Realistic agent-based simulation of infection dynamics and percolation

Kai Nagel, Christian Rakow, Sebastian Müller

Transport Systems Planning and Transport Telematics, Institute for Land and Sea Transport Systems, TU Berlin, Germany

Abstract

We present an agent-based epidemiological model that is based on an agent-based model for traffic and mobility. The model consists of individual agents that follow individual daily activity plans, which include, for each activity, locations, start times, and end times. Evidently, one can place a virus spreading dynamic on top of this, by infecting one or more agents, and then track the resulting virus dynamics through the model.

Normally, the model is used to investigate non-pharmaceutical interventions. In the present paper, we undertake steps to better understand the infection graph. It becomes clear that the typical infection graph representation that connects individual people is an even more expensive representation than our original, already expensive data-driven mobility model. We then undertake first steps towards analysing the model with respect to a possible percolation transition.

Keywords: Simulation, Mobility, Epidemiology, COVID-19, Percolation

1. Introduction

The project that motivates the present paper came out of the COVID-19 pandemics that started in spring 2020. Perceiving a lack of realistic virus spreading dynamics models in Germany, and recognizing that we could quickly build such a model based on our experience in transport modelling, we built a prototype in about two weeks [1]. Subsequently, we received funding to continue our research and to regularly report to the ministry of research (e.g. [2, 3]). Some of the research promised within that project is to investigate the infection graph that is induced by the mobility model. Given that the progress of an infection in a graph has something to do with percolation for which Dietrich Stauffer is well known [4, 5], this seems like a good fit for the present special issue.

Published in Physica A, http://dx.doi.org/10.1016/j.physa.2021.126322 .

The present paper will first present some possible models for virus spreading, from mathematical models via agent-based models to percolation, with a particular focus on our own agent-based model. It will then discuss different graph interpretations of our data, in particular showing that the common representations are, other than we originally assumed, even more expensive than the full mobility model from which we start. Another section will show results with respect to cluster size distributions, where a cluster is defined as how far a single initial seed spreads. Additionally, clusters near the percolation threshold are shown and discussed. The paper ends with a discussion and a conclusion.

2. Models for virus spreading

2.1. Compartmental models for virus spreading

The mainstay of epidemiological modelling are compartmental models. A useful starting point for the modelling of SARS-2 are the SEIR models, with states *susceptible*, *exposed* (or latently infected = infected but not contagious), *infectious*, and *recovered*. The transitions from one compartment to another are described by rate equations. For a simple SIR model:

$$\dot{I} = \beta \cdot S \cdot I - \mu \cdot I \ ,$$

where β is the probability to become infected given numbers of S and I, and μ is the probability to move into the R department.

At the onset of an epidemic, S can be taken as large and constant, and in consequence then $I(t) \approx I_0 \cdot e^{(\beta S - \mu) \cdot t}$. Evidently, if $\beta S - \mu > 1$, then there is an epidemic, otherwise not. So-called herd immunity is reached by the infection dynamics depleting S, and thus eventually reducing $\beta S - \mu$ to below one.

A shortcoming of compartmental models is that individuals do not have state. For example, it is difficult to attach age or geographical location to entities. The basic model does not even contain delay: If there is a sudden increase in the compartment of exposed, this will *immediately* lead to larger numbers of transitions to *infectious*.

2.2. Agent based models for virus spreading

An alternative are agent based models. In agent based models, each individual of the real world is represented by some synthetic avatar that follows certain rules. Examples of such an approach are by Virginia Biotechnology Institute [6, 7], Imperial College [8, 7], and by the Center for Statistics and Quantitative Infectious Diseases [9, 7]. An example for a similar approach on a global level is [10]. Groups that started more recently include [11, 12, 13, 14, 15, 16]. This section will concentrate on describing our own model; more information can be found in [17].

2.2.1. *Mobility model*

Since we come from the modelling of human mobility behavior and traffic flow [18], we use human movement patterns as starting point. The mobility model has a record of each synthetic person's locations and movement patterns over the day, including the activity types at the locations, and the vehicle types during movement. Locations are called *facilities*. Person go about their daily movements; when they spend time at the same facility, the virus can move from one person to another (Fig. 1). Our first version of the model just repeated a typical day; the present version repeats the same pattern from Mondays to Fridays, but has different movements for the same persons on Saturdays and Sundays.

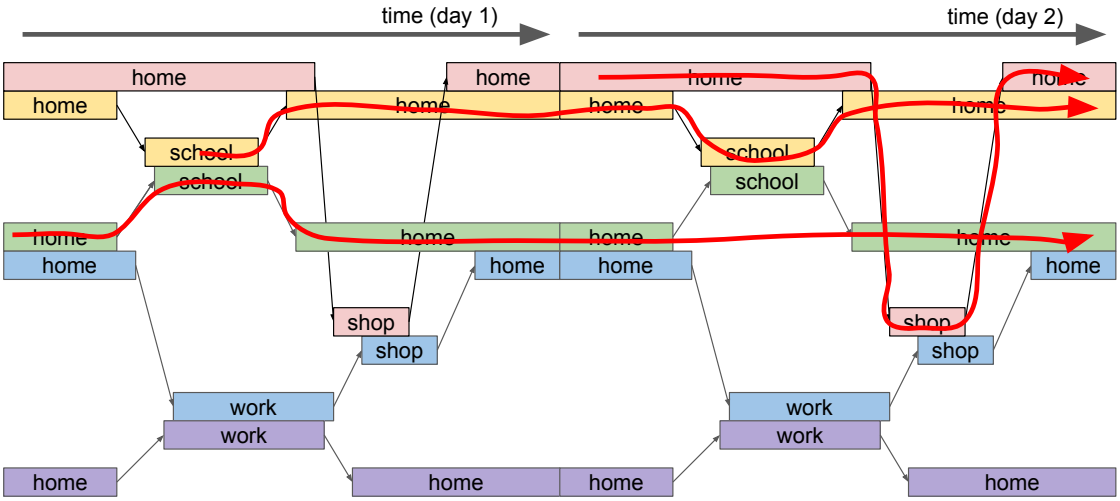


Figure 1: Infection process.

It would be easiest, or at least easiest to explain, if the movement patterns were directly taken from mobile phone data. In Germany, that is not possible for privacy reasons, and in consequence we use synthetic mobility patterns that are consistent with privacy requirements. The approach is to use some information from mobile phone data (but not the full trajectories), and process them together with information about the transport system and with statistical information from other surveys [19]. That approach leads to synthetic movement trajectories for the complete population. One advantage of that approach is that the synthetic trajectories are available for all of Germany, and the model can thus easily be set up for different parts of Germany. Also, given access to the right data sources, the model could also be built for other countries. For the time being, we concentrate our simulations on the metropolitan

area of Berlin, although we have run simulations for Munich and for the small city of Gangelt near Dusseldorf and Cologne (an early COVID-19 hotspot in Germany). The mobility data for Berlin which is the input to our simulations is available; see "Availability of data and materials".

2.2.2. Further sub-models

Given the information from the mobility model, one needs the following additional sub-models for infection modelling:

Disease progression model Once a synthetic person becomes infected, it progresses through states. We use states *susceptible*, *exposed*, *contagious* (but not showing symptoms), *showing symptoms*, *seriously sick* (= should be in hospital), *critical* (= needs intensive care and/or breathing support), and possibly *deceased*. None of these additional states have a strong influence on the infection dynamics, but they are important to (1) compare to hospital statistics for model calibration/validation, and (2) to predict hospital demand, which is important to assess the criticality the situation. Since disease progression is age dependent [8], the same levels of infection may lead to different hospital demands.

We use lognormal distributions, taken from the literature, for disease progression, as detailed in Fig. 2. We use the same age-dependent transition probabilities as [8], shown in the table in Fig. 2.

