Time-dependent heterogeneity leads to transient suppression of the COVID-19 epidemic, not herd immunity

Alexei V. Tkachenko[a,1], Sergei Maslov[b,c,d,1], Ahmed Elbanna[e,f], George N. Wong[b], Zachary J. Weiner[b], and Nigel Goldenfeld[b,d]

[a]Center for Functional Nanomaterials, Brookhaven National Laboratory, Upton, NY 11973; [b]Department of Physics, University of Illinois at Urbana–Champaign, Urbana, IL 61801; [c]Department of Bioengineering, University of Illinois at Urbana–Champaign, Urbana, IL 61801; [d]Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana–Champaign, Urbana, IL 61801; [e]Department of Civil Engineering, University of Illinois at Urbana–Champaign, Urbana, IL 61801; and [f]Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL 61801

Edited by Andrea Rinaldo, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, and approved March 2, 2021 (received for review July 28, 2020)

Epidemics generally spread through a succession of waves that reflect factors on multiple timescales. On short timescales, superspreading events lead to burstiness and overdispersion, whereas long-term persistent heterogeneity in susceptibility is expected to lead to a reduction in both the infection peak and the herd immunity threshold (HIT). Here, we develop a general approach to encompass both timescales, including time variations in individual social activity, and demonstrate how to incorporate them phenomenologically into a wide class of epidemiological models through reparameterization. We derive a nonlinear dependence of the effective reproduction number R_e on the susceptible population fraction S. We show that a state of transient collective immunity (TCI) emerges well below the HIT during early, high-paced stages of the epidemic. However, this is a fragile state that wanes over time due to changing levels of social activity, and so the infection peak is not an indication of long-lasting herd immunity: Subsequent waves may emerge due to behavioral changes in the population, driven by, for example, seasonal factors. Transient and long-term levels of heterogeneity are estimated using empirical data from the COVID-19 epidemic and from real-life face-to-face contact networks. These results suggest that the hardest hit areas, such as New York City, have achieved TCI following the first wave of the epidemic, but likely remain below the long-term HIT. Thus, in contrast to some previous claims, these regions can still experience subsequent waves.

COVID-19 | heterogeneity | overdispersion | epidemic theory

The COVID-19 pandemic is nearly unprecedented in the level of disruption it has caused globally, but also, potentially, in the degree to which it will change our understanding of epidemic dynamics and the efficacy of various mitigation strategies. Ever since the pioneering works of Kermack and McKendrick (1), epidemiological models have been widely and successfully used to quantify and predict the progression of infectious diseases (2–6). More recently, the important role in epidemic spreading played by population heterogeneity and the complex structure of social networks has been appreciated and highlighted in multiple studies (7–25). However, integration of this conceptual progress into reliable, predictive epidemiological models remains a formidable task. Among the key effects of heterogeneity and social network structure are 1) the role played by superspreaders and superspreading events during initial outbreaks (12, 13, 16, 26–29) and 2) a substantial reduction of the final size of epidemic (FSE) as well as the herd immunity threshold (HIT) compared to the homogeneous case (9, 10, 19–21, 23, 30). The COVID-19 pandemic has reignited interest in the effects of heterogeneity of individual susceptibility to the disease, in particular, to the possibility that it might lower both HIT and FSE (31–35). In studying epidemics in heterogeneous populations, it is important to emphasize the qualitative nature of two important timescales. First, overdispersion is dominated by short-term patterns of behavior and even accidental events rather than persistent population behavioral heterogeneity. Second, short-term overdispersion is generally assumed to have a limited impact on the long-term epidemic dynamics, being important primarily during early outbreaks dominated by superspreading events. In this paper, we attempt to provide a multiscale theory for epidemic progression and show that both overdispersion and persistent heterogeneity affect the overall progression of the COVID-19 epidemic. The significance of this multiscale perspective is that it provides a natural formalism to predict the occurrence and nature of successive epidemic waves, even when it might seem that a first wave has attained a state which could be mistaken for herd immunity.

There are several existing approaches to model the effects of heterogeneity on epidemic dynamics, each focusing on a different characteristic and parameterization. In the first approach, one stratifies the population into several demographic groups (e.g., by age), and accounts for variation in susceptibility of these groups and their mutual contact probabilities (2). While

Significance

Epidemics generally spread through a succession of waves that reflect factors on multiple timescales. Here, we develop a general approach bridging across these timescales and demonstrate how to incorporate population heterogeneity into a wide class of epidemiological models. We demonstrate that a fragile state of transient collective immunity emerges during early, high-paced stages of the epidemic, leading to suppression of individual epidemic waves. However, this state is not an indication of lasting herd immunity: Subsequent waves may emerge due to stochastic changes in individual social activity. Parameters of transient collective immunity are estimated using empirical data from the COVID-19 epidemic in several US locations.

Author contributions: A.V.T., S.M., A.E., and N.G. designed research; A.V.T., S.M., A.E., G.N.W., Z.J.W., and N.G. performed research; A.V.T., S.M., A.E., G.N.W., Z.J.W., and N.G. analyzed data; and A.V.T., S.M., A.E., G.N.W., Z.J.W., and N.G. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: oleksiyt@bnl.gov or maslov@illinois.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2015972118/-/DCSupplemental.

Published April 8, 2021.

POPULATION BIOLOGY

this approach represents many aspects of population dynamics beyond the homogeneous and well-mixed assumption, it clearly does not encompass the whole complexity of individual heterogeneity, interpersonal communications, and spatial and social structures. These details can be addressed in a second approach, where one analyzes the epidemic dynamics on real-world or artificial social networks (8, 13, 23, 36, 37). Through elegant mathematics, it is possible to obtain detailed results in idealized cases, including the mapping onto well-understood models of statistical physics such as percolation (9, 38). As demonstrated in ref. 21, the FSE is a very robust property of the epidemic, insensitive to fine details of its dynamics (39) defined by 1) distribution of susceptibilities in the population (20, 30, 40) and 2) correlations between infectivity and susceptibility. Importantly, it does not depend on the part of heterogeneous infectivity that is not correlated with susceptibility. However, these approaches, so far, have been mostly limited to the analysis of the FSE and distribution of outbreak sizes on a static social network.

For practical purposes, it is desirable to predict the complete time-dependent dynamics of an epidemic, preferably by explicitly including heterogeneity into classical well-mixed mean-field compartmentalized models. This approach was developed some time ago in the context of epidemics on networks (10, 23), and the resulting mean-field theory effectively reproduces the structure of heterogeneous well-mixed models extensively studied in the applied mathematics literature (19–22, 24, 30). The overall impact of heterogeneity occurs because, as the disease spreads, it preferentially immunizes the more susceptible individuals, so the remaining population is less susceptible, and spread is inhibited. This effect is further enhanced by a positive correlation between infectivity and susceptibility. In the context of static network models, this correlation is perfect, since both factors are proportional to the degree of individual nodes. Ref. 24 studied a hybrid model in which social heterogeneity represented by network degree was combined with a biological one. These approaches have been recently applied in the context of COVID-19 (31, 32, 35, 41, 42). The conclusion of these studies was that the HIT may be well below that expected in classical homogeneous models.

These approaches to heterogeneity delineate end-members of a continuum of theories: overdispersion describing short-term, bursty dynamics (e.g., due to superspreader accidents), as opposed to *persistent heterogeneity*, which is a long-term characteristic of an individual and reflects behavioral propensity, for example, to socialize in large gatherings without prudent social distancing. Overdispersion is usually modeled in terms of a negative binomial branching process (12, 13, 16, 26–28). Strictly speaking, both persistent heterogeneity and short-term variations contribute to the overdispersion of individual reproduction number. However, we will see below that the former is likely to be a much weaker source of variation compared to the latter. It is also generally presumed that short-term overdispersion is uncorrelated in time and thus has no effect on epidemic dynamics. Indeed, large variations in an individual's infectivity would average out as long as they are not correlated with susceptibility. But, since the initial exposure and secondary infections are separated by a single generation interval (typically about 5 d for COVID-19), the levels of individual social activity at those times are expected to be correlated, and (at least partially) reflect short-term overdispersion. How, then, can one understand the epidemic progression across various timescales, from early stages of a fast exponential growth to the final state of the herd immunity?

Below, we present a comprehensive yet simple theory that accounts for both social and biological aspects of heterogeneity, and predicts how, together, they modify early and intermediate epidemic dynamics, as well as global characteristics of the epidemic such as its HIT. Importantly, early epidemic dynamics may be sensitive to both persistent heterogeneity and short-term overdispersion. As a result, the apparent early-stage heterogeneity is typically enhanced compared to its long-term persistent level. This may lead to a suppression of the first wave of the epidemic due to reaching a novel state that we call transient collective immunity (TCI) determined by a combination of short-term and long-term heterogeneity, whose threshold is lower than the eventual HIT. The implication is that the first wave of an epidemic ends due to a combination of both persistent heterogeneity and whatever mitigation constraints are imposed on the population. As the latter are relaxed by authorities or through behavioral changes associated with seasonal factors, subsequent waves can still occur. Thus, TCI is dramatically different from the idea of herd immunity due to heterogeneity.

Our starting point is a generalized version of the heterogeneous well-mixed theory in the spirit of ref. 10, but we use the age-of-infection approach (1) rather than compartmentalized susceptible, infectious, recovered (SIR)/susceptible, exposed, infectious, recovered (SEIR) models of epidemic dynamics (see, e.g., ref. 2). Similar to multiple previous studies, we first completely ignore any time dependence of individual susceptibilities and infectivities, focusing only on their long-term persistent components. This approach implicitly assumes that any short-term overdispersion (responsible for, e.g., the superspreading phenomenon) is uncorrelated in time and thus effectively averaged out. This is a valid assumption if the large short-term variations in individual infectivity are completely uncorrelated with an individual's susceptibility. However, this approximation is hard to justify in the case of COVID-19. Indeed, if the two are correlated on the timescale of a single generation interval (5 d), this will strongly affect the overall epidemic dynamics. Therefore, our initial analysis is eventually expanded to a more general theory accounting for the nonnegligible effects of short-term overdispersion. In the case of persistent heterogeneity, we demonstrate how the model can be recast into an effective homogeneous theory that can readily encompass a wide class of epidemiological models, including various versions of the popular SIR/SEIR approaches. Specific innovations that emerge from our analysis are the nonlinear dependence of the effective reproduction number R_e on the overall population fraction S of susceptible individuals, and another nonlinear function S_e that gives an effective susceptible fraction, taking into account preferential removal of highly susceptible individuals.

A convenient and practically useful aspect of this approach is that it does not require extensive additional calibration in order to be applied to real data. In the effort to make quantitative predictions from epidemic models, accurate calibration is arguably the most difficult step, but is necessary due to the extreme instability of epidemic dynamics in both growth and decay phases (43, 44). We find that, with our approach, the entire effect of heterogeneity is, in many cases, well characterized by a single parameter which we call the *immunity factor* λ. It is related to the statistical properties of heterogeneous susceptibility across the population and to its correlation with individual infectivity. The immunity factor determines the rate at which R_e drops during the early stages of the epidemic as the pool of susceptibles is being depleted: $R_e \approx R_0(1-\lambda(1-S))$. Beyond this early linear regime, for an important case of gamma-distributed individual susceptibilities, we show that the classical proportionality, $R_e = R_0 S$, transforms into a power law scaling relationship $R_e = R_0 S^{\lambda}$. This leads to a modified version of the result for the HIT, $1-S_0 = 1-R_0^{-1/\lambda}$.

Heterogeneity in the susceptibility of individual members of the population has several different contributions: 1) biological, which takes into account differences in factors such as strength of immune response, genetics, age, and comorbidities; and 2) social, reflecting differences in the number and frequency of close contacts of different people. The immunity factor λ in our model combines these sources of heterogeneous susceptibility as well as its correlation with individual infectivity. As we demonstrate, under certain assumptions, the immunity factor is simply a product of social and biological contributions: $\lambda = \lambda_s \lambda_b$. In our study, we leverage existing studies of real-life face-to-face contact networks (13, 19, 37, 45–48) to estimate the social contribution to heterogeneous susceptibility, and the corresponding immunity factor λ_s. The biological contribution, λ_b, is expected to depend on specific details of each infection.

To test this theory, we use empirical data for the COVID-19 epidemic to independently estimate the immunity factor λ. In particular, we apply our previously described epidemic model that features multichannel Bayesian calibration (43) to describe epidemic dynamics in New York City (NYC) and Chicago from the start of the epidemic in mid-March until the end of the first wave around June 15, 2020. This model uses high-quality data on hospitalizations, intensive care unit (ICU) occupancy, and daily deaths to extract the underlying $R_e(S)$ dependence in each of two cities. In addition, we perform a similar analysis of data on individual states in the United States, using data generated by the model in ref. 49. Using both approaches, we find that the locations that were severely impacted by the COVID-19 epidemic show a more pronounced reduction of the effective reproduction number. This effect is much stronger than predicted by classical homogeneous models, suggesting a significant role of heterogeneity. The estimated immunity factor ranges between four and five. Importantly, this represents a transient value of the parameter λ observed on intermediate timescales and dependent on both persistent and short-term heterogeneity. Our estimates of the long-term value of the immunity factor defined by persistent heterogeneity only is considerably lower, about two. This difference explains why achieving the state of TCI after the first wave of the epidemic does not imply long-term herd immunity.

Finally, we integrate the persistent heterogeneity theory into our time-of-infection epidemiological model (43), and project possible outcomes of the second wave of the COVID-19 epidemic during the summer months in NYC and Chicago, using data up to the end of June 2020. By considering the worst-case scenario of a full relaxation of any currently imposed mitigation, we find that the results of the heterogeneity-modified model significantly modify the results from the homogeneous mode. In particular, based on our estimate of the immunity factor, our model predicts no second wave in NYC immediately after release of mitigations in June and up to September 2020, indicating that the TCI has likely been achieved there. Chicago, on the other hand, has not passed the TCI threshold that we infer, but the effects of heterogeneity would still result in a substantial reduction of the magnitude of the second wave there, even under the worst-case scenario. This, in turn, suggests that the second wave can be completely eliminated in such medium-hit locations, if appropriate and economically mild mitigation measures are adopted, including, for example, mask wearing, contact tracing, and targeted limitation of potential superspreading events, through limitations on indoor bars, dining, and other venues. We further investigate the issue of fragility of collective immunity in heterogeneous populations. By allowing rewiring of the social network with time, we demonstrate that the TCI may wane, much like an individual's acquired immunity may wane due to biological factors. This phenomenon would result in the emergence of secondary epidemic waves after a substantial period of low infection counts.

Theory of Epidemics in Populations with Persistent Heterogeneity

Following in the footsteps of refs. 10, 19, 22–24, 30, and 32, we consider the spread of an epidemic in a population of individuals who exhibit significant persistent heterogeneity in their susceptibilities to infection α. This heterogeneity may be biological or social in origin, and we assume these factors are independent: $\alpha = \alpha_b \alpha_s$. Effects of possible correlations between α_b and α_s have been discussed in ref. 24. The biologically driven heterogeneous susceptibility α_b is shaped by variations of several intrinsic factors such as the strength of individuals' immune responses, age, or genetics. In contrast, the socially driven heterogeneous susceptibility α_s is shaped by extrinsic factors, such as differences in individuals' social interaction patterns (their degree in the network of social interactions). Furthermore, individuals' different risk perceptions and attitudes toward social distancing may further amplify variations in socially driven susceptibility heterogeneity. We only focus on susceptibility that is a persistent property of an individual. For example, people who have elevated occupational hazards, such as health care workers, typically have higher, steady values of α_s. Similarly, people with low immune response, highly social individuals (hubs in social networks), or scofflaws would all be characterized by above-average overall susceptibility α.

