scientific reports

Check for updates

OPEN

Tracing contacts to evaluate the transmission of COVID-19 from highly exposed individuals in public transportation

Caio Ponte[1], Humberto A. Carmona[2], Erneson A. Oliveira[1,3,4]✉, Carlos Caminha[1], Antonio S. Lima Neto[5,6], José S. Andrade Jr.[2] & Vasco Furtado[1,7]

We investigate, through a data-driven contact tracing model, the transmission of COVID-19 inside buses during distinct phases of the pandemic in a large Brazilian city. From this microscopic approach, we recover the networks of close contacts within consecutive time windows. A longitudinal comparison is then performed by upscaling the traced contacts with the transmission computed from a mean-field compartmental model for the entire city. Our results show that the effective reproduction numbers inside the buses, Re^{bus}, and in the city, Re^{city}, followed a compatible behavior during the first wave of the local outbreak. Moreover, by distinguishing the close contacts of healthcare workers in the buses, we discovered that their transmission, Re^{health}, during the same period, was systematically higher than Re^{bus}. This result reinforces the need for special public transportation policies for highly exposed groups of people.

Human mobility is crucial to understanding the COVID-19 pandemic since the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) is disseminated individual to individual via droplet and airborne transmissions[1]. Considering that non-pharmacological interventions, such as social distancing and isolation, still represent fundamental measures to control the COVID-19 outbreak, the nature of SARS-CoV-2 dissemination also unveils the need to understand the role of the space where interaction between people occurs. There is consensus that superspreading events, which are usually investigated through contact tracing models[2–7], are more likely to happen in indoor environment, as substantiated by previous studies of indoor contagion in hospitals[8,9], restaurants[10], offices[11], and even on cruise ships[12,13]. However, the relation between the microscopic level of contagion in indoor environments and the macroscopic observables, such as the numbers of cases and deaths at the city scale or the populations of compartmental models[14–19], remains unclear.

Public transportation is one of the main forms of commuting, playing an important role in the pace of life in cities[20], specially in epidemics[21–26]. In spite of the fact that some cities have adopted social distancing and sanitary protocols on public transportation to control the COVID-19 outbreak, it is common that buses or subways get crowded at rush hour, mainly in the developing countries. In recent months, some studies have been proposed to establish the safety of public transportation regarding the indoor COVID-19 contagion[27–33]. To the best of our knowledge, however, none of these studies have considered the possibility of a comparative analysis based on a two-fold perspective, namely, the dynamic of people's movement in a city and the dynamic of the virus dissemination within vehicles of public transport.

Here, we define two data-driven mathematical models based on concepts of complex networks and non-linear dynamics in order to foster the understanding of the role public transportation plays in the COVID-19 pandemic. We use data about people's movement on buses and COVID-19 infections in the city of Fortaleza, a large Brazilian metropolis with a population of 2.67 million people. From March to December 2020, Fortaleza

[1]Programa de Pós Graduação em Informática Aplicada, Universidade de Fortaleza, Fortaleza, Ceará 60811-905, Brasil. [2]Departamento de Física, Universidade Federal do Ceará, Fortaleza, Ceará 60455-760, Brasil. [3]Laboratório de Ciência de Dados e Inteligência Artificial, Universidade de Fortaleza, Fortaleza, Ceará 60811-905, Brasil. [4]Mestrado Profissional em Ciências da Cidade, Universidade de Fortaleza, Fortaleza, Ceará 60811-905, Brasil. [5]Célula de Vigilância Epidemiológica, Secretaria Municipal da Saúde, Fortaleza, Ceará 60810-670, Brasil. [6]Centro de Ciências da Saúde, Universidade de Fortaleza, Fortaleza, Ceará 60811-905, Brasil. [7]Empresa de Tecnologia da Informação do Ceará, Governo do Estado do Ceará, Fortaleza, Ceará 60130-240, Brasil. ✉email: erneson@unifor.br

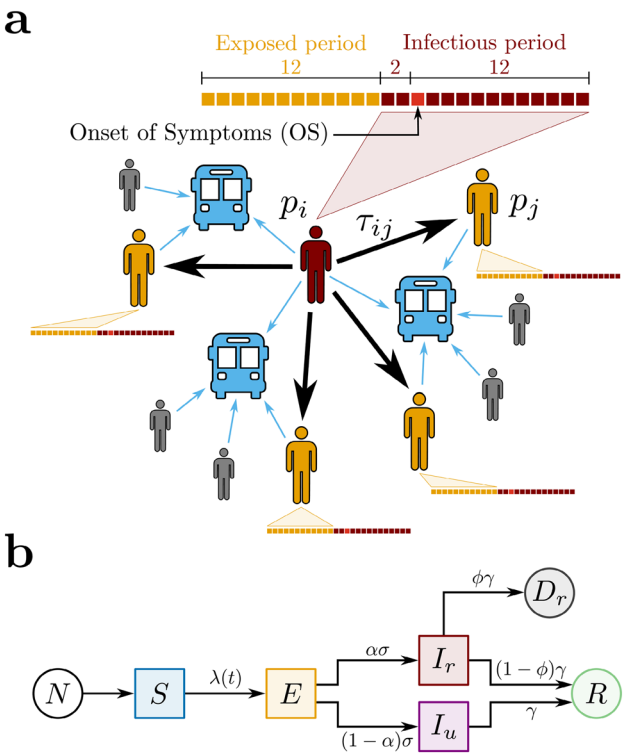

Figure 1. Proposed models for COVID-19 and spreading scenarios. (**a**) Potentially Infectious Contacts (PICs). We define a PIC when an infectious passenger p_i (in red) and an exposed passenger p_j (in yellow) share the same bus. The weight τ_{ij} is the estimated value of the ride time shared by p_i and p_j. The time lines show the infectious (in red) and the exposed periods (in yellow) of each passenger, where each square represents one day. The time lines are built based on the Onset of Symptoms (OS). Precisely, the infectious period begins 2 days before OS and ends 12 days after OS, while the exposed period begins 14 days before OS and ends 2 days before OS. Other passengers (in gray), even though they have shared the same bus with p_i, either were not notified as COVID-19 cases or, however notified, they were not considered as PICs because they were not in their exposed period. (**b**) The SEIIR model. The total population of size N provides the susceptible population S (in blue). The susceptible individuals become exposed E (in yellow) at a time-dependent rate $\lambda(t)$. The exposed individuals become infectious at a time rate σ. A fraction α of the infectious population is reported I_r (in red), while a fraction $(1-\alpha)$ is unreported I_u (in purple). The infectious individuals that recover, reported or not, become recovered R (in green) at a time rate γ. Finally, it is assumed that a fraction ϕ of the removed population γI_r deceases D_r (in dark gray).

recorded officially 88,983 cases and 4,620 deaths of COVID-19, achieving the peaks of 3,414 cases on May 1 and of 230 deaths on May 11[34]. At the microscopic scale, we define a contact tracing model to estimate the transmission within city buses and, at the macroscopic scale, a compartmental model is employed to estimate the transmission in the entire city. The main contribution of our study is a comparative analysis between these two distinct modeling approaches through the combination of daily epidemiological and mobility data during the first 9 months of the local COVID-19 outbreak, and through different social distancing restriction regimes. One specially relevant aspect of this work is the fact that we are able to trace within the public transportation vehicles (i.e., indoor environments) two groups of people, one of them with a higher exposure to the virus in comparison to the other. This allows us to shed light on potentially dangerous superspreading events in public transportation.

