Inferring pathogen superspreading potential using early spatial spread patterns

Qing Yao[1], Renquan Zhang[2], Tom Britton[3] and Sen Pei[1*]

1 Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University, New York, USA

2 School of Mathematical Sciences, Dalian University of Technology, Dalian, China

3 Department of Mathematics, Stockholm University, Stockholm, Sweden

*Corresponding author: sp3449@cumc.columbia.edu

Abstract

Superspreading driven by individual variation in transmissibility shapes novel pathogen emergence and the effectiveness of control measures. Current approaches to estimating superspreading typically rely on cluster size distributions informed by contact tracing or genomic data, limited by data availability and potential sampling bias. Here, we examined the impact of superspreading on the early spatial spread of novel pathogens using a branching process model incorporating inter-county mobility in the United States (US). To represent individual transmission heterogeneity, we modeled the number of secondary infections using a negative binomial distribution with a mean reproduction number R_0 and a dispersion parameter r that quantifies superspreading potential. Simulations suggest that stronger superspreading tends to slow early spatial invasion, with a larger variation in epidemic growth between different realizations. Using a graph neural network designed for epidemic inference, we demonstrated that r can be reliably inferred from early spatial spread patterns, robust to the spatiotemporal variation of R_0 and case underreporting. Application to early COVID-19 data in the US revealed strong superspreading prior to nationwide lockdown ($r = 0.50$, 95% CI: [0.23 - 1.20]), followed by weaker superspreading afterward ($r = 1.3$, [0.64 - 3.18]). Our study offers a new approach to quantifying pathogen superspreading potential using population-level observations.

Introduction

Superspreading, where a small proportion of individuals infect the majority of secondary infections, is a common feature of disease spread caused by individual variation in transmissibility[1–5]. Superspreading events have been reported for a range of pathogens including SARS[6–9], measles[10,11], Middle East respiratory syndrome coronavirus (MERS)[3,12], and more recently, SARS-CoV-2[5,13,14]. Previous studies indicate that superspreading has a profound impact on disease transmission dynamics, particularly during the early stage of outbreaks[15–17]. Given the same reproduction number, diseases with strong superspreading are less likely to maintain sustainable transmission chains; however, once established in a population, outbreaks tend to be more explosive[1]. Such epidemiological characteristics can greatly impact the effectiveness of intervention measures[18–20] and herd immunity thresholds[21–24]. As a result, estimating the superspreading potential of pathogens is crucial for epidemic control.

The branching process model is a convenient mathematical framework for modeling superspreading[25]. In this model, the number of secondary infections, or offspring, generated by one infectious person is modeled by a negative binomial distribution $NB(R_0, r)$, where R_0 is the basic reproduction number (the mean of the distribution) and r is the dispersion parameter. The parameter r quantifies the overdispersion of the offspring distribution: the variance of the distribution is $\sigma^2 = R_0 + R_0^2/r, r \in (0, \infty)$. A smaller dispersion parameter r represents a more heterogeneous distribution of secondary infections, thus a stronger superspreading potential[1,26].

Much effort has been made to infer the dispersion parameter r using outbreak data. One of the mainstream approaches relies on the distribution of transmission cluster size[27–29], estimated using either individual-level contact tracing records[1,30,31] or the classification of local versus imported cases[32]. More recent studies have utilized genomic data, reconstructing transmission trees to infer offspring variation[33,34]. However, high-quality contact tracing and dense sequencing are not always available during the early phase of an epidemic, and both may suffer observational bias towards large-scale transmission clusters and miss smaller ones. In addition, the geographical disparity in genomic data collection may limit the representativeness of those datasets[35]. Other studies attempt to use incidence-based likelihoods, where observed cases are modeled as negative binomial random numbers[36,37]. While more broadly applicable, this approach still faces several limitations, such as overlooking the spatiotemporal correlation of incidence in the likelihood and numerical instability, especially in settings with low case counts and short time window[36,37]. These challenges motivate the development of alternative approaches that leverage spatiotemporal patterns in more accessible population-level data.

The early spatial spread of a novel pathogen could be shaped by superspreading as transmission heterogeneity can impact the probability of stochastic extinction (or self-sustained transmission) when the pathogen is introduced into a new region. If validated, spatial spread patterns can offer an alternative dynamical signature to estimate dispersion parameters using population-level statistics. Metapopulation models informed by human

mobility are widely used to simulate the spatial spread of infectious diseases[14,24,38–41]. However, integrating individual-level variation into these population-level models is not straightforward. In addition, metapopulation models are typically represented by a set of ordinary differential equations with exponentially distributed infectious period, which deviates from the actual distributions estimated using real-world disease data.

In this study, we developed a branching process model with a spatial structure connecting over 3000 counties in the United States (US). Through extensive model simulations, we showed that superspreading can introduce substantial uncertainty in the early spatial spread of novel pathogens and impact the invasion speed across locations. To solve the inverse problem of inferring superspreading potential from early spatial spread patterns, we designed a Bayesian last-layer graph neural network (BLL-GNN)[42–44] that couples a deterministic graph-convolutional encoder with a Bayesian output layer. This approach is physics-informed[45–47]: inter-county mobility is encoded as the graph structure in the GNN, branching-process dynamics provide the prior knowledge that constrains the learning process, and the Bayesian layer quantifies parameter uncertainty. The inference framework was validated using extensive simulations of outbreaks under various transmission conditions. Results indicate that this method can accurately estimate pathogen superspreading potential using population-level observations, robust to the spatiotemporal variation of the basic reproduction number R_0 and case underreporting. We applied the inference framework to early COVID-19 data in the US and estimated that SARS-CoV-2 possessed strong superspreading during the early outbreak, in agreement with findings from other studies[5]. Additionally, the estimated superspreading dropped substantially after the nationwide lockdown, suggesting the potential effect of interventions on reducing transmission heterogeneity.

Results

Spatial spread patterns with superspreading

We developed a branching process model with a spatial structure to simulate the invasion of a novel pathogen across US counties. This model jointly incorporated a susceptible-exposed-infectious-recovered (SEIR) dynamics, individual variation in transmissibility, and human mobility between US counties (Methods). Specifically, the number of secondary infections produced by one infectious person was drawn from a negative binomial distribution $NB(R_0, r)$. Note, in statistics literature, the negative binomial distribution is usually parameterized using p, the probability of success in each Bernoulli trial, and the dispersion parameter r. In the context of epidemiology, the basic reproduction number R_0 is linked to the parameter p through $R_0 = r(1-p)/p$ (Methods). For interpretation convenience, we used R_0 and r to parameterize negative binomial distributions throughout this study. An illustration of the model structure is shown in Fig. 1. By varying the basic reproduction number R_0 and the dispersion parameter r, we can explore the impact of intrinsic disease transmissibility and superspreading potential on the early spatial spread patterns.

We first examined the geographical dispersion of a novel pathogen after a certain period of transmission. We simulated outbreaks originating from New York County (Manhattan, NY)

with various R_0 and r configurations for 60 days, starting from 100 infected individuals (Methods). In each experiment, we assigned the same R_0 and r in all counties. To capture the stochastic dynamics, 300 independent simulations were performed. We defined a county as "infected" if local daily new infections exceed 10 per 100,000 people. Taking average over all simulations, the number of infected counties by day 60 increased with growing R_0 (Fig. 2a). For a fixed R_0, stronger superspreading (i.e., smaller r) led to a lower number of infected counties on day 60, averaged across 300 independent simulations. Additional to lower average number of infected counties on day 60 for superspreading, another difference is that superspreading also increases variation (e.g. standard deviation) between simulations. (Fig. 2b).

We quantified the spatial invasion speed of pathogens using the daily number of counties that have first local infections. For instance, an invasion speed of 5 counties per day means that the outbreak produced at least one infection in 5 new counties each day. Averaged over the 60-day period, the mean invasion speed increased with growing transmissibility (R_0) and decreased for higher superspreading (Fig. 2c). To measure the temporal variation of the invasion speed, we computed the coefficient of variation (COV) of daily invasion speed within the 60-day simulation window, defined as the standard deviation of invasion speed divided by its mean. The COV quantifies the relative variation of invasion speed around the mean speed during the 60-day period. Pathogens with a lower R_0 had a larger COV of invasion speed, suggesting a more pronounced acceleration of spatial invasion over time (Fig. 2d). Given the same R_0, stronger superspreading led to a larger COV, possibly due to the higher uncertainty in simulation outcomes.

To further compare the temporal evolution of spatial spread at different superspreading levels, we compared the number of infected counties over the 60-day period for $R_0 = 2.5$, using the same parameter settings and initial conditions but different dispersion rates (Fig. 3a). To capture the variation of spatial spread generated using the stochastic branching process model, we performed 300 independent realizations for each parameter setting. In general, the models with smaller r (stronger superspreading) produced slower spatial dispersal during the early phase of the outbreak. For $r = 20$, the realization with a median number of infected counties by day 60 infected 67 counties. In contrast, for $r = 0.025$, the median trajectory only infected 48 counties, 28.3% lower than the median outcome for $r = 20$.

For branching process models, higher superspreading potential ($r = 0.025$) led to highly stochastic outcomes, indicated by the wide spread of trajectories generated from independent realizations, especially during the onset period between 15 and 25 days (Fig. 3a). For an outbreak with stronger superspreading potential, the initial spatial spread speed was more variable across different model realizations – some realizations could be faster due to occasionally more explosive infections, and some could be slower as transmission chains may die out. For $r > 1$, the effect of superspreading became less pronounced. We visualize the geographical range of infected counties in northeastern US on day 60 generated by branching process models with different r values (Fig. 3b). The branching model with large r, whose offspring distribution is closer to a Poisson distribution, produced a faster spatial

invasion for the pathogen originating from New York County during the 60-day period. Additional simulations with a higher $R_0 = 5.5$ further revealed pronounced difference in spatial spread at the national scale among different dispersion parameters. (Methods, Fig. S1).