In this work, we group individuals into subpopulations with similar values of α and describe the heterogeneity of the overall population by the probability density function (pdf) of this parameter, $f(\alpha)$. Since α is a relative measure of individual susceptibilities, without loss of generality, we set $\langle\alpha\rangle \equiv \int_0^\infty \alpha f(\alpha)\, d\alpha = 1$. Each person is also assigned an individual reproduction number R_i, which is an expected number of people that this person would infect in a fully susceptible population with $\langle\alpha\rangle = 1$. Accordingly, from each subpopulation with susceptibility α, there is a respective mean reproductive number R_α to which we refer as infectivity throughout this study. Any correlations between individual susceptibility and infectivity will significantly impact the epidemic dynamics. Such correlations are an integral part of most network-based epidemiological models, due to the assumed reciprocity in underlying social interactions, which leads to $R_\alpha \approx \alpha$ (9, 10, 23). In reality, not all transmissions involve face-to-face contacts, and biological susceptibility need not be strongly correlated with infectivity. Therefore, it is reasonable to expect only a partial correlation between α and R_α.

Let $S_\alpha(t)$ be the fraction of susceptible individuals in the subpopulation with susceptibility α, and let $j_\alpha(t) = -\dot{S}_\alpha$ be the corresponding daily incidence rate per capita in that subpopulation. At the start of the epidemic, we assume everyone is susceptible to infection: $S_\alpha(0) = 1$. The course of the epidemic is described by the following age-of-infection model:

$$-\frac{dS_\alpha}{dt} = j_\alpha(t) = \alpha S_\alpha(t) J(t) \qquad \textbf{[1]}$$

$$J(t) = \int_0^\infty \langle R_\alpha K(\tau) j_\alpha(t-\tau) \rangle \, d\tau. \qquad \textbf{[2]}$$

Here t is the physical time, and τ is the time since infection for an individual; $\langle \ldots \rangle$ represents averaging over α with pdf $f(\alpha)$. $J(t)$ is the force of infection, that is, per capita incidence rate in a fully susceptible subpopulation with $\alpha = 1$. R_α is the previously introduced infectivity, that is, the mean reproductive number of the subpopulation with susceptibility α. $K(\tau)$ is the pdf of the generation interval, which we assume to be independent of α, for the sake of simplicity. The homogeneous version of the age-of-infection model was introduced in the early days of mathematical epidemiology in the classical paper by Kermack and McKendrick (1). It is based on the observation that the force of infection, $J(t)$, is defined by the number of previously infected individuals. The contribution of each individual depends on his/her time since infection τ and is weighted by the infectivity profile $K(\tau)$. As shown in ref. 50, the rate of the exponential growth of the epidemic can be inferred from the Laplace transform of $K(\tau)$.

Well-known compartmentalized models correspond to specific functional forms of $K(\tau)$, such as the exponential distribution for the SIR model.

According to Eq. **1**, the fraction of susceptibles in the subpopulation with any given α can be expressed as

$$S_\alpha(t) = \exp(-\alpha Z(t)). \qquad \textbf{[3]}$$

Here $Z(t) \equiv \int_0^t J(t')dt'$. The total susceptible fraction of the population is related to the moment generating function M_α of the distribution $f(\alpha)$ (i.e., the Laplace transform of $f(\alpha)$) according to

$$S(t) = \int_0^\infty f(\alpha)e^{-\alpha Z(t)}d\alpha = M_\alpha(-Z(t)). \qquad \textbf{[4]}$$

Similarly, the effective reproductive number R_e can be expressed in terms of the parameter Z,

$$R_e(t) = \frac{1}{\langle\alpha\rangle}\int_0^\infty \alpha R_\alpha f(\alpha)e^{-\alpha Z(t)}d\alpha. \qquad \textbf{[5]}$$

Note that, for $Z = 0$, this expression gives the basic reproduction number $R_0 = \langle\alpha R_\alpha\rangle/\langle\alpha\rangle$. This result is reminiscent of the well-known result (23) $R_0 = \langle k_{in}k_{out}\rangle/\langle k_{in}\rangle$ for epidemic spread on a directed network with in and out degrees k_{in} (analogue of our susceptibility α) and k_{out} (analogue of R_α). Note that, due to our choice of normalization $\langle\alpha\rangle = 1$, the prefactor in Eq. **5** can be omitted. Since both S and R_e depend on time only through $Z(t)$, Eqs. **4** and **5** establish a parametric relationship between these two important quantities during the time course of an epidemic. In contrast to the classical case when these two quantities are simply proportional to each other, that is, $R_e = SR_0$, the relationship in the present theory is nonlinear due to heterogeneity. Eq. **5** was derived by substituting Eq. **1** into Eq. **2**. This leads to the renewal equation for $J(t)$ of the same form as in a homogeneous problem,

$$J(t) = \int_0^\infty d\tau K(\tau)R_e(t-\tau)J(t-\tau). \qquad \textbf{[6]}$$

Furthermore, by averaging Eq. **1** over all values of α, one arrives at the following heterogeneity-induced modification to the relationship between the force of infection and incidence rate:

$$\frac{dS}{dt} = -S_e J. \qquad \textbf{[7]}$$

Here

$$S_e(t) = \int_0^\infty \alpha f(\alpha)e^{-\alpha Z(t)}d\alpha = -\frac{dM_\alpha(-Z(t))}{dZ} \qquad \textbf{[8]}$$

is the effective susceptible fraction of the population, which is less than S, due to the disproportionate removal of highly susceptible individuals. Just as with R_e, it is a nonlinear function of S, defined parametrically by Eqs. **4** and **8**. Further generalization of this theory for the time-modulated age-of-infection model is presented in *SI Appendix*. There, we also discuss the adaptation of this approach for the important special case of a compartmentalized SIR/SEIR model.

One of the striking consequences of the nonlinearity of $R_e(S)$ is that the effective reproduction number could be decreasing at the early stages of an epidemic significantly faster than predicted by homogeneous models. Specifically, for $1 - S(t) \simeq Z(t) \ll 1$, one can linearize the effective reproduction number as

$$R_e \approx R_0(1 - \lambda(1 - S)). \qquad \textbf{[9]}$$

We named the coefficient λ the *immunity factor* because it quantifies the effect that a gradual buildup of population immunity has on the spread of an epidemic. The classical value of λ is one, but it may be significantly larger in a heterogeneous case. By linearizing Eq. **5** in terms of $1 - S \simeq Z \ll 1$ and dividing the result by $R_0 = \langle\alpha R_\alpha\rangle$, one gets

$$\lambda = \frac{\langle\alpha^2 R_\alpha\rangle}{\langle\alpha R_\alpha\rangle}. \qquad \textbf{[10]}$$

As one can see, the value of the immunity factor thus depends both on the statistics of susceptibility α and on its correlation with infectivity R_α.

We previously defined the overall susceptibility as a combination of biological and social factors: $\alpha = \alpha_s\alpha_b$. Here α_s is a measure of the overall social connectivity or activity of an individual, such as the cumulative time of close contact with other individuals averaged over a sufficiently long time interval (known as node strength in network science). Since the contribution of interpersonal contacts to an epidemic spread is mostly reciprocal, we assume $R_\alpha \approx \alpha_s$. On the other hand, in our analysis, we neglect a correlation between biological susceptibility and infectivity, as well as between α_b and α_s. Under these approximations, the immunity factor itself is a product of biological and social contributions, $\lambda = \lambda_b\lambda_s$. Each of them can be expressed in terms of leading moments of α_b and α_s, respectively,

$$\lambda_b = \frac{\langle\alpha_b^2\rangle}{\langle\alpha_b\rangle^2} = 1 + CV_b^2 \qquad \textbf{[11]}$$

$$\lambda_s = \frac{\langle\alpha_s^3\rangle}{\langle\alpha_s\rangle\langle\alpha_s^2\rangle} = 1 + \frac{CV_s^2(2 + \gamma_s CV_s)}{1 + CV_s^2}. \qquad \textbf{[12]}$$

These equations follow from Eq. **10** in the limit $R_\alpha = \text{const}$ and $R_\alpha \approx \alpha$, respectively. Although these equations resemble classical results for R_0 in heterogeneous networks (8–10, 23), here they describe a completely different effect of suppression of R_e in response to depletion of susceptible population S. That is why λ_s in Eq. **12** is proportional to the third moment of α_s instead of the second moment in the case of $R_0 = \langle\alpha R_\alpha\rangle \approx \alpha_s^2$. Note that the biological contribution to the immunity factor depends only on the coefficient of variation CV_b of α_b. On the other hand, the social factor λ_s depends on both the coefficient of variation CV_s and the skewness γ_s of the distribution of α_s. Due to our normalization, $\langle\alpha_s\rangle\langle\alpha_b\rangle \approx \langle\alpha_s\alpha_b\rangle = \langle\alpha\rangle = 1$.

The relative importance of biological and social contributions to the overall heterogeneity of α may be characterized by a single parameter χ. For a log-normal distribution of α_b, χ appears as a scaling exponent between infectivity and susceptibility: $R_\alpha \approx \alpha^\chi$ (see *SI Appendix* for details). The corresponding expression for the overall immunity factor is $\lambda = \langle\alpha^{2+\chi}\rangle/\langle\alpha^{1+\chi}\rangle$. The limit $\chi = 0$ corresponds to a predominantly biological source of heterogeneity, that is, $\lambda \approx \lambda_b = 1 + CV_\alpha^2$, where CV_α is the coefficient of variation for the overall susceptibility. In the opposite limit $\chi = 1$, population heterogeneity is primarily of social origin; hence $\lambda \approx \lambda_s$ is affected by both CV_α and the skewness γ_α of the pdf $f(\alpha)$. The biological contribution λ_b depends on specific biological details of the disease and thus is unlikely to be as universal and robust as the social one. For the COVID-19 epidemic, there is no strong evidence of a wide variation in attack rates unrelated to social activity, geographic location, or socioeconomic status. For instance, there is very little age variability in COVID-19 prevalence as reported by the NYC Department of Health (51) based on the serological survey that followed the first wave of the epidemic. Therefore, below, we will largely ignore possible biological heterogeneity, and focus on social heterogeneity.

So far, our discussion has focused on the early stages of epidemics, when the $R_e(S)$ dependence is given by a linearized expression Eq. **9**. To describe the nonlinear regime, we consider a gamma-distributed susceptibility: $f(\alpha) \approx \alpha^{1/\eta-1}\exp(-\alpha/\eta)$, where $\eta = CV_\alpha^2$. In this case, according to Eqs. **4** and **5**, R_e, S_e, and S are related by scaling relationships (see *SI Appendix*),

$$S_e(S) = S^{1+\eta} \qquad [13]$$

and

$$R_e(S) = R_0 S^\lambda. \qquad [14]$$

The exponent $\lambda = 1 + (1+\chi)CV_\alpha^2 = 1 + (1+\chi)\eta$ coincides with the early-epidemics immunity factor defined in Eqs. **9** and **10** for a general case. Note that, without correlation ($\chi = 0$), both scaling exponents would be the same; this result has been previously obtained for the SIR model in ref. 30 and more recently reproduced in ref. 41 in the context of COVID-19. The scaling behavior $R_e(S)$ is shown in Fig. 1 for $\lambda = 3 \pm 1$. While this range is arbitrary, it includes the empirical values of λ estimated below. This function is dramatically different from the classical linear dependence $R_e = SR_0$. To emphasize the importance of this difference, Eq. **14** immediately leads to a major revision of the classical result for the HIT $1 - S_0 = 1 - 1/R_0$. S_0 is the fraction of susceptible population at which the growth stops, while $1 - S_0$ is the relative size of the epidemic at that time. By setting $R_e = 1$ in Eq. **14**, we obtain

$$1 - S_0 = 1 - \left(\frac{1}{R_0}\right)^{1/\lambda}. \qquad [15]$$

Nonlinear modifications to homogeneous epidemiological models similar to Eqs. **13** and **14** have been proposed in the past as plausible descriptions of heterogeneous populations in various contexts. Specifically, they were used as empirical fits to simulations of the SIR model on small-world networks (19), as well as to the behavior of the Agent-Based Model on realistic urban contact networks (18). A conceptual explanation of the origin of a nonlinear relation between S and Re was proposed in refs. 19 and 30. However, the scaling law similar to Eqs. **13** and **14** has not been derived except in a special case of the SIR model without correlation between susceptibility and infectivity (30). As we were finalizing this paper for public release, a preprint by Aguas et al. (52) appeared online that independently obtained our Eqs. **14** and **15** for gamma-distributed susceptibilities. The same result has also been recently obtained in ref. 42. Our approach is more general: It provides the exact mapping of a wide class of heterogeneous well-mixed models onto homogeneous ones, and provides a specific relationship between the underlying statistics of α and R_α and the nonlinear functions $R_e(S)$ and $S_e(S)$. Of course, our methodology has the same limitations as the original heterogeneous well-mixed approximation (10). This approximation was shown to provide an adequate description for many classes of networks (19). Additional corrections may still arise, for example, due to clustering and other network structure not captured in its degree (α) distribution.

Our focus on the gamma distribution is well justified by the observation that the social strength α_s is approximately exponentially distributed, that is, it is a specific case of the gamma distribution with $\eta = CV_\alpha^2 = 1$ (see more discussion of this in the next section). A moderate biological heterogeneity would lead to an increase of the overall CV_α, but the pdf $f(\alpha)$ will still be close to the gamma distribution family. From the conceptual point of view, it is nevertheless important to understand how the function $R_e(S)$ would change if $f(\alpha)$ had a different functional form. In *SI Appendix*, we present analytic and numerical calculations for two other families of distributions: 1) an exponentially bounded power law $f(\alpha) \approx e^{-\alpha/\alpha_+}/\alpha^q$ ($q \geq 1$, with an additional cutoff at lower values of α), and 2) the log-normal distribution. In addition, we give an approximate analytic result that generalizes Eq. **14** for an arbitrary skewness of $f(\alpha)$. This generalization works remarkably well for all three families of distributions analyzed in this work. As suggested by Eqs. **11** and **12**, as the distribution becomes more skewed, the range between the $\chi = 0$ and $\chi = 1$

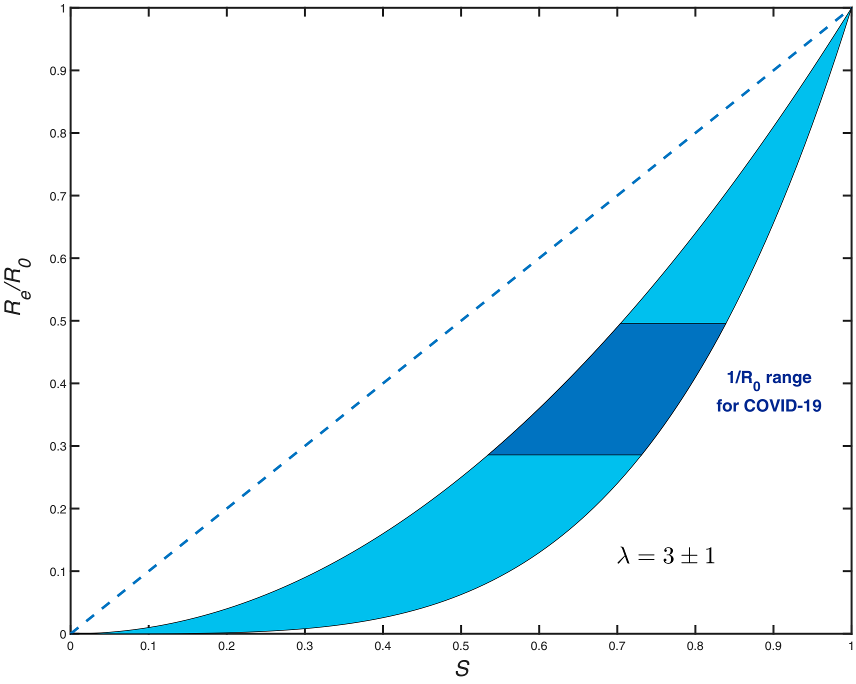

Fig. 1. R_e/R vs. S dependence for gamma-distributed susceptibility with $\lambda = 3 \pm 1$ (light blue area). The dashed line shows the classical homogeneous result, $R_e = R_0 S$. The dark blue region corresponds to the range $2 < R_0 < 3.5$ representative for COVID-19.

curves broadens. For instance, for distributions dominated by a power law, $f(\alpha) \approx 1/\alpha^q$, with $3 < q < 4$ and $\chi = 1$, λ diverges even though CV_α remains finite. This represents a crossover to the regime of so-called scale-free networks ($2 \leq q \leq 3$), which are characterized by zero epidemic threshold yet strongly self-limited dynamics: The epidemic effectively kills itself by immunizing the hubs on the network (10, 23, 53).