Modeling approach

Contact tracing model. We propose a contact tracing model using two datasets that relate bus validations to COVID-19 confirmed cases during the periods of social isolation, lockdown, and economic reopening in the city of Fortaleza, Ceará, Brazil (see "Methods"). Our model is a network based on Potentially Infectious Contacts (PICs), in which bus passengers during their infectious period—according to subsequent diagnosis of COVID-19—have shared the transport for a certain amount of time with other passengers, the latter in their exposed period—also according to subsequent COVID-19 diagnosis. Precisely, the proposed network is composed of vertices p_i that represent the passengers diagnosed with COVID-19, and weighted directed edges $c_k = (p_i, p_j, \tau_{ij})$ that represent PICs. For each edge, the direction is assigned from an infectious passenger p_i to an exposed passenger p_j, and the weight τ_{ij} is defined as the estimated value of the ride time shared by p_i and p_j on the same bus, as shown in Fig. 1a. We calculate τ_{ij} by superimposing the estimated ride times from p_i and p_j, considering the different moments of their boarding. Here, the epidemiological profile for COVID-19 transmission is characterized by the dates of the passengers' Onset of Symptoms (OS). The infectious period corresponds to the days in

Parameter	Mean/value	Variance
Transmission rate ($\beta^{(0)}$, days^{-1})	1.0	0.3
Relative transmission rate ($\mu^{(0)}$)	0.5	0.1
Fraction of reported ($\alpha^{(0)}$)	0.15	0.05
Exposed to infectious rate ($\sigma^{(0)}$, days^{-1})	0.23	0.05
Removal rate ($\gamma^{(0)}$, days^{-1})	0.28	0.05
Mortality ratio ($\phi^{(0)}$)	0.08	0.01
Number of particles (P)	300	
Cooling factor (a)	0.93	
Cooling factor (b)	0.93	

Table 1. Initial model parameters on March 24.

which a passenger diagnosed with COVID-19 can transmit the virus, initiating 2 days before OS and ending 12 days after OS. The exposed period refers to the time window during which the passenger can get the virus and maintain it latent until the infectious period. In this context, the exposed period begins 14 days before OS and ends 2 days before OS, i.e., the infectious and the exposed periods have a width of 14 and 12 days, respectively, and they do not overlap[35–37]. Furthermore, if there is more than one PIC related to an exposed passenger p_j, we consider solely the edge with the largest value of τ_{ij}. It is important to notice that, by crossing the datasets of bus validations and confirmed cases of COVID-19 in Fortaleza during the period from March to December 2020, we are able to identify 5159 pairs of infectious and exposed passengers that rode the same bus on the same day. However, their associated values of τ_{ij} could only be computed for 3023 (58.6%), due to missing information in the dataset of bus validations. From these pairs, we obtain that the network of PICs corresponds to a forest composed of 213 trees with a total of 530 vertices (infectious passengers) and 317 edges (PICs). From all vertices found, 97 were identified as healthcare workers (see "Methods"). The Centers for Disease Control and Prevention (CDC) recommends that any contact tracing strategy for COVID-19 should consider the concept of Close Contacts (CCs)[38], i.e., anybody who has been for at least 15 min within 6 ft ($\approx$ 2 meters) of an infectious person. Since buses are small, enclosed, and they have a great tendency to get crowded at rush hours, we define the CCs in the network of PICs only considering the time condition $\tau_{ij} > \tau_c$, where the threshold $\tau_c = 15$ min. Applying this criterion to the network of PICs, we find that the network of CCs is composed of 154 trees with a total of 360 vertices (infectious passengers) and 206 edges (CCs). In this case, 75 vertices were identified as healthcare workers. In order to understand the COVID-19 spreading in public transportation, we define the effective reproduction number for the contact tracing model, Re^{bus}, as the expected number of secondary cases produced by a single (typical) infection. Precisely, it accounts for two contributions in relation to who is spreading the disease: one due to reported infectious individuals, Re_r^{bus}, and another due to unreported infectious individuals Re_u^{bus}. Here, we assume that the fraction of newly reported to newly unreported cases generated by a typical reported infectious individual remains invariant during time. This is equivalent to consider the value of Re_r^{bus} proportional to the average number of outdegrees from the vertices in the network of CCs during a given time window, $\langle d_{out}^{CCs} \rangle$,

$$Re_r^{bus} = \chi \langle d_{out}^{CCs} \rangle. \tag{1}$$

The constant of proportionality χ involved in this relation will be explicitly computed through the calibration between the contact tracing and the compartmental models. Each consecutive time window has a width of 22 days and a step size of 5 days. We emphasize that our model has an intrinsic time delay regarding the consolidation of Re_r^{bus} that can reach $\approx$ 53 days. This value is associated to the time delay in the consolidation of COVID-19 dataset ($\approx$ 15 days) and to the superposition of the maxima of two infectious periods and one exposed period.

Compartmental model. We also adopt a compartmental model to describe the transmission of COVID-19 in order to estimate the levels of infection of the pathogen in Fortaleza. Here, we propose a SEIIR model that distinguishes the populations of Susceptible, Exposed, Infectious (reported or unreported), and Removed (recovered or deceased) individuals, as shown in Fig. 1b. Our model is inspired by the SEIIR model proposed by Li et al.[16]. The reported infectious population I_r corresponds to the number of individuals that had the SARS-CoV-2 infection confirmed by the health system. The unreported infectious population I_u comprises the complement of I_r, i.e., individuals that were infected with COVID-19 but remained unknown to health authorities. We assume that the large majority of the reported infectious individuals are symptomatic cases, in contrast to the population of unreported infectious individuals—of which the large majority is assumed to be of asymptomatic cases. Given this fundamental assumption and considering the recent finding that asymptomatic people are 42% less likely to transmit the SARS-CoV-2 than symptomatic ones[39], we define that the transmission rate for the unreported infectious population I_u is reduced by a dimensionless factor of μ in relation to the parameter β that represents the transmission rate for the reported infectious population I_r. In this context, the time-dependent rate at which the susceptible population S becomes the exposed population E is given by

$$\lambda(t) = \beta \frac{(I_r + \mu I_u)}{N}, \tag{2}$$

where N is the total population of Fortaleza, taken as constant, being approximately equal to 2.67 million people. A fraction α of the exposed individuals is presumed to become reported infectious at a rate σ, and the complementary fraction $(1-\alpha)$ to evolve to unreported infected at the same rate. Also, both reported and unreported infectious population are assumed to become part of the removed population at the same rate γ. We also keep track of the fraction ϕ of the removed reported infectious population evolving to death, so that the reported deceased population D_r increases at a rate of $\phi\gamma I_r$. The following system of coupled differential equations rules our model:

$$\frac{dS}{dt} = -\lambda S, \tag{3}$$

$$\frac{dE}{dt} = \lambda S - \sigma E, \tag{4}$$

$$\frac{dI_r}{dt} = \alpha\sigma E - \gamma I_r, \tag{5}$$

$$\frac{dI_u}{dt} = (1-\alpha)\sigma E - \gamma I_u, \tag{6}$$

$$\frac{dR}{dt} = (1-\phi)\gamma I_r + \gamma I_u, \tag{7}$$

$$\frac{dD_r}{dt} = \phi\gamma I_r. \tag{8}$$

The total population $N = S + E + I_r + I_u + R + D_r$ is conserved. Furthermore, it can be readily shown[16] that the effective reproduction number Re^{city} is given by

$$Re^{city} = \left[\alpha\frac{\beta}{\gamma} + (1-\alpha)\mu\frac{\beta}{\gamma}\right]\frac{S}{N}. \tag{9}$$