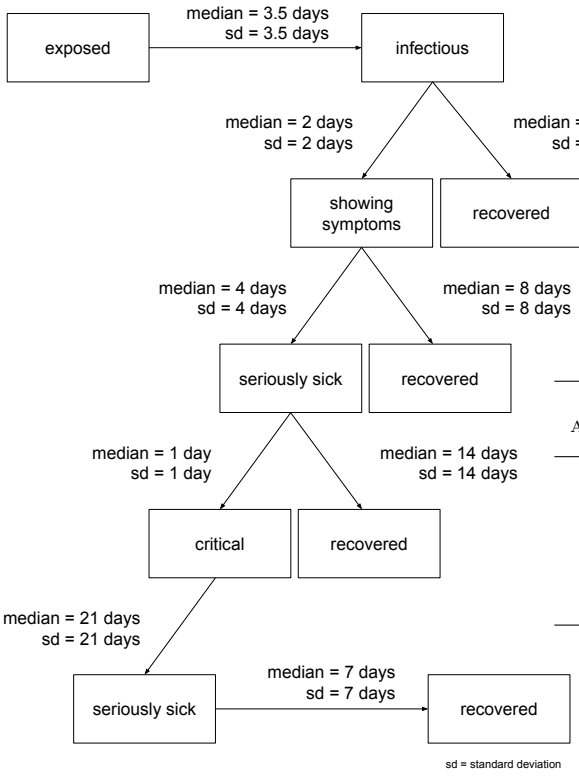
All of the states *contagious* to *critical* are in principle infectious. In practice, for SARS-2 most of the infections seem to happen from 2 days before to 2 days after starting to show symptoms [26], which is why we currently cut off infectiousness 4 days after becoming contagious.

Infection model Infection can happen if a susceptible and an infectious person are in the same facility or the same vehicle. Our infection model given contact is also taken from the literature [27, 28]:

$$p(infect|contact) = 1 - \exp\left(-\Theta \cdot sh \cdot ci \cdot in \cdot \tau\right) \ , \tag{1}$$

where sh is the shedding rate, in is the intake (reduced, e.g., by a mask) ci the contact intensity, and τ the duration of interaction between the two individuals. Θ is a calibration parameter.

For aerosol infections, which seems to be the main infection pathway for SARS-2 [29], it is plausible to parameterize the contact intensity as a function of room



Age-group	infectious cases becoming symptomatic	symptomatic cases requiring hospitalisation	hospitalised case requiring critical care
0 to 9	80%	0.05%	5.0%
10 to 19	80%	0.15%	5.0%
20 to 29	80%	0.6%	5.0%
30 to 39	80%	1.6%	5.0%
40 to 49	80%	2.45%	6.3%
50 to 59	80%	5.1%	12.2%
60 to 69	80%	8.3%	27.4%
70 to 79	80%	12.15%	43.2%
80+	80%	13.65%	70.9%

Figure 2: LEFT: Disease progression model [20, 21, 22, 23, 24, 25]. The transitions are described by lognormal distributions with median and standard deviation as given in the figure. The transition probabilities, where branches are possible, are given in the table on the RIGHT.

size rs and air exchange ae [30, 31]:

$$ci = const \cdot \frac{1}{rs \cdot ae} \ . \tag{2}$$

This ignores transients, and simply assumes that viral material mixes uniformly into the available space. If room size rs is twice as large, the resulting virus concentration will be half as large.

An air exchange ae of $1/h$ means that the air in the room is fully replaced once per hour. If one imagines virus accumulation for one hour and then complete replacement by window opening, one obtains a sawtooth function where the virus concentration in the average is half of what it is at the maximum. Twice as much air exchange in consequence reduces average virus concentration by

half. Note that this also implies that the same *ae* for a room twice as large means twice as much air volume exchange per hour.

The constant *const* is not needed since it is absorbed into Θ during model calibration.

Contact model One also needs a model for contact if simultaneously in the same facility/vehicle. One simple option is to assume contact between everybody in the same facility or vehicle.

Additional information about the model can be found in [17].

When all other parameters are given and fixed, then the question of how fast an initial infection grows depends on the parameter Θ: large Θ leads, in the average, to stronger growth and larger numbers of synthetic persons that eventually become infected.

Because of fluctuations, however, the results are more fragmented than for the compartmental models: It can well be the case that an initial seed dies out even when the system as a whole is super-critical. This asks for an interpretation along the lines of percolation, where super-criticality is defined as having a larger than zero probability of an initial seed growing to infinity, i.e. in that definition there is always a chance that an initial infection dies out.

2.3. Percolation

As is well known, the percolation problem can be defined as follows: Consider a grid of d dimensions. Occupy its cells with probability p. Cells are defined as connected if they are immediately adjacent, i.e. in d dimensions each cell has $2d$ neighbors. Now search for the critical density p_c where there is a non-zero probability that an occupied cell is connected via other occupied cells to cells that are infinitely far away.

As is also well known, that definition needs to be operationalized for computer simulations with finite space and finite time. One option is to define a grid of size L^d, and then search for the probability that two opposing boundary hyperplanes of size L^{d-1} are connected. That probability is found by averaging over many Monte Carlo simulations with different random seeds, where a different seed determines a different population of occupied cells given probability p. One then runs such simulations for different sizes of L, and finds

- If $p < p_c$, then the probability that the two hyperplanes are connected converges to zero with growing L.
- If $p > p_c$, then that probability converges to some finite value.

2.4. Cluster size distributions

At the percolation threshold p_c, one obtains

$$n(s) \sim s^{-\tau} \ , \tag{3}$$

where $\tau \approx 2.055$ in two dimensions, and $\tau = 2.5$ in 6 or more dimensions. Away from the percolation threshold, these distributions have exponential cut-offs:

- Below the percolation threshold, there are no large clusters.
- Above the percolation threshold, most cells are included in some system-wide network, and only small clusters exist as "islands".

2.5. Percolation and epidemics

Percolation has been investigated for epidemics decades ago [32], showing that many different model formulations fall into the same universality class [33]. In particular, it does not make a difference if, for bond percolation, the existence of links between nodes is computed beforehand (and thus a property of the substrate), or decided probabilistically on the fly. Clearly, our model falls into the second class, where infections are decided probabilistically on the fly according to Eq. (1).

3. Graph interpretation

Evidently, one can investigate percolation and/or epidemics on a graph. There is, in fact, considerable literature on this, for example two chapters in the book by Newman [34], or a long review article by Pastor-Satorras et al. [35]. In the following, we will present different graph representations of our model, and place them into the context of that review.

3.1. Person-centric infection graph

It is common to investigate the infection graph, which contains only the persons, and to draw edges between persons if they are able to infect each other [34, 35]. Unfortunately, that representation becomes too large for our simulation:

- Our original data has, for an area of 5 million inhabitants, about 46 million events. Since each facility/vehicle is entered and left, this corresponds to 23 million true edges.

- We have relatively large group sizes. For example, a public transit train easily contains 1000 persons. These 1000 persons do not interact on a single day, but given that the simulation runs many days, it is plausible to assume that they *could* interact. The same holds for large office buildings, large leisure facilities, etc. Our own representation of such a facility/vehicle with 1000 persons needs 1000 true edges in the sense defined above. If one would encode directly the connections between persons, this would increase to 1000^2 edges. That is, a person-based graph representation of our model might have about 1000 times more edges than the representation that we are currently using. In practice, we find "only" a little more than 100 million edges, i.e. four times as many as our original representation.

This makes the person-centric graph of our system too large for a typical desktop computer and graph analysis packages that we have tried, such as Gephi [36] or Graphia [37].

3.2. Static representation with facilities

An alternative representation is Fig. 3. This leaves the facilities/vehicles as intermediaries between the different persons, and thus the sparse representation in place. In the review by Pastor-Satorras et al. this is called "particle-network framework" (Section IX in [35]).

