Crowding and the shape of COVID-19 epidemics

Benjamin Rader[1,2,*], Samuel V. Scarpino[3,4,5,*,$], Anjalika Nande[6], Alison L. Hill[6,7], Ben Adlam[6], Robert C. Reiner[8,9], David M. Pigott[8,9], Bernardo Gutierrez[10,11], Alexander Zarebski[10], Munik Shrestha[3], John S. Brownstein[1,12], Marcia C. Castro[13], Christopher Dye[10], Huaiyu Tian[14], Oliver G. Pybus[9,15,$], Moritz U.G. Kraemer[10,$]

1. Computational Epidemiology Lab, Boston Children's Hospital, Boston, United States
2. Department of Epidemiology, Boston University School of Public Health, Boston, United States
3. Network Science Institute, Northeastern University, Boston, United States
4. ISI Foundation, Turin, Italy
5. Santa Fe Institute, Santa Fe, United States
6. Program for Evolutionary Dynamics, Harvard University, Cambridge, United States
7. Institute for Computational Medicine, Johns Hopkins University, Baltimore, United States
8. Department of Health Metrics, University of Washington, Seattle, United States
9. Institute for Health Metrics and Evaluation, University of Washington, Seattle, United States
10. Department of Zoology, University of Oxford, Oxford, United Kingdom
11. School of Biological and Environmental Sciences, Universidad San Francisco de Quito USFQ, Quito, Ecuador
12. Harvard Medical School, Boston, United States
13. Department of Global Health and Population, Harvard T.H. Chan School of Public Health, Boston, United States
14. State Key Laboratory of Remote Sensing Science, College of Global Change and Earth System Science, Beijing Normal University, Beijing, China
15. Department of Pathobiology and Population Science, The Royal Veterinary College, London, United Kingdom

*contributed equally as first authors

[$]correspondence should be addressed to moritz.kraemer@zoo.ox.ac.uk, oliver.pybus@zoo.ox.ac.uk and s.scarpino@northeastern.edu

[#]Members of the Open COVID-19 Data Working Group are listed at the end of the manuscript

Abstract

The COVID-19 pandemic is straining public health systems worldwide and major non-pharmaceutical interventions have been implemented to slow its spread[1–4]. During the initial phase of the outbreak, dissemination of SARS-CoV-2 was primarily determined by human mobility from Wuhan[5,6]. Yet empirical evidence on the effect of key geographic factors on local epidemic transmission is lacking[7]. We analyse highly-resolved spatial variables in cities together with case count data in order to investigate the role of climate, urbanization, and variation in interventions. We show that the degree to which cases of COVID-19 are compressed into a short period of time (peakedness of the epidemic) is strongly shaped by population aggregation and heterogeneity, such that epidemics in crowded cities are more spread over time, and crowded cities have larger total attack rates than less populated cities. Observed differences in the peakedness of epidemics are consistent with a metapopulation model of COVID-19 that explicitly accounts for spatial hierarchies. We pair our estimates with globally-comprehensive data on human mobility and predict that crowded cities worldwide could experience more prolonged epidemics.

Main

Predicting the epidemiology of the COVID-19 pandemic is a priority for guiding epidemic responses around the world. China has undergone its first epidemic wave and, remarkably, cities across the country are now reporting few or no locally-acquired cases[8]. Analyses have indicated that that the spread of COVID-19 from Hubei to the rest of China was driven primarily by human mobility from Wuhan[6,9], and that the stringent measures to restrict human movement and public gatherings within and among cities in China were associated with bringing local epidemics under control[5]. Key uncertainties remain as to which geographic factors drive the local transmission dynamics of COVID-19 and initial analysis suggests a limited role of climate in determining epidemic growth[10].

Spatial heterogeneity in infectious disease transmission can be influenced by local differences in population or human movements, such that high local population densities might catalyse the spread of novel pathogens due to higher contact rates with susceptible individuals[11,12]. For respiratory pathogens, the temporal clustering of cases in an epidemic (*i.e.,* the shortest period during which the majority of cases are observed) varies with increased indoor crowding and socio-economic and climatic factors[13–18]. The temporal concentration of cases is minimized when incidence is spread evenly across time and increases as incidence becomes more concentrated in particular days, as has been observed for influenza[13]. In any given location, a higher temporal concentration of cases may require a larger surge capacity in the public health system[19], especially for an emerging respiratory pathogen such as COVID-19[20].