Note that in the particular case that $S \approx N$ and all the infectious population are reported, meaning $\alpha = 1$, the value Re^{city} reduces to the traditional value $R_0 = \beta/\gamma$. From Eq. (9), we can identify $Re_r^{city} = (\beta/\gamma)(S/N)$ as the average number of secondary infections due to contagion with reported infectious individuals, while $Re_u^{city} = \mu(\beta/\gamma)(S/N)$ is the effective reproduction number due to contagion with unreported infectious individuals. Finally, the SEIIR model is used here as a core model within the Iterative Ensemble Kalman Filter (IEnKF) framework (see "Methods"). This approach allows us to investigate the time evolution of the effective reproduction number Re^{city} by inferring the mean parameters of the SEIIR model and initial populations (see Figs. S1 and S2 of the Supplementary Information). The IEnKF framework is systematically applied to running windows of 22 days, with step size of 5 days, starting from March 24 to November 9, 2020. We use as observable the cumulative number of deaths by SARS-CoV-2 reported daily by the health authorities. For the first window, the reported values of daily cases, $C^{(0)}$, and confirmed deaths by SARS-CoV-2 infections, $D^{(0)}$, are used to estimate the mean value for the exposed $E^{(0)} \approx C^{(0)}/(\alpha\sigma) \approx 4,982$ and deceased populations, $D^{(0)} \approx 1$. The mean and variance of the initial values for the model parameters adopted for the first window are listed in Table 1 along with the corresponding variances. These values are similar to the best-fit model posterior estimates in reference[16]. In order to minimize the sensibility from the initial conditions, for each window, we run 10 different trials with parameters and subpopulations drawn from normal distributions with the corresponding variance. After using IEnKF to estimate the values of all model parameters for the first window, the factor Re^{city} is calculated at its center. These parameters and all populations obtained by numerical integration of Eqs. (3)–(8) are then used as initial guesses for the second window, except for the deceased population, $D^{(0)}$, for which the mean value is estimated from the reported confirmed deaths by SARS-CoV-2 infections at the beginning of each window. The same procedure is then repeated for the third and subsequent windows.

Results and Discussion

Figure 2a shows the normalized moving averages of bus validations of individuals that got COVID-19, including healthcare workers. We note that healthcare workers that came into contact with SARS-CoV-2 during the studied period did not reduce their bus rides as much as other passengers. In addition, their normalized moving averages of bus validations are getting closer to each other again as the economic reopening progresses. The inset of Fig. 2a shows the daily bus validations, which gradually started to increase in the economic reopening. Figure 2b,c show the daily numbers of cases and deaths, respectively, following the same previous normalization and stratification.

The representativeness of the dataset of COVID-19 confirmed cases on buses is assessed comparing the daily numbers of infectious individuals within those vehicles and in the entire city, as shown in Fig. 3. The daily number of infectious individuals is computed taking into account the 14 days that the individuals remain infectious, i.e., each individual who tested positive for COVID-19 counts up to 14 times, once per day, for the infectious curve. Figure 3a,b show the daily numbers regarding all infectious passengers and those infectious passengers who are healthcare workers, respectively. Similarly, the daily numbers of infectious individuals and

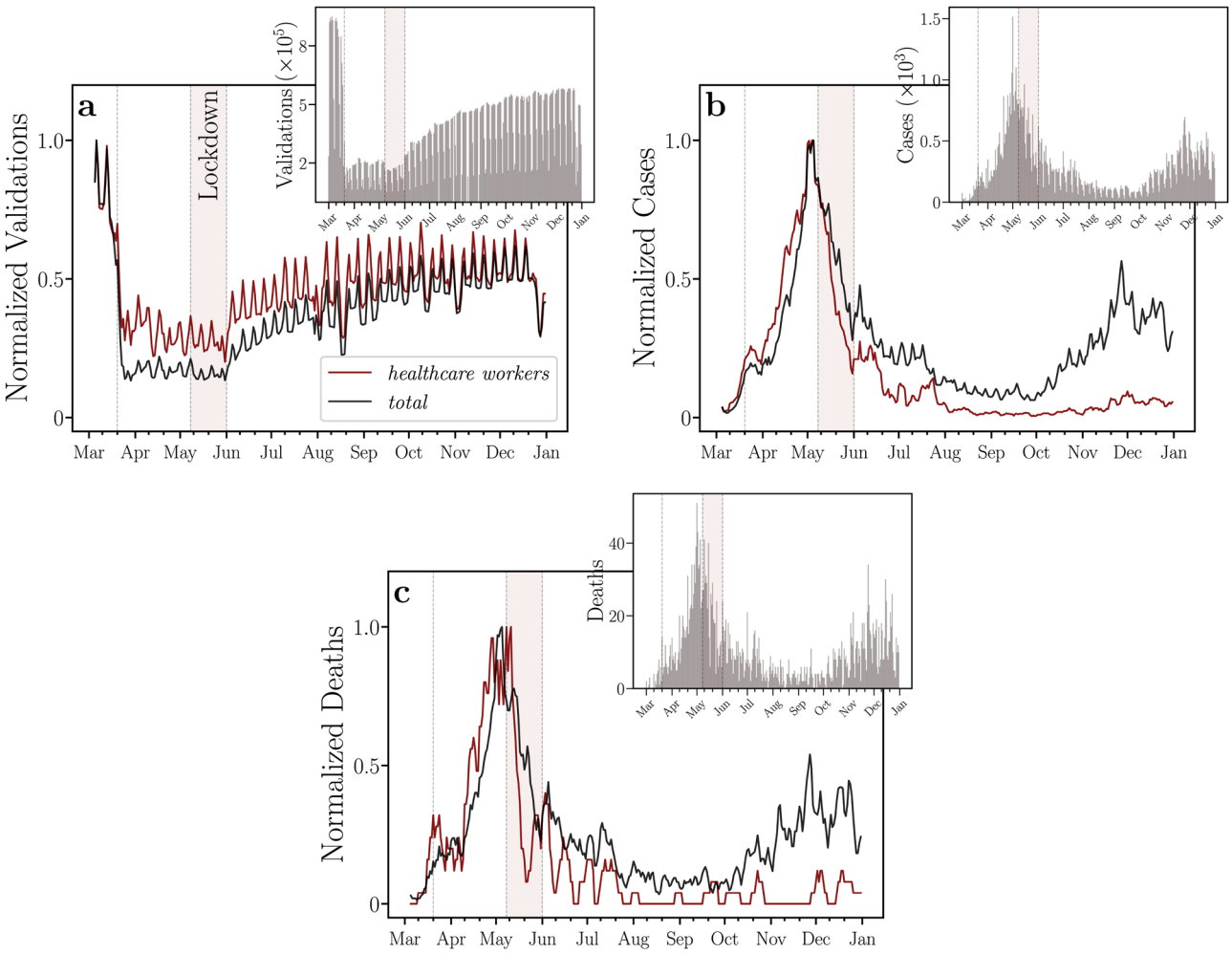

Figure 2. Time series and moving averages of bus validations, COVID-19 cases, and COVID-19 deaths. Time evolutions of the normalized moving averages of (**a**) bus validations, (**b**) COVID-19 cases, and (**c**)deaths, for healthcare workers (in red) and all individuals (in black). The insets show their corresponding daily numbers. In (**a**), we note that healthcare workers that came into contact with SARS-CoV-2 during the studied period did not reduce their bus rides as much as other passengers. In addition, the normalized moving averages of bus validations of healthcare workers and of all individuals are getting closer to each other again as the economic reopening progresses. In (**b**) and (**c**), we find that both the normalized moving averages of cases and of deaths, respectively, for healthcare workers increased before those of all individuals until the lockdown regime. The windows of moving averages have 5 days of width for all curves. We normalized each moving average by its maximum. The vertical dotted lines represent the beginning of social isolation (State Decree 33,519), lockdown (State Decree 33,574), and economic reopening (State Decree 33,608) regimes imposed on March 20, May 8, and June 1, 2020, respectively. We also highlight, in light red, the lockdown period in the city of Fortaleza.