The effects of stochastic extinction and superspreading events

Our simulations indicate that the effect of transmission heterogeneity on spatial spread has two folds. First, on average, a pathogen with strong transmission heterogeneity tends to spread more slowly into new areas during early outbreaks. Such slower spatial spread is primarily caused by the high extinction probability of an infection. For instance, for $R_0 = 2.5$ and a dispersion parameter $r = 0.025$, the probability of stochastic extinction for one infection is 0.96 (Supplementary Information. Note, stochastic extinction probability is different from the probability of generating zero offspring, which is 0.89 here). In contrast, for $r = 20$, the extinction probability is only 0.12. For small dispersion parameters, the establishment of sustained transmission in a new location thus likely requires more introductions, which will take a longer time. Second, stronger superspreading potential leads to a larger variation in epidemic growth among independent model realizations. Some rare superspreading events may expedite spatial spread by producing a large number of seeds introduced into new regions, potentially offsetting the effect of high extinction probability of these introductions.

To better disentangle the effects of stochastic extinction and superspreading events, we performed simulations where the number of secondary infections was generated using a Poisson mixture distribution. Specifically, we replaced the negative binomial distribution in the branching process model with a Poisson mixture: $p_\lambda Pois(R_0/2) + (1 - p_\lambda)Pois(\lambda)$ for $\lambda > R_0$, where $Pois(\lambda)$ is a Poisson distribution with mean λ, and p_λ controls the mixture between the two Poisson distributions. To ensure the distribution mean to be R_0, p_λ needs to satisfy $p_\lambda = (\lambda - R_0)/(\lambda - R_0/2)$. By varying the parameter λ, we can adjust the extinction probability and the probability of superspreading events: a larger λ leads to a higher extinction probability and rarer but more explosive superspreading events. The effect of λ on the offspring distribution is thus similar to that of $1/r$ in the negative binomial distribution; however, its impact on extinction probability is less pronounced[48]. For $R_0 = 2.5$, a large value of $\lambda = 40$ $(p_\lambda = 0.97)$ leads to an extinction probability of 0.55 (Supplementary Information), much lower than the extinction probability for small r values in the negative binomial distribution. A comparison of extinction probabilities for negative binomial distributions and the Poisson mixture distributions is provided in Table S2.

We simulated outbreaks using Poisson mixture distributions with $R_0 = 2.5$ and various λ values (Fig. S2). With a reduced difference in extinction probability, the median trajectories exhibited similar spatial spread speed irrespective of amount of superspreading events. However, we still observed an increased variance in the number of infected counties for larger λ values (Fig. S2), suggesting that superspreading events mainly contribute to the stochasticity of early spatial invasion dynamics.

Inferring superspreading using Bayesian Last-Layer Graph Neural Network

As superspreading can affect the invasion dynamics of novel pathogens, the early spatial spread patterns may serve as a dynamical signature that allows the inference of superspreading potentials. However, this inference task is challenging for traditional approaches – for likelihood-based methods, it is difficult to handle the spatiotemporal correlation in case counts; for machine learning algorithms, it is hard to manually design features that capture superspreading potential. To address these challenges, we employed a graph neural network (GNN) framework, which can flexibly learn the nuanced spatiotemporal patterns associated with different levels of superspreading. GNN has been widely applied in spatiotemporal modeling of complex dynamical systems, including infectious disease transmission[49–51]. To further capture uncertainty in the inferred parameters, we extended the architecture with a Bayesian last-layer formulation. The resulting Bayesian Last-Layer Graph Neural Network (BLL-GNN) combines a deterministic graph convolutional encoder with a Bayesian output layer to estimate the dispersion parameter (Methods, Supplementary Information). This hybrid design enables efficient uncertainty quantification while maintaining scalability to large, time-varying mobility networks[42].

We validated the inference framework using simulated outbreaks generate with different dispersion parameters. For a given reproduction number R_0, we generated an outbreak using a ground-truth dispersion parameter r. Our goal is to estimate this ground-truth r using the simulated daily new infections obtained from one single realization. To train the BLL-GNN, we generated synthetic outbreaks using the same R_0 and different r values as a training dataset. For each (R_0, r) combination, we generated 180 independent outbreaks for training. The BLL-GNN was optimized using this simulated training data for the regression task to output dispersion parameters with the input of daily new infections in all counties. We then applied the optimized BLL-GNN to the simulated outbreak data to infer the ground-truth r.

To validate the performance of the neural network for diseases with different levels of transmissibility R_0, we performed experiments for three fixed R_0 values: 1.5, 2.5, and 5.5. For reference, a table for the basic reproduction numbers R_0 and dispersion parameters r for a list of common infectious diseases are provided in Table S1. For each dispersion parameter, we generated 60 realizations of outbreaks and performed inference for each realization independently. In general, the graph neural network accurately inferred the dispersion parameter across a broad range of r values (Fig. 4a–c), except when r exceeded 10. This limitation is expected, as the negative binomial distribution converges to a Poisson distribution for large r, producing nearly indistinguishable spatiotemporal transmission patterns. Consistent with previous findings, estimation accuracy decreased once r surpassed unity, reflecting reduced overdispersion in the offspring[36,37]. The overall coverages of true r within the 95% CIs were 89.6%, 93.8% and 98.9% for $R_0 = 1.5$, 2.5, and 5.5, respectively.

In real-world settings, only a single realization of the stochastic epidemic process is typically observed, with unknown reproduction numbers varying over time and across locations. To validate our approach under such conditions, we generated synthetic outbreaks with county-specific, time-varying reproduction numbers and tested the model's ability to recover the true dispersion parameter (Fig. S3). For each county, we selected an initial R_0 linked to population size and then gradually decreased R_0 over time (Methods). To capture the

general pathogen transmissibility, instead of estimating the reproduction number in each county, we estimated the crude daily effective reproduction number R_t at the national level using EpiEstim[52], a statistical method based on renewal processes to estimate time-varying reproduction numbers. As the population infected during the early outbreak is negligible compared with the total population, we used the estimated effective reproduction number R_t as the proxy for the basic reproduction number R_0. The estimated reproduction number matched the temporal trend of R_0 in different counties (Fig. S4). We then used the estimated daily national reproduction numbers and different dispersion parameters to generate synthetic outbreaks as training datasets for the BLL-GNN (Methods), where the inputs are the daily case counts in all counties and the output is the dispersion parameter with associated uncertainty. To validate the trained BLL-GNN, we generated synthetic outbreaks using seven representative, ground-truth r values, among which three values were not used in model training and served as out-of-sample validation targets. Across the seven r values, the BLL-GNN accurately recovered the true dispersion parameters (Fig. 5a), despite that the reproduction numbers used in generating training datasets were different from the actual county-specific R_0.

In the above experiments, we assumed all infections were captured by the surveillance system. In reality, underreporting poses a major challenge for real-world inference[24,53]. To examine the robustness of the inference to underreporting, we further incorporated underreporting in the simulated outbreaks. Specifically, we assigned an initial reporting rate for each county in the range of 10%-15% and then gradually increased it over time to 15%-20% (Methods). Using the underreported case counts, we applied EpiEstim to estimate the daily national reproduction number and found that the results captured the temporal trend of R_0 (Fig. S4). For validation, we simulated outbreaks using seven ground-truth r values. Then we generated training datasets using the estimated reproduction numbers and various dispersion parameters, trained the BLL-GNN, and performed inference with the underreported case counts to estimate r (Methods). Even with severe, time-varying underreporting, the trained BLL-GNN was able to recover the true dispersion parameters (Fig. 5b).

Estimating superspreading of SARS-CoV-2 in the US

We applied the inference framework to the early phase of the COVID-19 pandemic in the US, focusing on two periods: February 23rd–March 14th, 2020 (before the nationwide lockdown) and March 15th–April 18th, 2020 (after lockdown). Studies have shown that changes in population mobility are strongly associated with changes in disease transmission[54]. To better capture changes in population mobility during this period, we incorporated time-varying inter-county mobility constructed from anonymized foot-traffic data provided by Advan Research Corporation[55] on Dewey Data platform[56]. In the inference, we used the daily new infections (both reported and unreported cases) estimated using a dynamic data-driven transmission model that accounted for time-varying ascertainment rates and reporting delays[24]. The estimated infection numbers were validated independently by serological data collected at different locations and time points[24].

During the pre-lockdown phase, we estimated a dispersion parameter of $r = 0.50$ (95% CI: 0.23 – 1.20) (Fig. 6), consistent with a pooled estimate of SARS-CoV-2 superspreading in literature[5]. Such a low r value indicates a strong superspreading potential, as observed in several early COVID-19 clusters worldwide[5,13,14]. After the implementation of mobility restrictions and behavioral interventions, the inferred dispersion parameter increased to $r = 1.3$ (95% CI: 0.64–3.18), suggesting a noticeable reduction in transmission heterogeneity. This shift implies that control measures not only decreased the overall pathogen transmissibility but also reduced opportunities for large-scale superspreading events, possibly achieved by social distancing and limiting large gatherings.

Discussion

Transmission heterogeneity can impact the probability of establishing sustained transmission when a novel pathogen is introduced into a new region. Understanding how superspreading shapes the invasion dynamics of emerging outbreaks across locations is critical for epidemic control. We developed a branching process model with a spatial structure to explore how superspreading potential affects the spatial spread patterns of infectious diseases. Our simulations and the comparison with a Poisson mixture distribution reveal the separate effects of high extinction probability and rare superspreading events on spatial spread: a pathogen with a high extinction probability tends to spread more slowly into new areas during early outbreaks, while rare superspreading events lead to a large variation in the epidemic growth between different simulations. Predicting the spatial invasion speed for pathogens with strong superspreading potential is therefore challenging for a single model realization given this high stochasticity.

Our finding of the impact of transmission heterogeneity motivated us to infer the dispersion parameter using spatial spread patterns. Using population-level observations to infer individual-level epidemiological characteristics is advantageous in resource-limited settings where contact tracing and genomic data are not available. This offers an alternative approach to estimating dispersion parameters that complements existing tools. We demonstrated that the inference framework is robust to the spatiotemporal variation of pathogen transmissibility and underreporting, suggesting its potential practical utility in real-world settings. The inference results may support early assessment of controllability of emerging pathogens. Understanding this key epidemiological property is critical in informing effective interventions during the early stage of outbreaks.