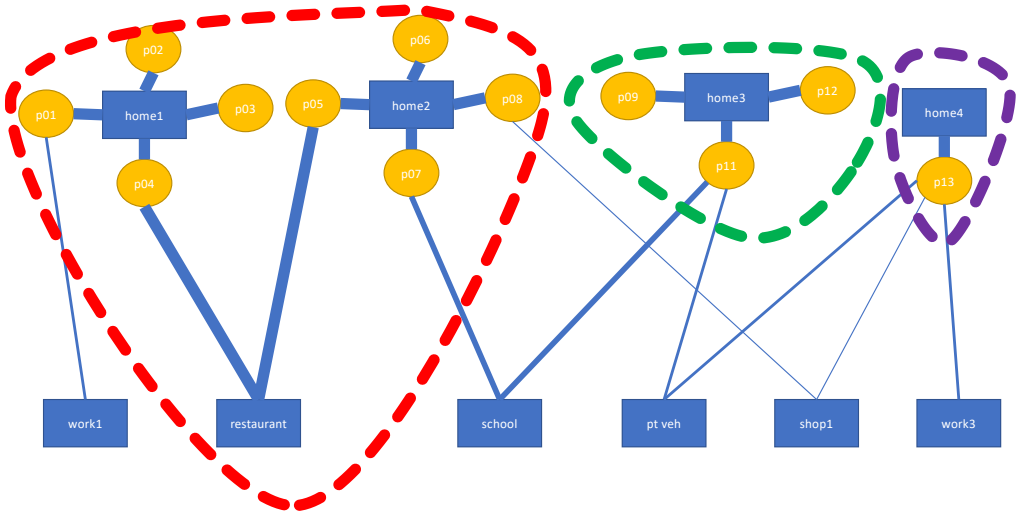


Figure 3: Graph interpretation: Static representation with facilities.

One could now use different edge weights, plotted as line thicknesses in the picture, between the persons and the facilities. This might even have different weights in different directions, e.g. a person easily bringing an infection into a facility/vehicle, but not having a high probability of contracting it. (This could, for example, be the case when wearing a mask with a valve.) However, (only) when making the simplification of assuming the same edge weight in both directions, it becomes clear that one could identify sub-clusters, as denoted by the dashed lines, similar to the metapopulation model of [38]. It seems improbable that one could stop the infection dynamics in those sub-clusters without rather drastic measures (such as completely dropping the "restaurant" activity denoted in the image, or separating family members from each other). This clarifies that the substrate on which the infection is progressing is far from homogeneous, and thus one needs a more general definition of percolation.

4. Methods and results

Neither of the two graph interpretations seem a useful starting point for investigation: The first one makes the graph even larger than it originally was; the second omits all time-dependent information. We therefore progress by simulating the original epidemic spreading model, i.e. the one described in Sec. 2.2, where persons move between and spend time in facilities and vehicles, and there can infect/get infected from other persons at the same location. We will return to the point of graph representation in Sec. 5.2 in the discussion.

4.1. Full model

As a first attempt to better understand the structure of our model, we use the full production model, with no interventions. Additionally, to obtain a more homogeneous situation, we do not include the separate models for saturdays and sundays, i.e. the simulation runs the same day over and over again.

We now set the calibration parameter Θ such that the model is near the percolation threshold, i.e. where initial seeds only grow in some of the cases. This is a counter-factual situation, since we have effectively reduced the transmissibility of SARS-2 until it barely spreads in our synthetic population. The motivation for this is to obtain clusters near the percolation threshold, which are easier to analyse than the full infection network.

We infect one randomly drawn person in the model, and wait until the infection dies out. It will always die out, potentially after a sufficient number of persons has been infected and who are, thus, immune. This is run over and over again, and the sizes s of the resulting clusters are noted. The resulting cluster size distributions, for

different values of Θ, are displayed in Fig. 4. The data is summed up into bins of increasing size, implying a multiplication with s. In addition, the probability to hit a cluster of size s is proportional to s. In consequence, overall we plot $s^2 \cdot n_s$ vs. s, where n_s is the cluster size distribution of Eq. (3).

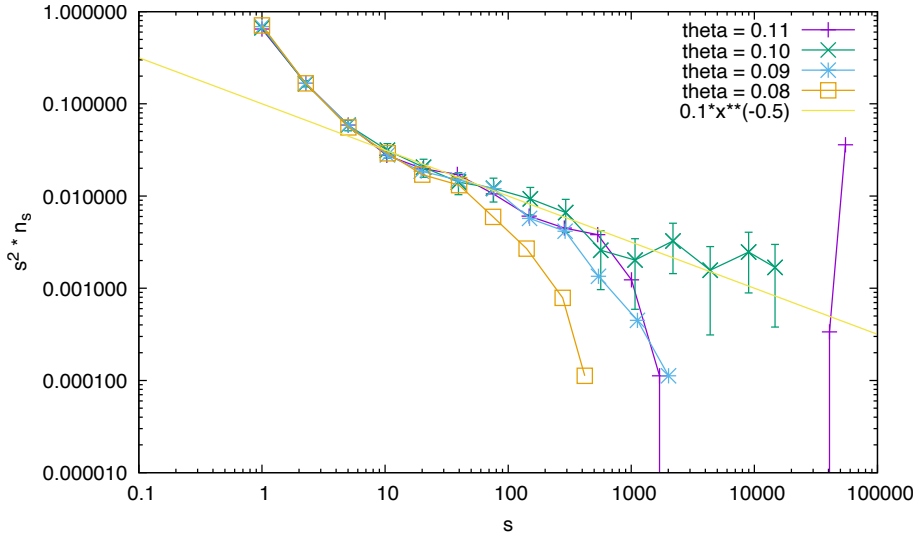


Figure 4: Cluster size distributions of Berlin model near the percolation threshold. The data looks consistent with a scaling law starting at $s \approx 10$ and a percolation threshold near $\Theta = 0.1$. – The straight line with slope -0.5 is added for comparison. The errorbars, given only for the curve closest to percolation to not overload the plot, denotes 3σ under the assumption of a Poisson distribution for the number of clusters of size s in each aggregation bin. The curve for each value of Θ is based on 8910 runs, i.e. 8910 clusters, many of them of size one.

The data displays the typical picture known from a percolation transition:

- For small values of Θ, the cluster size distribution ends early.
- For large values of Θ, the cluster size distribution also ends early, but there are also large clusters. This means that some infections die out, but some percolate through the population, implying the existence of a giant cluster.
- In between, the distribution becomes longer and longer.

Additionally, there is a clear break around $s = 10$. Possibly, infections within households are above criticality, and what we see is a variant of a metapopulation model [38]. Also see Sec. 5.2.

4.2. A more realistic situation

Fig. 4 and the related text were generated with a model where the Θ parameter was set such that the model barely percolates. This corresponds to a reproduction number R of approximately one, i.e. every infected person infects about one other person. Such a situation normally does not occur naturally, but it occurs quite often when the epidemic spreading is managed. A possible argument is as follows:[1] If the epidemic disease is dangerous enough, then it fills up hospitals quickly. For that reason, all countries that had the institutional capability to do so in the end reduced the spreading of the disease down to an R of approximately one. This also holds for countries such as Sweden, the USA, or UK.

Extensive simulations of possible interventions can be found on our project web page, `https://covid-sim.info`, in related publications [39, 40], and in our regular reports.[2] Here, in order to investigate such a controlled state further, we take our production model from the middle of January 2021 as a starting point. This includes the following interventions:

1. The general level of out-of-home activities was reduced to 64% percent.
2. Day care and schools were closed except for "emergency" cases. We model that as 0.2 participation at those activities.
3. Universities were closed.
4. Masks were obligatory in public transport and daily shopping.

To make the model more homogeneous for the investigations that will follow, we omitted weekend days, i.e. the model replays the same day over and over again. We also omitted outdoors temperature effects, i.e. it is assumed that all leisure activities occur indoors. Contact tracing, vaccinations and rapid tests have been omitted as well.

We consider this situation a useful starting point, since the situation in Germany at that point in time was such that infection numbers of the original variant of SARS-2 went down, while they went up for the new "Alpha" variant (B.1.1.7). This means that the Θ for the original variant was below the critical point, while for the alpha variant it was above the critical point. In consequence, it is plausible to investigate the behavior around the critical point.

[1] Dirk Brockmann, personal communication.