Role of Short-Term Variations in Social Activity

Short-term overdispersion in transmission is commonly presumed to have no effect on the overall epidemic dynamics, aside from the early outbreak often dominated by superspreaders. This would indeed be the case if overdispersed transmission were completely uncorrelated with individual susceptibility. But, since the timescale for an individual's infectivity (about 2 d) is comparable to a single generation interval (about 5 d) for the COVID-19 epidemics, ignoring such correlations appears unreasonable. We therefore developed a generalization of the theory presented in the previous section, that takes into account a time dependence of individual susceptibilities and infectivities, as well as temporal correlations between them. The theory is presented, using a path integral formulation, in *SI Appendix*. Here we present several important results directly related to the transient suppression of an epidemic and differentiate these effects from herd immunity.

Since fast variations are primarily caused by bursty dynamics of social interactions (54–57), and since heterogeneous biological susceptibility appears subdominant in the context of COVID-19, we set $\alpha_b = 1$ for all individuals. So α has a purely social origin. Let $a_i(t) = \alpha_i + \delta a_i(t)$ be the time-dependent susceptibility of a person, which we associate with a variable level of social activity. Here $\delta a_i(t)$ represents the time-dependent deviation of $a_i(t)$ from its persistent long-term average α_i. Note that index i labels individuals rather than population groups. The level of social activity quantified by $a_i(t)$ also determines individual infectivity $a(t)R$ around time t. Interestingly, even the classical result for the basic reproduction number in a heterogeneous system, $R_0 = R\langle\alpha^2\rangle$, needs to be modified due to correlated short-term variations in social activity,

$$R_0 = R\left(\langle\alpha^2\rangle + \overline{\delta a_i^2}\right). \qquad [16]$$

Here the bar $\overline{\ldots}$ represents averaging over individual members of the population indexed by i, in contrast with $\langle\ldots\rangle$, averaging over all subgroups with various values of persistent heterogeneity α.

In the time-dependent generalization of our theory, R_e and S no longer have a fixed functional relationship between them. Instead, this relationship becomes nonlocal in time. For instance, our result for the suppression of R_e at the early stages of the epidemic is still formally valid, but the effective value of immunity factor λ becomes time dependent, and Eqs. **9** and **10** become

$$\lambda_{\text{eff}}(t) = \lambda_\infty + \frac{1}{1 - S(t)} \int_0^\infty \delta\lambda(t, t')J(t - t')dt' \qquad [17]$$

$$\lambda_\infty = \frac{\langle\alpha^3\rangle + \overline{\alpha_i \delta a_i^2}}{\langle\alpha^2\rangle + \overline{\delta a_i^2}} \qquad [18]$$

$$\delta\lambda(t, t') = \frac{\overline{\delta a_i^2(t)\delta a_i(t - t')}}{\langle\alpha^2\rangle + \overline{\delta a_i^2}}. \qquad [19]$$

Constant λ_∞ reflects suppression of R_e due to the buildup of the long-term collective immunity. On the other hand, the time-dependent term $\delta\lambda(t')$ leads to an additional suppression of R_e over intermediate timescales. This term has likely played a significant role in shaping the transient self-limiting dynamics during the first wave of COVID-19 epidemic in some hard-hit locations.

Note that, according to Eq. **17**, $\delta\lambda(t')$ is being averaged with the weight proportional to the force of infection $J(t - t')$, since $1 - S(t) \approx \int_0^\infty J(y - t')dt'$. Since $\delta\lambda(t, t')$ decreases with time difference t', its effect on λ_{eff} should be the strongest during the initial period of fast exponential growth. The initial suppression of the epidemic is caused by the combined effect of mitigation measures and both terms in λ_{eff}. Since $\lambda_{\text{eff}} > \lambda_\infty$, the population may reach the state of TCI earlier than the actual long-term herd immunity determined by persistent heterogeneity. However, this state is fragile and may wane with time. Specifically, as $J(t)$ drops after the first wave, the second term in Eq. **17** gradually decays, bringing $\lambda_{\text{eff}}(t)$ closer to λ_∞. According to Eq. **19**, it is the correlation time of bursty social activity $\delta a(t)$ that sets the timescale over which this TCI state deteriorates, and the new epidemic wave may get ignited. The relationship between this relaxation time and the duration of a single epidemic wave also determines the typical value of λ_{eff} during that wave.

Despite a large number of empirical studies of contact networks (54–57), information about the temporal correlations in $\alpha(t)$ or its proxies remains limited. On the other hand, much more is known about parameters of persistent heterogeneity. Recently, real-world networks of face-to-face communications have been studied using a variety of tools, including Radio-frequency identification (RFID) devices (45), Bluetooth and Wi-Fi wearable tags, and smartphone apps (46, 47), as well as census data and personal surveys (13, 37, 48). Despite coming from a wide variety of contexts, the major features of contact networks are remarkably robust. In particular, pdfs of both the degree (the number of contacts per person) and the node strength plotted in log–log coordinates appear nearly constant, followed by a sharp fall after a certain upper cutoff. This behavior is generally consistent with an exponential distribution in $f_s(\alpha_s)$ (19, 46, 48), $f(\alpha) \approx e^{-\alpha/\langle\alpha\rangle}$. That sets the value of $\eta = CV_\alpha^2 \approx 1$. If not for short-term overdispersion, that would yield $\lambda = \langle\alpha_s^3\rangle/\langle\alpha_s^2\rangle = 3!/2! = 3$ according to Eq. **12**. However, with temporal effects taken into account, the buildup of long-term collective immunity is determined by $\lambda_{\text{eff}}(\infty) = \lambda_\infty$. In order to estimate it, we make a simple model assumption that the short-term overdispersion for a particular individual is proportional to the persistent value of that person's social activity: $\overline{\delta a_i^2} \approx \alpha_i$. This leads to

$$\lambda_\infty = 1 + \eta(1 + \chi^*). \qquad [20]$$

Here $\chi^* = \langle\alpha^2\rangle/\overline{a_i(t)^2}$ is a parameter that measures the relative strength of persistent heterogeneity and the overdispersion on the timescale of a single generation interval. Note that, formally, we recover our original result for λ in the purely persistent case, with χ^* replacing the parameter χ that originally quantified the correlation between infectivity and susceptibility. By assuming the limit of strong short-term overdispersion ($\chi^* \ll 1$), we obtain $\lambda_\infty \approx 2$. This estimate is consistent with numerical simulations of the agent-based epidemiological model on urban contact networks carried out in ref. 18.

As shown in *SI Appendix*, the very same value of λ_∞ should be used as a scaling exponent for long-term behavior of $R_e(S)$. Therefore, HIT is set by Eq. **15** with $\lambda = \lambda_\infty \approx 2$. Its value is plotted vs. R_0 in Fig. 2, along with the homogeneous result, and the estimated threshold of TCI. To estimate the corresponding transient immunity factor λ_{eff}, we analyze the empirical data for the first wave of the COVID-19 epidemic, below.

Application to the COVID-19 Epidemic

The COVID-19 epidemic reached the United States in early 2020, and, by March, it was rapidly spreading across multiple

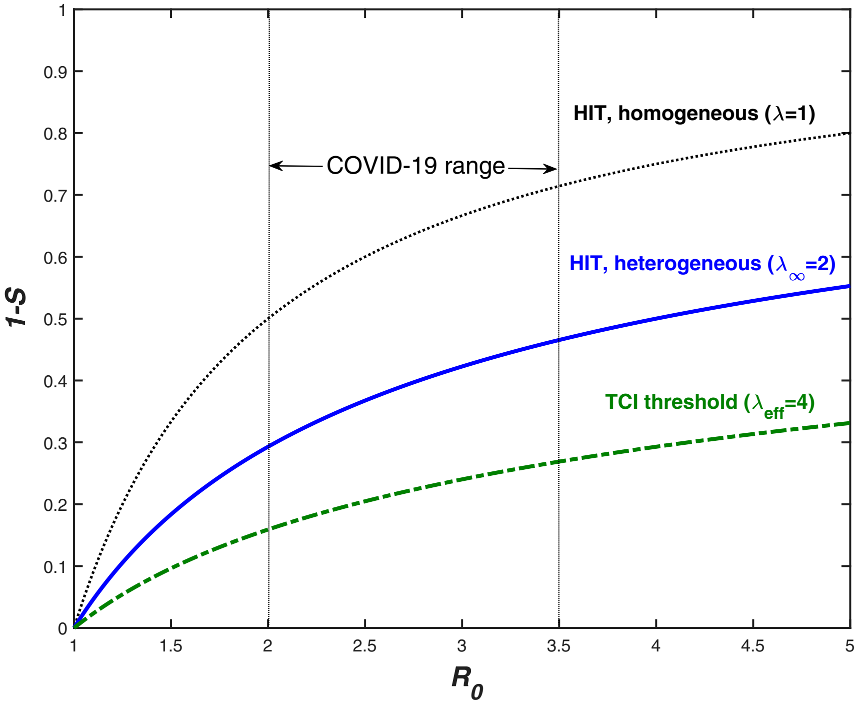

Fig. 2. TCI threshold (dot-dashed), long-term heterogeneous HIT (solid), and homogeneous HIT (dotted) for various values of R_0 (x axis). HIT (solid line) is determined by persistent heterogeneity. The corresponding immunity factor $\lambda_\infty \approx 2$ was estimated from Eq. **20** assuming strong short-term overdispersion ($\chi^* \ll 1$) and the exponential distribution of α ($\eta = 1$). For transient behavior, $\lambda_{\text{eff}} \approx 4$ is assumed based on analysis of empirical data for COVID-19 epidemic in select locations.

states. The early dynamics was characterized by a rapid rise in the number of cases, with doubling times as low as 2 d. In response to this, the majority of states imposed a broad range of mitigation measures including school closures, limits on public gatherings, and stay-at-home orders. In many regions, especially the hardest-hit ones like NYC, people started to practice some degree of social distancing even before government-mandated mitigation. In order to quantify the effects of heterogeneity on the spread of the COVID-19 epidemic, we apply the Bayesian age-of-infection model described in ref. 43 to NYC and Chicago (see *SI Appendix* for details). For both cities, we have access to reliable time series data on hospitalization, ICU room occupancy, and confirmed daily deaths due to COVID-19 (51, 58–60). We used these data to perform multichannel calibration of our model (43), which allows us to infer the underlying time progression of both $S(t)$ and $R_e(t)$. The fits for $R_e(S)$ for both cities are shown in Fig. 3*A*. In both cases, a sharp drop of R_e that occurred during the early stage of the epidemic is followed by a more gradual decline. For NYC, there is an extended range over which $R_e(S)$ has a constant slope in logarithmic coordinate. This is consistent with the power law behavior predicted by Eq. **14**, with the slope corresponding to transient immunity factor $\lambda_{\text{eff}} = 4.5 \pm 0.05$. Chicago exhibits a similar behavior but over a substantially narrower range of S. This reflects the fact that NYC was much harder hit by the COVID-19 epidemic. Importantly, the range of dates we used to estimate the immunity factor corresponds to the time interval after state-mandated stay-at-home orders were imposed, and before the mitigation measures began to be gradually relaxed. The signatures of the onset of the mitigation and of its partial relaxation are clearly visible on both ends of the constant-slope regime. To examine the possible effects of variable levels of mitigation on our estimates of λ_{eff}, in *SI Appendix*, we repeated our analysis in which $R_e(t)$ was corrected by Google's community mobility report in these two cities (61) (see *SI Appendix*). Although the range of data consistent with the constant slope shrank somewhat, our main conclusion remains unchanged. This provided us with a lower-bound estimate for the transient immunity factor: $\lambda_{\text{eff}} = 4.1 \pm 0.1$.

To test the sensitivity of our results to details of the epidemiological model and choice of the region, we performed an alternative analysis based on the data reported in ref. 49. In that study, the COVID-19 epidemic was modeled in each of the 50 US states and the District of Columbia. Because of the differences in population density, level of urbanization, use of public transport, etc., different states were characterized by substantially different initial growth rates of the epidemic, as quantified by the basic reproduction number R_0. Furthermore, the time of arrival of the epidemic also varied a great deal between individual states, with states hosting major airline transportation hubs being among the earliest ones hit by the virus. As a result of these differences, at any given time, the infected fraction of the population differed significantly across the United States (49). We use state level estimates of $R_e(t)$, R_0, and $S(t)$ as reported in ref. 49 to construct the scatter plot $R_e(t_0)/R_0$ vs. $S(t_0)$ shown in Fig. 3*B*, with t_0 chosen to be the last reported date in that study, May 17, 2020. By performing the linear regression on these data in logarithmic coordinates, we obtain the fit for the slope $\lambda_{\text{eff}} = 5.3 \pm 0.6$ and for $S = 1$ intercept around 0.54. In *SI Appendix*, Fig. S3, we present an extended version of this analysis for the 10 hardest-hit states and the District of Columbia, which takes into account the overall time progression of $R_e(t)$ and $S(t)$, and gives similar estimate $\lambda_{\text{eff}} = 4.7 \pm 1.5$. Both estimates of the immunity factor based on the state data are consistent with our earlier analysis of NYC and Chicago. In light of our theoretical picture, this value of this transient immunity factor, $\lambda_{\text{eff}} \simeq 4$, is set by the pace of the first epidemic wave in the United States. As expected, it exceeds our estimate of $\lambda_\infty \approx 2$ associated with persistent heterogeneity and responsible for the long-term herd immunity.

We can now incorporate this transient level of heterogeneity into our epidemiological model, and examine how future

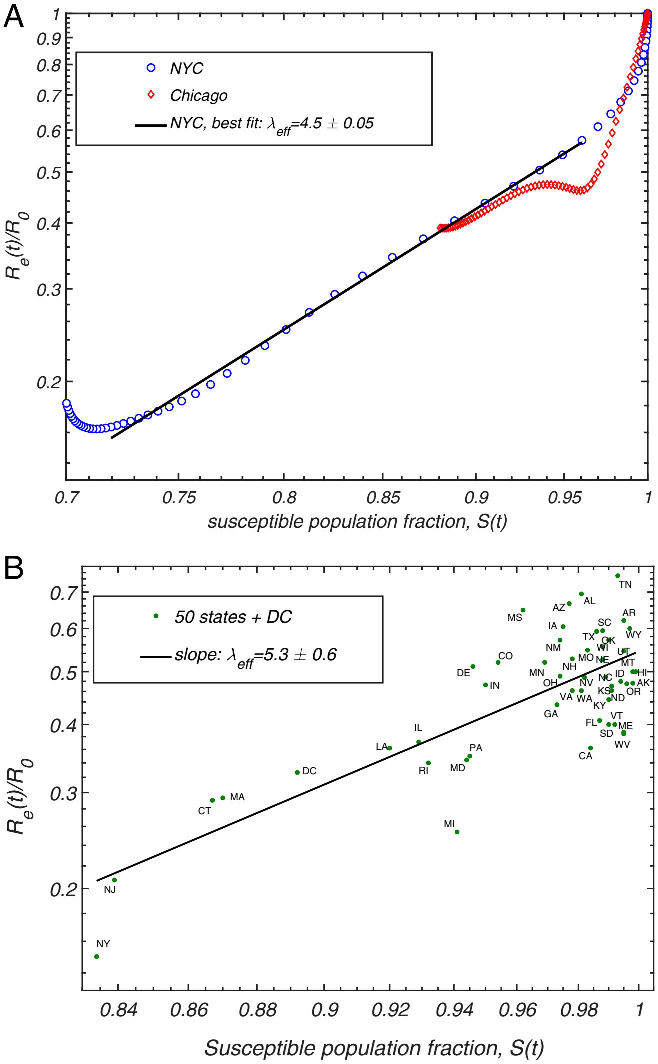

Fig. 3. Correlation between the relative reduction in the effective reproduction number $R_e(t)/R_0$ (y axis) with the susceptible population $S(t)$. (*A*) The progression of these two quantities for NYC and Chicago, as given by the epidemiological model described in ref. 43. (*B*) The scatter plot of $R_e(t_0)/R_0$ and $S(t_0)$ in individual states of the United States, evaluated in ref. 49 (t_0 is the latest date covered in that study).

projections change as a result of this modification. This is done by plugging scaling relationships given by Eqs. **13** and **14** into the force of infection and incidence rate equations of the original model. These equations are similar to Eqs. **6** and **7**, but also include time modulation due to the mitigation and a possible seasonal forcing (see *SI Appendix* for more details). After calibrating the model by using the data streams on ICU occupancy, hospitalization, and daily deaths up to the end of May, we explore a hypothetical worst-case scenario in which any mitigation is completely relaxed as of June 15, in both Chicago and NYC. In other words, the basic reproduction number R_0 is set back to its value at the initial stage of the epidemic, and the only factor limiting the second wave is the partial or full TCI, $R_e = R_0 S^\lambda$. The projected daily deaths for each of the two cities under this (unrealistically harsh) scenario are presented in Fig. 4 for various values of λ. For both cities, the homogeneous model ($\lambda = 1$, blue lines) predicts a second wave which is larger than the first one, with an additional death toll of around $35,000$ in NYC and $12,800$ in Chicago. The magnitude of the second wave is greatly reduced by heterogeneity, resulting in no second wave in either of the two cities for $\lambda = 5$ (black lines). Even for a modest value $\lambda = 3$ (red lines), which is less than our estimate, the second wave is dramatically reduced in both NYC and Chicago (by about 90% and 70%, respectively).