Our study builds on recent advances in deep learning for stochastic physical systems and applies these ideas to epidemiology. Conceptually aligned with physics-informed and simulation-based machine learning frameworks, our approach leverages mechanistic knowledge from stochastic branching processes to guide neural-network inference. By representing inter-county mobility as a dynamic graph and integrating the statistical structure of branching processes into the learning framework, we demonstrate how deep learning can infer latent mechanistic parameters, such as the dispersion parameter, directly from observed epidemic data. This work illustrates how these principles and simulation-based inference can be applied to stochastic, high-dimensional systems such as epidemic spread.

Several limitations exist in this study. First, the branching process model does not fully represent the heterogeneity in real-world epidemic process. Multiple factors can contribute to superspreading, such as individual variation in susceptibility, infectiousness, and number of close contacts. In our model, all individuals are assumed equally susceptible, and the number of secondary infections for each infectious individual follows the same negative binomial distribution. This formulation represents heterogeneity in infectiousness rather than in social activity or contact structure, which can affect the risk of both infecting others and getting infected. Consequently, the model only captures random variation in how many secondary infections each case causes – for example, due to differences in viral load or chance attendance at large gatherings – but does not account for individuals who are consistently more socially active or at higher risk of exposure. Therefore, while we refer to strong overdispersion as superspreading, it should be interpreted as reflecting variability in infectiousness rather than the super spreader behavior. Future studies can further incorporate other factors contributing to superspreading. Second, to generate training datasets, we estimated the national reproduction number and ignored the variation of R_0 across locations. The reproduction number was inferred using EpiEstim, which may yield overconfident estimates in the presence of overdispersion[57] and has higher bias when cases are few[58]. A promising direction for future work is the joint inference of both transmissibility and superspreading within a unified framework. Lastly, the computational cost for running the branching process model increases substantially as the number of infections grows. Future studies should explore more efficient simulation techniques to speed up the algorithm.

Methods

Data

The U.S. Census cross-county commuting data were obtained from the US Census Bureau[59]. We used the 5-year commuting data to extract the population flow across the 3142 counties in the US. Population in each county was downloaded from the US Census Bureau Population and Housing Unit Estimates Datasets[60]. The mobile phone-derived foot-traffic data was provided by Advan Research Corporation[55] through Dewey Data platform[56]. The estimated daily numbers of COVID-19 infections (both reported and unreported) in US counties were obtained from Pei et al.[24]

Mobility networks

The static cross-county mobility network is constructed using the US Census cross-county commuting data. This dataset provides commuting patterns between counties, denoted as n_{ij}, representing the daily number of people commuting from county j to county i. To remove negligible human mobility, we retained only commuting flows of more than 100 people per day. Subsequently, we normalized these flows based on the population of the originating county, ensuring they represent a proportion of the county's total population. The resulting mobility matrix is denoted by $M = \{m_{ij}\}$. The size of M is the number of all counties and county-equivalents in the US. Each entry m_{ij} for $i \neq j$ represents the probability of an individual commuting from county j to county i; m_{ii} represents the probability of an

individual staying in the same county i.

The time-varying cross-county mobility networks in 2020 were built using anonymized foot-traffic data derived from mobile devices. The original dataset recorded the number of visits that anonymized users made from their home census block groups to various points of interest (POI) such as workplaces, restaurants and bars, and other public venues. To obtain county-level movement patterns, we aggregated these visits according to the county of residence and the county of destination POI. This produced a sequence of daily mobility matrices representing the number of trips between every pair of origin and destination counties from February 23 to April 18, 2020. Each matrix was then normalized so that every entry $m_{ij}(t)$ $(i \neq j)$ indicates the probability an individual commuting from county j to county i on a given day. The resulting series of normalized matrices describes how inter-county connectivity changed over time during the early COVID-19 outbreak.

The branching process model

To represent individual variation in disease transmissibility, we used a negative binomial distribution to generate the number of secondary infections (or offspring) caused by an infectious person. Denote $P_n(X = x)$ as the probability of an infectious person n generating x secondary infections. We define

$$P_n(X = x) = \frac{\Gamma(x+r)}{\Gamma(x+1)\Gamma(r)}(1-p)^r p^x, \quad (1)$$

$$p = \frac{r}{R_0 + r}, \quad (2)$$

where R_0 is the basic reproduction number of the disease (i.e., the average number of infections caused by one infected individual in a fully susceptible population) and $r \in (0, \infty)$ is the dispersion parameter, characterizing the transmission heterogeneity among individuals. For a given basic reproduction number R_0, a smaller r leads to a distribution P_n with a long tail, indicating a higher probability of producing superspreading events. Conversely, a larger r yields a distribution with a more homogeneous transmission. For $r \to \infty$, the negative binomial distribution converges to a Poisson distribution. The distributions P_n with varying r values are shown in Fig. S5.

Model simulations proceeded as follows. For each exposed individual who contracted the disease on day t, we tracked the progress of disease states. Specifically, the individual became infectious on day $t + z$, where z is the latency period drawn from a Gamma distribution $\Gamma_z(Z_a, Z_b)$, and then recovered on day $t + z + d$, where d is the infectious period drawn from a Gamma distribution $\Gamma_D(D_a, D_b)$. The infected individual generated k secondary infections during the entire course of infection, computed as a random number drawn from $NB(R_0, r)$ multiplied by the fraction of susceptible population (considering the depletion of susceptible population). For each secondary infection, we assigned their infection time as day $t + z + vd$, where v follows a uniform distribution $U[0,1]$. That is, the timing of secondary infections was uniformly distributed within the infectious period of the primary infection. We further distributed these secondary infections spatially into different counties using commuting data. For instance, secondary infections generated in

county i were distributed to county j with a probability m_{ji}.

In our simulations, the latency period distribution was set as $\Gamma_Z(Z_a = 3, Z_b = 1)$ and the infectious period distribution was set as $\Gamma_D(D_a = 5, D_b = 1)$. For each combination of R_0 and r, 300 independent simulations were performed. For simulations with a fixed reproduction number across space and time, we initialized the outbreak with 100 infected individuals in New York County, New York. For scenarios with spatiotemporally varying reproduction numbers, we distributed the initial infections across multiple locations to mimic geographically dispersed seeding events: 50 individuals in New York County, New York, 50 in Santa Clara County, California, 20 in Suffolk County, Massachusetts, and 20 in King County, Washington.

Configuration of the BLL-GNN

We implemented a Bayesian Last-Layer Graph Neural Network (BLL-GNN) to infer the dispersion parameter r from simulated epidemic data. Each input sample contained a window of daily new infections across all counties, together with the mobility matrix. The encoder consisted of three graph convolutional layers (feature dimensions: input [number of time steps]→128→64→32), each followed by Exponential Linear Unit (ELU) activations and dropout (p = 0.5). Node embeddings were aggregated via global mean pooling, normalized using LayerNorm, and projected through a fully connected bottleneck (8 dimensions). The final output was computed through a Bayesian linear layer (implemented using the Blitz Bayesian framework[61]) to capture parameter uncertainty. The prior was defined as a two-component Gaussian mixture (wide component $\sigma_1 = 3.0$, narrow component $\sigma_2 = 0.1$, mixture weight $\pi = 0.8$).

The model was trained to predict log(r) using simulated outbreaks covering 21 dispersion parameter values ranging from 0.025 to 20. For each r, 300 realizations of outbreak data were generated. Optimization used the Adam optimizer and a fixed KL divergence weight ($\beta = 3 \times 10^{-3}$) during variational inference. All networks were implemented in PyTorch[62] and PyTorch Geometric[63]. Further mathematical details and inference equations are provided in the Supplementary Information.

Time-varying Bayesian graph neural network

In the real-world SARS-CoV-2 data, mobility across locations changed over time. To incorporate temporal heterogeneity in mobility changes, we extended the Bayesian GNN framework to handle time-varying mobility graphs. Each input sample consisted of daily new infections at all locations (node features, dimension N×T) and a sequence of adjacency matrices, $m_{ij}(t)$, describing mobility between counties across time (t =1,…,T). We implemented a temporal graph convolutional encoder, where at each time step node features were linearly transformed, propagated through normalized mobility matrices, and activated using the Exponential Linear Unit (ELU) function. Three layers of this operation, with output dimensions of 64, 64, and 32, were stacked, followed by a single LayerNorm and dropout layer. Temporal aggregation was also achieved by mean pooling across time. Then, similarly to the non-time-varying synthetic model, node embeddings were averaged to produce a

graph-level summary vector and that vector was passed through a fully connected bottleneck and finally a Bayesian linear layer with the default Gaussian mixture prior ($\sigma_1 = 1.0,\ \sigma_2 = 0.002,\ \pi = 0.5$), yielding a predictive posterior over $\log(\mathrm{r})$.

Generating synthetic outbreaks with county-specific R_0 and underreporting

We generated synthetic outbreaks with county-specific, time-varying basic reproduction numbers. The initial reproduction number $R_{0,\ell} \in [2.5, 4]$, determined by the population size pop_ℓ of county ℓ. Over time, the reproduction number $R_{t,\ell}$ decreases toward a floor value R_{floor} following a logistic function:

$$R_{0,\ell} = 2.5 + \frac{\mathrm{pop}_\ell - \mathrm{pop}_{min}}{\mathrm{pop}_{max} - \mathrm{pop}_{min}} \times (4 - 2.5), \tag{3}$$

$$R_{t,\ell} = R_{floor} + (R_{0,\ell} - R_{floor}) \frac{1}{1 + \exp\left(k(t - t_{mid})\right)}, \tag{4}$$

where pop_{min} and pop_{max} denote the minimum and maximum population size of all counties, $R_{floor} = 1.5$ is the lower bound, $k = 0.2$ is the steepness of the decay, and $t_{mid} = T/2$ is the midpoint of the simulation period T.

To evaluate the robustness of our inference framework under imperfect surveillance, we incorporated county-specific reporting rates $\rho_{t,\ell}$ for each county ℓ and day t. $\rho_{t,\ell}$ increased monotonically over time, mimicking improving case ascertainment during the early phase of the pandemic - from 10–15% at the start to 15–20% by the end of the 60-day window. For each county ℓ, the reporting rates were defined as

$$\rho_{t,\ell} = 0.1 + 0.05 q_\ell + 0.05 s_t, \tag{5}$$

where $q_\ell \sim \mathcal{U}(0, 1)$ represents the relative initial case reporting in county ℓ and s_t is a normalized sigmoid curve representing the nonlinear increase of reporting rate:

$$s_t \propto \frac{1}{1 + \exp\left(-0.6\left(\frac{t - \frac{T-1}{2}}{T-1}\right)\right)}, \tag{6}$$

where $s_0 = 0$ and $s_T = 1$.

