significance of spatial distance between agents. To obtain this, we generate the network by choosing two random agents, i and j, and calculate the distance between them, d_{ij}. We assign a connection between them with a distance-dependent probability p given by $p(d) = e^{-\rho \cdot d_{ij}}$. The choice of exponential function reflects the diminishing probability of connecting with increased distance. Here ρ is a parameter with units of inverse distance, such that $\rho = 0.1\,\text{km}^{-1}$ corresponds to a typical connection distance of 10 km. We choose this value based on the average distance travelled by labour force (Statistics Denmark: https://www.statistikbanken.dk/statbank5a/default.asp?w=1680). In a fraction of the cases (denoted ϵ_ρ), the two agents are connected without spatial dependence to include the effect of long distance connections. We choose $\epsilon_\rho = 4\,\%$ based on the fraction of workers travelling longer than 50 km to work (Statistics Denmark, https://www.statistikbanken.dk/statbank5a/default.asp?w=1680). Using this metric, we obtain a spatially distributed network, where one is more inclined to infect those in the vicinity than those at a distance, but it does not rule out spread at longer distances, especially for those living in rural areas. Based on this we obtained a distribution of connections in the network, and found that this had a group of highly connected nodes, that are not found in the distance independent network (Figure 3A).

Simulating the epidemic on this system with 100 initially infected agents, $N_{\text{init}} = 100$, we found a significant difference in the areas affected (Figure 3B). Now entire regions could obtain a local herd-immunity (green arrow Figure 3B) at the same time that other cities of similar size could be highly infected (red arrow Figure 3B) and finally other districts could be almost unaffected by the disease (grey arrow Figure 3B). The typical distance is in-built in this algorithm, and quickly leads to the urban

Variable	Description	Value	Range	Units
N_0 :	Population size	$5.8 \cdot 10^6$	$10^5 - 10^7$	–
N_{init} :	Number of agents initially infected	100	$1 - 10^4$	–
μ :	Average number of contacts	40	$10 - 100$	–
β :	Typical interaction strength	0.01	$0.001 - 0.1$	day^{-1}
λ_E :	Rate to move through ¼ of latency period	1	$0.5 - 4$	day^{-1}
λ_I :	Rate to move through ¼ of infectious period	1	$0.5 - 4$	day^{-1}
ρ :	Population clustering rate	0.1	$0 - 0.5$	km^{-1}
ϵ_ρ :	Fraction of distance-independent contacts	0.04	$0 - 1$	–

TABLE I: Overview of the seven parameters applied in this study, their typical value and the ranges we have considered. The first five parameters will lead to standard SEIR behaviour (Figure 2), whereas the bottom two parameters, ρ and ϵ_ρ, will induce specific spatial behaviour (Figure 3).

population being much more interconnected than the rural population. We noticed how agents with many contacts are more likely to be infected and observed how the mean value of connections was changed to a much larger degree in skewed distribution than in the distance independent network (Figure 3A). To quantify the effect of populaton clustering on the spread of an epidemic, we chose $\rho = 0.1\,\text{km}^{-1}$ and compared the ABM result to the SEIR model of similar parameters (Figure 3C). Here we found that the ABM has a significantly higher peak which is reached much earlier compared to the SEIR model, revealing that the number of infected at the same time is much higher than what one would obtain from standard assumptions. This fundamental observation is explained by the clusters in cities and this means that some people are more active and have more contacts than the average of the population. These agents are more likely to be infected, and this generates a high peak, since these agents can distribute the disease quickly in the early stages. However, once these are removed from the system, the remaining part of the population will have a lower distribution of connections and therefore the infection pressure is lowered. If we consider the total number of people having been affected by the disease, the ABM yields a lower value than the SEIR model (Figure 3D). This is again explained by the fact that given the population has a heterogeneous distribution of connectivity, a large number of people will avoid the infection due to the fact that they live in a local environment where the effective infection pressure is low.