[2] `https://depositonce.tu-berlin.de/simple-search?query=modus-covid`

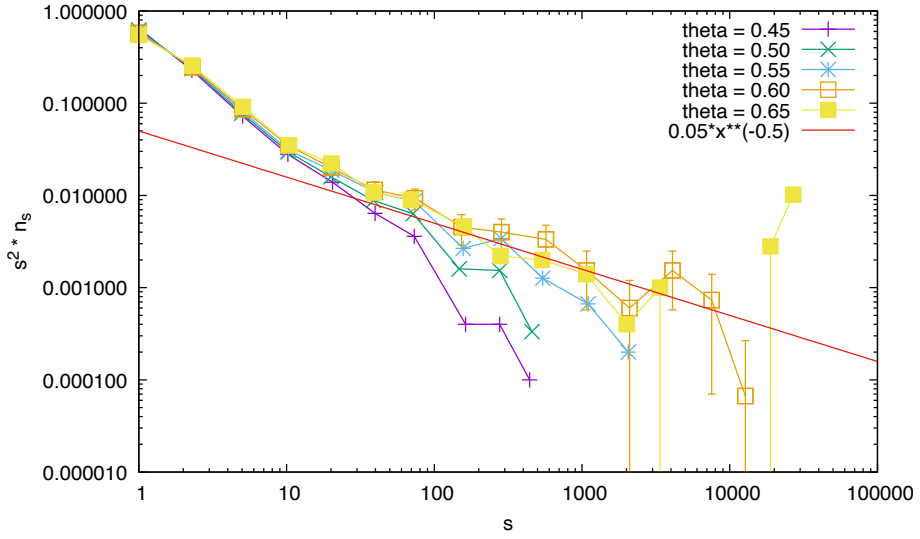


Figure 5: Cluster size distribution for the more realistic model from the first week of January 2021. – The straight line with slope -0.5 is added for comparison. The prescription for the errorbars is the same as in Fig. 4. The curves for for $\Theta = 0.45$ and 0.65 are based on 5000 runs. The curves for the other values of Θ are based on 15000 runs.

Fig. 5 presents the cluster size distribution for this situation, generated in the same way as Fig. 4. It behaves in the same way, showing a phase transition, this time at $\Theta \approx 0.5$. The Θ-value now is near its realistic value of our production model; the near-critical behavior is, as explained, generated by the interventions.

We find the same break for small cluster sizes s as before, although the break seems to have shifted to larger values of s than before. Also, there seems to be an additional break around $s = 2000$, and maybe one around $s = 150$. Possibly, this model is more fragmented, because many activity types have been fully suppressed (schools), or essentially removed from the infection dynamics because of mask obligations (public transport, shopping). Also, having all these breaks in the data makes it difficult to decide on a possible slope for $s^2 \times n_s \sim s^{-\tau+2}$; at best, one can say that the data is not inconsistent with a slope of 0.5, implying $\tau = 2.5$. This would need to be investigated further, but the present method is computationally too expensive (Sec. 5.1).

4.3. Super-spreading

We now add the feature that some persons are more infectious than others. Recent work [26] has confirmed that there are large differences in maximum virus load

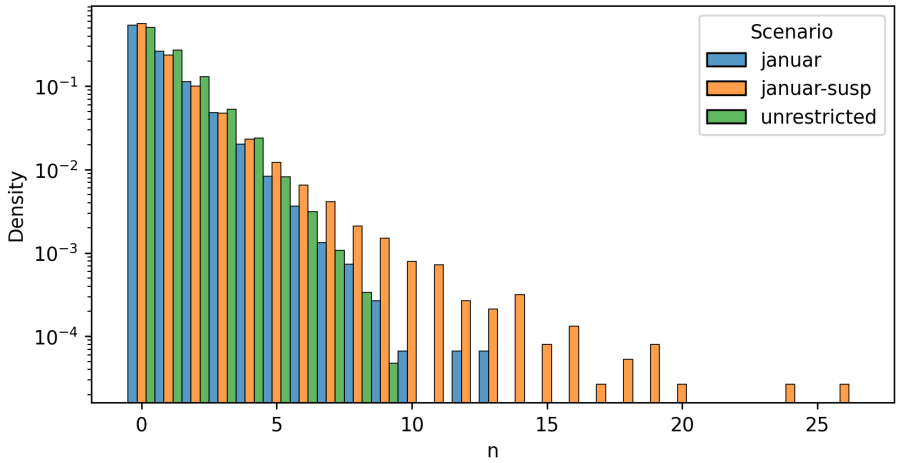


Figure 6: Distribution over the number of secondary cases per infected persons for each scenario. The superspreading scenario is showing an over-dispersion with a longer tail. "unrestricted" = original model where Θ was reduced to near the percolation threshold; "januar" = model where the interventions have reduced the model to near the percolation threshold; "januar-susp" = version of the "januar" model with super-spreading.

between persons, and that these result in large differences of breeding infectious material from swabs of these persons. The last link, to true real world infectiousness, is missing, but it is plausible to assume that it is there.

We model this by extending the infection model as described in Sec. 2.2.2 with an individual parametrization for each person. Each person p is assigned a value for their infectiousness and susceptibility $\text{inf}_p, \text{sus}_p \sim LogNormal(1, \sigma^2)$. To show the effects of super-spreading, we choose $\sigma = 1$ for subsequent runs. Fig. 6 shows the distribution of the number of reproductions per infected person, indeed confirming that there is more dispersion for this model variant.

Fig. 7 shows the cluster size distribution around the percolation threshold for the model including super-spreading. Evidently, the break that separates small cluster sizes remains, and roughly at the same position as for the model without super-spreading. The data for larger s is quite unstable, making it difficult to make any statement. Again, it is not inconsistent with $\tau = 2.5$.

4.4. Clusters

One advantage of simulation near the percolation threshold is that one obtains clusters of manageable size which are also relevant for the dynamics. Fig. 8 shows a typical larger infection cluster, with all activity types open but Θ reduced to the

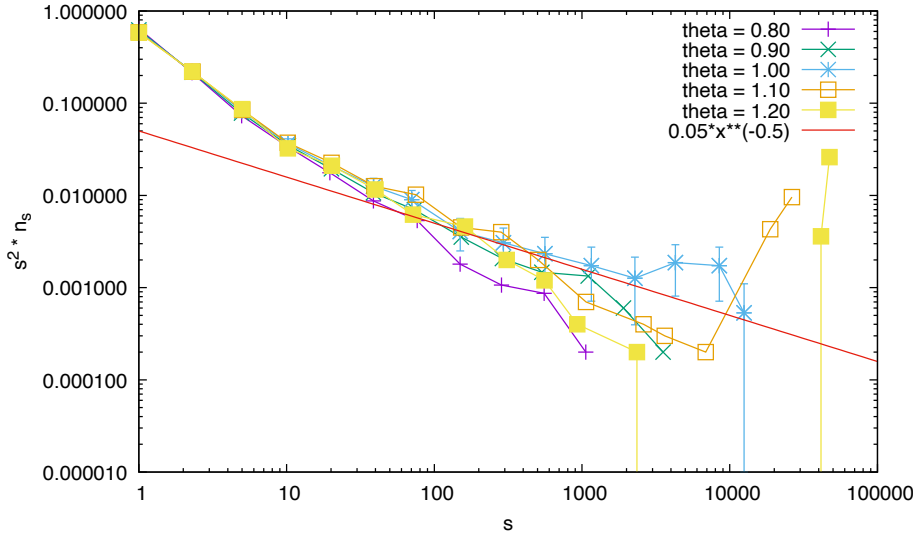


Figure 7: Cluster size distribution with superspreading for first week of January 2021. – The straight line with slope -0.5 is added for comparison. The prescription for the errorbars is the same as in Fig. 4. The curves for $\Theta = 0.8$, 0.9 and 1.0 are based on 15000 runs, those for $\Theta = 1.1$ on 10000 runs, and those for 1.2 on 5000 runs.

percolation threshold, as in Sec. 4.1. One finds that there are multiple small clusters in different settings, e.g. school, leisure, etc. Those clusters mostly remain in their setting, but the infections are cascaded, i.e. the first person infects a small number of others, those in turn infect a typically somewhat larger number of others, but the infection in that group then comes to an end when (presumably) the susceptible persons are exhausted.

Recall that the model replays the same day over and over again. This is what makes such cascading invasions of such a group possible. This behavior is, presumably, more realistic for the school and the work activities than for the leisure activities.