infectious healthcare workers of the entire city are shown in Fig. 3c,d, respectively. While the first and second waves of the epidemic can be clearly identified in both curves shown in Fig. 3a (infectious passengers) and Fig. 3c (infectious individuals), only highly attenuated peaks during the second wave period can be visualized in the corresponding curves for healthcare workers, as shown in Fig. 3b,d. We conjecture that the explanation for this behavior may be twofold. First, due to the high contagion of healthcare workers during the first wave, this group of people may have achieved a large percentage of immunity, as compared to the rest of the population. Second, efficient Personal Protective Equipment (PPE) became more available in hospitals after the first wave. The results in Fig. 3e show that the percentage of infectious passengers with respect to all infectious individuals in Fortaleza was higher than 1% during most of the epidemic period. Finally, the evolution in time of the fraction between infectious passengers and infectious individuals in the city who are both healthcare workers is shown in Fig. 3f.

The histogram of the values of τ_{ij} for the network of PICs is shown in Fig. 4a. The obtained distribution is characterized by the average $\langle \tau_{ij} \rangle^{PICs} \approx 28$ min. We find that CCs, defined by $\tau_{ij} > \tau_c = 15$ min, represent about 62% of the PICs, as shown by the Complementary Cumulative Distribution Function (CCDF) in the inset of Fig. 4a. For the network of CCs, the average of the shared ride times is $\langle \tau_{ij} \rangle^{CCs} \approx 39$ min. Figure 4b shows the network of CCs taking into consideration the periods of social isolation, lockdown, and economic reopening. As depicted, it is composed of several trees, where the vertices represent bus passengers that were diagnosed with

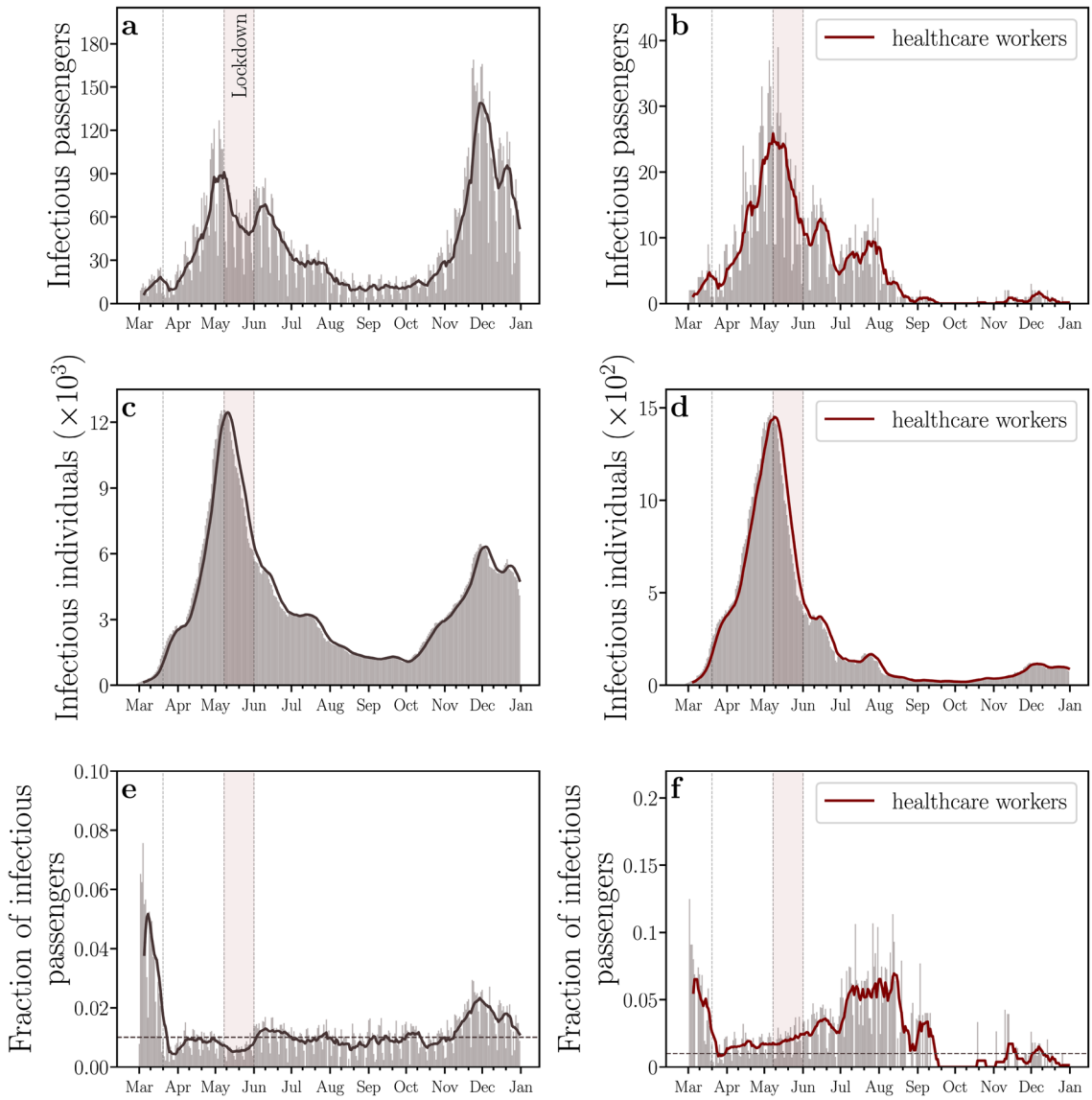

Figure 3. Representativeness of the dataset of COVID-19 confirmed cases on buses. The daily numbers of (**a**) all infectious passengers, (**b**) infectious passengers who are healthcare workers, (**c**) all infectious individuals in the entire city, and (**d**) infectious individuals who are healthcare workers in the entire city. For healthcare workers, we highlight an unexpected emergence of a single peak in the daily numbers of infectious individuals within buses and in the entire city, which contrasts to the first and the second waves of COVID-19. We conjecture that the explanation for this behavior may be the lack of Personal Protective Equipment (PPE) in hospitals during the first wave or the herd immunity of healthcare workers during the second wave. (**e**) The fraction of all infectious passengers. (**f**) The fraction of infectious passengers who are healthcare workers. These results show that the percentage of infectious passengers with respect to all infectious individuals in Fortaleza was higher than 1% during most of the epidemic period (dashed gray line). All solid lines represent moving averages with windows of 7 days.

COVID-19 and the edges correspond to CCs. Bus passengers identified as healthcare workers in the network are highlighted in red. The size of the vertices is proportional to their outdegrees.