We tested these results as a function of the distance parameter ρ (Figure 3E). Here we found including the spatial dependence leads to a sudden rise in the I_{max} and that this peaks for $\rho \approx 0.075\,\text{km}^{-1}$ which corresponds to a typical interaction distance of around 15 km. The fact that this curve has an extremum is surprising, but at this point, we have a very heterogeneous distribution of connectivity, but local interaction networks still not so small that the newly infected agents are surrounded by other infected agents. For extreme values of $\rho \geq 0.2\,\text{km}^{-1}$, corresponding to a typical distance of $\leq 5\,\text{km}$, the spread of the disease will move as a one-dimensional frontier which significantly slows down the spread. Considering the R_∞ measure, we do not find any extrema, and, for decreasing ρ, the total number of infected individuals will decay (Figure 3F). This is as one would expect, since a still larger fraction will live in quite isolated regions where the disease will have difficulty in spreading to.

To test the robustness of these results, we fixed $\rho = 0.1\,\text{km}^{-1}$ and tested how this was affected by introducing distance independent contacts modelled by ϵ_ρ. Of course, at $\epsilon_\rho = 1$ we should obtain the same results as we did in Figure 2, but we found that for a large range of values for ϵ_ρ, the height I_{max} is unaffected, and first after more than 50 % of all contacts being distance independent, we find a significant effect (Figure 3G). This shows that our results are quite robust and that long distance connections are