Note that our predictions about the second wave in NYC and Chicago have been made based on the data up to June 10, 2020 and extended up to early September 2020. The real epidemic dynamic in both cities during this time interval was consistent with the "no second wave" scenario shown in Fig. 4. However, one must be warned against using it to put form bounds on λ_{eff}, since we considered the worst-case scenario of full release of mitigation to prepandemic levels. In reality, some mitigation measures, for example, mask wearing and social distancing, stayed in place. Ultimately, second waves broke out in both cities in the late fall. The mechanisms leading to gradual degradation of the TCI state are described in the next section.

Fragility of TCI

One of the consequences of the bursty nature of social interactions is that the state of TCI gradually wanes due to changes of individual social interaction patterns on timescales longer than a single generation interval. This may be viewed as a slow rewiring of social networks. In the context of the COVID-19 epidemic, individual responses to mitigation factors such as stay-at-home orders may differ across the population. When mitigation measures are relaxed, individual social susceptibility α_s inevitably changes. The impact of these changes on collective immunity depends on whether each person's α_s during and after the mitigation are sufficiently correlated. For example, the TCI state would be compromised if people who practiced strict self-isolation compensated for it by an above-average social activity after the first wave of the epidemic has passed.

To illustrate the effects of postmitigation rewiring of social networks, we consider a simple modification of the heterogeneous model with no persistent heterogeneity ($\alpha = 1$ for everyone) and exponentially distributed instantaneous levels of social activity $a_i(t)$. This corresponds to $\lambda_{\mathrm{eff}}(0) = 3$ and $\lambda_{\mathrm{eff}}(\infty) = \lambda_\infty = 1$. In this model, each individual completely changes the set of his/her social connections at some timescale τ_s. These changes destroy heterogeneity, giving rise to gradual relaxation of susceptible fraction S_a toward its overall mean value S. To model this, we modify Eq. **1** to include a simple relaxation term,

$$\dot{S}_a = -\alpha S_a J - \frac{1}{\tau_s}(S_a - S) \quad . \qquad \textbf{[21]}$$

Epidemiological models with rewiring of underlying social networks have been studied before (62) (see ref. 63 for a review), but under a constraint that the individual level of social activity quantified by network degree is preserved. In contrast, the dynamics described above stems from the individual level of social activity α_s changing in time.

We simulate the full heterogeneous SIR model (see *SI Appendix* for details) in which Eq. **1** was replaced with Eq. **21**. Fig. 5 shows the results of this simulation, where the first wave of the epidemic is mitigated, thereby reducing the effective reproduction number $R_0 = 2.5$. During the course of the mitigation, R_0 is multiplied by $\mu = 0.7$. After the mitigation measures have been lifted at the end of the first wave, the population is positioned slightly below the TCI threshold, preventing the immediate start of the second wave. However, gradual rewiring of the social network with time constant $\tau_s = 150$ d ultimately results in the second and even the third wave of the epidemic (Fig. 5). Fig. 5, *Inset* shows $R_e(t)$ plotted as a function of $S(t)$ in this epidemic. Note that each of the waves follows the power law relationship between $R_e(t) \approx S(t)^\lambda$ predicted by Eq. **14**.

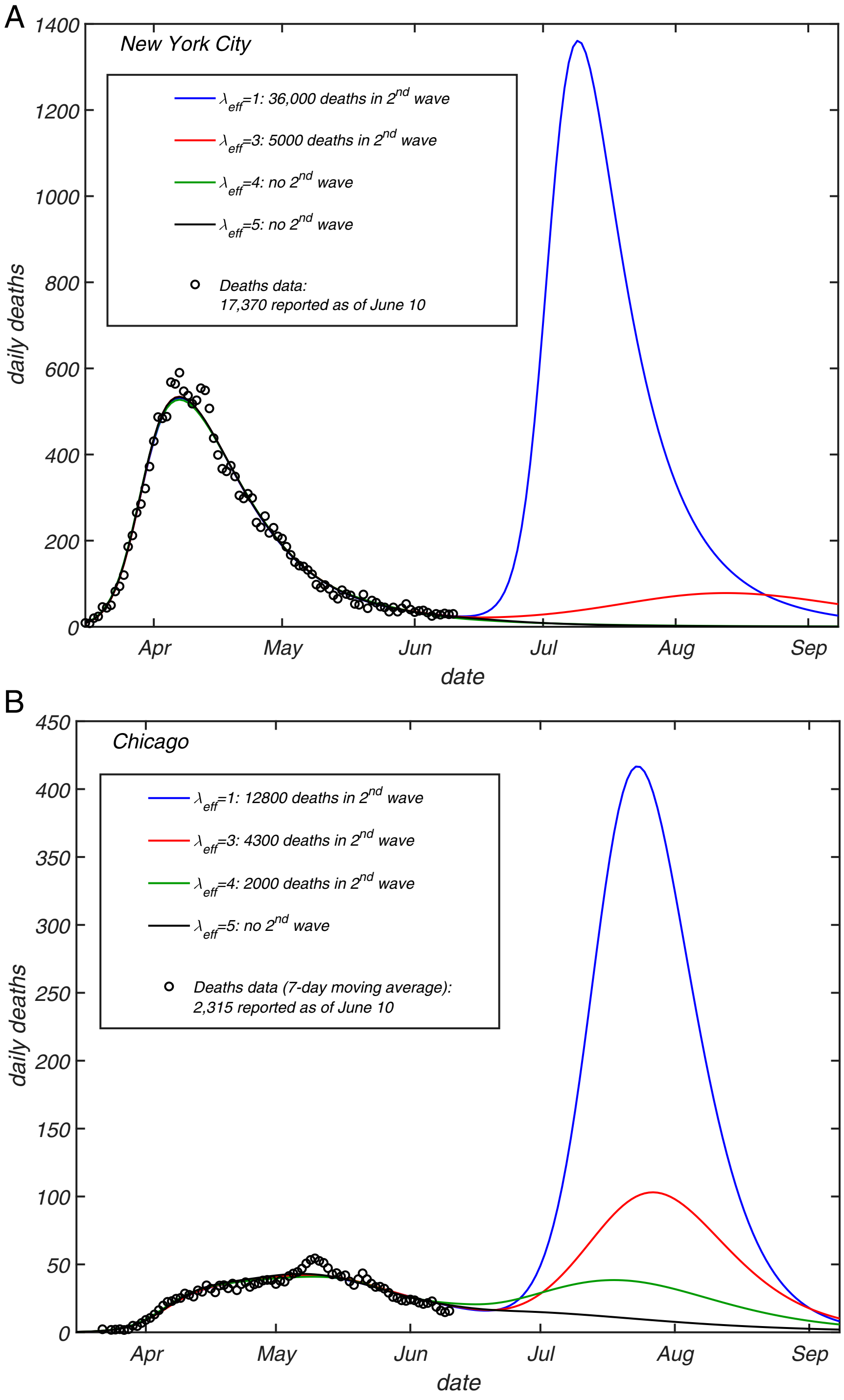

Fig. 4. Projections of daily deaths under the hypothetical scenario in which any mitigation is completely eliminated as of June 15, 2020, for (*A*) NYC and (*B*) Chicago. Different curves correspond to different values of the transient immunity factor $\lambda_{\text{eff}} = 1$ (blue), 3 (red), 4 (green), and 5 (black lines). The model described in ref. 43 was fully calibrated on daily deaths (circles), ICU occupancy, and hospitalization data up to the end of May. See *SI Appendix* for additional details, including CIs.

Since constant rewiring eliminates correlations in individual social activity on scales longer than τ_s, the epidemic stops after multiple waves bring the total fraction of infected individuals close to the unmodified (homogeneous) HIT $1/R_0$. Note, however, that, in this case, there is almost no overshoot, and thus the final size of the epidemic is reduced compared to the case of a purely homogeneous and unmitigated epidemic.

Discussion

In this work, we have demonstrated how the interplay between short-term overdispersion and persistent heterogeneity in a population leads to dramatic changes in epidemic dynamics on multiple timescales, transient suppression of the epidemic during its early waves all the way up to the state of long-term herd immunity. First, we developed a general approach that allows

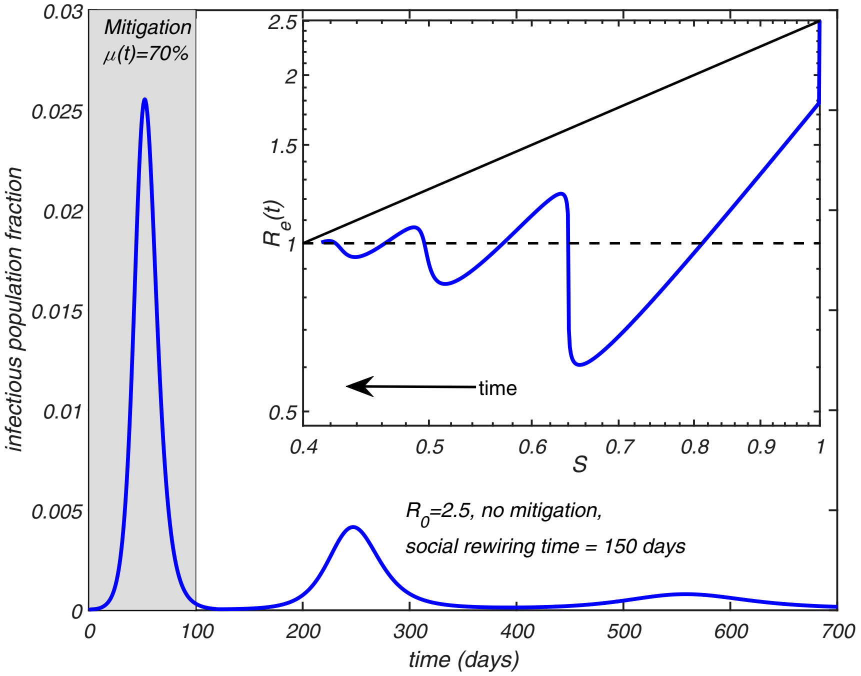

Fig. 5. Effect of social rewiring on the epidemic dynamics. The time course of an epidemic in a heterogeneous SIR model with $R_0 = 2.5$ and $\lambda = 3$. During the first 100 d, a mitigation factor $\mu = 0.7$ is applied. Social networks gradually rewire with a time constant $\tau_s = 150$ d. The figure shows multiple waves. (*Inset*) $R_e(t)$ plotted as a function of *S*(*t*). Solid black line shows the homogeneous limit reached after multiple waves.

for the persistent heterogeneity to be easily integrated into a wide class of traditional epidemiological models in the form of two nonlinear functions $R_e(S)$ and $S_e(S)$, both of which are fully determined by the statistics of individual susceptibilities and infectivities. Furthermore, $R_e(S)$ is largely defined by a single parameter, the immunity factor λ, introduced in our study. Like susceptibility itself, λ has two contributions: biological and social (Eqs. **11** and **12**).

We then expanded our approach to include effects of time dependence of individual social activity, and, in particular, of likely correlations over the timescale of a single generation interval. While our results for purely persistent heterogeneity confirmed and corroborated that HIT would be suppressed compared to the homogeneous case, addition of temporal variations led to a dramatic revision of that simple narrative. Both persistent heterogeneity and short-term overdispersion contributions lead first to a slowdown of a fast-paced epidemic, and to its medium-term stabilization. However, this state of TCI is fragile and does not constitute long-term herd immunity. HIT is indeed suppressed, but only due to the persistent heterogeneity. This suppression is significantly weaker than the initial stabilization responsible for the TCI state reached after the first wave of a fast-paced epidemic.

Among other implications of the TCI phenomenon is the suppression of the so-called overshoot. Namely, it is well known that most models predict that an epidemic will not stop once HIT is passed, ultimately reaching a significantly larger cumulative attack rate, FSE. Multiple prior studies (9, 10, 20, 21, 30, 34) have shown that FSE would be suppressed by persistent heterogeneity, similarly to HIT. In *SI Appendix*, we present a simple result that unifies several previously studied limiting cases, and gives an explicit equation for the FSE for the gamma-distributed susceptibility and variable level of its correlation with infectivity. However, because of the transient suppression of the early waves of the epidemic discussed in this work, the overshoot effect would be much weaker or essentially eliminated. For instance, our simple rewiring model demonstrates how the epidemic, after several waves, ultimately reaches HIT level, but does not progress much beyond it (Fig. 5). The FSE result may still be used, but primarily as an estimate for the size of the first wave of an (unmitigated) epidemic. In that case, the transient value of immunity factor λ_{eff} should be assumed.

By applying our theory to the COVID-19 epidemic, we found evidence that the hardest-hit areas, such as NYC, have likely passed TCI threshold by the end of the first wave, but are less likely to have achieved real long-term herd immunity. Other places that had intermediate exposure, such as, for example, Chicago, while still below the TCI threshold, have their effective reproduction number reduced by a significantly larger factor than predicted by traditional epidemiological models. This gives a better chance of suppressing the future waves of the epidemic in these locations by less disruptive measures than those used during the first wave, for example, by using masks, social distancing, contact tracing, control of potential superspreading events, etc. However, similar to the case of NYC, transient stabilization of the COVID-19 epidemic in Chicago will eventually wane. As for the permanent HIT, although suppressed compared to classical value, it definitely has not yet been passed in those two locations.

In a recent study (35), the reduction of HIT due to heterogeneity has been illustrated using a toy model. In that model, 25% of the population was assumed to have their social activity reduced by 50% compared to a baseline, while another 25% had their social activity elevated twofold. The rest of the population was assigned the baseline level of activity. According to Eq. **12**, the immunity factor in that model is $\lambda = 1.54$. For this immunity factor, Eq. **15** predicts HIT at $S_0 = 64\%$, 55%, and 49%, for $R_0 = 2$, 2.5, and 3, respectively. Despite the fact that the model distribution is not gamma shaped, these values are in a very good agreement with the numerical results reported in ref. 35: $S_0 = 62.5\%$, 53.5%, and 47.5%, respectively.