At this point, we show that it is possible to perform a direct comparison between the computed values of Re_r^{bus} obtained from the contact tracing model for different time windows and the corresponding effective reproduction numbers Re^{city} estimated from the compartmental model. First, it is reasonable to assume that $Re_r^{bus} = Re_r^{city}$, as long as the population traveling by public buses can be considered as statistically equivalent, from an epidemiologic point of view, to the rest of the city. As a consequence of this assumption and using Eq. (9), we can write that

$$Re_r^{bus} = \psi Re^{city}, \tag{10}$$

where the parameter $\psi = [\alpha + (1-\alpha)\mu]^{-1}$ depends on the time window used for model inference with the IEnKF technique. We now proceed with the comparison between contact tracing and compartmental models. In practical terms, this is achieved by upscaling $\langle d_{out}^{CCs} \rangle$ to the numerical values obtained for Re^{city} during the

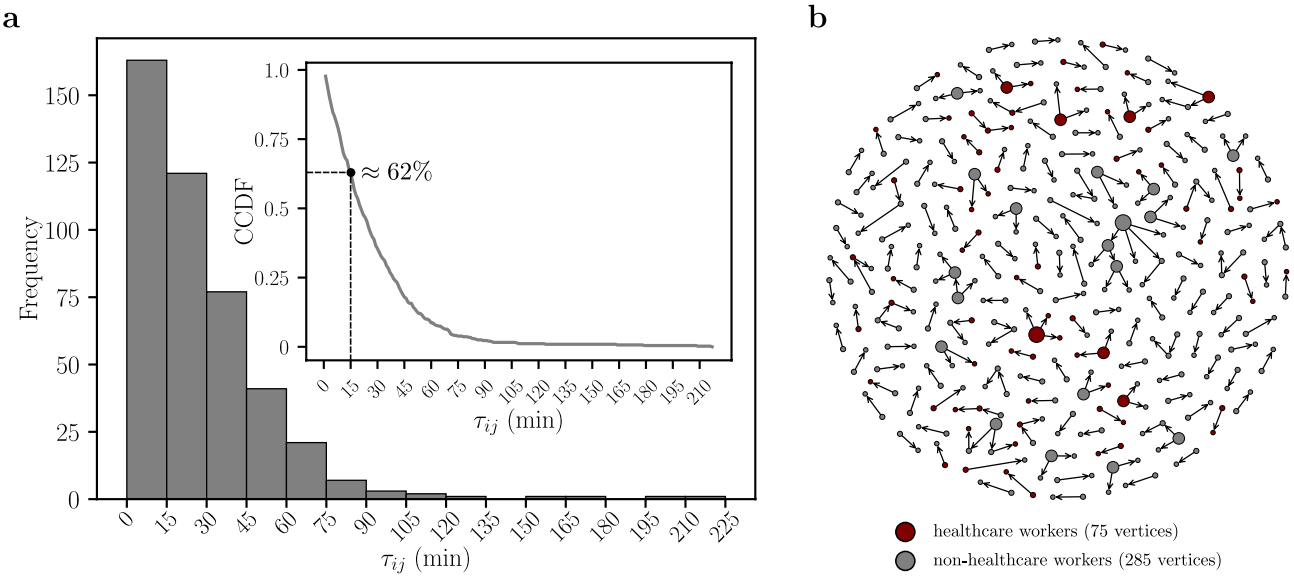

Figure 4. Shared Ride Time Histogram τ_{ij} and Network of Close Contacts (CCs). (**a**) We show the weight distribution τ_{ij} of the network of Potentially Infectious Contacts (PICs) in 15-minute-width bins. The average of the shared ride times of PICs is $\langle \tau_{ij} \rangle^{PICs} \approx 28$ min. Applying the threshold $\tau_c = 15$ min in the network of PICs, we define a network of Close Contacts (CCs). For the network of CCs, the average of the shared ride times is $\langle \tau_{ij} \rangle^{CCs} \approx 39$ min. The inset shows the Complementary Cumulative Distribution Function (CCDF) of τ_{ij}. We find that the percentage of the edges with τ_{ij} greater than $\tau_c = 15$ min is $\approx 62\%$ (dashed line), i.e., most part of PICs are CCs. (**b**) The vertices represent the bus passengers that were diagnosed with COVID-19 and the edges corresponds to the CCs each passenger had using the public transportation system in Fortaleza. The healthcare and non-healthcare workers are represented by red and gray vertices, respectively. The size of the vertices is proportional to their outdegrees.

early period of the SARS-CoV-2 epidemic, before the restrictions of isolation and social distancing imposed by the State Government took effect. Considering Eqs. (1) and (10), we use the relation $\chi \langle d_{out}^{CCs} \rangle = \psi Re^{city}$ and the numerical values of $\langle d_{out}^{CCs} \rangle$, ψ and Re^{city} on the day that corresponds to the maximum of Re^{city} during the first wave (April 8, 2020) to calculate $\chi \approx 37$. This constant combined with the values of ψ from the inference with the compartmental model, and the values of $\langle d_{out}^{CCs} \rangle$ from the contact tracing, both calculated for all time windows, are then used to obtain the entire curve of $Re^{bus} = (\chi/\psi)\langle d_{out}^{CCs} \rangle$ (see the time evolution of χ/ψ in Fig. S3 of the Supplementary Information). The value of χ can be understood as the product of two factors, $\chi = \chi_{rr}\chi_{ru}$. Assuming the equality between the proportions of CCs in the pairs of infectious and exposed passengers with existing and missing values of τ_{ij}, we estimate $\chi_{rr} \approx 1/0.586 \approx 1.70$ as a balance factor for possible missing CCs. The value of the remaining factor $\chi_{ru} \approx 21.76$ expresses the sub-notification of the confirmed cases as well as our lack of knowledge on the transmission from reported to unreported infectious passengers, for which the factor $(1-\alpha)/\alpha$ could be a lower bound (see Fig. S2 of the Supplementary Information for a sensitivity analysis of the model with parameter α), approximately between 5 and 9[16]. In an entirely similar fashion, by considering only reported infectious passengers that can be identified as healthcare workers, we can estimate their particular effective reproduction number as $Re^{health} = (\chi/\psi)\langle d_{out}^{CCs} \rangle^{health}$ with the same upscaling factor χ/ψ used for all infectious passengers and $\langle d_{out}^{CCs} \rangle^{health}$ is the average of the vertices outdegrees for healthcare workers.

In Fig. 5, we show the comparison between the estimates of Re^{bus} and Re^{city} from March to November 2020. Although the contact tracing and compartmental models are defined on different scales, the former on a microscopic scale and the latter on a macroscopic scale, the two curves capture the same decreasing trend associated to both social isolation and lockdown periods. We note that Re^{bus} consistently follows Re^{city} during the local COVID-19 outbreak, except for a three-month period between the first and the second waves of daily cases. In this period, the Re^{bus} decayed to undetectable standards despite the fact that the number of daily bus validations has increased (see Fig. 2a), i.e., our contact tracing model did not find any CC due to the low number of cases after the first wave. As also shown in Fig. 5, Re^{health} was systematically higher than Re^{bus}, which unveils that the healthcare workers played an important role in the transmission within buses during the first wave of COVID-19 in Fortaleza. Furthermore, Re^{health} remained undetectable even in the beginning of the second wave, in contrast to Re^{bus} and notwithstanding the increase of the number of daily bus validations of healthcare workers, as shown in Fig. 2a. As shown in the inset of Fig. 5, the maximum ratio Re^{health}/Re^{bus} occurred soon after the lockdown period, since the hospitals were still overloaded due to the peak of cases at the beginning of May and the new daily infections were low in the beginning of the reopening period. We emphasize that the complement of Re^{health}, due to non-healthcare workers, behaves similar to Re^{bus}. We also emphasize that our results do not suffer any influence from the Brazilian vaccination program, since the first shots were applied in Brazil at the end of January 2021, i.e., after the period studied here.