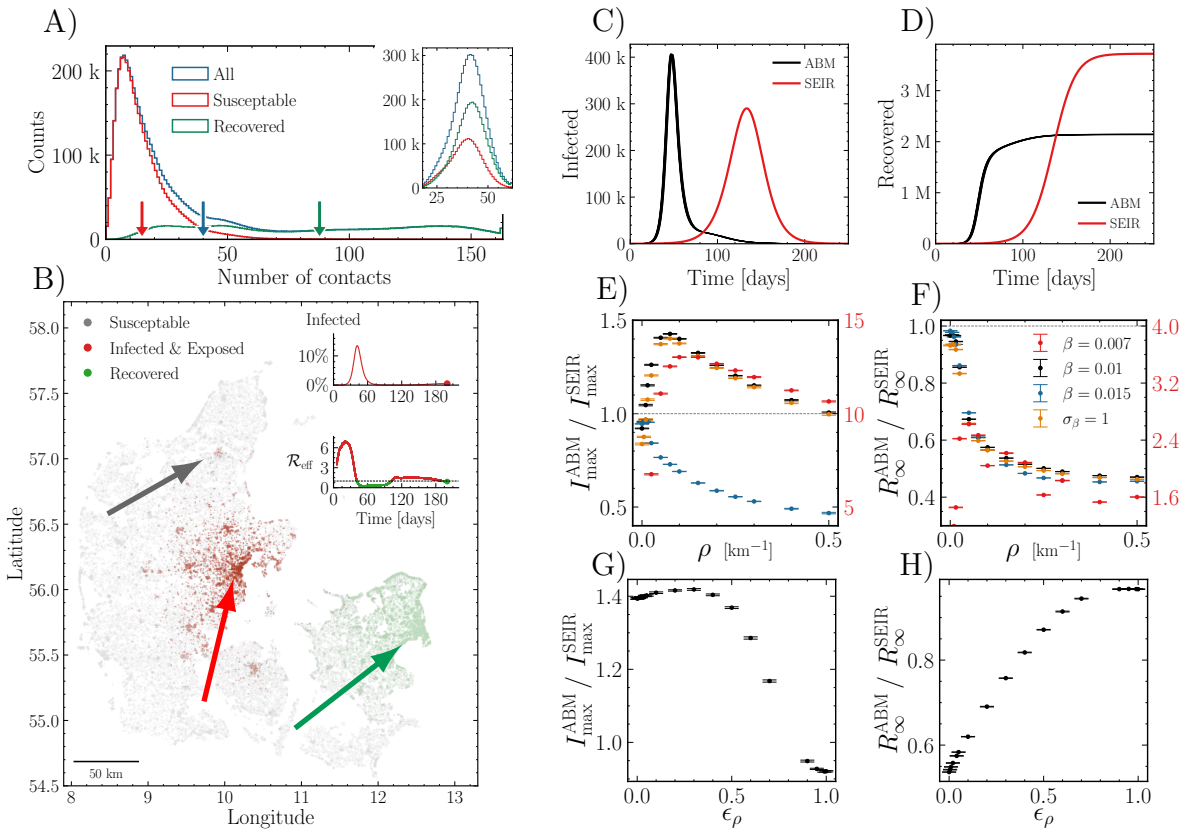

FIG. 3: **A)** The histograms shows how many agents are respectively susceptible (red) or recovered (green) at the end of an epidemic for a skewed distribution in the number of contacts. The arrows show the mean of each of the tree distributions. The inset shows the same for a un-skewed distribution in the number of contacts. **B)** Visualisation of the spatial position of infected agents during the infection. Green arrow: Largest city in Denmark (Copenhagen): recovered. Red Arrow: Second largest city in Denmark (Aarhus): Highly infectious. Gray Arrow: Third largest city of Denmark (Aalborg): Unaffected. **C)** Number of infected agents 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 agents 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_{max}, between deterministic (SEIR) and ABM for different values of ρ. The data are shown for $\beta = 0.01$ (standard value) in black, $\beta = 0.015$ in blue, and with infection rate heterogeneity (notated as $\sigma_\beta = 1$) in orange. The data for $\beta = 0.007$ are shown in red with a factor 10 scaling (see the right y-axis). **F)** Relative difference in total number of infected at the end of the epidemic, R_∞, between deterministic (SEIR) and ABM for different values of ρ. Colors similar to **E**, however, here only a factor 4 scaling is used for $\beta = 0.007$. **G)** Same as **E**, but as a function of ϵ_ρ. **H)** Same as **F**, but as a function of ϵ_ρ.

not a crucial element in the modelling of disease spreads on a network at a national scale. Finally, we also tested this effect on the R_∞ measure, and here we found that ϵ_ρ leads to a higher number of infected agents, and that this increase is linear for the range of realistic parameters (Figure 3H), but that the R_∞ measure does not grow above 1, i.e. that the total number of infected agents at the end of the epidemic in the agent based model is never higher than in the SEIR model. Based on these results, we moved on to test how these results affected early predictions for the spread of the disease.

V. OVERESTIMATING THE PROPOSED CURVE IN THE PRESENCE OF SPATIAL CLUSTERS

In the early phase of a pandemic (and in particular the current COVID–19 pandemic), estimates of the pandemic size was needed. These were typically predicted by SIR-like models. However, these models are based on several assumptions of homogeneity and the predictions relating to the number of infected as a function of time (the curve to be "flattened") were biased. In this section we explore the size of this bias. To shed light on this, we generated a pandemic with our spatial ABM (see section IV). We considered the initial phase of the pandemic, using the number of infected every day, and fitted the SEIR model, eq. (1), to this From the fit parameters, we extrapolated the pandemic size predicted by the SEIR fit and compared this to the actual realisation of the pandemic in the ABM. In this way, we estimated the size of the bias, obtained by using an early SEIR-fit, to predict the pandemic and understand its relation to the homogeneity assumption. Specifically, we defined the early phase to be the period of time starting when 0.1 % of the population were infected until the time when 1 % of the population were infected (blue lines Figure 4A). We assumed that we knew the number of infected per day (i.e that we roughly knew the number of non-registered infected which was possible to estimate based on the first outbreaks). We then fitted β and a time delay, τ, to the SEIR model with a χ^2–fit (assuming Poissonian statistics) and kept λ_E and λ_I fixed to the true numbers (used in the simulation). The initial number of infected, N_{init}, was also fixed to the true numbers and and assumed evenly spread out on the four exposed states. The fit parameters were then inserted into the SEIR model, and $I_{\text{max}}^{\text{fit}}$ and R_{∞}^{fit} were then extracted from the pandemic predicted by the fitted model. Finally, we compared these to the actual $I_{\text{max}}^{\text{ABM}}$ and R_{∞}^{ABM} from the ABM simulation. We repeated this procedure to reduce the effect of statistical fluctuations.