Validation of the inference framework

To validate the inference framework, we simulated one realization of an outbreak using the branching process model with a given ground-truth dispersion parameters r and the time-varying $R_{t,\ell}$, producing the daily new infections $\mathrm{Inf}^s_{t,\ell}$ across all counties during T days. To estimate the crude reproduction number, we aggregated the daily county-level infections to the national level and used EpiEstim to reconstruct the national time-varying reproduction number $\hat{R}_t$ (Fig.S4). The inferred $\hat{R}_t$ were used to generate a set of synthetic training datasets spanning a broad range of r values. Here we used 21 r values selected from 0.025 to 20. Then, we trained the BLL-GNN on these synthetic datasets and applied it to that the single realization $\mathrm{Inf}^s_{t,\ell}$ to estimate r. In total, we performed inference for seven ground-truth dispersion parameters. Among them, three values were not used in the training datasets,

serving as out-of-sample validation targets.

When incorporating underreporting rates, we multiplied the daily infections by $\rho_{t,\ell}$ to obtain the corresponding observed cases $O_{t,\ell}$. The effective reproduction number $\hat{R}_t^r$ was inferred using the national observed cases using the EpiEstim. The model was then trained on the training dataset generated using the estimated $\hat{R}_t^r$. To account for the underreported infection numbers in the inputs to the BLL-GNN, we rescaled the observed cases using $O_{t,\ell}/0.15$, assuming a uniform, fixed reporting rate of 15% across all locations and time. This rescaling keeps the magnitude of input infections similar to that in the training datasets.

Data availability

Extracted mobility network data and population data used in this paper are available at our project's GitHub repository (https://github.com/Qing1011/branchingmodel).

Code availability

Custom code and data supporting the statistical analysis are publicly available at GitHub (https://github.com/Qing1011/branchingmodel).

Acknowledgements

This study was supported by funding from National Institutes of Health R35GM156799 (S.P.) and National Science Foundation DMS-2229605 (S.P.). We thank Chester Tan for discussion on graph neural network methods.

Author contributions

S.P., Q.Y., T.B., and R.Z. conceived the study, Q.Y. performed the analysis, S.P. and Q.Y. curated the data, Q.Y., R. Z., T.B. and S.P. investigated the results. Q.Y. and S.P. drafted the manuscript, all authors revised and reviewed the manuscript.

Competing interests

All authors declare no competing interests.

References

1. 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).
2. Woolhouse, M. E. *et al.* Heterogeneities in the transmission of infectious agents: implications for the design of control programs. *Proc. Natl. Acad. Sci.* **94**, 338–342 (1997).
3. Kucharski, A. & Althaus, C. L. The role of superspreading in Middle East respiratory syndrome coronavirus (MERS-CoV) transmission. *Eurosurveillance* **20**, 21167 (2015).
4. Su, Z. *et al.* Emergence of nonlinear crossover under epidemic dynamics in heterogeneous networks. *Phys. Rev. E* **102**, 052311 (2020).
5. Du, Z. *et al.* Systematic review and meta-analyses of superspreading of SARS-CoV-2 infections. *Transbound. Emerg. Dis.* **69**, e3007–e3014 (2022).
6. Riley, S. *et al.* Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. *Science* **300**, 1961–1966 (2003).
7. Lipsitch, M. *et al.* Transmission Dynamics and Control of Severe Acute Respiratory Syndrome. *Science* **300**, 1966–1970 (2003).
8. Shen, Z. *et al.* Superspreading SARS Events, Beijing, 2003. *Emerg. Infect. Dis.* **10**, 256–260 (2004).
9. Yu, I. T. S. *et al.* Evidence of Airborne Transmission of the Severe Acute Respiratory Syndrome Virus. *N. Engl. J. Med.* **350**, 1731–1739 (2004).
10. Jansen, V. A. *et al.* Measles outbreaks in a population with declining vaccine uptake. *Science* **301**, 804–804 (2003).
11. De Serres, G. *et al.* Largest measles epidemic in North America in a decade—Quebec, Canada, 2011: contribution of susceptibility, serendipity, and superspreading events. *J. Infect. Dis.* **207**, 990–998 (2013).
12. Kim, K. H., Tandi, T. E., Choi, J. W., Moon, J. M. & Kim, M. S. Middle East respiratory syndrome coronavirus (MERS-CoV) outbreak in South Korea, 2015: epidemiology, characteristics and public health implications. *J. Hosp. Infect.* **95**, 207–213 (2017).
13. Lemieux, J. E. *et al.* Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. *Science* **371**, eabe3261 (2021).
14. Li, R. *et al.* Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science* **368**, 489–493 (2020).
15. Zhao, S. *et al.* Characterizing superspreading potential of infectious disease: Decomposition of individual transmissibility. *PLoS Comput. Biol.* **18**, e1010281 (2022).
16. 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**, 2000058 (2020).
17. Kucharski, A. J. *et al.* Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *Lancet Infect. Dis.* **20**, 553–558 (2020).
18. Althouse, B. M. *et al.* Superspreading events in the transmission dynamics of SARS-CoV-2: Opportunities for interventions and control. *PLOS Biol.* **18**, e3000897 (2020).
19. Susswein, Z. & Bansal, S. *Characterizing Superspreading of SARS-CoV-2 : From Mechanism to Measurement*. http://medrxiv.org/lookup/doi/10.1101/2020.12.08.20246082 (2020) doi:10.1101/2020.12.08.20246082.
20. Sneppen, K., Nielsen, B. F., Taylor, R. J. & Simonsen, L. Overdispersion in COVID-19

increases the effectiveness of limiting nonrepetitive contacts for transmission control. *Proc. Natl. Acad. Sci.* **118**, e2016623118 (2021).
21. Randolph, H. E. & Barreiro, L. B. Herd Immunity: Understanding COVID-19. *Immunity* **52**, 737–741 (2020).
22. Tavori, J. & Levy, H. Super-Spreaders Out, Super-Spreading In: The Effects of Infectiousness Heterogeneity and Lockdowns on Herd Immunity. https://doi.org/10.48550/ARXIV.2101.09188 (2021) doi:10.48550/ARXIV.2101.09188.
23. Britton, T., Ball, F. & Trapman, P. A mathematical model reveals the influence of population heterogeneity on herd immunity to SARS-CoV-2. *Science* **369**, 846–849 (2020).
24. Pei, S., Yamana, T. K., Kandula, S., Galanti, M. & Shaman, J. Burden and characteristics of COVID-19 in the United States during 2020. *Nature* **598**, 338–341 (2021).
25. Becker, N. Estimation for discrete time branching processes with application to epidemics. *Biometrics* 515–522 (1977).
26. Grassly, N. C. & Fraser, C. Mathematical models of infectious disease transmission. *Nat. Rev. Microbiol.* **6**, 477–487 (2008).
27. Kwok, K. O. *et al.* Inferring super-spreading from transmission clusters of COVID-19 in Hong Kong, Japan, and Singapore. *J. Hosp. Infect.* **105**, 682–685 (2020).
28. Blumberg, S. & Lloyd-Smith, J. O. Comparing methods for estimating R0 from the size distribution of subcritical transmission chains. *Epidemics* **5**, 131–145 (2013).
29. Blumberg, S. & Lloyd-Smith, J. O. Inference of R0 and Transmission Heterogeneity from the Size Distribution of Stuttering Chains. *PLoS Comput. Biol.* **9**, e1002993 (2013).
30. Bi, Q. *et al.* Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. *Lancet Infect. Dis.* **20**, 911–919 (2020).
31. Wallinga, J. & Teunis, P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Am. J. Epidemiol.* **160**, 509–516 (2004).
32. Endo, A., Abbott, S., Kucharski, A. J., Funk, S., & others. Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. *Wellcome Open Res.* **5**, 67 (2020).
33. Gómez-Carballa, A., Pardo-Seco, J., Bello, X., Martinón-Torres, F. & Salas, A. Superspreading in the emergence of COVID-19 variants. *Trends Genet.* **37**, 1069–1080 (2021).
34. Tran-Kiem, C. & Bedford, T. Estimating the reproduction number and transmission heterogeneity from the size distribution of clusters of identical pathogen sequences. *Proc. Natl. Acad. Sci.* **121**, e2305299121 (2024).
35. Brito, A. F. *et al.* Global disparities in SARS-CoV-2 genomic surveillance. *Nat. Commun.* **13**, 7003 (2022).
36. Zhang, Y., Britton, T. & Zhou, X. Monitoring real-time transmission heterogeneity from incidence data. *PLOS Comput. Biol.* **18**, e1010078 (2022).
37. Ho, F. *et al.* Accounting for the potential of overdispersion in estimation of the time-varying reproduction number. *Epidemiology* **34**, 201–205 (2023).
38. Colizza, V. & Vespignani, A. Invasion threshold in heterogeneous metapopulation networks. *Phys. Rev. Lett.* **99**, 148701 (2007).
39. Balcan, D. *et al.* Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc. Natl. Acad. Sci.* **106**, 21484–21489 (2009).