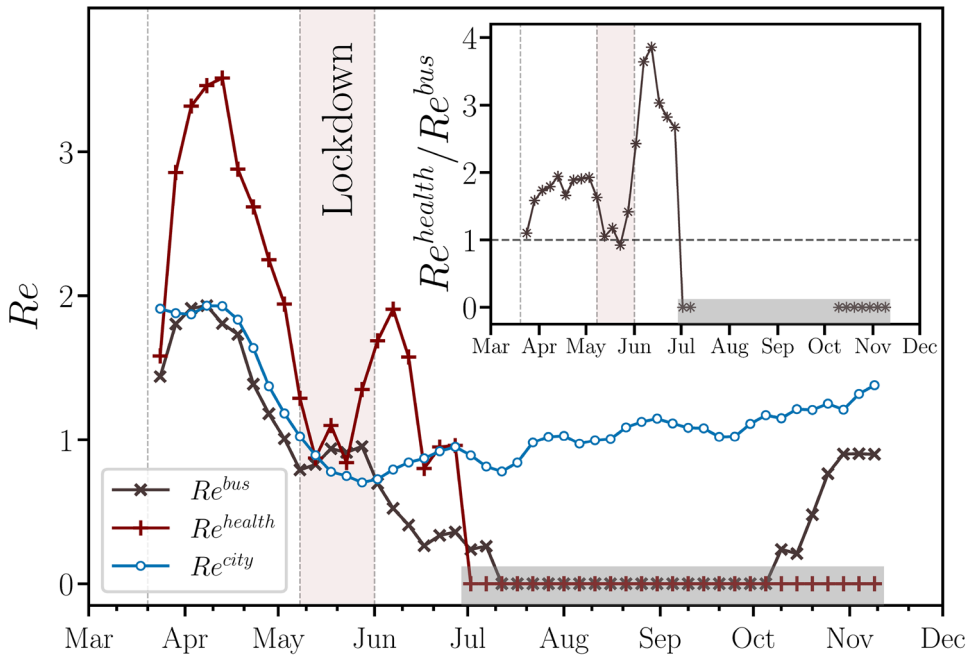

Figure 5. Time evolution of the effective reproduction numbers. Moving averages of the effective reproduction number for the entire city, Re^{city} (blue ○), for buses, Re^{bus} (gray ×), and for healthcare workers in the buses, Re^{health} (red +). We find that Re^{bus} consistently follows Re^{city} during the local COVID-19 outbreak, except for a three-month period between the first and the second waves of daily cases. We also show that Re^{health} was systematically higher than Re^{bus}, which unveils that the healthcare workers played an important role in the transmission within buses during the first wave of COVID-19 in Fortaleza. The inset shows that the maximum ratio Re^{health}/Re^{bus} occurred soon after the lockdown period. The windows of moving averages have 22 days of width with step size of 5 days for all curves. In the period indicated by the shaded regions in the main plot and its inset, both Re^{bus} and Re^{health} decayed to undetectable standards, i.e., no CCs could be identified under the framework of our contact tracing approach. The vertical dotted lines represent the beginning of social isolation (State Decree 33,519), lockdown (State Decree 33,574), and economic reopening (State Decree 33,608) regimes imposed on March 20, May 8, and June 1, 2020, respectively. We also highlight, in light red, the lockdown period in the city of Fortaleza.

Conclusions

In summary, two epidemiological models have been used in this work to understand the transmission on public transportation during the COVID-19 outbreak in Fortaleza, Ceará, Brazil. Whilst the compartmental model accounts for the transmission in the entire city (macroscopic scale), the contact tracing model has been used to estimate the transmission inside city buses (microscopic scale) through the concept of CCs. Both models were fed with real data of bus validations and of COVID-19 confirmed cases and deaths. Our results show that Re^{bus} consistently follows Re^{city} during the local COVID-19 outbreak, except for a three-month period between the first and the second waves of daily cases. We conjecture that similar behaviors would be obtained if the indoor environment was another place with some degree of homogeneous mixing, e.g., offices, gyms, or schools. Furthermore, the transmission from healthcare workers within buses until the end of July is characterized by a value of Re^{health} persistently greater than Re^{bus}. In other words, we have found that healthcare workers had transmitted more the disease than usual passengers on the buses of the city of Fortaleza during the first wave of COVID-19 cases. Healthcare workers, even the non-frontline professionals, are more likely to get and, consequently, spread the pathogen because their social network distances to individuals that tested positive for COVID-19 are very short compared to non-highly exposed workers. Despite being more tested, healthcare workers may not even know that they are infectious when they board a bus due to eventual time delays of the result of a COVID-19 test. Other groups of highly exposed people may affect the dynamics of dissemination of the virus in a similar way, e.g., education workers and police officers. We emphasize that finding a CC (even considering reported cases of COVID-19 and their epidemiological profiles) on a bus is not sufficient to ensure that the transmission indeed happened within it. However, it is the best measure that we can perform in order to infer the transmission rate. Another limitation that can be pointed out in our study is the fact that no mechanism of reinfection is considered in the compartmental model. Therefore, our results reinforce the worldwide claim that it is imperative to propose special policies to support displacement (or to avoid it) of highly exposed groups of people. Finally, we suggest that the intensity and the necessity of using public transportation by highly exposed groups must be seriously considered as a criterion to prioritize their vaccination.

Methods

Datasets. *Bus validations.* Most part of bus passengers in Fortaleza (≈ 94%) pay their bus fares with a smart card. Every time a passenger passes their card on a ticket gate of a bus, a validation record is created. The Fortaleza City Hall compiled and made available an anonymized dataset of bus validations with the following information: a citizen's ID (a hash code), a vehicle ID (another hash code), the date and time of the validation record and the estimated ride time. The dataset ranges from March to December 2020, totaling 107,488,528 validation registers that refers to 1,426,569 different passengers.

COVID-19 confirmed cases and deaths. The dataset of COVID-19 confirmed cases and deaths is an anonymized list of all individuals diagnosed with the disease in Fortaleza from March to December 2020. These data were also processed and made available by the Fortaleza City Hall. Such dataset is organized in columns as follows: a citizen's ID (the same hash code used in the previous dataset), the date of OS, a confirmed death flag, the date of death and a healthcare worker flag. In the period of time ranged by the data, there are 85, 553 confirmed cases (5960 of healthcare workers) and 3075 confirmed deaths (227 of healthcare workers). These numbers are slightly different from those found in oficial records[34] because they only considered cases with dates of OS filled in. We emphasize that these healthcare workers are not only the frontline professionals but also people whose jobs are related to the health field. Finally, we found that 9032 people (721 healthcare workers) were diagnosed with COVID-19 and used their smart card on buses at least once from March to December 2020.