Thus there is a crucial distinction between the persistent heterogeneity, short-term variations correlated over the timescale

Table 1. Effects of heterogeneity on suppression of the effective reproduction number R_e in selected locations

		Attack rate, $1 - S$, %		R_e suppression	
Location	Deaths, per 1,000 (refs.)	Estimated	Seroprevalence (refs.)	Transient	Long-term
NYC	2.1 (51)	30	23 (68)	0.24 to 0.33	0.50 to 0.58
Lombardy	1.7 (69)	24	23 (70, 71)	0.33	0.58
London	0.9 (72)	13	13 (73)	0.58	0.75
Chicago	0.9 (58)	13	20 (74)	0.41 to 0.58	0.64 to 0.75
Stockholm	0.9 (75, 76)	13	12 (76)	0.58	0.75

The transient and long-term suppression coefficients R_e/R_0 are calculated using $\lambda_{\text{eff}} = 4$ and $\lambda_\infty = 2$, respectively. Fraction of susceptible population S as of early June 2020 is estimated from the cumulative reported death count per capita, assuming the infection fatality rate of 0.7% (67). This estimate is supplemented by seroprevalence data for late spring–early summer 2020.

of a single generation interval, and overdispersion in transmission statistics associated with short-term superspreading events (12, 13, 16, 26–28). In our theory, a personal decision to attend a large party or a meeting would only contribute to persistent heterogeneity if it represents a recurring behavioral pattern. On the other hand, superspreading events are shaped by short-time variations in individual infectivity (e.g., a person during the highly infectious phase of the disease attending a large gathering). Hence, the level of heterogeneity inferred from the analysis of such events (12, 27) would be significantly exaggerated and should not be used to estimate the TCI threshold and HIT. Specifically, the statistics of superspreading events is commonly described by the negative binomial distribution with dispersion parameter k estimated to be in the range 0.1 to 0.3 for COVID-19 (28, 64, 65). This is much stronger overdispersion than the value $k = 1$ estimated from persistent heterogeneity, based on the exponential distribution of α. Thus, persistent heterogeneity is a weaker source of variation compared to short-term variations. According to ref. 12, this is consistent with the expected value of the individual-level reproduction number R_i drawn from a gamma distribution with the shape parameter $k \simeq 0.1 \ldots 0.3$. This distribution has a very high coefficient of variation, $CV^2 = 1/k \simeq 3 \ldots 10$. In the case of a perfect correlation between individual infectivity and susceptibility α, this would result in an unrealistically high estimate of the immunity factor: $\lambda = 1 + 2CV^2 = 1 + 2/k \simeq 7 \ldots 20$. For this reason, according to our perspective and calculation, the final size of the COVID-19 epidemic may have been substantially underestimated in ref. 31. Similarly, the degree of heterogeneity assumed in other recent studies (32, 52) is considerably larger than our estimates. Based on our analysis, the value of the immunity factor λ depends on the pace of the epidemic and on the timescale under consideration. We estimated its long-term value (responsible for the permanent HIT) as $\lambda_\infty \approx 2$. However, the transient values are expected to be higher, especially during the first several waves of COVID-19 in select locations, characterized by large growth rates. Our analysis of the empirical data in NYC and Chicago indicates that the slowdown of the epidemic dynamics in those locations was consistent with $\lambda \approx 4$. In Table 1, we present our estimates of the factor by which R_e is transiently suppressed as a result of depletion of susceptible population in selected locations in the world, as of early June 2020, as well as the predicted long-term suppression related to acquisition of a partial herd immunity.

Finally, we summarize the assumptions and limitations of our study. First, we assume a long-lasting biological immunity of recovered individuals. Second, our approach is based on the well-mixed approximation in which geographic heterogeneity as well as nontrivial properties of the contact network (clustering, degree–degree correlations, etc.) are ignored. In addition, our description of transient epidemic dynamics is based on the approximation of a constant λ_{eff}, while gradual degradation of TCI with time is illustrated using a simplified model, Eq. **21**. A generalization of the present model including explicit description of stochastic social activity is needed (see ref. 66). Furthermore, additional calibration based on long-term empirical data is required before our approach can be used for reliably guiding policy decisions during an epidemic.

Population heterogeneity manifests itself at multiple scales. At the most coarse-grained level, individual cities or even countries can be assigned some level of susceptibility and infectivity, which inevitably vary from one location to another, reflecting differences in population density and its connectivity to other regions. Such spatial heterogeneity will result in self-limiting epidemic dynamics at the global scale. For instance, hard-hit hubs of the global transportation network, such as NYC during the COVID-19 epidemic, would gain full or partial TCI, thereby limiting the spread of infection to other regions during future waves of the epidemic. This might be a general mechanism that ultimately limits the scale of many pandemics, from the Black Death to the 1918 influenza.

Data Availability. All study data are included in the article and *SI Appendix*.

ACKNOWLEDGMENTS. We gratefully acknowledge discussions with Mark Johnson at Carle Hospital. The calculations we have performed would have been impossible without the data kindly provided by the Illinois Department of Public Health through a Data Use Agreement with Civis Analytics. This work was supported by the University of Illinois System Office, the Office of the Vice-Chancellor for Research and Innovation, the Grainger College of Engineering, and the Department of Physics at the University of Illinois at Urbana–Champaign. Z.J.W. is supported in part by the US Department of Energy (DOE) Computational Science Graduate Fellowship, provided under Grant DE-FG02-97ER25308. A.E. acknowledges partial support by NSF CAREER Award 1753249. This work made use of the Illinois Campus Cluster, a computing resource that is operated by the Illinois Campus Cluster Program in conjunction with the National Center for Supercomputing Applications and which is supported by funds from the University of Illinois at Urbana–Champaign. This research was partially done at, and used resources of, the Center for Functional Nanomaterials, which is a US DOE Office of Science Facility, at Brookhaven National Laboratory under Contract DE-SC0012704.

1. W. O. Kermack, A. G. McKendrick, A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. Ser. A* **115**, 700–721 (1927).
2. M. J. Keeling, P. Rohani, *Modeling Infectious Diseases in Humans and Animals* (Princeton University Press, 2011).
3. K. Rock, S. Brand, J. Moir, M. J. Keeling, Dynamics of infectious diseases. *Rep. Prog. Phys.* **77**, 026602 (2014).
4. J. Ma, Estimating epidemic exponential growth rate and basic reproduction number. *Infect. Des. Model.* **5**, 129–141 (2020).
5. C. Fraser, Estimating individual and household reproduction numbers in an emerging epidemic. *PloS One* **2**, e758 (2007).
6. G. Chowell, Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts. *Infect. Dis. Model.* **2**, 379–398 (2017).

7. A. L. Lloyd, R. M. May, Epidemiology - How viruses spread among computers and people. *Science* **292**, 1316–1317 (2001).
8. R. M. May, A. L. Lloyd, Infection dynamics on scale-free networks. *Phys. Rev. E* **64**, 066112 (2001).
9. M. E. J. Newman, Spread of epidemic disease on networks. *Phys. Rev. E* **66**, 016128 (2002).
10. Y. Moreno, R. Pastor-Satorras, A. Vespignani, Epidemic outbreaks in complex heterogeneous networks. *Euro. Phys. J. B* **26**, 521–529 (2002).
11. Z. Dezső, A. L. Barabási, Halting viruses in scale-free networks. *Phys. Rev. E* **65**, 055103 (2002).
12. J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, W. M. Getz, Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, 355–359 (2005).
13. L. A. Meyers, B. Pourbohloul, M. E. J. Newman, D. M. Skowronski, R. C. Brunham, Network theory and SARS: Predicting outbreak diversity. *J. Theor. Biol.* **232**, 71–81 (2005).
14. M. J. Keeling, K. T. D. Eames, Networks and epidemic models. *J. R. Soc. Interface* **2**, 295–307 (2005).
15. M. J. Ferrari, S. Bansal, L. A. Meyers, O. N. Bjornstad, Network frailty and the geometry of herd immunity. *Proc. Biol. Sci.* **273**, 2743–2748 (2006).
16. M. Small, C. Tse, D. M. Walker, Super-spreaders and the rate of transmission of the sars virus. *Phys. Nonlinear Phenom.* **215**, 146–158 (2006).
17. M. Roy, M. Pascual, On representing network heterogeneities in the incidence rate of simple epidemic models. *Ecol. Complex.* **3**, 80–90 (2006).
18. P. D. Stroud *et al.*, Semi-empirical power-law scaling of new infection rate to model epidemic dynamics with inhomogeneous mixing. *Math. Biosci.* **203**, 301–318 (2006).
19. S. Bansal, B. T. Grenfell, L. A. Meyers, When individual behaviour matters: Homogeneous and network models in epidemiology. *J. R. Soc. Interface* **4**, 879–891 (2007).
20. G. Katriel, The size of epidemics in populations with heterogeneous susceptibility. *J. Math. Biol.* **65**, 237–262 (2012).
21. J. C. Miller, A note on the derivation of epidemic final sizes. *Bull. Math. Biol.* **74**, 2125–2141 (2012).
22. S. Bansal, L. A. Meyers, The impact of past epidemics on future disease dynamics. *J. Theor. Biol.* **309**, 176–184 (2012).
23. R. Pastor-Satorras, C. Castellano, P. Van Mieghem, A. Vespignani, Epidemic processes in complex networks. *Rev. Mod. Phys.* **87**, 925 (2015).
24. W. Gou, Z. Jin, How heterogeneous susceptibility and recovery rates affect the spread of epidemics on networks. *Infect. Des. Model.* **2**, 353–367 (2017).
25. Y. Kim, H. Ryu, S. Lee, Agent-based modeling for super-spreading events: A case study of MERS-CoV transmission dynamics in the Republic of Korea. *Int. J. Environ. Res. Publ. Health* **15**, 2369 (2018).
26. A. P. Galvani, R. M. May, Epidemiology - Dimensions of superspreading. *Nature* **438**, 293–295 (2005).
27. Z. Shen *et al.*, Superspreading SARS events, Beijing, 2003. *Emerg. Infect. Dis.* **10**, 256–260 (2004).
28. A. Endo, S. Abbott, A. Kucharski, S. Funk, Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. *Wellcome Open Res.* **5**, 67 (2020).
29. J. C. Miller, A primer on the use of probability generating functions in infectious disease modeling. *Infect. Dis. Model.* **3**, 192–248 (2018).
30. A. S. Novozhilov, On the spread of epidemics in a closed heterogeneous population. *Math. Biosci.* **215**, 177–185 (2008).
31. L. Hébert-Dufresne, B. M. Althouse, S. V. Scarpino, A. Allard, Beyond R0: Heterogeneity in secondary infections and probabilistic epidemic forecasting. medRxiv [Preprint] (2020). 2020.02.10.20021725. Accessed 14 February 2021.
32. M. G. M. Gomes *et al.*, Individual variation in susceptibility or exposure to SARS-CoV-2 lowers the herd immunity threshold. medRxiv [Preprint] (2020). https://doi.org/10.1101/2020.04.27.20081893. Accessed 14 February 2021.
33. P. V. Brennan, L. P. Brennan, Susceptibility-adjusted herd immunity threshold model and potential R0 distribution fitting the observed COVID-19 data in Stockholm. medRxiv [Preprint] (2020). https://doi.org/10.1101/2020.05.19.20104596. Accessed 14 February 2021.
34. F. Ball, Deterministic and stochastic epidemics with several kinds of susceptibles. *Adv. Appl. Probab.* **17**, 1–22 (1985).
35. T. Britton, F. Ball, P. Trapman, A mathematical model reveals the influence of population heterogeneity on herd immunity to SARS-CoV-2. *Science* **369**, 846–849 (2020).
36. J. C. Stack, S. Bansal, V. S. A. Kumar, B. Grenfell, Inferring population-level contact heterogeneity from common epidemic data. *J. R. Soc. Interface* **10**, 20120578 (2013).
37. S. Eubank *et al.*, Modelling disease outbreaks in realistic urban social networks. *Nature* **429**, 180–184 (2004).
38. P. Grassberger, On the critical behavior of the general epidemic process and dynamical percolation. *Math. Biosci.* **63**, 157–172 (1983).
39. D. Ludwig, Final size distribution for epidemics. *Math. Biosci.* **23**, 33–46 (1975).
40. R. Hickson, M. Roberts, How population heterogeneity in susceptibility and infectivity influences epidemic dynamics. *J. Theor. Biol.* **350**, 70–80 (2014).
41. C. Rose *et al.*, Heterogeneity in susceptibility dictates the order of epidemiological models. arXiv [Preprint] (2020). https://arxiv.org/abs/2005.04704. Accessed 14 February 2021.
42. J. Neipel, J. Bauermann, S. Bo, T. Harmon, F. Jülicher, Power-law population heterogeneity governs epidemic waves. *PloS One* **15**, e0239678 (2020).
43. G. N. Wong *et al.*, Modeling COVID-19 dynamics in Illinois under non-pharmaceutical interventions. *Phys. Rev. X* **10**, 041033 (2020).
44. M. Castro, S. Ares, J. A. Cuesta, S. Manrubia, The turning point and end of an expanding epidemic cannot be precisely forecast. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 26190–26196 (2020).
45. L. Isella *et al.*, What's in a crowd? Analysis of face-to-face behavioral networks. *J. Theor. Biol.* **271**, 166–180 (2011).
46. B. F. Nielsen, K. Sneppen, L. Simonsen, J. Mathiesen, Heterogeneity is essential for contact tracing. medRxiv [Preprint] (2020). https://doi.org/10.1101/2020.06.05.20123141. Accessed 14 February 2021.
47. M. Starnini *et al.*, "Robust modeling of human contact networks across different scales and proximity-sensing techniques" in *Social Informatics*, G. L. Ciampaglia, A. Mashhadi, T. Yasseri, Eds. (Springer International, Cham, Switzerland, 2017), pp. 536–551.
48. L. Danon, J. M. Read, T. A. House, M. C. Vernon, M. J. Keeling, Social encounter networks: Characterizing Great Britain. *Proc. Biol. Sci.* **280**, 20131037 (2013).
49. H. J. T. Unwin *et al.*, "State-level tracking of COVID-19 in the United States" (Rep. 23, World Health Organization, 2020).
50. J. Wallinga, M. Lipsitch, How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc. Biol. Sci.* **274**, 599–604 (2006).
51. NYC Department of Health and Mental Hygiene, Data from "Covid-19 County Cases, Tests, and Death by Day." GitHub. https://github.com/nychealth/coronavirus-data. Accessed 20 June 2020.
52. R. Aguas *et al.*, Herd immunity thresholds for SARS-CoV-2 estimated from unfolding epidemics. medRxiv [Preprint] (2020). https://doi.org/10.1101/2020.07.23.20160762. Accessed 14 February 2021.
53. R. Pastor-Satorras, A. Vespignani, Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* **86**, 3200–3203 (2001).
54. A. L. Barabasi, The origin of bursts and heavy tails in human dynamics. *Nature* **435**, 207–211 (2005).
55. G. Kossinets, D. J. Watts, Empirical analysis of an evolving social network. *Science* **311**, 88–90 (2006).
56. D. Rybski, S. V. Buldyrev, S. Havlin, F. Liljeros, H. A. Makse, Scaling laws of human interaction activity. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 12640–12645 (2009).
57. J. Saramäki, E. Moro, From seconds to months: An overview of multi-scale dynamics of mobile telephone calls. *Euro. Phys. J. B* **88**, 164 (2015).
58. Illinois Department of Public Health, Data from "COVID-19 statistics." Illinois Department of Public Health. https://www.dph.illinois.gov/content/covid-19-county-cases-tests-and-deaths-day. Accessed 20 June 2020.
59. The City, Data from "Covid-19 New York City Data." GitHub. https://github.com/thecityny/covid-19-nyc-data. Accessed 20 June 2020.
60. The City, Data from "Coronavirus in New York City." https://projects.thecity.nyc/2020_03_covid-19-tracker/. Accessed 20 June 2020.
61. J. Fitzpatrick, K. DeSalvo, "Helping public health officials combat COVID-19." *The Keyword* (2020). https://www.blog.google/technology/health/covid-19-community-mobility-reports?hl=en. Accessed 14 February 2021.
62. E. Volz, L. A. Meyers, Susceptible–infected–recovered epidemics in dynamic contact networks. *Proc. Biol. Sci.* **274**, 2925–2934 (2007).
63. S. Bansal, J. Read, B. Pourbohloul, L. A. Meyers, The dynamic nature of contact networks in infectious disease epidemiology. *J. Biol. Dynam.* **4**, 478–489 (2010).
64. Z. Susswein, S. Bansal, Characterizing superspreading of SARS-CoV-2: From mechanism to measurement. medRxiv [Preprint] (2020). https://doi.org/10.1101/2020.12.08.20246082. Accessed 14 February 2021.
65. K. Sun *et al.*, Transmission heterogeneities, kinetics, and controllability of SARS-CoV-2. *Science* **371**, eabe2424 (2020).
66. A. V. Tkachenko *et al.*, How dynamic social activity shapes an epidemic: Waves, plateaus, and endemic state. medRxiv [Preprint] (2021). https://doi.org/10.1101/2021.01.28.21250701. Accessed 24 February 2021.
67. G. Meyerowitz-Katz, L. Merone, A systematic review and meta-analysis of published research data on COVID-19 infection-fatality rates. medRxiv [Preprint] (2020). https://doi.org/10.1101/2020.05.03.20089854. Accessed 14 February 2021.
68. Centers for Disease Control and Prevention, Data from "Commercial laboratory seroprevalence surveys." Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/commercial-lab-surveys.html. Accessed 15 December 2020.
69. Presidenza del Consiglio dei Ministri - Dipartimento della Protezione Civile, Data from "COVID-19." GitHub. https://github.com/pcm-dpc/COVID-19. Accessed 20 June 2020.
70. E. Percivalle *et al.*, Prevalence of SARS-CoV-2 specific neutralising antibodies in blood donors from the Lodi Red Zone in Lombardy, Italy, as at 06 April 2020. *Euro Surveill.* **25**, 2001031 (2020).
71. G. Pagani *et al.*, Seroprevalence of SARS-CoV-2 significantly varies with age: Preliminary results from a mass population screening. *J. Infect.* **81**, e10–e12 (2020).
72. London Assembly, Data from "Coronavirus (COVID-19) Death." https://data.london.gov.uk/dataset/coronavirus–covid-19–deaths. Accessed 15 December 2020.
73. H. Ward *et al.*, Antibody prevalence for SARS-CoV-2 following the peak of the pandemic in england: REACT2 study in 100,000 adults. medRxiv [Preprint] (2020). https://doi.org/10.1101/2020.08.12.20173690. Accessed 14 February 2021.
74. A. R. Demonbreun *et al.*, Patterns and persistence of SARS-CoV-2 IgG antibodies in a US metropolitan site. medRxiv [Preprint] (2020). https://doi.org/10.1101/2020.11.17.20233452. Accessed 14 February 2021.
75. M. Runesson, Data from "Covid-19/Corona data from Stockholm county project." GitHub. https://github.com/mrunesson/covid-19. Accessed 15 December 2020.
76. X. C. Dopico *et al.*, Seropositivity in blood donors and pregnant women during 9-months of SARS-CoV-2 transmission in Stockholm, Sweden. medRxiv [Preprint] (2020). https://doi.org/10.1101/2020.12.24.20248821. Accessed 14 February 2021.