Results

Spatial population structure predicts the shape of epidemics of COVID-19

China and Italy provide detailed epidemiological time series for COVID-19 [2,21,22] across a wide range of geographic contexts, hence the outbreaks in these countries provide an opportunity to evaluate the role of local factors in shaping epidemic behaviour. We use daily epidemiological data from Chinese cities[23,24] and Italian provinces, climate and population data, and the response to local interventions as measured by human mobility data from Baidu Inc[25] and COVID-19 Aggregated Mobility Research Dataset (https://www.google.com/covid19/mobility/), to identify drivers of transmission, with a focus on how the temporal clustering of cases differs among prefectures in China and provinces in Italy. A summary of the main findings, limitations and policy implications of our study is shown in Table 1.

We used daily incidence data of confirmed COVID-19 cases aggregated at the prefectural level (n = 293) in China (**Figure 1a**) and provinces in Italy (n = 108). Prefectures and provinces are administrative units that typically have one urban center (**Figure 1b**). We aggregate daily individual-level data collected from official government reports[22]. Epidemiological data in each prefecture were truncated to exclude dates before the first and after the last day of reported cases during the first epidemic. Cases reported after March 1, 2020 that were imported from outside China were excluded from the analysis. All epidemiological data from Hubei province were excluded because of the lack of prefecture-level epidemiological data and issues with consistent reporting prior to January 20th, 2020. The shape of epidemic curves varied between prefectures, with some showing a rapid rise and decline in reported cases and others showing more prolonged epidemics (**Figure 1a, Extended Data Figure 1**).

To characterize the temporal clustering of cases for each prefecture and province we calculated the Shannon diversity index of the distribution of incident cases[13]. We defined the incidence distribution p_{ij} for a given city to be the proportion of COVID-19 cases during the first epidemic wave j that occurred on day i. The Shannon index of incidence for a given prefecture and year is given by $v_j = \left(-\sum_i p_{ij} \log p_{ij}\right)^{-1}$. Because v_j is a function of the disease incidence curve in each location, rather than of absolute incidence values, it is less sensitive to varying reporting rates among cities. The Shannon index is maximal when all cases occur on the same day and minimal when each day of the epidemic has the same number of incident cases (e.g., 'flat' epidemic curves). It is highly correlated with alternative measures of epidemic peakedness, such as the proportion of cases that occur at the peak +/- one day

(**Extended Data Figure 2**). The total attack rate of reported COVID-19 cases in each prefecture is strongly negatively correlated with the Shannon index in China (**Figure 1c**), hence less peaked epidemics have a larger total attack rate (Pearson's r = -0.67, 95% CI: -0.73 – -0.59, p-value < 0.01; for Italy R^2 = 0.33, p-value < 0.01). We hypothesize that this variation among cities in total attack rate and the temporal clustering of cases is the result of the spatial organization of human populations.

To test this hypothesis we used Lloyd's index of mean crowding[13,26], treating the population count of each spatial grid cell as an individual unit (**Figure 1**). The term 'mean crowding' used here is a specific geographic metric that summarizes both population density and how density is distributed across a prefecture (*i.e.,* patchiness, **Figure 1**). Higher values of Lloyd's index suggest a spatially aggregated population structure. For example, Xi'an has high values of crowding whilst Bozhou has a comparable population density but a population that is more evenly distributed across the prefecture (**Figure 1b**). We performed log-linear regression modeling to determine the association between the temporal clustering of cases with socio-economic and environmental variables, including reductions in population flows during the outbreak period (for details, see **Methods**).

We found that the temporal clustering of cases is significantly negatively correlated with the mean number of contacts (p-value < 0.01) but positively correlated with mean population density (p-value < 0.01) and varies widely across China and Italy (**Figure 2**, **Supplementary Table 1**). This observation contrasts with the expectations of simple and classical epidemiological models, which predict higher peakedness in crowded areas due to the increased availability of susceptible individuals[27,28]. The spatial scale at which this relationship is best explained was 10x10km but results were statistically significant at all spatial scales between 1-50km^2 (**Extended Data Figure 3**, p-value < 0.01). Mean specific humidity and population mobility remained significantly negatively correlated with epidemic peakedness when included in a multivariate model with crowding (**Supplementary Table 1**, p-values < 0.01).

Using weekly human mobility data, we find that within-city human mobility during the outbreak is correlated with the temporal clustering of cases (*i.e.*, prefectures that have larger reductions in mobility also have lower epidemic peakedness, **Extended Data Fig. 4, Supplementary Table 1,** p-value < 0.01). When we combined mobility reduction in a model with crowding and humidity we found that these variables each remained significant predictors of the temporal clustering of cases (Extended Data Table 1, p-value < 0.01). These results suggest that although measures to reduce mobility can successfully lead to a flattening of the epidemic curve, population crowding is an independent contributor to the shape of epidemics in these two countries.