Fig. 9 shows the same plot for the model of January, where a behavior close to criticality was reached by policy interventions, cf. Sec. 4.2. Here, one finds that there are mostly leisure clusters, with interdispersed infections at home. We know from other analysis [17] that in our simulations of that regime, infections at leisure and infections at home carry approximately the same weight (also see Sec. 4.5 below). Given that knowledge, note that their dynamical behavior is quite different: leisure comes in small clusters, as described above for the unrestricted model, but the infections at home rarely form clusters that are larger than three.

Figure 8: Cutout of an infection cluster in the unrestricted regime. The colors denot infection contexts: blue home, orange leisure, green schools, brown work, pink public transport, gray other.

In consequence, the suspicion that the break in the cluster size distribution towards small sizes of s has something to do with households is wrong. Instead, it seems to be caused by the group sizes of the other activities.

Super-spreading (Fig. 10) changes the picture towards some persons infecting many other persons, while most persons infect few other persons. This is interesting to note, but it is unclear if this makes a big difference for the present model type, since also without super-spreading the infection is able to work its way through groups. Possibly, it makes a difference for test trace and isolate (TTI) strategies,

where clusters may be easier to find, especially when there are many asymptomatic carriers [41].

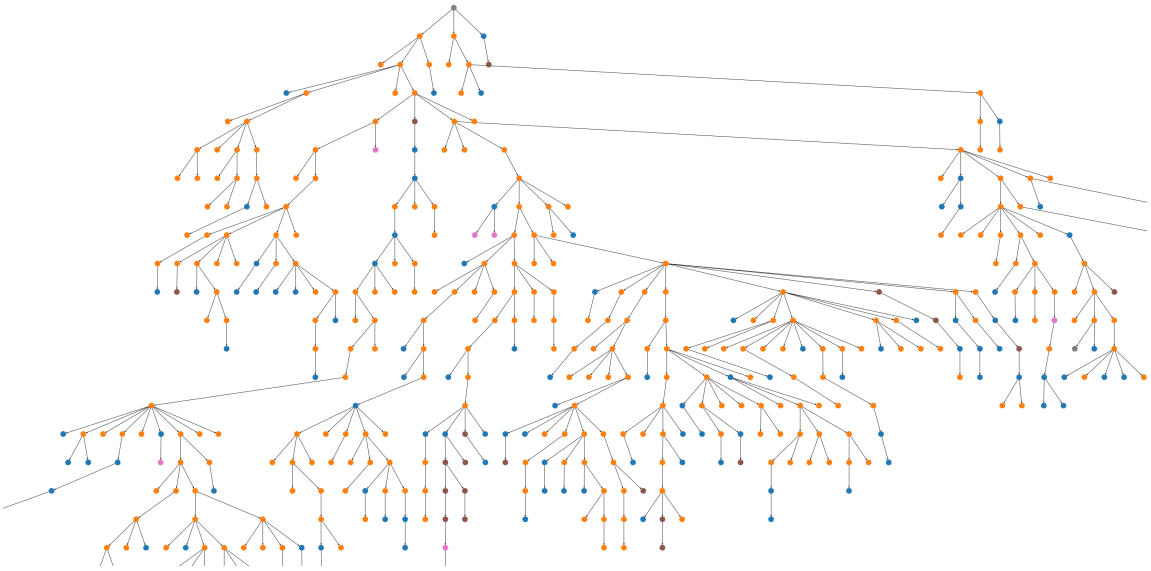


Figure 9: Cutout of an infection cluster in the regime of Jan'21. The colors are the same as in Fig. 8.

4.5. Change of infection contexts with cluster size

Finally, we show the distribution of the infection contexts as a function of cluster size for the three "realistic" models of Secs. 4.1 and 4.2. The corresponding plot for model of 4.3 looks similar to the latter, i.e. one finds that super-spreading does not make a difference for this type of analysis. In contrast, the model with all activity types of Sec. 4.1 behaves quite differently from the restricted model of Sec. 4.2. In both models, the initial seeds cause infections at home, i.e. within the own household. However, the infection dynamics then moves to the other contexts, mostly leisure in the model of January (bottom plot), and first leisure and then schools in the model with all activity types (top plot).

This implies that statistics taken from randomly drawn initial infections (such as in the present paper) need to be treated with care. Evidently, an infection that has run for many generations has a different behavior than one where the seed was randomly drawn. This can also be noted in our real-world situations, where, say, the opening of schools results in a transient phase of about two weeks until the contribution of schools to the infection dynamics has grown to its stationary level. This

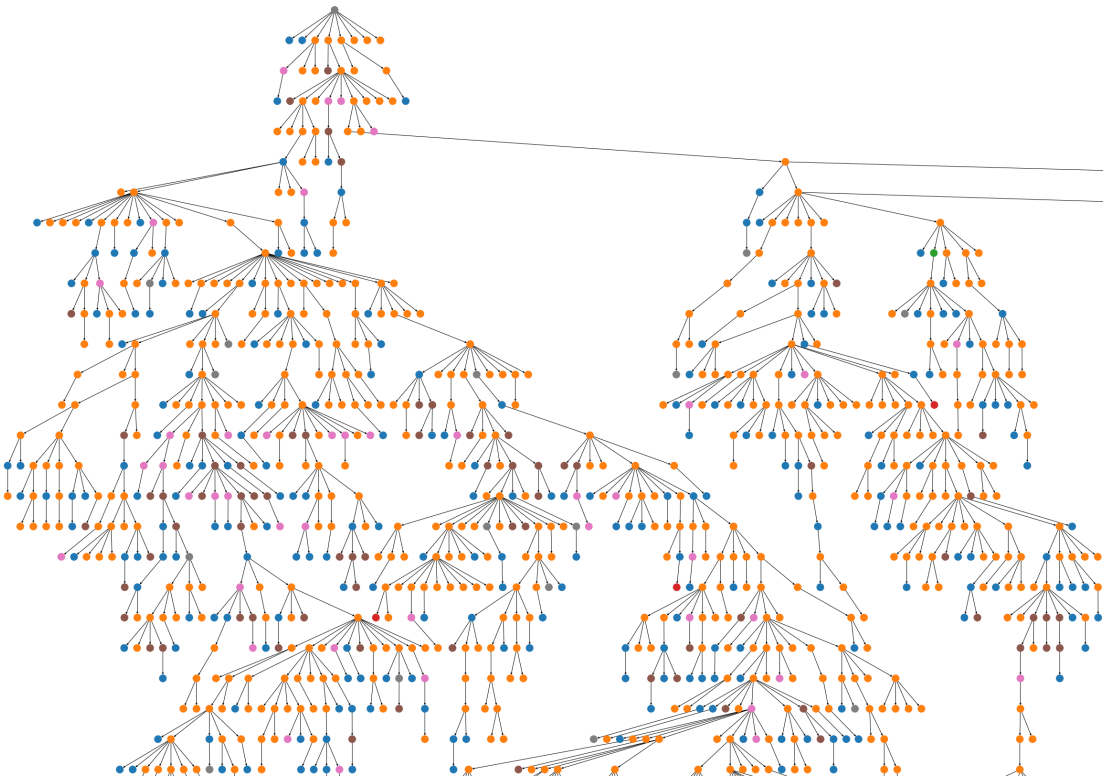


Figure 10: Cutout of a cluster showing the effects of super-spreading. The colors are the same as in Fig. 8.

also makes first-order analysis of the effects of interventions [42, 43, 44] incomplete, and confirms the necessity of models that complement statistical analysis.

5. Discussion

5.1. Computational issues

The computational model used for this paper moves all synthetic persons through their daily plans, i.e. entering and leaving facilities and vehicles as they are in the plan. This is an inefficient implementation for the type of investigation here, where we start with a single seed, and normally only have small clusters.