Doing so we found that without population clustering, the estimated predicted curves estimated the number of infected individuals quite well (Figure 4A), and the total number of infected individuals was estimated to a good precision as well. However, when introducing population clustering using the spatial ABM ($\rho = 0.1\,\text{km}^{-1}$), we found a severe overestimation of the disease based on the early curves (Figure 4B). The timing of the peak number of infected matched the actual curves quite well, but a severe overestimation of the height was found. This result can be interpreted by the fact that in societies where population density and individual contact number is clustered, the early phase will be driven by people with many contacts which happen in cities where population densities are high. We tested this effect as a function of the distance parameter ρ and found that the overestimation increases significantly even for very small spatial heterogeneities. It grows monotonically but with the highest steepness in the beginning (Figure 4C). This pattern turned out to be similar for the total number of infected predicted (Figure 4D), which is expected since the total number of infected can be directly calculated in a standard SIR model based on the basic parameters. Since our results are dependent on spatiality, we again tested if the presence of long-distance connections, ϵ_ρ, would affect the overall conclusions. Here we found that if the fraction of long-distance connections was below 10 %, the size of the peak bias was constant, and if this fraction was larger than this, a linear decay was observed until it reaches 1 (which it should). This happens when around 90 % of all contacts in the network are distance independent (Figure 4E). Again we found a similar pattern in the predictions of the total number of infected by the disease, even though the decay started as soon as distance independent contacts were introduced (Figure 4F).

These curves indicate the same thing; that when we fit a SEIR model (which does not take population clustering and heterogeneity of contact numbers into account) to infection numbers during the beginning