Our multivariate-model can explain a large fraction of the variation in epidemic peakedness among Chinese cities and Italian provinces and sensitivity analyses confirm the robustness of our results to potential noise in location-specific incidence distributions ($R^2 = 0.638$, **Extended Data Fig. 2, Supplementary Table 1, Extended Data Fig. 5**). To evaluate the out-of-sample performance of our model we (i) performed n-fold cross validation at the prefecture-level in China (Spearman's rho = 0.61, 95% bootstrap CI: 0.52 – 0.68, p-value < 0.01), (ii) used the fitted model in China to estimate peak intensity at the corresponding administrative level 2 locations, i.e., province-level, in Italy (Spearman's rho = 0.57, 95% bootstrap CI: 0.41 – 0.69, p-value < 0.01), and (iii) performed n-fold cross validation at the province-level in Italy (Spearman's rho = 0.65, 95% bootstrap CI: 0.52 – 0.76, p-value < 0.01). These results suggest that the model is robust to both within- and between-country out-of-sample testing (**Extended Data Figure 6**).

To evaluate the potential impact of the temporal clustering of cases on the peak attack rate and total attack rate we performed a simple linear regression (**Supplementary Table 2**). For locations that have a single peak, the attack rate at the peak is highest in two settings: i) in crowded locations with high population size (prefectures that also experience high total attack rates), ii) in locations that have lower population and lower crowding and therefore high temporal clustering of cases (**Extended Data Figure 7**). Other prefectures that have low population and low crowding sometimes experience very short outbreaks with small peak attack rate suggesting local stochastic extinction possibly due to limited mixing between populations. We hypothesize that the observation that high peak attack rates can sometimes be found in low crowding areas is related to rare superspreading events as observed in Bergamo, Italy or Mulhouse, France.

Simulation of COVID-19 epidemics in hierarchically structured populations

We hypothesize that the mechanism underlying our central observation (that more crowded cities experience less peaked outbreaks) is that crowding enables sustained transmission among households and through a city's population, leading incidence to be widely distributed through time. To explore this proposed mechanism, we simulated stochastic epidemic dynamics in two types of populations. Simple, well-mixed transmission models in which contact rates are high in crowded regions were not consistent with our findings, because they predict crowded regions would have more temporally-clustered outbreaks. To capture realistic contact patterns, we created hierarchically-structured populations[29] in which individuals had high rates of contact within their social units (which are defined broadly and could represent households, care homes, hospitals, prisons, etc.), lower rates with individuals from other units but within the same neighbourhoods, and relatively rare contact with other individuals in other neighbourhoods within the same prefecture (**Figure 3a**). These assumptions are consistent with reports that the majority of onward transmission after lockdowns were implemented, occurred in households or in other close contact situations[2,30]. In this scenario, less crowded prefectures often had more peaked and shorter outbreaks that were isolated to specific neighborhoods, while more crowded prefectures could sustain drawn-out outbreaks of larger final size, which jumped among the more highly-connected neighborhoods (**Figures 3b and c**). Further, if the reproduction number of COVID-19 is over-dispersed [31–33] then crowding could enable local outbreaks to spread more widely due to the availability of contacts[34].

We also simulated outbreak dynamics under extensive social distancing measures, as observed in Chinese prefectures (75% reduction in contact rates[35,36]). If social distancing reduces non-household contacts by the same relative amount in all locations, there will be more contacts remaining in crowded areas, since baseline contact rates are higher. Consequently, outbreaks in crowded regions could be larger and take longer to end after intervention (**Figure 3d, Figure 1c, Extended Data Fig. 1**).

Using the fitted model from China paired with globally comprehensive covariates we extrapolate our results to cities across the world (**Figure 4**). Human mobility data from Baidu Inc were not available for locations outside of China. Therefore, we use aggregated human mobility data from Google's COVID Mobility Research Dataset (Methods) to capture relative differences in human mobility through time. At the global scale, cities in yellow are predicted to have concentrated and peaked epidemics, and cities in blue are predicted to have more prolonged outbreaks (**Figure 4b**, a full list is provided in **the Supplementary Information**). In general, the epidemics in coastal cities were less peaked and were larger and more prolonged, which could be attributable to high levels of population crowding in coastal cities. These predictions rely on fitted relationships of the first epidemic curves from Chinese and Italian cities and therefore should be interpreted very cautiously when generalizing to other settings.