An alternative would be to grow the population as it is needed. Tadić and Melnik [12] present a model that does that. In their model, however, the new persons that are added into the model are created with randomly drawn properties, which is different from our approach, where the synthetic persons have attributes including

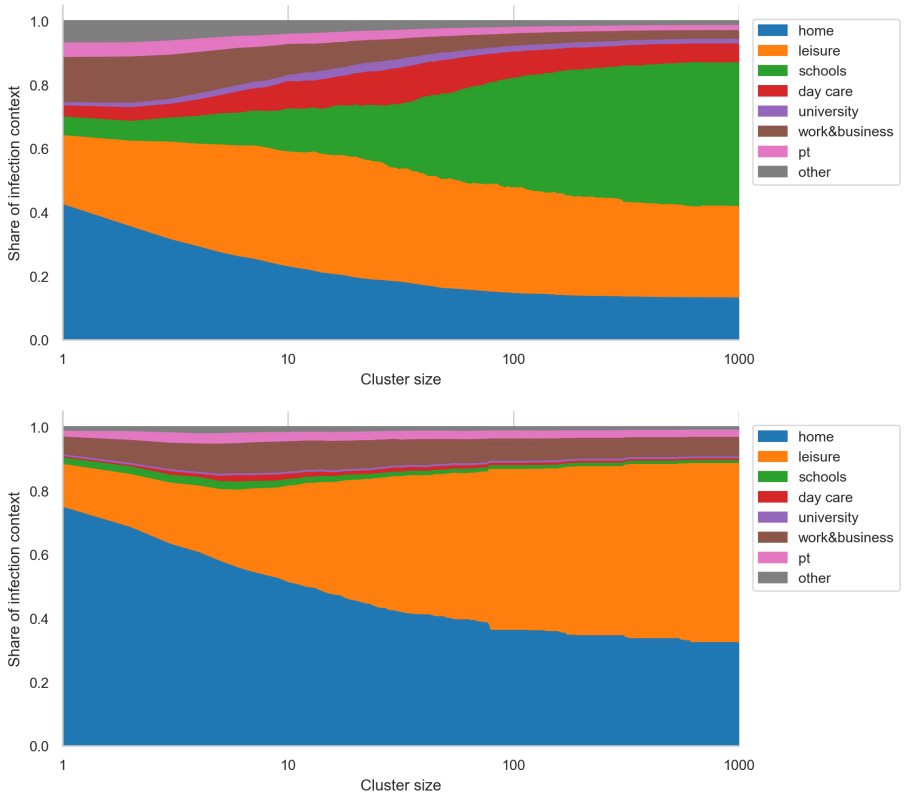


Figure 11: Distribution of infection context depending on the cluster size for unrestricted (top) and the January model (bottom). The super-spreading model looks similar to the January model. Small clusters show a larger share of household infections, while in big clusters the infection share shifts to the most infectious contexts.

daily plans that are known from the beginning, and which make up the structure of the model. Possibly, a data structure would need to be found and implemented that could identify those synthetic persons that use the same facilities or vehicles as the contagious persons, and then compute the interaction only with those. This implies, however, a major implementation effort.

5.2. Network analysis

Much is known about epidemics on synthetic graphs [35]. Our work starts "at the other end", i.e. from a realistic mobility model. The mobility model induces where people are co-located in the same facilities or vehicles and thus can infect each other, and how the infection is moved from one place to another.

For any given day, that process could be mapped onto a graph, as explained in Sec. 3.1: Persons who are co-located in a facility or a vehicle are connected; the strength of the connection is given by the infection probability described by the model (Eq. (1) and the following text); connections might not be symmetric since, e.g., one person may wear a mask and the other one not.

Since the material analyzed in the present paper replays the same day over and over again, this graph would be static, and one would indeed have a spreading process in that graph, where edges are traversed with probabilities given by link strengths, and nodes that have become immune cannot be infected a second time. According to [33], this would fall into the same universality class as percolation where the links are pre-computed from the probabilities, and in consequence the percolation picture is a useful starting point.

However, investigating that full graph for our model of Berlin is beyond reach for standard computational tools. In contrast, the introduced approach of investigating infection clusters near the percolation threshold leads to cluster sizes that can be handled, and from there to a number of insights. While the full model (Sec. 4.1) does not show discernible scale breaks beyond the one around $s = 10$, the model with reduced activity participation (Sec. 4.2) has additional such breaks. This intuits that the large number of activity types in the full model washes over the possibly resulting scale breaks, and therefore opens an avenue for analysis: To reduce the model to one that consists only of home and one out-of-home activity type, and obtain a better understanding of that set-up before moving on. Fig. 12 shows a first result in this direction, using an illustrative scenario where 10'000 persons live in single-person households, and simultaneously go to a joint activity for one hour per day. Evidently, the model still displays a percolation-like phase transition, but there are no more discernible scale breaks.

All of our results imply that an important element of the dynamics is that once the infection is inserted into a sub-group, it does *not* infect that sub-group in a single super-spreading event, but rather needs multiple generations inside that sub-group. At the same time, members of that sub-group may carry it into other sub-groups. Cliques, which denote fully connected subsets of nodes, are not the right language to describe this, since cliques imply a zero-or-one connectedness, whereas here the strengths of the edges inside the sub-group are important. Possibly, the language of simplicial complexes [45, 46] may be useful here.

5.3. Multi-day trajectories

The model considered in the present paper plays back the same day over and over again. Our current production model has separate submodels for Saturdays and

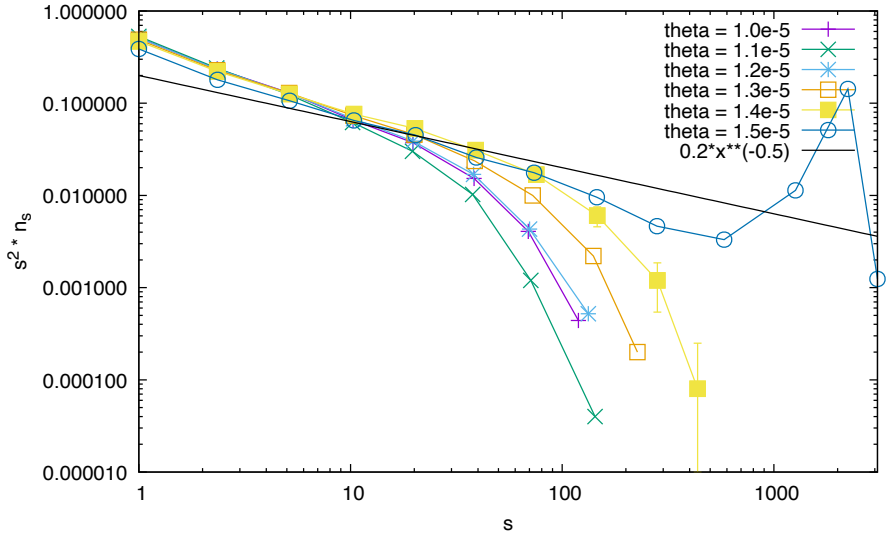


Figure 12: Cluster size distributions of the illustrative model near the percolation threshold. – The straight line with slope -0.5 is added for comparison. The prescription for the errorbars is the same as in Fig. 4. Each curve is based on 25000 runs.

Sundays, thus being able to represent the effect of special events on the weekends [17]. Clearly, it would be good to have even longer mobility trajectories, over several weeks or even months [47, 48]. Evidently, that will mean that the repeated visits to certain sub-groups (such as home, school or work) will be overlayed with singular events such as weddings or other large gatherings, and it remains to be investigated how these two elements interact. For Covid-19, clearly, the singular events were mostly suppressed, so that the investigation of the repeated visits is a useful starting point.

6. Conclusion

Much is known about epidemics on synthetic graphs [35]. Our work starts "at the other end", i.e. from realistic mobility models. It is not possible to map, in a simple way, the infection graphs that our model generates back to the synthetic graph models treated in the literature. Therefore, the insight drawn from those synthetic models cannot be taken directly to the real situation.

Our results show that the model displays elements of a percolation transition, in particular a cluster size distribution near the critical threshold that behaves similar to a percolation cluster size distribution in the sense that it has only small clusters below the threshold, clusters of all sizes at the threshold, and small clusters plus

one large cluster above the threshold. It is also demonstrated that this behavior is a result of the full model, since a simplified model where persons from single-person households join in one group once per day do not show the same behavior.