40. Pei, S., Kandula, S., Yang, W. & Shaman, J. Forecasting the spatial transmission of influenza in the United States. *Proc. Natl. Acad. Sci.* **115**, 2752–2757 (2018).
41. Chang, S. *et al.* Mobility network models of COVID-19 explain inequities and inform reopening. *Nature* **589**, 82–87 (2021).
42. Harrison, J., Willes, J. & Snoek, J. Variational Bayesian last layers. *ArXiv Prepr. ArXiv240411599* (2024).
43. Ovadia, Y. *et al.* Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Adv. Neural Inf. Process. Syst.* **32**, (2019).
44. Fiedler, F. & Lucia, S. Improved uncertainty quantification for neural networks with bayesian last layer. *IEEE Access* **11**, 123149–123160 (2023).
45. Karniadakis, G. E. *et al.* Physics-informed machine learning. *Nat. Rev. Phys.* **3**, 422–440 (2021).
46. Meng, C., Griesemer, S., Cao, D., Seo, S. & Liu, Y. When physics meets machine learning: A survey of physics-informed machine learning. *Mach. Learn. Comput. Sci. Eng.* **1**, 20 (2025).
47. Willard, J., Jia, X., Xu, S., Steinbach, M. & Kumar, V. Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Comput. Surv.* **55**, 1–37 (2022).
48. Andersson, H. & Britton, T. *Stochastic Epidemic Models and Their Statistical Analysis*. vol. 151 (Springer Science & Business Media, 2012).
49. Fritz, C., Dorigatti, E. & Rügamer, D. Combining graph neural networks and spatio-temporal disease models to improve the prediction of weekly COVID-19 cases in Germany. *Sci. Rep.* **12**, 3930 (2022).
50. Tomy, A., Razzanelli, M., Di Lauro, F., Rus, D. & Della Santina, C. Estimating the state of epidemics spreading with graph neural networks. *Nonlinear Dyn.* **109**, 249–263 (2022).
51. Panagopoulos, G., Nikolentzos, G. & Vazirgiannis, M. Transfer graph neural networks for pandemic forecasting. in *Proceedings of the AAAI Conference on Artificial Intelligence* vol. 35 4838–4845 (2021).
52. Cori, A. *et al.* Package 'epiestim'. *CRAN Vienna Austria* **13**, (2020).
53. Lloyd-Smith, J. O. Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases. *PloS One* **2**, e180 (2007).
54. Nouvellet, P. *et al.* Reduction in mobility and COVID-19 transmission. *Nat. Commun.* **12**, 1090 (2021).
55. Advan Research. Foot Traffic / Weekly Patterns.
56. Dewey Data. Weekly Patterns. https://doi.org/10.82551/x1pp-1f65.
57. Němcová, B., Goldstein, I. H., Sebastian, J., Minin, V. M. & Bracher, J. Unjustified Poisson assumptions lead to overconfident estimates of the effective reproductive number. *medRxiv* 2025–07 (2025).
58. Parag, K. V. Improved estimation of time-varying reproduction numbers at low case incidence and between epidemic waves. *PLoS Comput. Biol.* **17**, e1009347 (2021).
59. US Census Bureau Commuting. https://www.census.gov/topics/employment/commuting/guidance/commuting.html.
60. Population and Housing Unit Estimates Datasets. https://www.census.gov/programs-surveys/popest/data/data-sets.html.

61. Esposito, P. BLiTZ - Bayesian Layers in Torch Zoo (a Bayesian Deep Learing library for Torch). *GitHub repository* (2020).
62. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. https://doi.org/10.48550/ARXIV.1912.01703 (2019) doi:10.48550/ARXIV.1912.01703.
63. Fey, M. & Lenssen, J. E. Fast Graph Representation Learning with PyTorch Geometric. https://doi.org/10.48550/ARXIV.1903.02428 (2019) doi:10.48550/ARXIV.1903.02428.

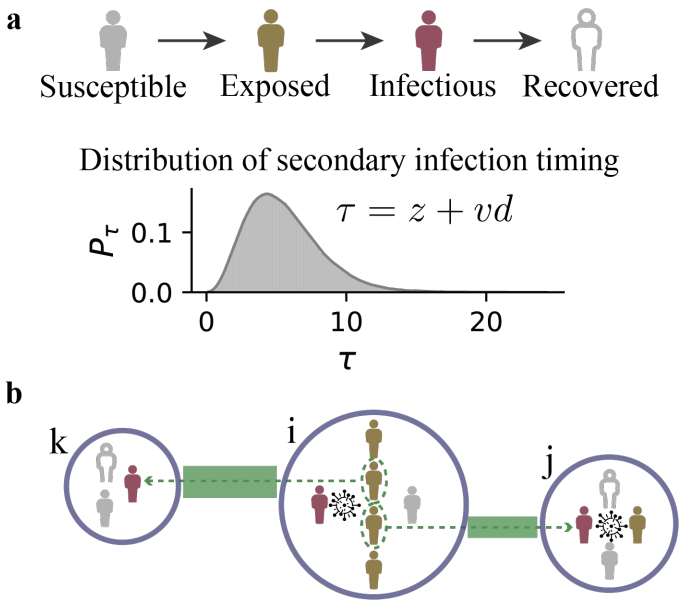

Fig. 1 Illustration of the branching process model dynamics. (**a**). Transition of individual states in the model. A person becomes infected when transitioning from susceptible to exposed state. Each individual has a random latent period of z days following a Gamma distribution, after which it becomes infectious and recovers after an additional d days drawn from a Gamma distribution. An infectious individual infects k individuals, each uniformly distributed over the infectious period. The number of secondary infections k was generated by multiplying a negative binomial random number by the fraction of susceptible population in that location. The timing for generating a secondary infection is $\tau = z + vd$, where v is drawn from a uniform distribution $U[0,1]$. The inset shows the probability distribution of P_τ. (**b**). A network diagram illustrating the spread of infectious disease across counties. In this diagram, each circle represents a county, and edges represent commuting flows between them. The grey, yellow, red, and hollow human icons represent susceptible, exposed, infectious, and recovered individuals, respectively. In county i, an infectious person (red) leads to four exposed individuals (yellow). Of these, two commute to counties j and k, respectively. In county j, the exposed individual becomes infectious (red) and further infects another person (yellow). In contrast, in county k, the exposed commuter becomes infectious (red) but does not result in any new infections.

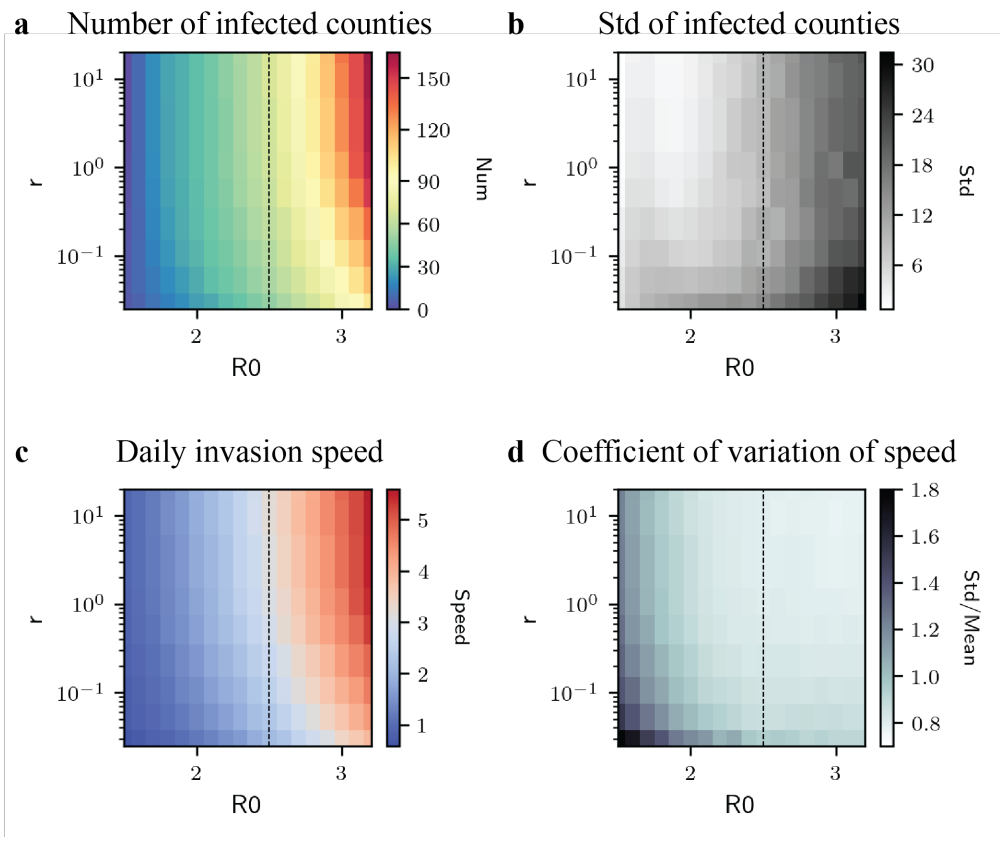

Fig. 2 Spatial spread patterns of the branching process model. (**a**). The number of infected counties (with daily new infections at least 10 per 100,000 people) by day 60 of simulations using various combinations of R_0 $(1.5 \leq R_0 \leq 3.3)$ and r $(0.025 \leq r \leq 20)$. For each parameter combination, we performed 300 independent simulations and presented the mean number of infected counties. The vertical dash line marks the parameter R_0 used in simulations shown in Fig. 3. (**b**). The standard deviation of the number of infected counties by day 60 for 300 independent simulations. (**c**). The daily invasion speed (i.e., the daily number of new counties with at least one infection) averaged over the 60-day period for outbreaks generated using different combinations of R_0 and r. (**d**). The coefficient of variation of daily invasion speed within the 60-day simulation window, defined as the standard deviation divided by the mean of speed. The coefficient of variation quantifies the relative variation of invasion speed around the mean speed during the 60-day period. Results were computed using 300 independent simulations.