PNAS
www.pnas.org

Supplementary Information for

Time-dependent heterogeneity leads to transient suppression of the COVID-19 epidemic, not herd immunity

Alexei V. Tkachenko, Sergei Maslov, Ahmed Elbanna, George N. Wong, Zachary J. Weiner and Nigel Goldenfeld

Alexei V. Tkachenko
E-Mail: oleksiyt@bnl.gov
Sergei Maslov
E-mail: maslov@illinois.edu

This PDF file includes:

Supplementary text
Figs. S1 to S6 (not allowed for Brief Reports)
SI References

Supporting Information Text

Derivation of quasi-homogeneous model.

Age-of-infection model. We start with the same age-of-infection model as described in the main text, but include additional time-dependent modulation of the force of infection :

$$J(t) = \mu(t)\left\langle \int_0^\infty d\tau R_\alpha K(\tau) j_\alpha(t-\tau)\right\rangle \tag{S1}$$

Here, the modulation factor $\mu(t)$ can be due (e.g.) to mitigation measures or seasonal forcing. Due to this modification, Eq. (5) should be rewritten as follows:

$$\frac{R_e(t)}{\mu(t)R_0} \equiv S_R(t) = \frac{1}{R_0}\int_0^\infty \alpha R_\alpha f(\alpha) e^{-\alpha Z(t)} d\alpha \tag{S2}$$

Here $R_0 = \int_0^\infty \alpha R_\alpha f(\alpha) d\alpha$ is the basic reproduction number. As a reminder, we set $\langle\alpha\rangle = 1$. Now one can write the renewal equation for force of infection which is formally identical to the one for a homogeneous case:

$$J(t) = \mu(t) R_0 \int_0^\infty d\tau K(\tau) S_R(t-\tau) J(t-\tau) \tag{S3}$$

As discussed in the main text, equations for incidence rate $\dot{S}$, S_e and S complete our quasi-homogeneous description:

$$\frac{dS(t)}{dt} = -S_e(t)J(t) \tag{S4}$$

$$S = \int_0^\infty f(\alpha) e^{-\alpha Z(t)} d\alpha \tag{S5}$$

$$S_e = \int_0^\infty \alpha f(\alpha) e^{-\alpha Z(t)} d\alpha = -\frac{\partial S(Z)}{\partial Z} \tag{S6}$$

The set of Eqs.(S3)-(S4) completely describes the epidemic dynamics, as long S_R and S_e are specified as functions of fraction of susceptible population, S.

Compartmentalized SIR/SEIR models. The basic SIR and SIER models can be viewed as particular cases of the age-of infection model discussed above. However, because of their great importance and wide use, we present our construction for a specific case of SEIR:

$$\dot{S}_\alpha = -\alpha S_\alpha J \tag{S7}$$

$$\dot{E}_\alpha = \alpha S_\alpha J - \gamma_E E_\alpha \tag{S8}$$

$$\dot{I}_\alpha = \gamma_E E_\alpha - \gamma_I I_\alpha \tag{S9}$$

Here, $J(t) = \mu(t)\gamma_I \int_0^\infty R_\alpha I_\alpha f(\alpha) d\alpha$ is force of infection. We define infectivity-weighted "Exposed" and "Infectious" fractions as

$$E(t) = \frac{1}{R_0}\int_0^\infty R_\alpha E_\alpha f(\alpha) d\alpha \tag{S10}$$

$$I(t) = \frac{J}{\gamma_I \mu(t) R_0} = \frac{1}{R_0}\int_0^\infty R_\alpha I_\alpha f(\alpha) d\alpha \tag{S11}$$

This leads to a complete description of epidemic dynamics with three ordinary differential equations, formally equivalent to those for the homogeneous case:

$$\dot{S}(t) = J(t)\int_0^\infty \alpha S_\alpha f(\alpha) d\alpha = -\gamma_I R_0 \mu(t) S_e(t) I(t) \tag{S12}$$

$$\dot{E}(t) = \frac{J}{R_0}\int_0^\infty \alpha R_\alpha S_\alpha d\alpha - \gamma_E E(t) = \gamma_I R_0 \mu(t) S_R(t) I(t) - \gamma_E E(t) \tag{S13}$$

$$\dot{I}(t) = \gamma_E E(t) - \gamma_I I(t) \tag{S14}$$

Eqs.(S2),S5 and (S6) relating functions S_R, S and S_e can be recovered by using exponential Anzats, $S_\alpha = e^{-\alpha Z}$.

Correlation parameter and scaling relationship between infectivity and susceptibility. Below we consider a model in which biological susceptibility α_b is correlated neither with infectivity nor with social strength α_s of an individual. On the other hand, both the overall susceptibility and infectivity are proportional to α_s. Let f_x and f_y be probability density functions (pdfs) of variables $x \equiv \ln\alpha_s$ and $y \equiv \ln\alpha_b$. It is reasonable to assume a log-normal distribution for α_b, since biological susceptibility can be modeled as a product of several random factors (due to age, gender, genetics, pre-existent conditions, etc). This corresponds to a Gaussian form for f_y with variance σ^2 and mean $-\sigma^2/2$ (assuming normalization $\langle\alpha_b\rangle = 1$). For a given value of α, this translates into Gaussian distribution of variable x with the same variance, and mean $\ln\alpha + \sigma^2/2$. This allows us to calculate the average α_s which is proportional to R_α:

$$R_\alpha \sim \langle\alpha_s\rangle \sim \frac{\int f_x(x)\exp\left(x - \frac{(x-\ln\alpha-\sigma^2/2)^2}{2\sigma^2}\right)dx}{\int f_x(x)\exp\left(-\frac{(x-\ln\alpha-\sigma^2/2)^2}{2\sigma^2}\right)dx} \qquad \text{[S15]}$$

This integral can be evaluated by the method of steepest descents: for most pdfs f_x and f_y, will be dominated by the vicinity of point x_0 defined by the condition $f'(x_0)/f(x_0) = (x_0/\sigma^2 - 1/2)$. By expanding $\ln f(x)$ in $x' = x - x_0$, we obtain $f_x(x') \approx f(x_\sigma)\exp(rx' - \kappa x'^2/2)$, where $r = f'(x_0)/f(x_0) = x_0/\sigma^2 - 1/2$ and $\kappa = -f''(x_0)/f(x_0) + r^2$. After substituting this Gaussian approximation for f_x back into the above equation, we obtain the scaling relationship between α and R_α

$$R_\alpha \sim \exp\left(\frac{(\sigma^2+\ln\alpha)^2 - (\ln\alpha)^2}{2\sigma^2(1+\kappa\sigma^2)}\right) \sim \alpha^\chi \qquad \text{[S16]}$$

Here $\chi = 1/(1+\kappa\sigma^2)$.

Functions $S_R(S)$ and $S_e(S)$. According to Eq.(S5), function $S(Z)$ is directly related to the moment generating function M_α for pdf $f(\alpha)$

$$S = \langle e^{-\alpha Z}\rangle_\alpha = M_\alpha(-Z) = 1 - Z + \frac{\langle\alpha^2\rangle Z^2}{2} - \frac{\langle\alpha^3\rangle Z^3}{6} + \cdots \qquad \text{[S17]}$$

This function also determines the effective fraction of susceptible population S_e, Eq.(S6):

$$S_e = \langle\alpha e^{-\alpha Z}\rangle_\alpha = -\frac{dS}{dZ} \qquad \text{[S18]}$$

Once effective susceptible fraction S_e is expressed as function of S, it completely determines how S_R (and hence R_e) behaves in both limiting cases of strong and weak correlations, respectively:

$$S_R^{(\chi)} = \begin{cases} \langle\alpha e^{-\alpha Z}\rangle_\alpha = -dS/dZ = S_e, & \chi = 0 \\ \frac{1}{\langle\alpha^2\rangle}\frac{dS^2}{dZ^2} = \frac{S_e}{\langle\alpha^2\rangle}\frac{dS_e}{dS}, & \chi = 1 \end{cases} \qquad \text{[S19]}$$

Application to specific distributions of susceptibility.

Gamma distribution. Consider the gamma distribution with $\langle\alpha\rangle = 1$ and $CV_\alpha^2 = \eta$:

$$f(\alpha) \sim \alpha^{1/\eta - 1}\exp(-\alpha/\eta) \qquad \text{[S20]}$$

By using Eqs.(S2),(S5),(S6), we obtain:

$$S = (1+\eta Z)^{-1/\eta} \qquad \text{[S21]}$$

$$S_e = (1+\eta Z)^{-1/\eta - 1} = S^{1+\eta} \qquad \text{[S22]}$$

$$S_R = (1+\eta Z)^{-(1+(\chi+1)/\eta)} = S^\lambda \qquad \text{[S23]}$$

This leads to the scaling relationship $R_e = R_0 S^\lambda$, Eq. (14).

Truncated power law distribution. We now consider power law distributed α, $f(\alpha) \sim 1/\alpha^{1+s}$ $(s > 0)$, with upper and lower cut-offs, $\epsilon\alpha_+$ and α_+, respectively. If the upper cut-off is implemented as an exponential factor $\exp(-\alpha/\alpha_+)$, we recover the functional form identical to the gamma distribution, Eq. (S20) discussed above, but with negative values of the shape factor:

$$f(\alpha) = \frac{\alpha_+^{q-1}\exp(-\alpha/\alpha_+)}{\alpha^q\Gamma(1-q,\epsilon)} \qquad \text{[S24]}$$

Due to the normalization $\langle\alpha\rangle = 1$,

$$\alpha_+ = \frac{\Gamma(1-q,\epsilon)}{\Gamma(2-q,\epsilon)}. \qquad \text{[S25]}$$

In the case of gamma distribution, the coefficient of variation CV_α would completely determine the overall shape of pdf. For power law with exponent $1 \leq q \leq 3$, the value of $\eta = CV^2$ sets the dynamic range the between upper and lower cut-offs, i.e. the parameter ϵ:

$$1 + \eta = \langle \alpha^2 \rangle = \frac{\Gamma(1-q,\epsilon)\Gamma(3-q,\epsilon)}{\Gamma(2-q,\epsilon)^2} \tag*{[S26]}$$

By using Eq. (S2) and (S5), we obtain exact results for S and S_R in terms of Z:

$$S = \frac{\Gamma(1-q,\epsilon(1+\alpha_+ Z))}{\Gamma(1-q,\epsilon)(1+\alpha_+ Z)^{1-q}} \tag*{[S27]}$$

$$S_R = \frac{\Gamma(\nu,\epsilon(1+\alpha_+ Z))}{\Gamma(\nu,\epsilon)(1+\alpha_+ Z)^{\nu}} \tag*{[S28]}$$

Here $\nu = 2 + \chi - q$. The resulting function $R_e/R_0 = S_R(S)$ is shown in Fig. S1 for several values of the exponent q.

For $\chi = 0$, the overall function $S_R(S) = S_e(S)$ can be very well fitted by an empirical analytic formula that depends only on $\lambda_0 = 1 + CV_\alpha^2$ and an additional shape parameter $\Delta_\lambda = CV_\alpha(\gamma_\alpha - 2CV_\alpha)$:

$$S_e \approx \frac{S}{(1+\Delta_\lambda(1-S))^{(\lambda_0-1)/\Delta_\lambda}} \tag*{[S29]}$$

According to Eq. (S19), this function completely defines behavior of S_R in both limits of the weak and strong correlation regimes :

$$S_R \approx \frac{(1+(\Delta_\chi-1)(1-S))\,S}{(1+\Delta_\lambda(1-S))^{(\lambda-\Delta_\chi)/\Delta_\lambda}} \tag*{[S30]}$$

Here $\Delta_\chi = (\Delta_\lambda + 1)/\lambda_0$, and $\lambda = \lambda_1$ for $\chi = 1$. For $\chi = 0$, δ_χ has to be set to 1.

Log-normal distribution. The log-normal distribution is a very natural candidate to describe statistics of α. It universally emerges for multiplicative random processes. Transmission of an infection involves a complex chain of random events, both social and biological, which can be conceptualized as such multiplicative process. For instance, it may depend on how likely a given person would be involved in a potential superspreading event, how likely that person would have a close contact with a potential infector, what would be the duration of their contact, how effective the individual immune system is in preventing and suppressing the infection.

For the log-normal distribution, the initial drop in R_e according to Eqs. (10), is noticeably faster than for a gamma distribution: $\lambda = (1 + CV_\alpha^2)(1 + \chi CV_\alpha^2)$. However, the initial linear regime is also much narrower. Figure S1 shows the dependence $R_e(S)$ for the log-normal distribution alongside with the above results for gamma and power law distributions computed for the same values of CV (specifically, $CV_\alpha^2 = 2$). As one can see from these plots, despite a stronger effect of heterogeneity at the early stage, the curves generated by log-normal distribution approach $R_e = 0$ significantly slower than those corresponding to the gamma distribution. Note that the overall behavior of $R_e(S)$ generated by the log-normal distribution closely matches the one obtained for the power law distribution with a certain scaling exponent q. This exponent would depend on CV and should approach 1 in the limit of sufficiently wide distribution when the log-normal pdf asymptotically approaches a power law $1/\alpha$ with upper and lower cut-offs.