Discussion

Our findings confirm previous work on the peakedness of epidemics transmission for influenza in cities[13]. Our work provides empirical support for the role of spatial organization in determining infectious disease dynamics[29,37] and, specifically, spatial variability in transmission parameters[38]. Furthermore, with lower total incidence in small cities compared with larger cities, the risk of resurgence could be elevated due to lower population immunity after the first wave of the epidemic. Higher seroprevalence for COVID-19 in urban areas[39] provides initial data to support these finding, however there remains an urgent need to expand serological data collection and provide a full picture of attack rates across cities worldwide[40]. Even though our model does not account for over-dispersion in COVID-19 transmission, there is a theoretical link between the reproduction number in heterogeneous environments and Lloyd's crowding index of aggregation[41], such that the reproduction number increases with higher aggregation[34]. We report that in dense cities reductions in mobility tend to be larger, which potentially elevates the effectiveness of non-pharmaceutical interventions in dense cities[42]. However, assessing the effect of within-city connectivity and its spatial heterogeneity on disease dynamics will be critical to further our understanding of how COVID-19 spreads in urban areas. We found that there is an association between climatic factors and the peakedness of epidemics but particular caution will need to be applied in interpreting these relationships outside the two studied countries (Italy, China). More work is needed to provide causal evidence for the effect of climatic factors on transmission dynamics of COVID-19 during the pandemic and post-pandemic phase[10].

Currently, non-pharmaceutical interventions are the primary control strategy for COVID-19. As a result, public health measures are often focused on 'flattening the curve' to lower the risk of essential services running out of capacity. We show that spatial context, especially crowding are important factors for assessing the shape of epidemic curves. Therefore, it will be critical to view non-pharmaceutical interventions through the perspective of crowding (*i.e.*, how does an intervention reduce the circle of contacts of an average individual) in cities across the world.

Acknowledgements: The authors thank Kathryn Cordiano for statistical assistance. We thank the Open COVID-19 Data Working Group members. BR acknowledges funding from Google.org. MUGK acknowledges funding from European Commission H2020 program (MOOD project) and a Branco Weiss Fellowship. OGP, MUGK and HT acknowledge funding from the Oxford Martin School. HT acknowledges funding from the Beijing Science and Technology Planning Project (Z201100005420010). ALH and AN acknowledge funding from the US National Institutes of Health (DP5OD019851). The funding bodies had no role in study design, data collection and analysis, preparation of the manuscript, or the decision to publish. All authors have seen and approved the manuscript.

Author contributions: MUGK, OGP, SVS conceived the research. BR, ALH, AN, BA, SVS, MUGK analysed the data. BR and SVS analysed human mobility data. CD, OGP, MUGK, SVS interpreted the data. MUGK wrote the first draft of the manuscript. All authors contributed to interpretation of results and manuscript writing.

Competing interests: The authors declare no competing interests.

References

1. Fraher, E. P. *et al.* Ensuring and Sustaining a Pandemic Workforce. *N. Engl. J. Med.* NEJMp2006376 (2020). doi:10.1056/NEJMp2006376
2. Leung, K., Wu, J. T., Liu, D. & Leung, G. M. First-wave COVID-19 transmissibility and severity in China outside Hubei after control measures, and second-wave scenario planning: a modelling impact assessment. *Lancet* **6736**, (2020).
3. Ji, Y., Ma, Z., Peppelenbosch, M. P. & Pan, Q. Potential association between COVID-19 mortality and health-care resource availability. *Lancet Glob. Heal.* **8**, e480 (2020).
4. Rosenbaum, L. Facing Covid-19 in Italy — Ethics, Logistics, and Therapeutics on the Epidemic's Front Line. *N. Engl. J. Med.* NEJMp2005492 (2020). doi:10.1056/NEJMp2005492
5. Tian, H. *et al.* An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. *Science* eabb6105 (2020). doi:10.1126/science.abb6105
6. 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 (2020).
7. Lipsitch, M., Swerdlow, D. L. & Finelli, L. Defining the Epidemiology of Covid-19 — Studies Needed. *N. Engl. J. Med.* **382**, 1194–1196 (2020).
8. World Health Organization (WHO). Coronavirus disease 2019 (COVID-19) Situation Report - 71. (2020).
9. Zhao, S. *et al.* Quantifying the association between domestic travel and the exportation of novel coronavirus (2019-nCoV) cases from Wuhan, China in 2020: a correlational analysis. *J. Travel Med.* **27**, 1–3 (2020).
10. Baker, R. E., Yang, W., Vecchi, G. A., Metcalf, C. J. E. & Grenfell, B. T. Susceptible supply limits the role of climate in the COVID-19 pandemic. *Science* **2535**, 2020.04.03.20052787 (2020).
11. Rocklöv, J. & Sjödin, H. High population densities catalyse the spread of COVID-19. *J. Travel Med.* 1–2 (2020). doi:10.1093/jtm/taaa038
12. Kraemer, M. U. G. *et al.* Big city, small world: density, contact rates, and transmission of dengue across Pakistan. *J. R. Soc. Interface* **12**, 20150468 (2015).
13. Dalziel, B. D. *et al.* Urbanization and humidity shape the intensity of influenza epidemics in U.S. cities. *Science* **362**, 75–79 (2018).
14. Shaman, J., Pitzer, V. E., Viboud, C., Grenfell, B. T. & Lipsitch, M. Absolute humidity and the seasonal onset of influenza in the continental United States. *PLoS Biol.* **8**, (2010).
15. Gog, J. R. *et al.* Spatial transmission of 2009 pandemic influenza in the US. *PLoS Comput. Biol.*

10, e1003635 (2014).