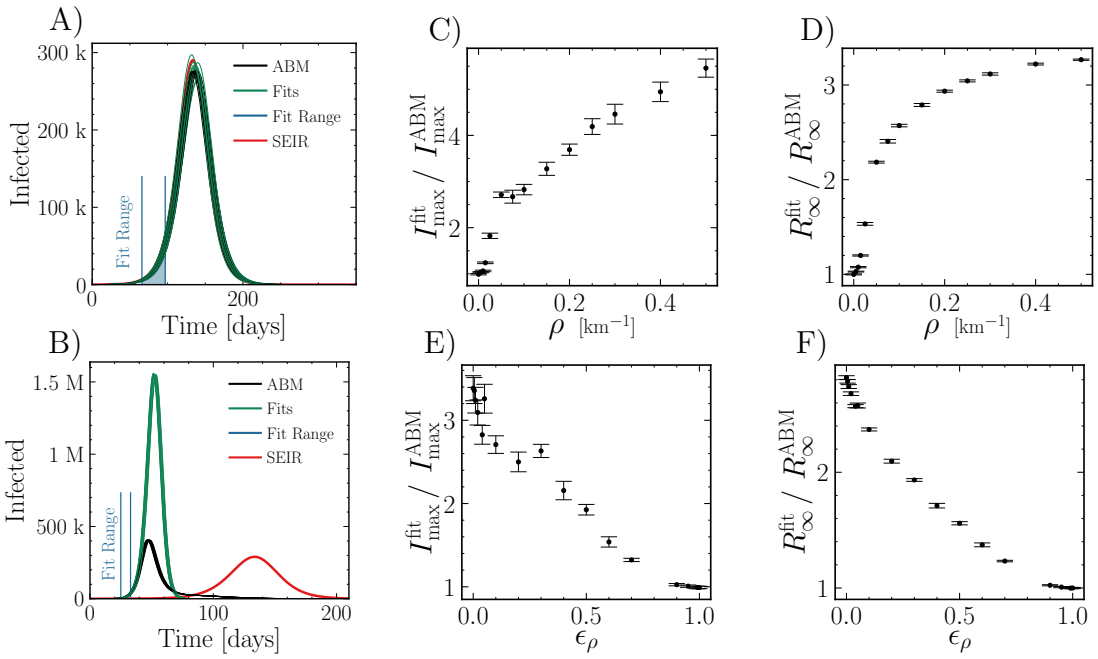

FIG. 4: **A)** Number of infected agents in the pandemic for the ABM in black, the SEIR model in red and the SEIR fits to the ABM data in green. Blue lines show the interval where parameters are fitted (also shown below the curves). Here $\rho = 0\,\text{km}^{-1}$, i.e. no population clustering. **B)** Same as **A** but with population clustering ($\rho = 0.1\,\text{km}^{-1}$). **C)** Relative difference in maximal number of infected, $I_{\max}$, between the fit and the ABM for different values of ρ. **D)** Relative difference in total number of infected at the end of the epidemic, R_∞, between the fit and the ABM for different values of ρ. **E)** Same as **C**, but as a function of ϵ_ρ. **F)** Same as **D**, but as a function of ϵ_ρ.

of a pandemic, we seem to overestimate the effect of the disease by at least a factor of 2. It should be noted that these results apply where inhomogeneities exist, eg. at a national level. In small societies, where population densities are roughly constant, predictions based on SEIR models are affected by variations in contact numbers of individual agents, even though this effect is largely enhanced by population clustering. The SEIR predictions of the entire population can only be considered a good approximation when this is also constant.

VI. DISCUSSION

During 2020 many countries have been faced with the task of making predictions on the spread and control of COVID–19. At the early stage, when it started spreading across europe, almost everybody in every country across education and social status were taught 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 we are talking about. During this early period, there was a broad agreement that the mean reproduction number of the disease was around $2.5 - 3$ (see review [31]), and based on this number, predictions and curves "to be flattened", were fundamentally generated using various SEIR models, These can be made more complex by including for instance different age groups, different sub-regions, but they always have the flaw, that in each compartment they assume well mixed populations and hence high connectivity is automatically assumed. It is of utmost importance to understand the implications of these assumptions, in order to develop realistic models for the future to support decision-making, as the economic and social consequences of these decisions are large. In

addition, COVID–19 prediction models have been heavily criticized ([29][30]). Therefore, it is important to develop realistic models in order to increase the confidence of the authorities and society, as models can actually be useful ([29]).

In this paper, we investigated how much an SEIR model differed from a more realistic ABM model, using a spectrum of parameters. Our research revealed that if agents are placed at realistic geographical positions, and that interactions are predominantly occurring between agents in a relative proximity to each other, then this is enough to create a discrepancy of approximately 40 % when comparing to a SEIR model of otherwise similar parameters. While this is not something critical, since the actual pandemic does not follow an SEIR model in the first place, it does get critical when one use this model to make predictions based on an early curve of the data. To highlight this, we used an ABM model, and made predictions using the SEIR model based on the initial slope of the pandemic. Here we found that the early predictions made by SEIR are significantly overestimating the impact of the disease both in terms of maximal number of infected and the total number of infected people.

Our findings are not directly transferrable to the data after lock-down of numerous countries, since we considered the disease without interventions. This was done in order to minimize the enormous number of considered parameters applied to lock-down and to focus on the issues when making the earliest predictions. Our results indicate that with the replication number of approximately $2.5 - 3$ that was measured initially, the point at which herd immunity is reached is much lower than what models usually predict, and thus we do not have to fear as many infected as the well-mixed models predict. This could have huge impact on the outbreaks that will occur during the fall of 2020, and second waves will be smaller than expected, especially in countries that have already measured a high number of infected.