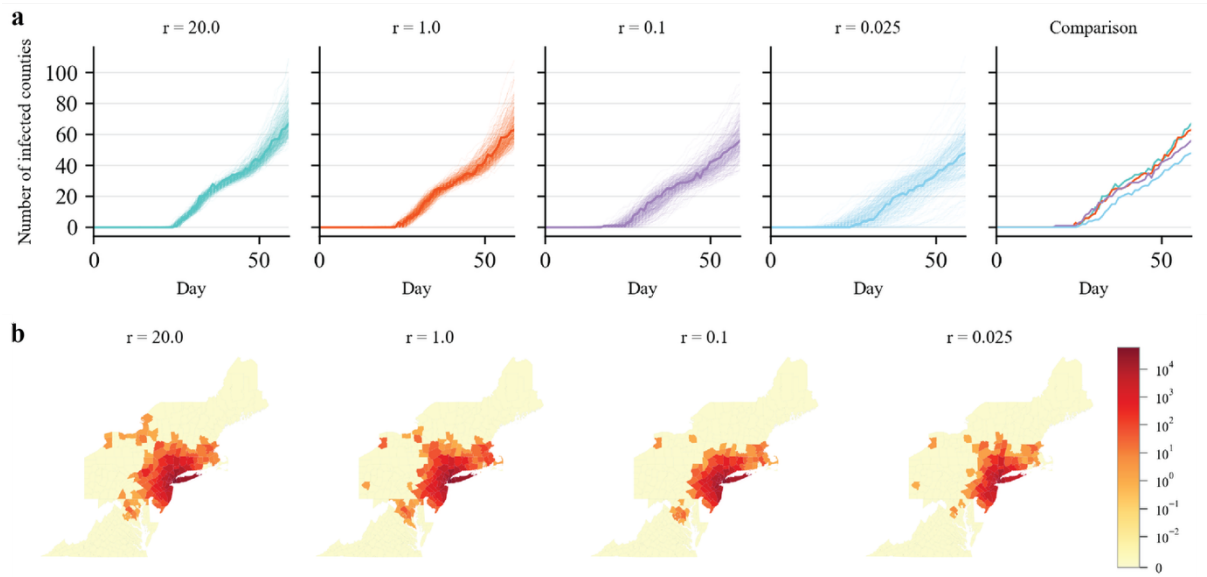

Fig. 3. Effect of superspreading on spatial spread. (a) We compare the number of infected counties (daily new infections exceeding 10 per 100,000 people) over time for branching process models with $R_0 = 2.5$ and varying dispersion parameters. A smaller r indicates stronger superspreading. For each r value, we performed 300 independent realizations. Thick solid lines show the trajectory for the realization that infected the median number of counties on T = 60 days. All the other realizations are plotted using thinner trajectories. The last panel compares the median trajectories for different dispersion parameters **(b)** Daily new infections in each county near the outbreak origin on day 60 of the simulation. Colors of counties indicate the daily new infections on day 60, generated using branching process models with different r. We visualize the outcome for the model realization that infected the median number of counties on T = 60 days (the thick solid lines in (a)). The outbreaks were initiated by 100 infections in New York County.

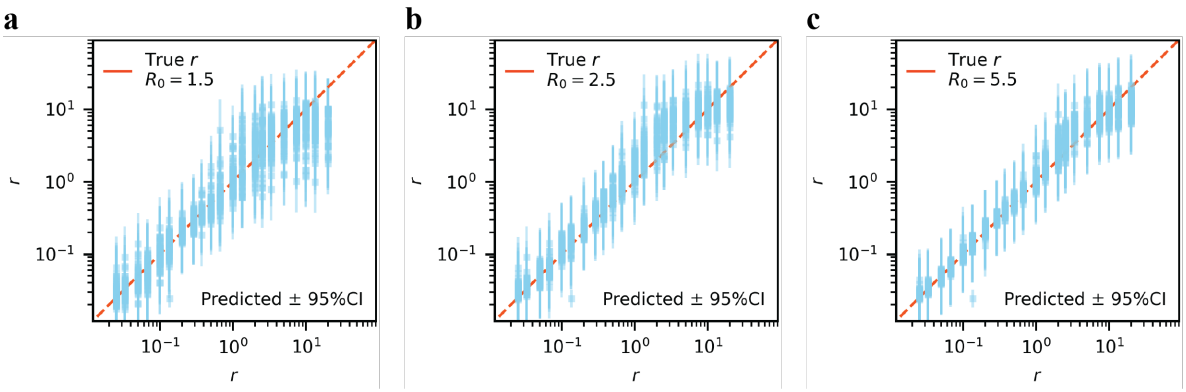

Fig. 4. Performance of the Bayesian graph neural networks for synthetic outbreaks with different R0. We generated three groups of synthetic outbreaks using $R_0 = 1.5$ **(a)**, 2.5 **(b)**, and 5.5 **(c)**. The x-axis shows the true values of the dispersion parameter r and the y-axis shows the inferred values for synthetic epidemics generated with different basic reproduction numbers R_0. A perfect inference would lie on the dashed red diagonal line. For each true r value, we generated 60 out-of-sample independent realizations of the outbreak and performed inference for each realization separately. For each parameter r, the plot overlays inference results from 60 independent inference. Blue dots indicate the median point estimates and vertical bars represent the 95% CIs from the Bayesian model.

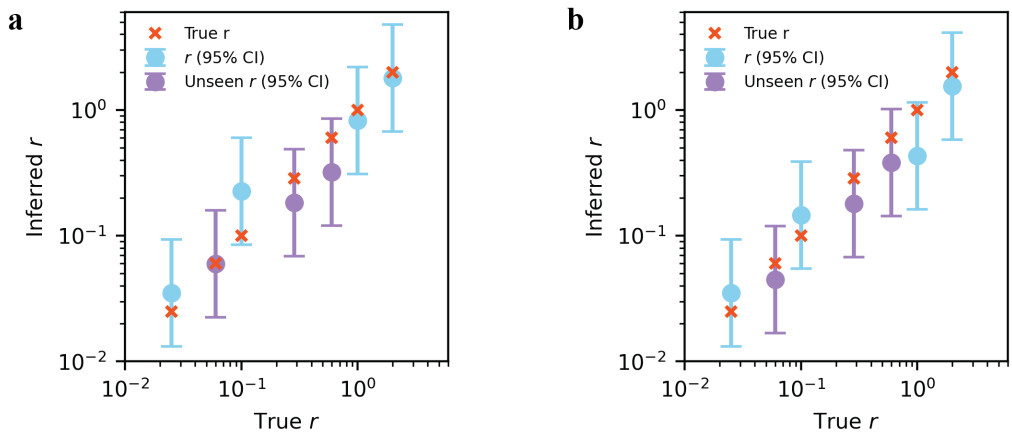

Fig. 5 Performance of the Bayesian graph neural networks for outbreaks with location-specific time-varying R_o and underreporting. We generated synthetic outbreaks using reproduction numbers that vary over space and time and used the daily new infections (without underreporting) to infer dispersion parameters **(a)**. The x-axis shows the ground-truth r values used in epidemic simulation and the y-axis shows the inference results. The red crosses show the true r values. The dots indicate the median point estimates, and the bars represent the 95% CIs. Blue color shows the r values included in the generation of training datasets, and purple color shows the values not used in model training (i.e., out-of-sample validation targets). We additional included underreporting rates that vary in different locations and time and used underreported case counts to perform inference. Results are shown in panel **(b)**.

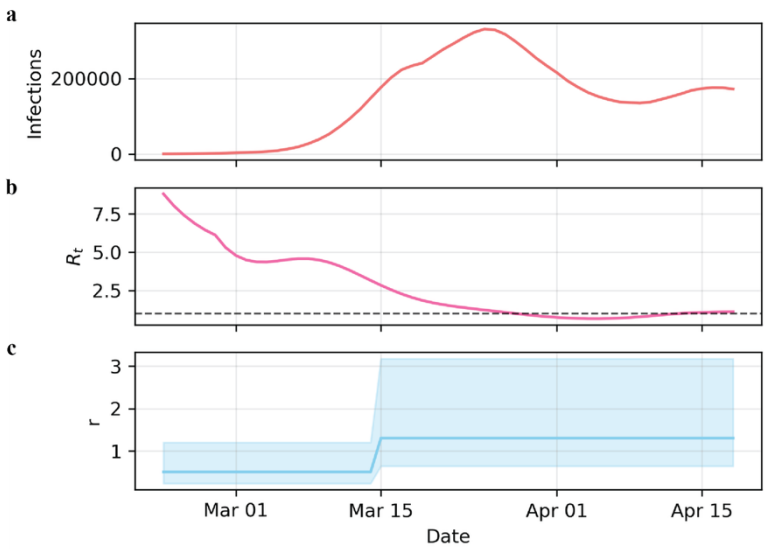

Fig. 6 Estimating superspreading during the early COVID-19 outbreak in the US. Panel **(a)** Estimated daily infections (both reported and unreported) across U.S. counties from 23 February to 18 April 2020, from Pei et al[24]. **(b)** The time-varying national effective reproduction number reconstructed using EpiEstim (dashed line marks $R_0 = 1$). **(c)** Estimates of the dispersion parameter r (smaller r indicates stronger superspreading) for the two periods: February 23rd–March 14th, 2020 (before the nationwide lockdown) and March 15th–April 18th, 2020 (after lockdown). Solid lines show posterior medians; shading areas show the estimated 95% CIs.

Supporting information for

Inferring pathogen superspreading potential using early spatial spread patterns

Qing Yao[1], Renquan Zhang[2], Tom Britton[3] and Sen Pei[1*]

1 Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University, New York, USA

2 School of Mathematical Sciences, Dalian University of Technology, Dalian, China

3 Department of Mathematic, Stockholm University, Stockholm, Sweden

*Corresponding author: sp3449@cumc.columbia.edu

Computation of stochastic extinction probability

The extinction probability q, is the probability that, starting from one infected person, the chain of transmission eventually dies out. It should be the smallest value satisfies $q = G(q)$, where $G(s)$ is the probability generating function (PGF) of a process and $s \in [0,1]$ is the dummy variable of the PGF. The shape of $G(s)$ depends on the basic reproduction number R_0. When $R_0 \leq 1$, the only solution is $q = 1$, meaning the extinction is certain. When $R_0 > 1$, a smaller solution $0 < q < 1$ exists, representing the probability that fewer individuals are infected and eventually no new infections.

In the case of Negative Binomial distribution, the corresponding PGF is:

$$G(s)^{\mathrm{NB}} = \sum_{k=0}^{\infty} P(X = k)s^k = \left(\frac{p}{1-(1-p)s}\right)^{-r},$$

When parameterized in R_0 and r,

$$G(s)^{\mathrm{NB}} = \left(1 + \frac{R_0}{r}(1-s)\right)^{-r}.$$

In the case of mixed Poisson distributions, the corresponding PGF is:

$$G(s)^{\mathrm{MP}} = \sum_{k=0}^{\infty} P(X = k)s^k = p_\lambda e^{\frac{R_0}{2}(s-1)} + (1-p_\lambda)e^{\lambda(s-1)}.$$

$s = 1$ is always a solution of the equation. To find other solutions for $q = G(q)^{\mathrm{NB}}$ and $q = G(q)^{\mathrm{MP}}$, we used bisection method to find the intersection of the two functions.