16. Shaman, J. & Kohn, M. Absolute humidity modulates influenza survival, transmission, and seasonality. *Proc. Natl. Acad. Sci.* **106**, 3243–3248 (2009).
17. Chetty, R. *et al.* The association between income and life expectancy in the United States, 2001-2014. *JAMA - J. Am. Med. Assoc.* **315**, 1750–1766 (2016).
18. Kissler, S. M., Tedijanto, C., Goldstein, E., Grad, Y. H. & Lipsitch, M. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science* **21**, 1–9 (2020).
19. Crawford, J. M. *et al.* Laboratory Surge Response to Pandemic (H1N1) 2009 Outbreak, New York City Metropolitan Area, USA. *Emerg. Infect. Dis.* **16**, 8–13 (2010).
20. Grasselli, G., Pesenti, A. & Cecconi, M. Critical Care Utilization for the COVID-19 Outbreak in Lombardy, Italy. *JAMA* **19**, NEJMoa2002032 (2020).
21. Li, Q. *et al.* Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. *N. Engl. J. Med.* NEJMoa2001316 (2020). doi:10.1056/NEJMoa2001316
22. Xu, B. *et al.* Epidemiological data from the COVID-19 outbreak, real-time case information. *Sci. Data* **7**, (2020).
23. Xu, B. *et al.* Epidemiological data from the COVID-19 outbreak, real-time case information. *figshare* (2020). doi:10.6084/m9.figshare.11949279
24. Xu, B. & Kraemer, M. U. G. Open access epidemiological data from the COVID-19. *Lancet Infect. Dis.* **3099**, 30119 (2020).
25. Aurora Big Data. 2017 Mobile Map App Research Report: Which of the Highest, the Baidu, and Tencent Is Strong? (2017). Available at: https://baijiahao.baidu.com/s?id=1590386747028939917&wfr=spider&for=pc. (Accessed: 17th March 2020)
26. Lloyd, M. `Mean Crowding'. *J. Anim. Ecol.* **36**, 1 (1967).
27. May, R. M. & Anderson, R. M. Spatial heterogeneity and the design of immunization programs. *Math. Biosci.* **72**, 83–111 (1984).
28. Anderson, R. M. & May, R. M. *Infectious diseases of humans: dynamics and control*. (Oxford University Press, 1991).
29. Watts, D. J., Muhamad, R., Medina, D. C. & Dodds, P. S. Multiscale, resurgent epidemics in a hierarchical metapopulation model. *Proc. Natl. Acad. Sci.* **102**, 11157–11162 (2005).
30. Aylward, Bruce (WHO); Liang, W. (PRC). Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19). **2019**, 16–24 (2020).
31. 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 (2005).

32. Kucharski, A. J. *et al.* Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *Lancet Infect. Dis.* **3099**, 1–7 (2020).

33. Riou, J. & Althaus, C. L. Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. *Eurosurveillance* **25**, 1–5 (2020).

34. Southwood, T. R. E. The Sampling Programme and the Measurement and Description of Dispersion. in *Ecological Methods* 7–69 (Springer Netherlands, 1978). doi:10.1007/978-94-015-7291-0_2

35. Zhang, J. *et al.* Changes in contact patterns shape the dynamics of the COVID-19 outbreak in China. *Science* **8001**, eabb8001 (2020).

36. Lai, S. *et al.* Effect of non-pharmaceutical interventions to contain COVID-19 in China. *Nature* (2020). doi:10.1038/s41586-020-2293-x

37. Sattenspiel, L. Simulating the Effect of Quarantine on the Spread of the 1918–19 Flu in Central Canada. *Bull. Math. Biol.* **65**, 1–26 (2003).

38. Meyers, L. A. Contact network epidemiology: Bond percolation applied to infectious disease prediction and control. *Bull. Am. Math. Soc.* **44**, 63–87 (2006).

39. Kissler, S. M. *et al.* Reductions in commuting mobility predict geographic differences in SARS-CoV-2 prevalence in New York City. *Harvard* (2020).