The general pandemic overestimation of the SEIR model originates in the fact that the reproduction number is a function of the average infection strength (i.e β) and the average number of contacts for the infected agents. In the early stage, the agents with high number of contacts are quickly infected and thereby they cause a higher reproduction number, which is not representative for the entire population. From our assumption that the probability to interact with people depends on the mutual distance, a heavy-tailed connection distrubution is generated automatically. Thereby, the measured reproduction number in the hearly phase, represents an upper bound for the for later stages of the epidemics, and the actual herd immunity level should be much smaller. To quantify this effect, we see in Figure 3E, that in our specific model, the herd immunity (defined as the fraction of recovered at the epidemic end without any interventions) is lowered to 58 % of the SEIR model estimate. This value depends on the distribution of number of contacts and thus spatiallity, but this general conclusion does not alter much with varying input values (for $rho = 0.1\,\mathrm{km}^{-1}$ we get $\approx 64\,\%$ smaller compared to SEIR, while $\rho = 0.3\,\mathrm{km}^{-1}$ we get $\approx 49\,\%$).

Early predictions of a pandemic will lay the foundations for the politics of counter-measures and for instance lockdowns. Since this is a multi-billion affair, it is important to precisely estimate which curve that might be expected for the pandemic without any political actions. This work seriously questions the validity of predictions using SEIR models, and suggest that the effect of population clustering should be much more closely investigated in future research. While it has absolutely no political agenda to minimize the amount of political interactions, it should serve as an input in the current debate of how to minimize

all the severe consequenses of a crisis like COVID–19.

---

[1] Chinazzi, M., Davis, J. T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., ... & Viboud, C. (2020). The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID–19) outbreak. Science, 368(6489), 395-400.

[2] WHO https://www.who.int/news-room/detail/27-04-2020-who-timeline—COVID–19

[3] WHO https://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/

[4] Chang, S. L., Harding, N., Zachreson, C., Cliff, O. M., & Prokopenko, M. (2020). Modelling transmission and control of the COVID–19 pandemic in Australia. arXiv preprint arXiv:2003.10218.

[5] Milne, G. J., & Xie, S. (2020). The effectiveness of social distancing in mitigating COVID–19 spread: a modelling analysis. medRxiv.

[6] Lau, M. S., Grenfell, B., Thomas, M., Bryan, M., Nelson, K., & Lopman, B. (2020). Characterizing superspreading events and age-specific infectiousness of SARS-CoV-2 transmission in Georgia, USA. Proceedings of the National Academy of Sciences, 117(36), 22430-22435.

[7] Kong, L., Wang, J., Han, W., & Cao, Z. (2016). Modeling heterogeneity in direct infectious disease transmission in a compartmental model. International journal of environmental research and public health, 13(3), 253.

[8] Fernandes, N. (2020). Economic effects of coronavirus outbreak (COVID–19) on the world economy. Available at SSRN 3557504.

[9] Anderson, R. M., Heesterbeek, H., Klinkenberg, D.,& Hollingsworth, T. D. (2020). How will country-based mitigation measures influence the course of the COVID–19 epidemic?. The Lancet, 395(10228), 931-934.

[10] Hellewell, J., Abbott, S., Gimma, A., Bosse, N. I., Jarvis, C. I., Russell, T. W., ... & Flasche, S. (2020). Feasibility of controlling COVID–19 outbreaks by isolation of cases and contacts. The Lancet Global Health.

[11] Ferguson, N., Laydon, D., Nedjati Gilani, G., Imai, N., Ainslie, K., Baguelin, M., ... & Dighe, A. (2020). Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand.

[12] Keeling, M. J., Hollingsworth, T. D., & Read, J. M. (2020). The Efficacy of Contact Tracing for the Containment of the 2019 Novel Coronavirus (COVID–19). medRxiv.

[13] Kuniya, T. (2020). Prediction of the epidemic peak of coronavirus disease in Japan, 2020. Journal of clinical medicine, 9(3), 789.

[14] Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., & Shaman, J. (2020). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). Science, 368(6490), 489-493.