Iterated Ensemble Kalman Filter. We use the Iterated Ensemble Kalman Filter (IEnKF) framework[16,41–43] to infer the compartmental model parameters and initial subpopulations. The algorithm is based on comparing predictions of the model $f(.)$ obtained by the numerical integration of Eqs. (3)–(7) of the main text with a set of T observations $\mathcal{O}_1, \ldots, \mathcal{O}_T$ taken at discrete times $t_1, \ldots, t_T$ within an observation window (see Fig. S4 of the Supplementary Information). The inference framework starts from an initial state vector $X^{(0)} = \{S, E, I_r, I_u, R, D_r\}^{(0)}$ (see Table 1), as well as an initial parameter vector $\theta^{(0)} = \{\beta, \mu, \sigma, \gamma, \alpha, \phi\}^{(0)}$ (see Table 1). To these vectors, uncertainties are attributed in terms of the variance matrices σ_X and σ_θ, respectively. For each iteration m, an ensemble of P "particles" is generated such that each particle has the initial state at time t_0 drawn from a multivariate normal distribution with mean $X^{(m-1)}$ and variance $a^{(m-1)}\sigma_X$, where $0 < a < 1$ is a "cooling factor". The initial state vector for particle i is denoted by $X(t_0, i) = \mathcal{N}(X^{(m-1)}, a^{(m-1)}\sigma_X)$. These states are also used to set $X_F(t_0, i)$, which define the posterior distribution at time t_0. Analogously, each particle i has an initial parameter vector $\theta(t_0, i) = \mathcal{N}(\theta^{(m-1)}, b^{(m-1)}\sigma_\theta)$, where $0 < b < 1$ is another cooling factor. The inference proceeds by numerically integrating the model from these initial conditions, such that the predicted vector state for each particle i at time t_n is obtained from the prior distribution, $X_P(t_n, i) = f(X_F(t_{n-1}, i), \theta(t_{n-1}, i))$. Based on these predictions, a weight $W(t_n, i)$ is assigned to each particle i, such that

$$W(t_n, i) = \exp\left(-\frac{|\mathcal{O}(t_n, i) - \mathcal{O}_n|}{\Theta}\right), \tag{11}$$

where $\mathcal{O}(t_n, i)$ is the predicted value for the observed quantities at time t_n for particle i, and Θ is a "temperature". In our case, $\mathcal{O}(t_n, i)$ is the prediction for the cumulative number of daily reported deaths $D_r(t_n, i)$. The filtering process is accomplished by keeping the particles with the largest weights with probability $\mathcal{P} = W(t_n, i)/\sum_j W(t_n, j)$. The states of the filtered particles will set the posterior distribution at time t_n, $X_F(t_n, i) = X_P(t_n, i_{best})$, where i_{best} is the set of the indexes of the filtered particles[42]. The parameter vector is updated at time t_n using $\theta(t_n, i) = \mathcal{N}(\theta(t_{n-1}, i_{best}), b^{(m-1)}\sigma_\theta)$. This filtering process continues until all the observations $\mathcal{O}_1, \ldots, \mathcal{O}_N$ are compared. The iterative process continues by setting the initial state vector $X^{(m)}$ and parameter vector $\theta^{(m)}$ for the next iteration with the same observation window. The next parameter vector is given by[41]:

$$\theta^{(m)} = \theta^{(m-1)} + V(t_1)\sum_{n=1}^{N} V^{-1}(t_n)\left(\bar{\theta}(t_n) - \bar{\theta}(t_{n-1})\right), \tag{12}$$

where $\bar{\theta}(t_n)$ is the sample mean of $\theta(t_n, i_{best})$ and $V(t_n)$ is the variance[41,42]. The next state vector is given by the sample mean,

$$X^{(m)} = \frac{1}{P}\sum_{j_{best}=1}^{P} X(t_0, j_{best}). \tag{13}$$

After each iteration m, the initial state vector $X^{(m)}$ and parameter vector $\theta^{(m)}$ are used to compute the evolution of the model for the whole observation window $1, \ldots, t_T$. The performance of the inferred model is computed by evaluating the error

$$\varepsilon^{(m)} = \frac{1}{T}\sum_{n=1}^{T}\left|\mathcal{O}_n^{(m)} - \mathcal{O}_n\right|^2. \tag{14}$$

The iteration continues until $\left|\varepsilon^{(m)} - \varepsilon^{(m-1)}\right| < \varepsilon_{max}$, where the threshold used here is $\varepsilon_{max} = 0.01$. The goodness of the fit is also checked by computing the Pearson coefficient, R^2, between the integrated model cumulative number of deaths and the corresponding observations. The value of $R^2 > 0.96$ for all observationwindows. The Table S1 of the Supplementary Information lists the inferred parameters for all windows. Figure S1 of the Supplementary Information shows the epidemiological curves obtained from the inference with the SEIIR

compartmental model. As depicted, good agreement can be observed between model and observations for the evolution of the cumulative deaths.

Ethics declarations. This study was approved by the Institutional Review Board (IRB) at Universidade de Fortaleza (UNIFOR). All methods were conducted in accordance with relevant guidelines and regulations. Two datasets were used with the approval and consent obtained by the Fortaleza City Hall, Ceará, Brazil. The first is a list of COVID-19 confirmed cases and deaths of patients in Fortaleza and the second consists of bus validations records from smart cards of passengers, both collected during the period from March to December, 2020.

Data availability

The data that support the findings of this study are available in Zenodo with the identifier http://doi.org/10.5281/zenodo.4763399.

Received: 4 June 2021; Accepted: 6 December 2021
Published online: 27 December 2021