40. Lipsitch, M., Swerdlow, D. L. & Finelli, L. Defining the Epidemiology of Covid-19 — Studies Needed. *N. Engl. J. Med.* NEJMp2002125 (2020). doi:10.1056/NEJMp2002125

41. Mat, N. F. C., Edinur, H. A., Razab, M. K. A. A. & Safuan, S. A Single Mass Gathering Resulted in Massive Transmission of COVID-19 Infections in Malaysia with Further International Spread. *J. Travel Med.* 1–9 (2020). doi:10.1093/jtm/taaa059

42. Flaxman, S. *et al.* Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries. *Imp. Coll. London* 1–35 (2020). doi:10.25561/77731

Figure legends

Figure 1: Maps of crowding in prefectures in China. *(a) Examples of epidemic curves that are normalized to show the percentage of cases across the whole epidemic that occur at each given day. Beijing and Shanghai (red) have less peaked epidemics than Wenzhou and Zhuhai. (b) Examples of prefectures in China with different levels of crowding and population size. The colour scale illustrates the estimated number of inhabitants per grid cell (1km x 1km). (c) Relationship*

between the Shannon index of the incidence curve and the final attack rate for prefectures in China.

Figure 2: Crowding and the temporal clustering of transmission of COVID-19 in China. *(a) negative association between log10 of epidemic peakedness, as measured by Shannon's diversity index, and log population crowding, as measure by Lloyd's mean crowding . The point sizes indicate the size of the population in each city, (b) Map of epidemic peakedness in China at the prefectural level. Blue and green colours indicate lower peakedness and red and yellow colours higher peakedness. Grey prefectures had either no reported cases or were not included in analyses due to potential inconsistencies in reporting of early cases (Hubei Province).*

Figure 3: Mechanisms generating less peaked epidemics in crowded populations. (*a) Schematic of a hierarchically-structured population model consisting of households and "neighborhoods" within a prefecture. Transmission is more likely among contacts connected at lower spatial levels. Crowded populations have greater number of contacts outside the household, and interventions reduce the number of these connections in both populations (pink dotted lines). (b - c) Simulated outbreak dynamics in the absence of interventions in crowded vs sparse populations. For the networks in (b), blue nodes are individuals who were eventually infected by the end of the outbreak. In (c), thin blue lines show individual realizations of the model, the average shown by the thick grey line. (d) Simulated outbreak dynamics in the presence of strong social distancing measures in crowded vs. sparse populations. The intervention was implemented at day 15 (vertical dotted line) and led to a 75% reduction in contacts similar to observed changes in contact rates in China[35,36]. Mean values of median log epidemic peakedness (Shannon index) are = -2.3 for low crowding and -2.8 for high crowding.*

Figure 4: Predicted epidemic peakedness across the world. *(a) Maps of cities and their population densities at a 1x1km scale. Madrid, Spain and Colombo, Sri Lanka have low predicted peakedness, whilst Novosibirsk, Russia and Ulaanbaatar, Mongolia which have high predicted peakedness. (b) Map of predicted epidemic peakedness for 310 cities across the world for which both human population data and mobility data were available for the study period.*

Table 1 Policy summary

Background	There are obvious differences in the geographic distribution of COVID-19 cases within and among countries. We hypothesise that some of these differences are due to spatial variability in population crowding. Using detailed case count data from COVID-19 among cities in China and Italy, we fit multiple regression models to explain variability in the shape of epidemics among them.
Main findings and limitations	We found that cities with higher crowding have longer epidemics and higher attack rates after the first epidemic wave. Using a metapopulation model that splits cities into neighborhood subunits is consistent with these findings, suggesting that the hierarchical structure and organization of cities are influential in defining their epidemics. We predict that comparatively rural areas may experience more peaked epidemics. As with all modelling studies, further data generated during the epidemic may change our parameter estimates and large-scale serological data would help verify our findings. Further, it will be important to evaluate whether cities that have greater peak incidence may be more prone to strained healthcare systems.
Policy implications	Our results have implications for assessing the drivers of transmission of SARS-CoV-2. Spatial factors such as crowding and population density may elevate the risk of sustained (longer) outbreaks, even after the implementation of lockdowns. Cities that are less crowded and have lower attack rates might be more susceptible to experiencing future outbreaks if SARS-CoV-2 is successfully re-introduced.

Methods

Epidemiological data

No officially reported line list was available for cases in China. We use a standardised protocol[43] to extract individual level data from December 1st, 2019 - March 30th, 2020. Sources are mainly official reports from provincial, municipal, or national health governments. Data included basic demographics (age, sex), travel histories, and key dates (dates of onset of symptoms, hospitalization, and confirmation). Data were entered by a team of data curators on a rolling basis and technical validation and geo-positioning protocols were applied continuously to ensure validity. A detailed description of the methodology is available[22]. Lastly, total numbers were matched with officially reported data from China and other government reports. Daily case counts from Italian provinces (n = 107) were extracted from the Presidenza del Consiglio dei Ministri Dipartimento della Protezione Civile (https://github.com/pcm-dpc/COVID-19).