[15] van Bunnik, B. A., Morgan, A. L., Bessell, P., Calder-Gerver, G., Zhang, F., Haynes, S., ... & Lepper, H. C. (2020). Segmentation and shielding of the most vulnerable members of the population as elements of an exit strategy from COVID–19 lockdown. medRxiv.

[16] Sneppen, K., & Simonsen, L. (2020). Impact of Superspreaders on dissemination and mitigation of COVID–19. medRxiv.

[17] Prem, K., Liu, Y., Russell, T. W., Kucharski, A. J., Eggo, R. M., Davies, N., ... & Abbott, S. (2020). The effect of control strategies to reduce social mixing on outcomes of the COVID–19 epidemic in Wuhan, China: a modelling study. The Lancet Public Health.

[18] Danon, L., Brooks-Pollock, E., Bailey, M., & Keeling, M. J. (2020). A spatial model of COVID–19 transmission in England and Wales: early spread and peak timing. MedRxiv.

[19] Al-Tawfiq, J. A., & Rodriguez-Morales, A. J. (2020). Super-spreading events and contribution to transmission of MERS, SARS, and COVID–19.

[20] Paull, S. H., Song, S., McClure, K. M., Sackett, L. C., Kilpatrick, A. M., & Johnson, P. T. (2012). From superspreaders to disease hotspots: linking transmission across hosts and space. Frontiers in Ecology and the Environment, 10(2), 75-82.

[21] Gopinath, S. C., Tang, T. H., Chen, Y., Citartan, M., & Lakshmipriya, T. (2014). Bacterial detection: From microscope to smartphone. Biosensors and Bioelectronics, 60, 332-342.

[22] Wang, Z. Q., Cui, J., & Xu, B. L. (2006). The epidemiology of human trichinellosis in China during 2000–2003. Acta tropica, 97(3), 247-251.

[23] Kang, D., Choi, H., Kim, J. H., & Choi, J. (2020). Spatial epidemic dynamics of the COVID–19 outbreak in China. International Journal of Infectious Diseases.

[24] Chase-Topping, M. E., McKendrick, I. J., Pearce, M. C., MacDonald, P., Matthews, L., Halliday, J., ... & Woolhouse, M. E. (2007). Risk factors for the presence of high-level shedders of Escherichia coli O157 on Scottish farms. Journal of clinical microbiology, 45(5), 1594-1603.

[25] Bansal, S., Grenfell, B. T., & Meyers, L. A. (2007). When individual behaviour matters: homogeneous and network models in epidemiology. Journal of the Royal Society Interface, 4(16), 879-891.

[26] Park, D. J., Dudas, G., Wohl, S., Goba, A., Whitmer, S. L., Andersen, K. G., ... & Winnicki, S. M. (2015). Ebola virus epidemiology, transmission, and evolution during seven months in Sierra Leone. Cell, 161(7), 1516-1526.

[27] Brauer, F., Castillo-Chavez, C., & Castillo-Chavez, C. (2012). Mathematical models in population biology and epidemiology (Vol. 2, p. 508). New York: Springer.

[28] Giuliani, D., Dickson, M. M., Espa, G., & Santi, F. (2020). Modelling and predicting the spread of Coronavirus (COVID–19) infection in NUTS-3 Italian regions. arXiv preprint arXiv:2003.06664.

[29] Holmdahl, I., & Buckee, C. (2020). Wrong but useful—what COVID–19 epidemiologic models can and cannot tell us. New England Journal of Medicine.

[30] Wynants, L., Van Calster, B., Bonten, M. M., Collins, G. S., Debray, T. P., De Vos, M., ... & Schuit, E. (2020). Prediction models for diagnosis and prognosis of COVID–19 infection: systematic review and critical appraisal. bmj, 369.

[31] Boldog, P., Tekeli, T., Vizi, Z., Dénes, A., Bartha, F. A., & Röst, G. (2020). Risk assessment of novel coronavirus COVID–19 outbreaks outside China. Journal of clinical medicine, 9(2), 571