Final Size of Epidemic. Here we derive a simple result for the final size of epidemic in a population with a persistent heterogeneity. To do this, we integrate Eq. (S3) over time t, assuming no mitigation, $\mu(t) = 1$. This yields a relation $Z_\infty = \int_0^\infty R_e(t)J(t)dt = \int_{S_\infty}^1 R_e(S)dS/S_e(S)$ for the final value of Z when the epidemic has run its course, and this in turn can conveniently be expressed in terms of the fraction of the susceptible population, S_∞:

$$S_\infty = M_\alpha\left(-\int_{S_\infty}^1 \frac{R_e(S)dS}{S_e(S)}\right) \tag*{[S31]}$$

This equation is valid for an arbitrary distribution of α, arbitrary correlation between susceptibility and infectivity, and for any statistics of the generation interval. This result can be also obtained as a solution to a general integral equation derived in Ref. (1) for the well-mixed case. Eq. S31 combines and generalizes several well-known results: (i) in the weak correlation limit ($R_\alpha = R_0$), when the integral in the r.h.s. is equal to $R_0(1 - S_\infty)$, Eq.(S31) reproduces results of Refs. (1–4), (ii) in the opposite limit of a strong correlation ($R_\alpha \sim \alpha$), the integration gives $R_0(1 - S_e(S_\infty))/\langle\alpha^2\rangle$, and one recovers the result for the FSE on a network (1, 5, 6).

For the case of gamma-distributed persistent susceptibility Eq. (S31), gives:

$$S_\infty = \left(1 + \frac{R_0\eta\left(1 - S_\infty^{\lambda-\eta}\right)}{\lambda - \eta}\right)^{-1/\eta} \tag*{[S32]}$$

It should be emphasized however that this result is of limited relevance to more realistic situations. Even if one assumes no government-imposed mitigation or societal response to the epidemic, the case of fully persistent heterogeneity is just an approximation. As we demonstrate in our paper, short-term correlations of time-dependent individual susceptibilities and infectivities lead to transient stabilization of a fast-pacing epidemic. Because of this effect, Eq.(S31)-Eq.(S32) should be interpreted as an estimate of the size of the first wave rather than the actual FSE.

Path-integral theory of epidemic with time-dependent heterogeneity. Here we present a generalization of the theory developed in the previous section that incorporates the effects of time variations of individual susceptibilities and infectivities, as well as temporal correlations between them. Since these fast variations are primarily caused by bursty dynamics of social interactions, and since heterogeneous biological susceptibility appears subdominant in the context of COVID-19, we set $\alpha_b = 1$ for all individuals, so that α has purely social origin. Let $a_i(t) = \alpha_i + \delta a_i(t)$ be the time-dependent susceptibility of a person. Because of the social nature of $a(t)$, one's individual infectivity is also proportional to it at any given time: $\beta_i(t) = R \cdot K(\tau) a_i(t)$. As before, τ is time from infection, $K(\tau)$ is the pdf of generation intervals. Accordingly R is individual reproductive number of an "average" person with social activity $a_i(t) = 1$, in the fully susceptible population. The state of an individual is described by a step function $s_i(t)$ which is 1 as long as the person is susceptible, and turns to 0 at the moment of infection. The time evolution of the epidemic follows a stochastic generalization of (Eqs.(1)-(2):

$$\mathrm{E}\left[\dot{s}_i(t)\right] = -a_i(t) s_i(t-0) J(t) \qquad \text{[S33]}$$

$$J(t) = -\int_0^\infty R \cdot K(\tau) \overline{a_i(t)\dot{s}_i(t-\tau)} d\tau \qquad \text{[S34]}$$

Here bar $\overline{\ldots}$ represents averaging over individual members of population (indexed by i), in contrast with $\langle \ldots \rangle$, averaging over all subgroups with various values of persistent heterogeneity α. E$[\ldots]$ stands for expected value.

The overall quasi-homogeneous description given by Eqs.S3),(S4), remains valid. It is obtained by averaging Eqs. (S33)-(S34) over the entire population. However, in contrast to the case of persistent heterogeneity, variables $S(t)$, $S_e(t)$ and R_e are no longer connected to each other by a simple functional relationship. To relate them we first note that the average probability that an individual is still susceptible at time t is given by $E[s_i(t)] = \exp\left(-\int_{-\infty}^t J(t')a_i(t')dt'\right)$. Therefore,

$$S(t, [J(t')]) \equiv \overline{s_i(t)} = \overline{\exp\left[-\int_{-\infty}^t J(t')a_i(t')dt'\right]} \qquad \text{[S35]}$$

In other words, S becomes a functional over the set of all possible epidemic trajectories $J(t)$. It still has the structure of a moment generating function for the field $a_i(t)$, and thus is a direct analogue of the partition function broadly used in statistical physics, stochastic calculus, and field theory. The specific form of this functional depends on probabilities assigned to different individual trajectories $a_i(t)$. As a natural generalization of the case of persistent heterogeneity, $S_e(t)$ and $R_e(t)$ can be obtained as, respectively, the first and the second variations of the functional S over $J(t)$:

$$S_e(t) = \overline{a_i(t)s_i(t)} = -\frac{\delta S(t, [J(t')])}{\delta J(t)} \qquad \text{[S36]}$$

$$R_e(t) = R\overline{a_i^2(t)s_i(t)} = R\frac{\delta^2 S(t, [J(t')])}{\delta J(t)^2} \qquad \text{[S37]}$$

For the sake of simplicity, in deriving Eq.(S37) we assumed $\alpha_i(t)$ to be smoothed over the timescale of a single generation interval. As a result, $\int_0^\infty \overline{a_i(t)a_i(t-\tau)}K(\tau)d\tau \approx \overline{a_i^2(t)}$. By applying Eq.(S37) to the initial state of fully susceptible population we obtain the result for R_0:

$$R_0 = R\overline{a_i^2} = R\left(\langle\alpha^2\rangle + \overline{\delta a_i^2}\right) \qquad \text{[S38]}$$

At the early stages of epidemic $E[s_i(t)] \approx 1 - \int_{-\infty}^t a_i(t')J(t')dt'$ for the entire population. After substituting this expression for $s_i(t)$ to Eq.(S37) one obtains a generalization of our previous result for the initial suppression of R_e, Eqs.(9)-(10):

$$R_e(t) \approx R_0\left(1 - \int_0^\infty \Lambda(t,t')J(t-t')dt'\right) \qquad \text{[S39]}$$

$$\Lambda(t,t') = \frac{\overline{\tilde{\alpha}_i^2(t)a_i(t-t')}}{\langle\alpha^2\rangle + \overline{\delta a_i^2}} = \lambda_\infty + \delta\lambda(t,t') \qquad \text{[S40]}$$

Here $\lambda_\infty = \Lambda(\infty)$ and $\delta\lambda(t,t')$ are the constant and time-dependent contributions to "immunity kernel" $\Lambda(t,t')$ which are discussed in the main text in the context of the definition of λ_{eff}. Note that since the empirical value of λ_{eff} for COVID-19 is relatively large (between 4 and 5), the attack rate at which the TCI state would be achieved is rather low (10%-15%) assuming R_0 between 2 and 3. In this case we are well within the range of our linearized regime. The long-term HIT is determined by a lower value of $\lambda_\infty \simeq 2$. In that case we derived the non-linear dependence of R_e on S without linearization.

To obtain a corrected result for HIT, we assume a very slow progression of the epidemic (e.g. due to a gradual relaxation of the level of mitigation). In this case, any intermediate-term correlations between time dependent variations $\delta\alpha_i(t)$ become negligible, and we largely recover the formalism developed for pure persistent heterogeneity. Belowm we make the same assumption that was used for the estimate of λ_∞ in the main text: $\overline{\delta a_i^2} \sim \alpha_i$. This relationship can be rewritten in terms of $\chi^* = \langle\alpha^2\rangle/\overline{a_i^2}$, as $\overline{\delta a_i^2} = \chi^*\alpha_i\langle\alpha^2\rangle$. It leads to the following modification to the result for S_R, Eq(S2):

$$S_R = \frac{R}{R_0}\int\left(\alpha^2 + (1-\chi^*)\overline{a_i^2}\alpha\right) f(\alpha)e^{-\alpha Z}d\alpha = \left(\chi^*\langle\alpha^2 e^{-Z\alpha}\rangle + (1-\chi^*)\langle\alpha e^{-Z\alpha}\rangle\right) \qquad \text{[S41]}$$

Here we used Eq. (S38) for R_0. As one can see, S_R is a linear combination of two earlier results $S_R^{(\chi)}$, for $\chi = 0$ and 1, respectively (both are given by Eq. (S19)):

$$S_R = \chi^* S_R^{(1)} + (1-\chi^*) S_R^{(0)} \qquad \text{[S42]}$$

For the case of gamma-distributed α, this leads to interpolative result

$$S_R = \chi^* S^{1+\eta} + (1-\chi^*) S^{1+2\eta} \approx S^{\lambda_\infty} \qquad \text{[S43]}$$

here $\lambda_\infty = 1 + (1+\chi^*)\eta$ is long-term immunity factor obtained in the main text within linearized approximation for $R_e(S)$.

A more detailed derivation of these and other results is at Ref. (7).

Age-of-infection model used to simulate the epidemic dynamics in NYC and Chicago. In our simulations of the COVID-19 epidemic dynamics in Chicago and NYC shown in Fig. 4 we used the age-of-infection model we previously developed for the state of Illinois (8). The daily incidence (i.e. the daily number of newly-infected individuals per capita) $j(t)$, determines the dynamics of susceptible individuals according to

$$\frac{dS(t)}{dt} = -j(t). \qquad \text{[S44]}$$

The incidence itself follows the renewal equation,

$$j(t) = R_e(t) \int_0^\infty d\tau \, K(\tau) j(t-\tau). \qquad \text{[S45]}$$

Here, $R_e(t)$ is time-dependent effective reproduction number, $K(\tau)$ is the probability density function (PDF) of generation intervals. We parameterize the effective reproduction number $R_e(t)$ according to Eq. 14

$$R_e(t) = R_0 \mu(t) S(t). \qquad \text{[S46]}$$

where R_0 is the basic reproduction number, $\mu(t)$ a mitigation factor, $S(t)$ is susceptible population fraction.

For both NYC and Chicago, we have access to reliable data (9–12) describing time series of the following variables

- $H(t)$, the total number of hospitalized (but not critical) patients
- $C(t)$, the number of critically ill patients currently in ICU beds
- $D(t)$, the cumulative number of daily deaths.

In our age-of-infection model changes in these variables are described by daily flux variables:

- $\sigma(t)$, the number of infected individuals who become symptomatic
- $h(t)$, the number of daily admissions to all hospitals
- $r(t)$, the daily number of patients discharged from all hospitals
- $c(t)$, the daily number of patients transferred from the main floor of a hospital to its ICU
- $v(t)$, the daily number of patients transferred from the ICU to the main floor of a hospital, and
- $d(t)$, the daily number of deaths in ICU rooms.

Figure S4 schematically depicts the topology of our model along with the names of all flux and cumulative variables. The dynamics of any flux variable $y(t)$ defined above may be obtained from the variable $x(t)$ directly preceding it in the chain of events shown in Fig. S4:

$$y(t) = p_y \int_0^\infty d\tau \, K_y(\tau) x(t-\tau). \qquad \text{[S47]}$$

Here, p_y is the proportion of individuals undergoing the transition $x \to y$ with time delays distributed according to a probability density function $K_y(t)$.

We fix the generation interval mean and standard deviation to 4 and 3.25 days respectively (13, 14), while our incubation time distribution has fixed mean 5.5 days and a standard deviation of 2 days (15, 16). We calibrate the remaining delay (τ_x) and fraction (p_x) parameters shown in Figure S4 to data downloaded from (9–12) by sampling over the high-dimensional model parameter space using a Markov chain Monte Carlo (MCMC) approach as described in details in Ref. (8). This procedure allowed us to determine the time evolution of $R_e(t)$ and $S(t)$ in NYC and Chicago, which was used in Fig. 3. Figure S2 shows the $R_e(t)$ divided by the mobility factor calculated from Google community mobility report, Ref. (17). We use the average of Retail, Grocery, Transit, and Workplaces categories. For NYC we use the average mobility of its five counties: New York county, Bronx county, Kings county, Richmond county, and Queens county, weighted by their population fractions.

To construct Fig. 4 we modified our simulations by replacing Eqs. (S44), (S45) with their quasi-homogeneous generalizations Eqs. (S4), (S3) and setting $S_R = S^{\lambda}$ and $S_e = S^{1+\eta}$, where $\eta = (\lambda - 1)/2$. After calibrating this model on data up to June 10, 2020 we predicted the effect of relaxing the mitigation factor $\mu(t)$ back to 1 on June 15, 2020. The results of these simulations are shown in Fig. 4. Figs. S5 (NYC) and S6 (Chicago) show our predictions along with 95% confidence intervals caused by parameter uncertainty. Our predictions go up to early September 2020 at which point confidence intervals become too wide.

References

1. Miller JC (2012) A note on the derivation of epidemic final sizes. *Bulletin of mathematical biology* 74(9):2125–2141.
2. Novozhilov AS (2008) On the spread of epidemics in a closed heterogeneous population. *Mathematical Biosciences* 215(2):177–185.
3. Katriel G (2012) The size of epidemics in populations with heterogeneous susceptibility. *Journal of Mathematical Biology* 65(2):237–262.
4. Ball F (1985) Deterministic and stochastic epidemics with several kinds of susceptibles. *Advances in Applied Probability* 17(1):1–22.
5. Newman MEJ (2002) Spread of epidemic disease on networks. *Physical Review E* 66(1):016128.
6. Moreno Y, Pastor-Satorras R, Vespignani A (2002) Epidemic outbreaks in complex heterogeneous networks. *European Physical Journal B* 26(4):521–529.
7. Tkachenko AV, et al. (2021) Stochastic social behavior coupled to covid-19 dynamics leads to waves, plateaus and an endemic state. *medRxiv* 2021.01.28.21250701.
8. Wong GN, et al. (2020) Modeling covid-19 dynamics in illinois under non-pharmaceutical interventions. *Physical Review X* 10.
9. (2020) Data were downloaded from https://www.dph.illinois.gov/covid19/covid19-statistics.
10. (2020) Data from https://github.com/thecityny/covid-19-nyc-data.
11. (2020) Data originally due to https://www.thecity.nyc/.
12. (2020) Data from https://github.com/nychealth/coronavirus-data.
13. Nishiura H, Linton NM, Akhmetzhanov AR (2020) Serial interval of novel coronavirus (COVID-19) infections. *International Journal of Infectious Diseases* 93:284–286.
14. Du Z, et al. (2020) Serial Interval of COVID-19 among Publicly Reported Confirmed Cases. *Emerg Infect Dis* 26(6):2020.02.19.20025452.
15. Lauer SA, et al. (2020) The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. *Annals of Internal Medicine* 172(9):577.
16. Linton NM, et al. (2020) Incubation Period and Other Epidemiological Characteristics of 2019 Novel Coronavirus Infections with Right Truncation: A Statistical Analysis of Publicly Available Case Data. *Journal of Clinical Medicine* 9(2):538.
17. (2020) https://www.blog.google/technology/health/covid-19-community-mobility-reports?hl=en.
18. Unwin HJT, et al. (2020) Report 23: State-level tracking of COVID-19 in the United States WHO Collaborating Centre for Infectious Disease Modelling MRC Centre for Global Infectious Disease Analytics.