References

1. The Lancet Respiratory Medicine. COVID-19 transmission-up in the air. *Lancet Respir. Med.*https://doi.org/10.1016/S2213-2600(20)30514-2 *(2020).*
2. Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E. & Getz, W. M. Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, 355–359. https://doi.org/10.1038/nature04153 (2005).
3. Lossio-Ventura, J. A. *et al.* DYVIC: DYnamic VIrus Control in Peru, in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 2264–2267 (2020). https://doi.org/10.1103/PhysRevResearch.3.013163
4. Serafino, M. *et al.* Superspreading k-cores at the center of COVID-19 pandemic persistence. *arXiv preprint*arXiv:2103.08685 *(2021). https://doi.org/10.1101/2020.08.12.20173476*
5. Reyna-Lara, A. *et al.* Virus spread versus contact tracing: Two competing contagion processes. *Phys. Rev. Res.*. https://doi.org/10.1103/PhysRevResearch.3.013163 (2021).
6. Hamner, L. *et al.* High SARS-CoV-2 attack rate following exposure at a choir practice—Skagit County, Washington, March 2020. *MMWR Morb. Mortal Wkly. Rep.* **69**, 606–610. https://doi.org/10.15585/mmwr.mm6919e6 (2020).
7. Majra, D., Benson, J., Pitts, J. & Stebbing, J. SARS-CoV-2 (COVID-19) superspreader events. *J. Infect.* **82**, 36–40. https://doi.org/10.1016/j.jinf.2020.11.021 (2021).
8. Liu, Y. *et al.* Aerodynamic analysis of SARS-CoV-2 in two Wuhan hospitals. *Nature* **582**, 557–560. https://doi.org/10.1038/s41586-020-2271-3 (2020).
9. Lednicky, J. A. *et al.* Viable SARS-CoV-2 in the air of a hospital room with COVID-19 patients. *Int. J. Infect. Dis.* **100**, 476–482. https://doi.org/10.1016/j.ijid.2020.09.025 (2020).
10. Kwon, K. *et al.* Evidence of long-distance droplet transmission of SARS-CoV-2 by direct air flow in a restaurant in Korea. *J. Korean Med. Sci.* https://doi.org/10.3346/jkms.2020.35.e415 (2020).
11. Böhmer, M. M. *et al.* Investigation of a COVID-19 outbreak in Germany resulting from a single travel-associated primary case: A case series. *Lancet Infect. Dis* **20**, 920–928. https://doi.org/10.1016/S1473-3099(20)30314-5 (2020).
12. Kakimoto, K. *et al.* Initial investigation of transmission of COVID-19 among crew members during quarantine of a cruise ship–Yokohama, Japan, February 2020. *MMWR Morb. Mortal Wkly. Rep.* **69**, 312–313. https://doi.org/10.15585/mmwr.mm6911e2 (2020).
13. Mizumoto, K. & Chowell, G. Transmission potential of the novel coronavirus (COVID-19) onboard the diamond Princess Cruises Ship, 2020. *Infect. Dis. Modell.* **5**, 264–270. https://doi.org/10.1016/j.idm.2020.02.003 (2020).
14. Keeling, M. J. & Rohani, P. *Modeling Infectious Diseases in Humans and Animals* (Princeton University Press, 2011).
15. Gómez-Gardeñes, J., Soriano-Panos, D. & Arenas, A. Critical regimes driven by recurrent mobility patterns of reaction–diffusion processes in networks. *Nat. Phys.* **14**, 391–395. https://doi.org/10.1038/s41567-017-0022-7 (2018).
16. Li, R. *et al.* Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science* **368**, 489–493. https://doi.org/10.1126/science.abb3221 (2020).
17. Arenas, A. *et al.* Modeling the spatiotemporal epidemic spreading of COVID-19 and the impact of mobility and social distancing interventions. *Phys. Rev. X* **10**, 041055. https://doi.org/10.1103/PhysRevX.10.041055 (2020).
18. Tian, T. *et al.* Evaluate the risk of resumption of business for the states of New York, New Jersey and Connecticut via a pre-symptomatic and asymptomatic transmission model of COVID-19. *J. Data Sci.* **19**, 178–196. https://doi.org/10.6339/21-JDS994 (2021).
19. Tian, T. *et al.* The effects of stringent and mild interventions for coronavirus pandemic. *J. Am. Stat. Assoc.* **116**, 481–491. https://doi.org/10.1080/01621459.2021.1897015 (2021).
20. Vuchic, V. R. *Urban Transit: Operations, Planning, and Economics* (Wiley, 2017).
21. Stoddard, S. T. *et al.* The role of human movement in the transmission of vector-borne pathogens. *PLoS Negl. Trop. Dis.* **3**, e481. https://doi.org/10.1371/journal.pntd.0000481 (2009).
22. Edelson, P. J. & Phypers, M. TB transmission on public transportation: A review of published studies and recommendations for contact tracing. *Travel Med. Infect. Dis.* **9**, 27–37. https://doi.org/10.1016/j.tmaid.2010.11.001 (2011).
23. Bomfim, R. *et al.* Predicting dengue outbreaks at neighbourhood level using human mobility in urban areas. *J. R. Soc. Interface* **17**, 20200691. https://doi.org/10.1098/rsif.2020.0691 (2020).
24. Kraemer, M. U. G. *et al.* The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* **368**, 493–497. https://doi.org/10.1126/science.abb4218 (2020).
25. Schlosser, F. *et al.* COVID-19 lockdown induces disease-mitigating structural changes in mobility networks. *Proc. Natl. Acad. Sci.* **117**, 32883–32890. https://doi.org/10.1073/pnas.2012326117 (2020).
26. Melo, H. P. *et al.* Heterogeneous impact of a lockdown on inter-municipality mobility. *Phys. Rev. Res.* https://doi.org/10.1103/PhysRevResearch.3.013032 (2021).
27. Di Carlo, P. *et al.* Air and surface measurements of SARS-CoV-2 inside a bus during normal operation. *PLoS ONE* **15**, e0235943. https://doi.org/10.1371/journal.pone.0235943 (2020).
28. Goscé, L. & Johansson, A. Analysing the link between public transport use and airborne transmission: Mobility and contagion in the London underground. *Environ. Health* **17**, 1–11. https://doi.org/10.1186/s12940-018-0427-5 (2018).
29. Jenelius, E. & Cebecauer, M. Impacts of COVID-19 on public transport ridership in Sweden: Analysis of ticket validations, sales and passenger counts. *Transport. Res. Interdiscip. Perspect.* https://doi.org/10.1016/j.trip.2020.100242 (2020).
30. Shen, Y. *et al.* Community outbreak investigation of SARS-CoV-2 transmission among bus riders in eastern China. *JAMA Intern. Med.* **180**, 1665–1671. https://doi.org/10.1001/jamainternmed.2020.5225 (2020).

31. Shen, J. *et al.* Prevention and control of COVID-19 in public transportation: Experience from China. *Environ. Pollut.* https://doi.org/10.1016/j.envpol.2020.115291 (2020).
32. Zhang, Z. *et al.* Disease transmission through expiratory aerosols on an urban bus. *Phys. Fluids* **33**, 015116. https://doi.org/10.1063/5.0037452 (2021).
33. Hu, M. *et al.* Risk of coronavirus disease 2019 transmission in train passengers: An epidemiological and modeling study. *Clin. Infect. Dis.* **72**, 604–610. https://doi.org/10.1093/cid/ciaa1057 (2021).
34. Integrasus, Secretaria de Saúde do Estado do Ceará. https://integrasus.saude.ce.gov.br Accessed on October 18, 2021.
35. He, X. *et al.* Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat. Med.* **26**, 672–675. https://doi.org/10.1038/s41591-020-0869-5 (2020).
36. Li, Q. *et al.* Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *N. Engl. J. Med.* **382**, 1199–1207. https://doi.org/10.1056/NEJMoa2001316 (2020).
37. Sanche, S. *et al.* High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2. *Emerg. Infect. Dis.* **26**, 1470–1477. https://doi.org/10.3201/eid2607.200282 (2020).
38. Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/php/contact-tracing/contact-tracing-plan/contact-tracing.html Accessed on May 14, 2021.
39. Byambasuren, O. *et al.* Estimating the extent of asymptomatic COVID-19 and its potential for community transmission: Systematic review and meta-analysis. *J. Assoc. Med. Microbiol. Infect. Dis. Canada* **5**, 223–234. https://doi.org/10.3138/jammi-2020-0030 (2020).
40. Le Marshall, J., Rea, A., Leslie, L., Seecamp, R. & Dunn, M. Error characterisation of atmospheric motion vectors. *Aust. Meteorol. Mag.* **53**, 123–131 (2004).
41. Ionides, E. L., Bretó, C. & King, A. A. Inference for nonlinear dynamical systems. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 18438–18443. https://doi.org/10.1073/pnas.0603181103 (2006).
42. King, A. A., Ionides, E. L., Pascual, M. & Bouma, M. J. Inapparent infections and cholera dynamics. *Nature* **454**, 877–880. https://doi.org/10.1038/nature07084 (2008).
43. Sakov, P., Oliver, D. S. & Bertino, L. An iterative EnKF for strongly nonlinear systems. *Mon. Weather Rev.* **140**, 1988–2004. https://doi.org/10.1175/MWR-D-11-00176.1 (2012).

Acknowledgements

We gratefully acknowledge CNPq, CAPES, FUNCAP, the National Institute of Science and Technology for Complex Systems in Brazil and the Edson Queiroz Foundation for financial support.

Author contributions

C.P., H.A.C., E.A.O., C.C., A.S.L., J.S.A. and V.F. designed research; C.P., H.A.C., E.A.O., C.C., A.S.L., J.S.A. and V.F. performed research; C.P., H.A.C., E.A.O., C.C., A.S.L., J.S.A. and V.F. analyzed data; and C.P., H.A.C., E.A.O., C.C., A.S.L., J.S.A. and V.F. wrote the paper. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-03998-y.

Correspondence and requests for materials should be addressed to E.A.O.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.