Last Layer Bayesian GCN architecture and inference

We implemented a Bayesian Graph Convolutional Network (GCN) to infer the dispersion parameter r from simulated epidemic data. The encoder consists of three graph convolutional layers that transform node features using the normalized mobility matrix:

$$H^{h+1} = \phi\left(\widetilde{\boldsymbol{M}}\boldsymbol{H}^h\boldsymbol{W}^h\right), \quad \mathrm{h} = 0, 1, 2,$$

where ϕ is the ELU nonlinearity and $\widetilde{M}$ is the normalized mobility matrix. After the final convolution, node embeddings are aggregated by global mean pooling:

$$z = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{H}_i^3.$$

The LayerNorm and dropout ($p = 0.5$) was applied to obtain the bottleneck representation $\tilde{z}$. The output layer is Bayesian linear regression layer, defined as:

$$f_w(\tilde{z}) = \boldsymbol{w}^T\tilde{z} + b,$$

where $\boldsymbol{w}$ is the vector of weights in the last Bayesian layer, with prior

$$p(\boldsymbol{w}) = \pi\mathcal{N}(0, \sigma_1^2 I) + (1-\pi)\mathcal{N}(0, \sigma_2^2 I).$$

The model predicted log dispersion rate y and assume a Gaussian likelihood for training labels.

During the training, variational inference is performed by maximizing the evidence lower bound (ELBO):

$$\mathcal{L}(\theta) = \sum_{n=1}^{M} \mathbb{E}_{q_\theta}[\log p(y_n|\tilde{z}_n, w, b)] - \beta \mathrm{KL}(q_\theta || p),$$

where q_θ is the approximate posterior distribution over the weight $\boldsymbol{w}$ and bias b. KL weight $\beta = 3 \times 10^{-3}$. To ensure numerical stability and to represent irreducible noise, a fixed variance floor ($\sigma_{\min} = 0.5$) was added to the Gaussian likelihood during training. This prevents the model from becoming overconfident.

For posterior prediction, we draw S Monte Carlo samples from q_θ and compute

$$y^{(s)} = f_{w^{(s)}}(\tilde{Z}^*) = \log(\mathrm{r}).$$

Then we get the posterior mean $\mu_{\log} = \frac{1}{S} \sum_S y^{(s)}$ and epistemic variance $v_{\mathrm{ep}} = \mathrm{Var}(y^{(s)})$. To ensure numerical stability, we added a minimum value a floor sigma $\sigma_{\min} = 0.5$ to predicted variances:

$$v_{\mathrm{tot}} = v_{\mathrm{ep}} + {\sigma_{\min}}^2.$$

Then we assume $\log r \sim \mathcal{N}(\mu_{\log}, v_{\mathrm{tot}})$, and report mean $\mathbb{E}[r] = \exp(\mu_{\log} + \frac{1}{2} v_{\mathrm{tot}})$, median $\exp(\mu_{\log})$, $\mathrm{sd}(\mathrm{r}) = \sqrt{(\exp(v_{\mathrm{tot}})\text{-}1)(\exp(2\mu_{\log}+v_{\mathrm{tot}}))}$ and 95% confidence interval $\mathrm{CI}_{0.95}(r) = [\exp(\mu_{\log} - z\sigma_{\mathrm{tot}}), \exp(\mu_{\log} + z\sigma_{\mathrm{tot}})]$, with $z = 1.9599639845$ and $\sigma_{\mathrm{tot}} = \sqrt{v_{\mathrm{tot}}}$.

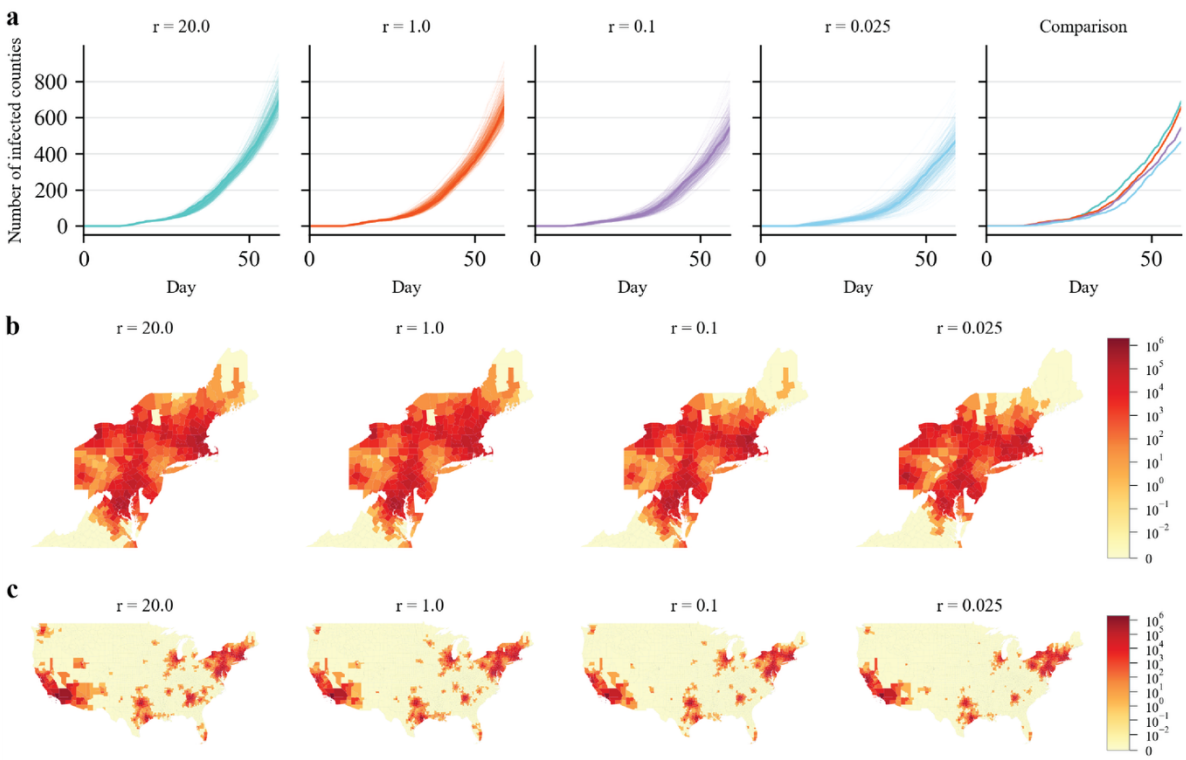

Fig. S1. Effect of superspreading on spatial spread for $R_0 = 5.5$. (a) shows the number of counties with at least 10 daily new infections per 100,000 people over 60 days. For each parameter setting, we performed 300 independent simulations. Thick solid lines show the trajectory for the realization that infected the median number of counties on T=60 days. All the other realizations are plotted using thinner trajectories. **(b)** and **(c)** show daily new infections in each county on day 60 in the northeastern and mainland US, respectively. We visualize the simulation whose number of infected counties on day 60 is the median of the ensemble.

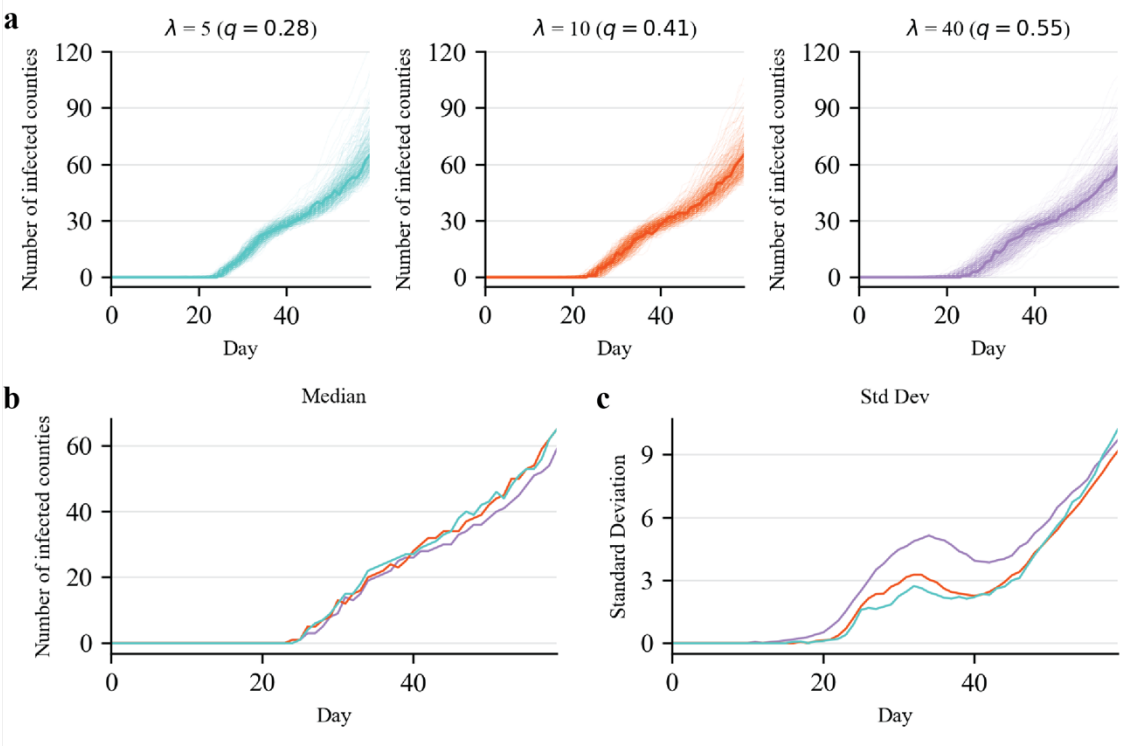

Fig. S2. Simulations using Poisson mixture distributions for the number of secondary infections. We simulate outbreaks with $R_0 = 2.5$ originating from New York County, starting from 100 initial infections. Simulations were performed for four different λ values. (**a**) The plots show the number of counties with at least 10 daily new infections per 100,000 people over 60 days. For each parameter setting, we performed 300 independent simulations. The q values in the parentheses are extinction probabilities for different λ values. Thick solid lines show the median trajectory among the 300 realizations (minimizes total L1 distance to others). All the other realizations are plotted using thinner trajectories. (**b**) Comparison of the median trajectories. (**c**) The standard deviation of number of infected counties on each day across 300 simulations.