It remains to understand the model better. Optimally, as a consequence, we would be able to move from the current simulation-driven approach to predicting the effect of interventions to one where we can design good interventions from better understanding the model and the reality that it represents.

Acknowledgments

Dietrich Stauffer was my (KN's) advisor when I wrote my master thesis, on the topic of cloud growth [49]. What emerged was a model that essentially populated the sky with "cells" of different humidity, and a randomly inserted cloud seed would trigger cloud growth in all neighboring cells of large enough humidity. Evidently, this would generate percolation clusters.

I have learned a lot from him, including his mantra that a computer program should not have more lines than the author has years of age. Still, I came from climate modelling, which uses complex modelling systems with many lines of code [50], and moved on to transport modelling, also with many lines of code [18]. The issue both with climate modelling and with transport modelling is that one needs to have enough aspects of reality in the simulation models in order to be useful for prediction, and just dealing with the corresponding input data, including the necessary additional modeling, makes up much of those modelling packages. However, Dietrich Stauffer's mantra has always been in the back of my head; we are now using version control systems with systematic regression testing; we are trying to keep the number of lines of code under control; etc. And still: To some extent one needs to hope that the structure of the computer code is robust enough so that small glitches that are presumably still in the code will not destroy the scientific results. Clearly, scientific results from simulation need to be corroborated by other, independent, simulations before they can become scientifically fully accepted.

The work on the paper was funded by the Ministry of research and education (BMBF) Germany (01KX2022A) and TU Berlin; regular reports can be found through this search: `https://depositonce.tu-berlin.de/simple-search?query=modus-covid`. Zuse Institute Berlin (ZIB) provided CPU time.

Availability of data and materials

For computer code see `https://github.com/matsim-org/matsim-episim-libs`. Simulations were computed with version 67f98d4c570fa89227cb3e1426849b964c20cc64

of the code, started with command

```
java -jar matsim-episim-*.jar runParallel \
     --setup org.matsim.run.batch.BerlinPercolation \
     --params org.matsim.run.batch.BerlinPercolation$Params
```

A small number of things needs to be set manually in those computer codes; please contact the authors. The necessary input data for those simulations is available here: `https://doi.org/10.14279/depositonce-11495`. This contains all the events when synthetic persons enter or leave facilities or vehicles. The simulation outputs used for this paper, in particular the infection clusters, are made available here: `http://dx.doi.org/10.14279/depositonce-12054`.

References

[1] S. A. Müller, M. Balmer, A. Neumann, K. Nagel, Mobility traces and spreading of COVID-19, VSP Working Paper 20-06, TU Berlin, Transport Systems Planning and Transport Telematics, 2020. URL `http://dx.doi.org/10.14279/depositonce-9835`.

[2] S. A. Müller, W. Charlton, R. Ewert, C. Rakow, T. Schlenther, K. Nagel, MODUS-COVID vorhersage vom 8.4.2020, 2020. doi:`10.14279/DEPOSITONCE-10016`.

[3] S. A. Müller, W. Charlton, N. D. Conrad, R. Ewert, C. Rakow, H. Wulkow, T. Conrad, K. Nagel, C. Schütte, MODUS-COVID bericht vom 02.10.2020, 2020. doi:`10.14279/DEPOSITONCE-10624.2`.

[4] D. Stauffer, Introduction to percolation theory, Taylor & Francis, London, GB, and Philadelphia, PA, 1985.

[5] D. Stauffer, A. Aharony, Introduction To Percolation Theory: Second Edition, CRC Press, 2018.

[6] S. Eubank, H. Guclu, V. S. A. Kumar, M. V. Marathe, A. Srinivasan, Z. Toroczkai, N. Wang, Modelling disease outbreaks in realistic urban social networks, Nature 429 (2004) 180–184. doi:`10.1038/nature02541`.

[7] M. E. Halloran, N. M. Ferguson, S. Eubank, I. M. Longini, Jr, D. A. T. Cummings, B. Lewis, S. Xu, C. Fraser, A. Vullikanti, T. C. Germann, D. Wagener, R. Beckman, K. Kadau, C. Barrett, C. A. Macken, D. S. Burke,

P. Cooley, Modeling targeted layered containment of an influenza pandemic in the united states, Proc. Natl. Acad. Sci. U. S. A. 105 (2008) 4639–4644. doi:10.1073/pnas.0706849105.

[8] N. Ferguson, D. Laydon, G. Nedjati Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, Z. Cucunuba Perez, G. Cuomo-Dannenburg, Others, Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand, 2020. doi:10.25561/77482.

[9] D. L. Chao, M. E. Halloran, V. J. Obenchain, I. M. Longini, Jr, FluTE, a publicly available stochastic influenza epidemic simulation model, PLoS Comput. Biol. 6 (2010) e1000656. doi:10.1371/journal.pcbi.1000656.

[10] L. Hufnagel, D. Brockmann, T. Geisel, Forecast and control of epidemics in a globalized world, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 15124–15129. doi:10.1073/pnas.0308344101.

[11] A. Davids, G. du Rand, C.-P. Georg, T. Koziol, J. Schasfoort, SABCoM: A spatial Agent-Based COVID-19 model, 2020. doi:10.2139/ssrn.3663320.

[12] B. Tadić, R. Melnik, Modeling latent infection transmissions through biosocial stochastic dynamics, PLoS One 15 (2020) e0241163. doi:10.1371/journal.pone.0241163.

[13] S. Chang, E. Pierson, P. W. Koh, J. Gerardin, B. Redbird, D. Grusky, J. Leskovec, Mobility network models of COVID-19 explain inequities and inform reopening, Nature (2020). doi:10.1038/s41586-020-2923-3.

[14] S. L. Chang, N. Harding, C. Zachreson, O. M. Cliff, M. Prokopenko, Modelling transmission and control of the COVID-19 pandemic in australia, Nat. Commun. 11 (2020) 5710. doi:10.1038/s41467-020-19393-6.

[15] A. Najmi, S. Nazari, F. Safarighouzhdi, C. R. MacIntyre, E. J. Miller, T. H Rashidi, Facemask and social distancing, pillars of opening up economies, PLoS One 16 (2021) e0249677. doi:10.1371/journal.pone.0249677.

[16] A. Aleta, D. Martín-Corral, A. Pastore Y Piontti, M. Ajelli, M. Litvinova, M. Chinazzi, N. E. Dean, M. E. Halloran, I. M. Longini, Jr, S. Merler, A. Pentland, A. Vespignani, E. Moro, Y. Moreno, Modelling the impact of testing, contact tracing and household quarantine on second waves of COVID-19, Nat Hum Behav (2020). doi:10.1038/s41562-020-0931-9.

[17] S. A. Müller, M. Balmer, W. Charlton, R. Ewert, A. Neumann, C. Rakow, T. Schlenther, K. Nagel, Predicting the effects of COVID-19 related interventions in urban settings by combining activity-based modelling, agent-based simulation, and mobile phone data, medRxiv (2021) 2021.02.27.21252583. doi:10.1101/2021.02.27.21252583.

[18] A. Horni, K. Nagel, K. W. Axhausen, The Multi-Agent Transport Simulation MATSim, Ubiquity Press, London, UK, 2016. doi:10.5334/baw.

[19] Senozon, Mobility Pattern Recognition (MPR) und Anonymisierung von Mobilfunkdaten, https://senozon.com/wp-content/uploads/Whitepaper_MPR_Senozon_DE.pdf, 2020. Accessed: 2020-7-21.

[20] WHO, Report of the WHO-China joint mission on coronavirus disease 2019 (COVID-19). 2020, https://www.who.int/publications/i/item/report-of-the-who-china-joint-mission-on-coronavirus-disease-2019-(covid-19), 2020.

[21] X. He, E. H. Y. Lau, P. Wu, X. Deng, J. Wang, X. Hao, Y. C. Lau, J. Y. Wong, Y. Guan, X. Tan, X. Mo, Y. Chen, B. Liao, W. Chen, F. Hu, Q. Zhang, M. Zhong, Y. Wu, L. Zhao, F. Zhang, B. J. Cowling, F. Li, G. M. Leung, Temporal dynamics in viral shedding and transmissibility of COVID-19, 2020. doi:10.1038/s41591-020-0869-5.