Estimating epidemic peakedness

Epidemic peakedness was estimated for each prefecture by calculating the inverse Shannon entropy of the distribution of COVID-19 cases. Inverse Shannon entropy was used to fit time series of other respiratory infections (influenza)[13]. The inverse Shannon entropy of incidence for a given prefecture in 2020 is then given by $v_j = \left(-\sum_i p_{ij} \log p_{ij}\right)^{-1}$. Because v_j is a function of incidence distribution in each location rather than raw incidence it is invariant under differences in overall reporting rates between cities or attack rates. We then assessed how peakedness $v \propto \sum_j v_j$ varied across geographic areas in China. As an alternative measure of temporal clustering of cases we estimated the proportion of cases at the peak +/- one day (**Extended Data Figure 2**).

Proxies for COVID-19 interventions using within city human mobility data from China

Estimates of within city reductions of human mobility between the period before and after the lockdown was implemented on January 23, 2020 were extracted from Lai et al. [36]. Daily measures of human mobility were extracted from the Baidu Qianxi web platform to estimate the proportion of daily movement within prefectures in China. Relative mobility volume was available from January 2, 2020 to January 25, 2020. For each city change in relative mobility was defined by $m_i = m_{il}(lockdown)/m_{ib}(baseline)$ where m_i is defined as mobility in prefecture i. Baidu's mapping service is estimated to have a 30% market share in China and more data can be found[5,6].

Data on drivers of transmission of COVID-19

Prefecture-specific population counts and densities were derived from the 2020 Gridded Population of The World, a modeled continuous surface of population estimated from national census data and the United Nations World Population Prospectus[44]. Population counts are defined at a 30 arc-second resolution (approximately 1 km x 1 km at the equator) and extracted within administrative-2 level cartographic boundaries defined by the National Bureau of Statistics of China. Lloyd's mean crowding, $\frac{[\sum_i (q_i - 1) q_i]}{\sum_i q_i}$, was estimated for each prefecture where q_i represents the population count of each non-zero pixel within a prefecture's boundary and the resulting value estimates an individual's mean number of expected neighbors[13,45]. When fitting the models, we consider the numerator $[\sum_i (q_i - 1) q_i]$, which we refer to as "contacts" and the denominator $\sum_i q_i$, i.e., population size, as separate predictors. We note that a negative slope for "contacts" and a positive slope for "population" supports a negative coefficient for Lloyd's mean crowding.

Daily temperature (°F), relative humidity (%) and atmospheric pressure (Pa) at the centroid of each prefecture was provided by The Dark Sky Company via the Dark Sky API and aggregated across a variety of data sources. Specific humidity (kg/kg) was then calculated using the R package, humidity[16]. Meteorological variables for each prefecture were then averaged across the entirety of the study period.

Statistical analysis

We normalized the values of epidemic peakedness between 0 and 1, and for all non-zero values fit a Generalized Linear Model (GLM) of the form:

$$\log(Y_j) \sim \beta_0 + \beta_1 \log(C_j) + \beta_2 q_j + \beta_3 \log(P_j) + \beta_4 \log(f_j) + \beta_5 log(t_j)$$

where for each prefecture j, Y is the scaled inverse Shannon entropy measure of epidemic peakedness derived from the COVID-19 time series, C is the mean number of contacts[26,46], q is the mean specific humidity over the reporting period in kg/kg, P is the estimated population density and f is the relative change in population flows within each prefecture and t is daily mean temperature.

Projecting epidemic peakedness in cities around the world

We selected 310 urban centers from the European Commission Global Human Settlement Urban Centre Database and their included cartographic boundaries[47]. To ensure global coverage, up to the five most populous cities in each country were selected from the 1,000 most populous urban centers recorded in the database. Population count, crowding, and meteorological variables were then estimated following identical procedures used to calculate these variables in the Chinese prefectures. Weather measurements were averaged over the 2-month period starting on February 1, 2020.

The parameters from the model of epidemic peakedness predicted by humidity, crowding and population size (see **Supplementary Table** 1, Model 6) were used to estimate relative peakedness in the 310 urban centers. A full list of predicted epidemic peakedness values can be found in **Supplementary Table** 3.

Global human mobility data

We used the Google COVID-19 Aggregated Mobility Research Dataset, which contains anonymized relative mobility flows aggregated over users who have turned on the Location History setting, which is off by default. This is similar to the data used to show how busy certain types of places are in Google Maps — helping identify when a local business tends to be the most crowded. The mobility flux is aggregated per week, between pairs of approximately $5km^2$ cells worldwide and for the purpose of this study aggregated for 310 cities worldwide. We calculated both, mobility within each city's shapefile and mobility coming into each city. For each city change in relative mobility was defined by $m_i = m_{il}(April)/m_{ib}(December)$ where m_i is defined as mobility in city i.