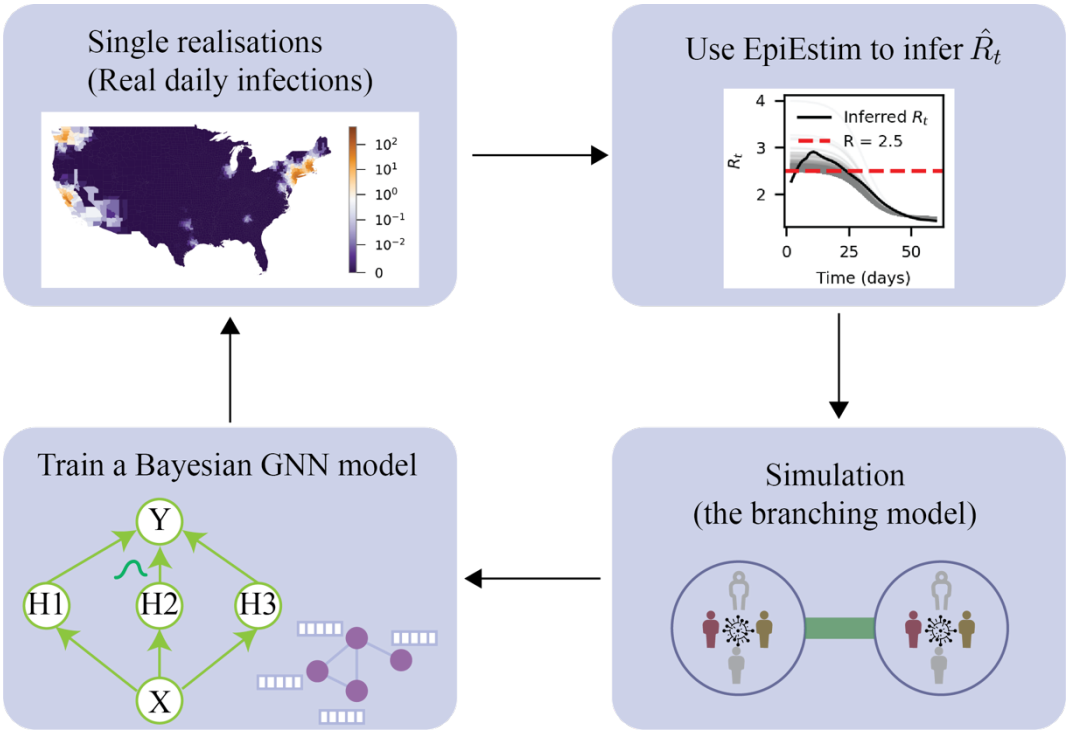

Fig. S3. The inference pipeline. Schematic overview of the workflow for inferring the dispersion parameter r. Realized daily infections across US counties are first used with EpiEstim to estimate the time-varying reproduction number $\hat{R}_t$. These estimates parameterize the branching-process simulation with various level of superspreading potentials. The simulated spatiotemporal infection data are then used to train a Bayesian GNN model, which learns to predict the dispersion parameter from simulated incidence data. The trained model is finally applied to the real daily infection data to infer superspreading potential.

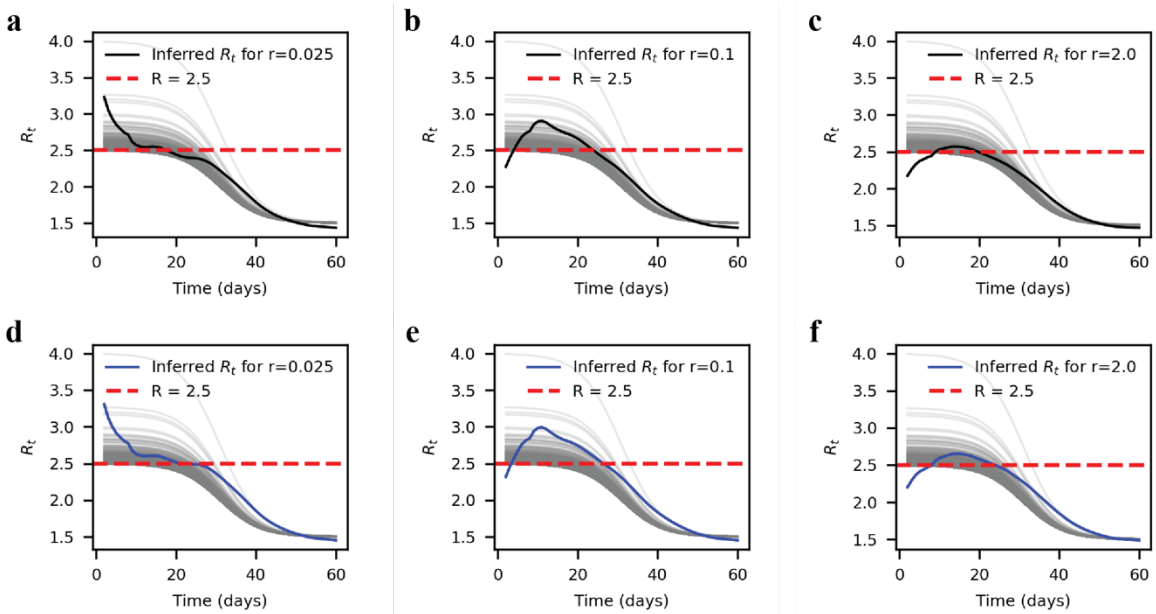

Fig. S4. Estimated national reproduction numbers compared with location-specific reproduction numbers. Panels **(a–c)** show the national reproduction numbers estimated using EpiEstim for simulated outbreaks with dispersion parameters $r = 0.025$, $r = 0.1$ and $r = 2.0$, respectively. Grey lines indicate the actual county-specific reproduction numbers $R_{t,\ell}$ used in generating the outbreaks. Solid black lines represent the estimated national-level reproduction number from the aggregated daily new infections across all counties. (**d-f**). We additionally incorporated county-specific, time-varying reporting rate to generate observed daily infection data with underreporting. Solid blue lines show the reproduction number inferred from underreported observed cases. The estimated national reproduction numbers obtained from EpiEstim captured the general temporal trend of $R_{t,\ell}$, regardless of underreporting.

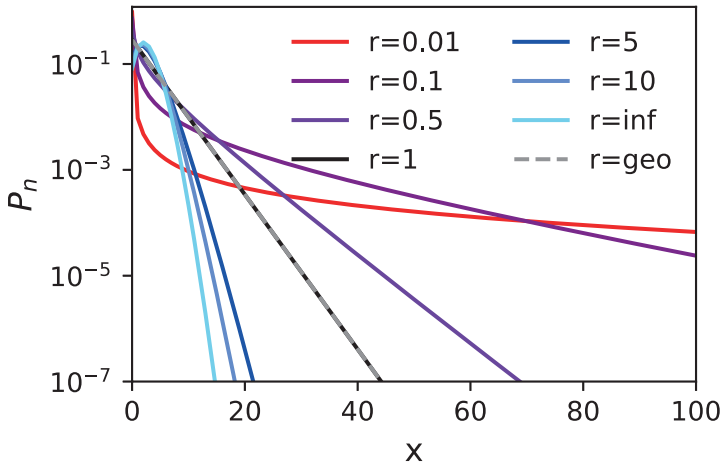

Fig. S5. The probability density function of negative binomial distributions. The distribution of a smaller r has a longer tail and indicates a higher probability of superspreading events.

Table S1. Estimated basic reproduction numbers and dispersion parameters for a list of common infectious diseases.

Pathogens	Estimated basic reproduction R_0	Estimated dispersion rate r	References
SARS-CoV-2	2.03–2.77	0.06–2.97	[1], [2]
SARS	0.94–1.88	0.11–0.64	[3]
MERS-CoV	0.24–1.3	0.045-0.14	[4], [5]
Smallpox	0.8–3.91	0.29–2.32	[3]
Mpox	0.21–0.42	0.14–3.57	[3], [6]
Measles	11-18	0.12-0.65	[3], [7]

Table S2. Extinction probabilities for negative binomial distributions and Poisson mixture distributions.

Distribution	Parameter	$P(X = 0)$	q
Negative Binomial	$r = 0.025$	0.89	0.96
	$r = 0.1$	0.72	0.86
	$r = 1.0$	0.29	0.4
	$r = 20.0$	0.095	0.13
Poisson mixture	$\lambda = 200$	0.28	0.61
	$\lambda = 40$	0.28	0.55
	$\lambda = 10$	0.25	0.41
	$\lambda = 5$	0.19	0.28

References

[1] Li, Ruiyun, et al. "Substantial undocumented infection facilitates the rapi0.045-0d dissemination of novel coronavirus (SARS-CoV-2)." Science 368.6490 (2020): 489-493.

[2] Du, Zhanwei, et al. "Systematic review and meta‐analyses of superspreading of SARS‐CoV‐2 infections." Transboundary and Emerging Diseases 69.5 (2022): e3007-e3014.

[3] Lloyd-Smith, James O., et al. "Superspreading and the effect of individual variation on disease emergence." Nature 438.7066 (2005): 355-359.

[4] Cauchemez, Simon, et al. "Middle East respiratory syndrome coronavirus: quantification of the extent of the epidemic, surveillance biases, and transmissibility." The Lancet infectious diseases 14.1 (2014): 50-56.

[5] Choe, Seoyun, Hee-Sung Kim, and Sunmi Lee. "Exploration of superspreading events in 2015 MERS-CoV outbreak in Korea by branching process models." International Journal of Environmental Research and Public Health 17.17 (2020): 6137.

[6] Blumberg, Seth, and James O. Lloyd-Smith. "Inference of R0 and transmission heterogeneity from the size distribution of stuttering chains." PLoS computational biology 9.5 (2013): e1002993.

[7] Guerra, Fiona M., et al. "The basic reproduction number (R0) of measles: a systematic review." The Lancet Infectious Diseases 17.12 (2017): e420-e428.