[22] R. Wölfel, V. M. Corman, W. Guggemos, M. Seilmaier, S. Zange, M. A. Müller, D. Niemeyer, T. C. Jones, P. Vollmar, C. Rothe, M. Hoelscher, T. Bleicker, S. Brünink, J. Schneider, R. Ehmann, K. Zwirglmaier, C. Drosten, C. Wendtner, Virological assessment of hospitalized patients with COVID-2019, Nature 581 (2020) 465–469. doi:10.1038/s41586-020-2196-x.

[23] M. Dreher, A. Kersten, J. Bickenbach, P. Balfanz, B. Hartmann, C. Cornelissen, A. Daher, R. Stöhr, M. Kleines, S. Lemmen, Others, Charakteristik von 50 hospitalisierten COVID-19-Patienten mit und ohne ARDS, Dtsch. Arztebl. Int. 117 (2020) 271–278.

[24] D. Wang, B. Hu, C. Hu, F. Zhu, X. Liu, J. Zhang, B. Wang, H. Xiang, Z. Cheng, Y. Xiong, Y. Zhao, Y. Li, X. Wang, Z. Peng, Clinical characteristics of 138 hospitalized patients with 2019 novel Coronavirus-Infected pneumonia in wuhan, china, JAMA (2020). doi:10.1001/jama.2020.1585.

[25] Robert Koch Institute, RKI - SARS-CoV-2 steckbrief zur Coronavirus-Krankheit-2019 (COVID-19), https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Steckbrief.html, 2020. Accessed: 2020-3-18.

[26] T. C. Jones, G. Biele, B. Mühlemann, T. Veith, J. Schneider, J. Beheim-Schwarzbach, T. Bleicker, J. Tesch, M. L. Schmidt, L. E. Sander, F. Kurth, P. Menzel, R. Schwarzer, M. Zuchowski, J. Hofmann, A. Krumbholz, A. Stein, A. Edelmann, V. M. Corman, C. Drosten, Estimating infectiousness throughout SARS-CoV-2 infection course, Science (2021). doi:10.1126/science.abi5273.

[27] T. Smieszek, A mechanistic model of infection: why duration and intensity of contacts should be included in models of disease spread, Theor. Biol. Med. Model. 6 (2009) 25. doi:10.1186/1742-4682-6-25.

[28] T. Smieszek, Models of epidemics: how contact characteristics shape the spread of infectious diseases, Ph.D. thesis, Ph.D. thesis, ETH Zurich, Switzerland, 2010. doi:10.3929/ETHZ-A-006109956.

[29] D. Lewis, COVID-19 rarely spreads through surfaces. so why are we still deep cleaning?, Nature 590 (2021) 26–28. doi:10.1038/d41586-021-00251-4.

[30] J. Lelieveld, F. Helleis, S. Borrmann, Y. Cheng, F. Drewnick, G. Haug, T. Klimach, J. Sciare, H. Su, U. Pöschl, Model calculations of aerosol transmission and infection risk of COVID-19 in indoor environments, Int. J. Environ. Res. Public Health 17 (2020). doi:10.3390/ijerph17218114.

[31] M. Kriegel, U. Buchholz, P. Gastmeier, P. Bischoff, I. Abdelgawad, A. Hartmann, Predicted infection risk for aerosol transmission of SARS-CoV-2, medRxiv (2020). doi:10.1101/2020.10.08.20209106.

[32] P. Grassberger, On the critical behavior of the general epidemic process and dynamical percolation, Math. Biosci. 63 (1983) 157–172. doi:10.1016/0025-5564(82)90036-0.

[33] J. L. Cardy, P. Grassberger, Epidemic models and percolation, J. Phys. A Math. Gen. 18 (1985) L267. doi:10.1088/0305-4470/18/6/001.

[34] M. Newman, Networks, Oxford University Press, 2018.

[35] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, A. Vespignani, Epidemic processes in complex networks, Rev. Mod. Phys. 87 (2015) 925–979. doi:10.1103/RevModPhys.87.925.

[36] M. Bastian, S. Heymann, M. Jacomy, Gephi: An open source software for exploring and manipulating networks, 2009. URL: http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

[37] T. C. Freeman, S. Horsewell, A. Patir, J. Harling-Lee, T. Regan, B. B. Shih, J. Prendergast, D. A. Hume, T. Angus, Graphia: A platform for the graph-based visualisation and analysis of complex data, bioRxiv (2020). URL: https://www.biorxiv.org/content/early/2020/09/03/2020.09.02.279349. doi:10.1101/2020.09.02.279349. arXiv:https://www.biorxiv.org/content/early/2020/09/03/2020.09.02.279349.full.pdf

[38] V. Colizza, A. Vespignani, Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: theory and simulations, J. Theor. Biol. 251 (2008) 450–467. doi:10.1016/j.jtbi.2007.11.028.

[39] S. A. Müller, M. Balmer, A. Neumann, K. Nagel, Mobility traces and spreading of COVID-19, 2020. doi:10.14279/DEPOSITONCE-9835.

[40] S. A. Müller, M. Balmer, B. Charlton, R. Ewert, A. Neumann, C. Rakow, T. Schlenther, K. Nagel, Using mobile phone data for epidemiological simulations of lockdowns: government interventions, behavioral changes, and resulting changes of reinfections, 2020. doi:10.1101/2020.07.22.20160093.

[41] S. Omi, H. Oshitani, Japan's COVID-19 response, https://www.mofa.go.jp/files/100061341.pdf, 2020. Accessed: 2020-8-20.

[42] J. M. Brauner, S. Mindermann, M. Sharma, D. Johnston, J. Salvatier, T. Gavenčiak, A. B. Stephenson, G. Leech, G. Altman, V. Mikulik, A. J. Norman, J. T. Monrad, T. Besiroglu, H. Ge, M. A. Hartwick, Y. W. Teh, L. Chindelevitch, Y. Gal, J. Kulveit, Inferring the effectiveness of government interventions against COVID-19, Science (2020). doi:10.1126/science.abd9338.

[43] M. Sharma, S. Mindermann, C. Rogers-Smith, G. Leech, B. Snodin, J. Ahuja, J. B. Sandbrink, J. T. Monrad, G. Altman, G. Dhaliwal, L. Finnveden, A. J. Norman, S. B. Oehm, J. F. Sandkühler, T. Mellan, J. Kulveit, L. Chindelevitch, S. Flaxman, Y. Gal, S. Mishra, J. M. Brauner, S. Bhatt, Understanding the effectiveness of government interventions in europe's second wave of COVID-19, 2021. doi:10.1101/2021.03.25.21254330.

[44] U. Berger, C. Fritz, G. Kauermann, Eine statistische analyse des effekts von verpflichtenden tests an schulen mit präsenzunterricht im vergleich zum distanzunterricht 238 (2021). doi:10.5282/ubm/epub.76005.

[45] D. Wang, Y. Zhao, H. Leng, M. Small, A social communication model based on simplicial complexes, Phys. Lett. A 384 (2020) 126895. doi:10.1016/j.physleta.2020.126895.

[46] M. Andjelković, B. Tadić, S. Maletić, M. Rajković, Hierarchical sequencing of online social graphs, Physica A: Statistical Mechanics and its Applications 436 (2015) 582–595. doi:10.1016/j.physa.2015.05.075.

[47] K. W. Axhausen, A. Zimmermann, S. Schönfelder, G. Rindsfüser, T. Haupt, Observing the rhythms of daily life: A six-week travel diary, Transportation 29 (2002) 95–124.

[48] IfV, German mobility panel, http://http://mobilitaetspanel.ifv.kit.edu/english/index.php, accessed 27 February 2019. URL: http://http://mobilitaetspanel.ifv.kit.edu/english/index.php.

[49] K. Nagel, E. Raschke, Self-organized criticality in cloud formation?, Physica A 182 (1992) 519.

[50] K. Nagel, Une Parametrisation du frottement des ondes de gravité dans le modèle de circulation générale du LMD (A parametrization of the gravity wave drag in the general circulation model of the LMD), Master's thesis, University Paris 6, Paris, France, 1989.