To produce this dataset, machine learning is applied to log data to automatically segment it into semantic trips[48]. To provide strong privacy guarantees, all trips were anonymized and aggregated using a differentially private mechanism[49] to aggregate flows over time (see https://policies.google.com/technologies/anonymization). This research is done on the resulting heavily aggregated and differentially private data. No individual user data was ever manually inspected, only heavily aggregated flows of large populations were handled.

All anonymized trips are processed in aggregate to extract their origin and destination location and time. For example, if users traveled from location a to location b within time interval t, the corresponding cell (a,b,t) in the tensor would be n∓err, where err is Laplacian noise. The automated Laplace mechanism adds random noise drawn from a zero-mean Laplace distribution and yields (ϵ, δ)-differential privacy guarantee of $\epsilon = 0.66$ and δ = 2.1 × 10−29. The parameter ϵ controls the noise intensity in terms of its variance, while δ represents the deviation from pure ϵ-privacy. The closer they are to zero, the stronger the privacy guarantees. Each user contributes at most one increment to each partition. If they go from a region a to another region b multiple times in the same week, they only contribute once to the aggregation count.

These results should be interpreted in light of several important limitations. First, the Google mobility data is limited to smartphone users who have opted in to Google's Location History feature, which is off by default. These data may not be representative of the population as whole, and furthermore their representativeness may vary by location. Importantly, these limited data are only viewed through the lens of differential privacy algorithms, specifically designed to protect user anonymity and obscure fine detail. Moreover, comparisons across rather than within locations are only descriptive since these regions can differ in substantial ways.

Simulating epidemic dynamics

We simulated a simple stochastic SIR model of infection spread on weighted networks created to represent hierarchically-structured populations. Individuals were first assigned to households using the distribution of household sizes in China (data from UN Population Division, mean 3.4 individuals). Households were then assigned to "neighborhoods" of ~100 individuals, and all neighborhood members were connected with a lower weight. A randomly-chosen 10% of individuals were given "external" connections to individuals outside the neighborhood. The total population size was N=1000. Simulations were run for 300 days and averages were taken over 20 iterations. The SIR model used a per-contact transmission rate of β=0.15/day and recovery rate γ=0.1/day. For the simulations without interventions, the weights were $w_{HH} = 1$, $w_{NH} = 0.01$, and $w_{EX} = 0.001$ for the crowded prefecture and $w_{EX} = 0.0001$ for the less crowded prefecture. For the simulations with interventions, the household and neighborhood weights were the same but we used $w_{EX} = 0.01$ for the crowded prefecture and $w_{EX} = 0.001$ for the "sparse" prefecture. The intervention reduced the weight of all connections outside the household by 75%.

Data availability: We collated epidemiological data from publicly available data sources (news articles, press releases and published reports from public health agencies) which are described in full here[22]. Epidemiological and spatial data used in this study is available via Github (https://github.com/Emergent-Epidemics/covid_hierarchy). The Google COVID-19 Aggregated Mobility Research Dataset used for this study is available with permission from Google, LLC.

Code availability: The code associated with the data analysis and statistics is available from https://github.com/Emergent-Epidemics/covid_hierarchy. The simulation code is available from here: https://github.com/alsnhll/SIRNestedNetwork

43. Ramshaw, R. E. *et al.* A database of geopositioned Middle East Respiratory Syndrome Coronavirus occurrences. *Sci. Data* **6**, 318 (2019).

44. Doxsey-Whitfield, E. *et al.* Taking Advantage of the Improved Availability of Census Data: A First Look at the Gridded Population of the World, Version 4. *Pap. Appl. Geogr.* **1**, 226–234 (2015).
45. Reiczigel, J., Lang, Z., Rózsa, L. & Tóthmérész, B. Properties of crowding indices and statistical tools to analyse parasite crowding data. *J. Parasitol.* **91**, 245–252 (2005).
46. Wade, M. J., Fitzpatrick, C. L. & Lively, C. M. 50-year anniversary of Lloyd's "mean crowding": Ideas on patchy distributions. *J. Anim. Ecol.* **87**, 1221–1226 (2018).
47. Florczyk, A. *et al.* GHS-UCDB R2019A - GHS Urban Centre Database 2015, multitemporal and multidimensional attributes. *Eur. Comm. Jt. Res. Cent.* (2019).
48. Bassolas, A. *et al.* Hierarchical organization of urban mobility and its connection with city livability. *Nat. Commun.* **10**, 4817 (2019).
49. Wilson, R. J. *et al.* Differentially Private SQL with Bounded User Contribution. 1–21 (2019).