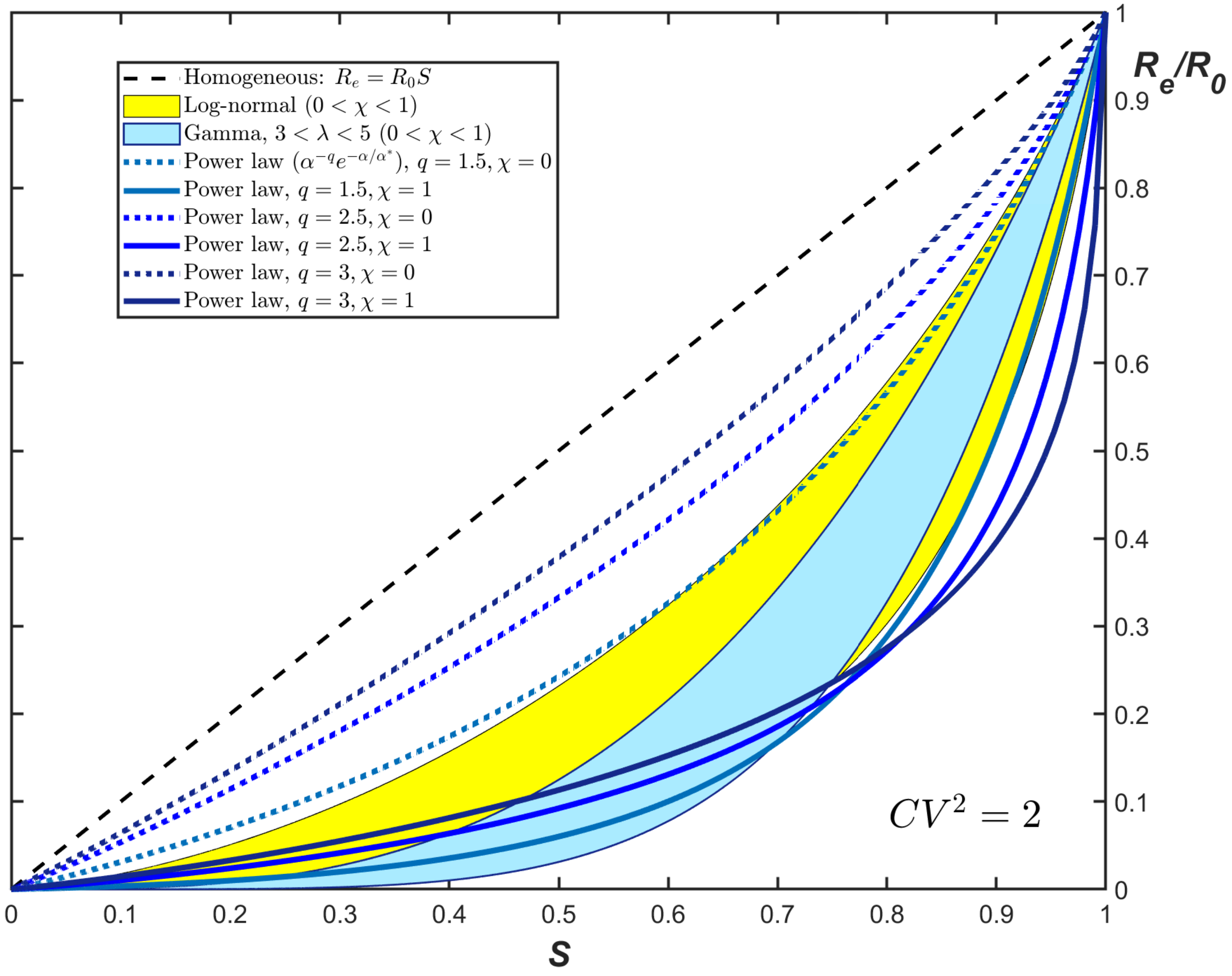

Fig. S1. R_e/R_0 vs S dependence for three different families of probability distribution $f(\alpha)$: Gamma (light blue), truncated power law (dashed lines), and log-normal (yellow). Different curves correspond to the same value of the coefficient of variation $CV_\alpha^2 = 2$, and two limiting values (0 and 1) of the correlation parameter χ.

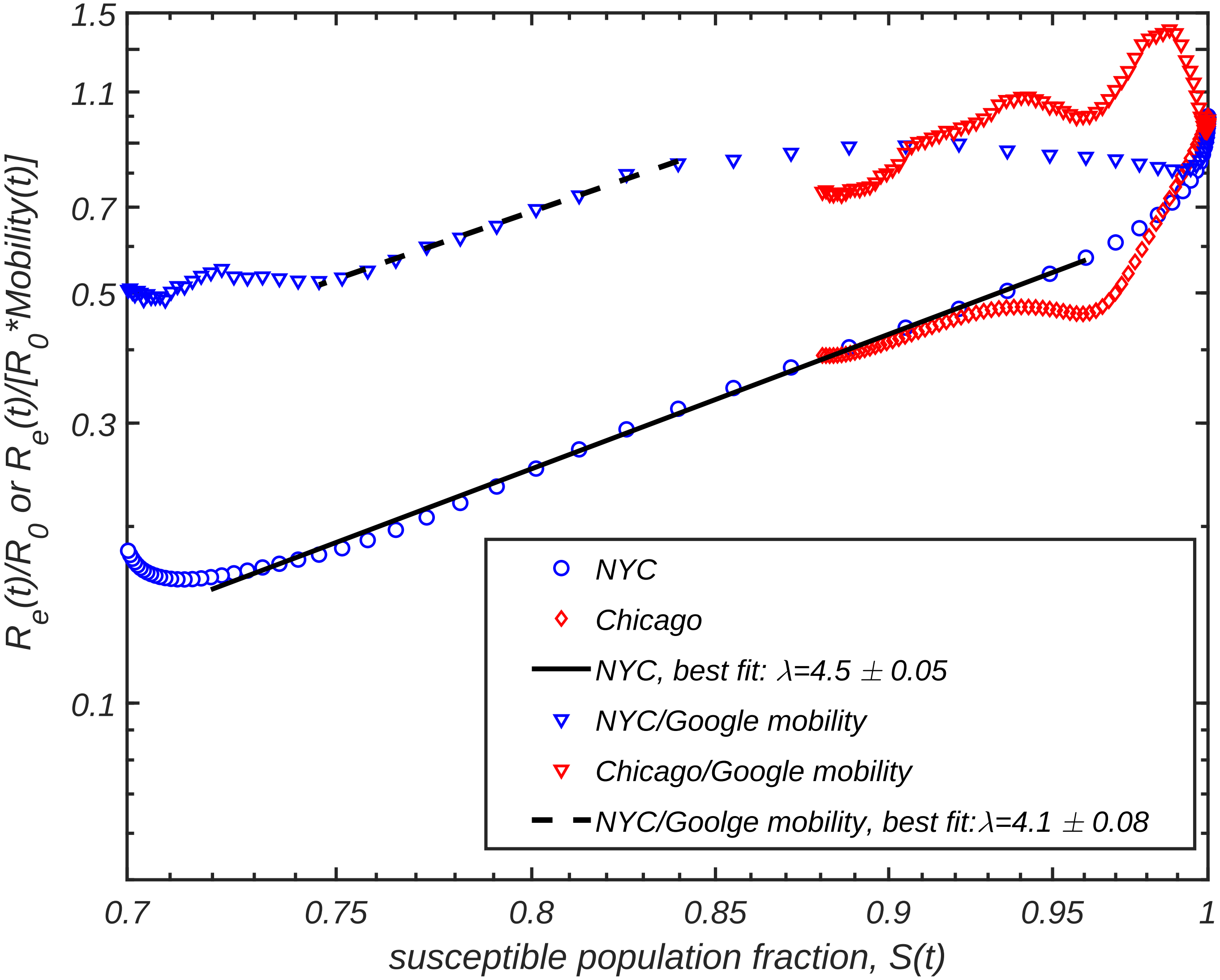

Fig. S2. Exploration of effect of mobility on data presented in Figure 3(A). Triangles represent data points for NYC and Chicago with $R_e(t)/R_0$ corrected by a mobility factor calculated from Google community mobility report, Ref. (17). We compute the mobility for NYC using average mobility of its five counties: New York county, Bronx county, Kings county, Richmond county, and Queens county, weighted by their population fraction.

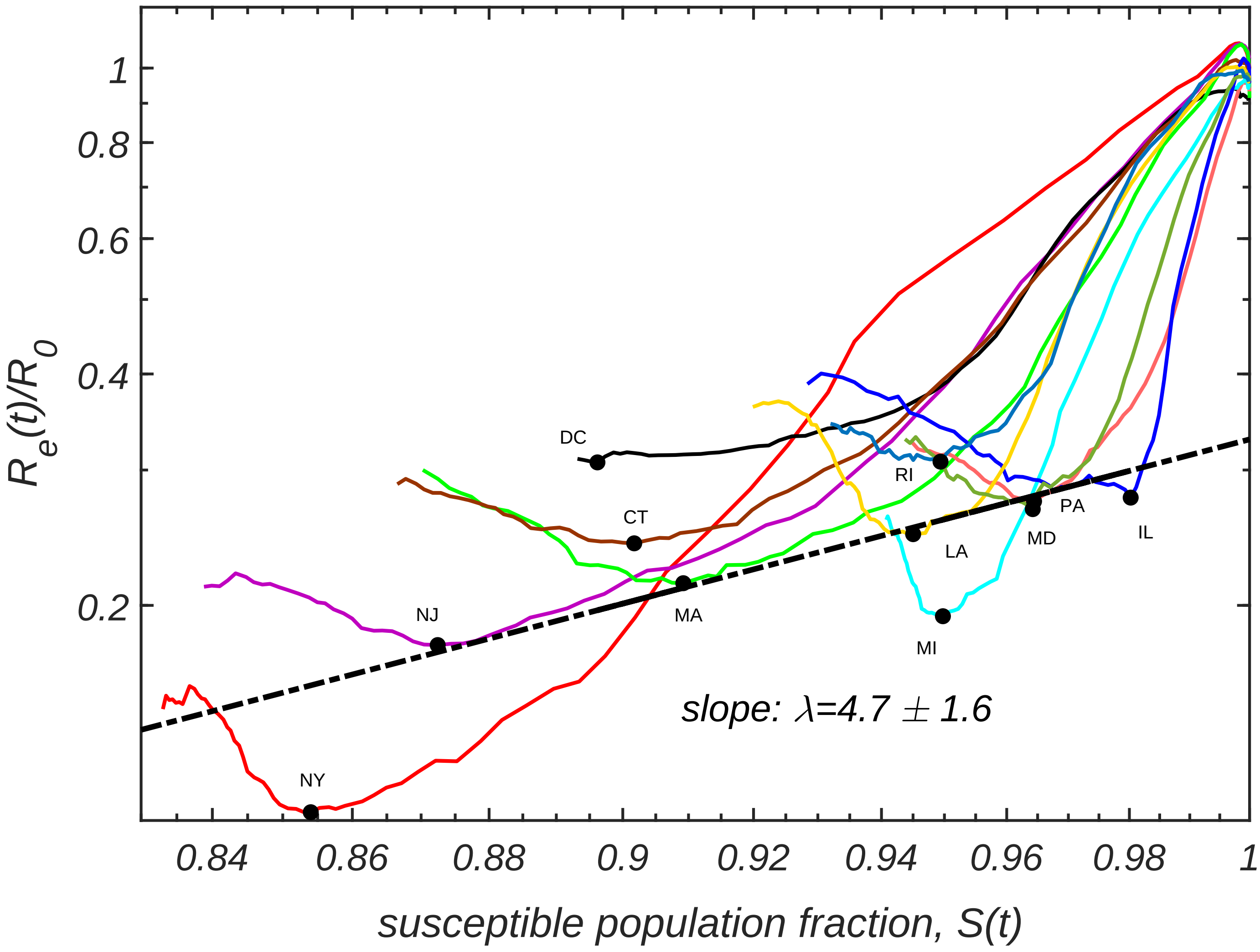

Fig. S3. Time progressions of $Re(t)/R_0$ and $S(t)$ for the hardest-hit US states and DC, as reported in Ref. (18). Black dots correspond to absolute minima of transmission and population susceptible fractions. The dashed line with slope $\lambda = 4.7 \pm 1.6$ is the best power law fit through these black dots.

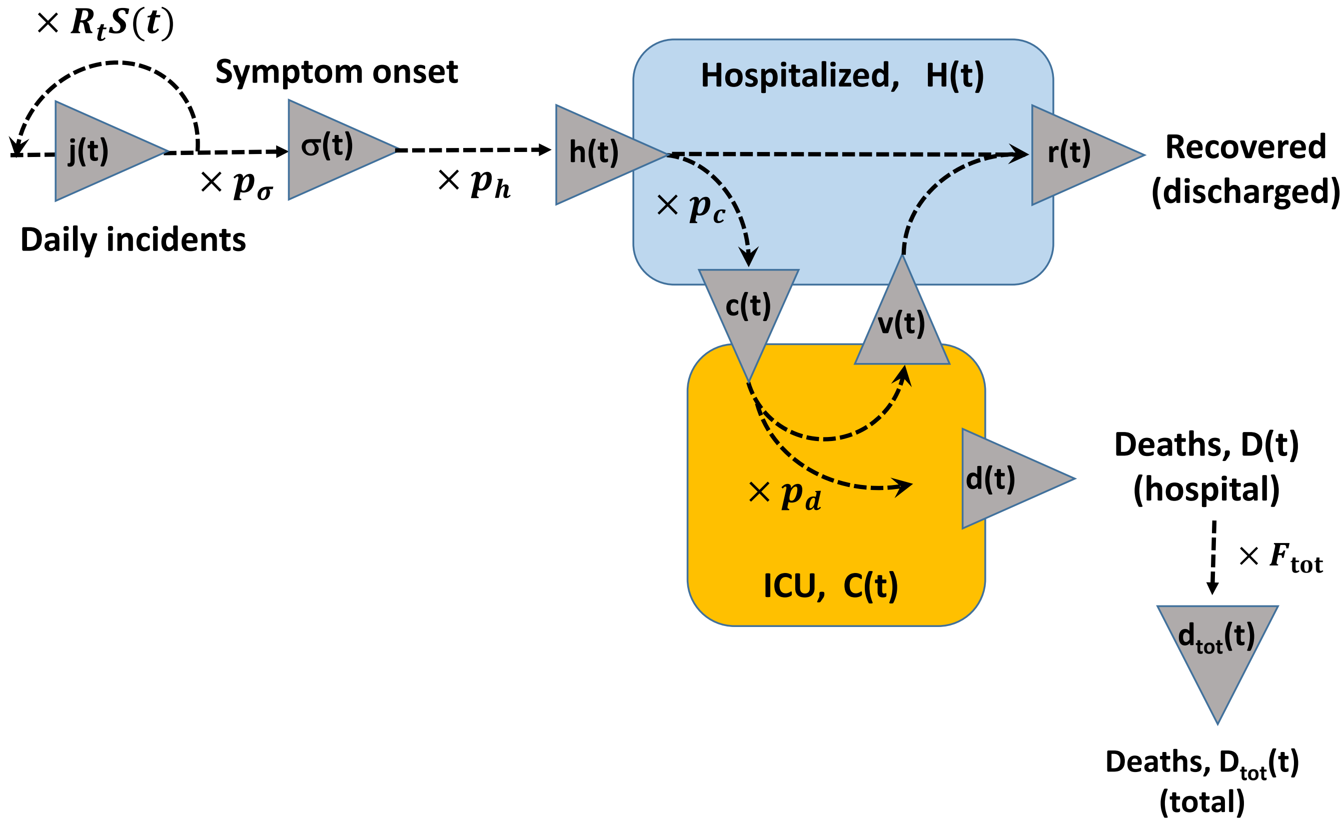

Fig. S4. The topology of our model along with the names of all flux and state variables: the daily incidence, $j_i(t)$; the daily number of newly symptomatic individuals, $\sigma_i(t)$; the number of daily admissions to all hospitals, $h_i(t)$; the daily number of patients discharged from all hospitals, $r_i(t)$; the daily number of patients transferred from the main floor of a hospital to its ICU, $c_i(t)$; the daily number of patients transferred from the ICU to the main floor of a hospital, $v_i(t)$; the daily number of deaths in hospitals, $d_i(t)$; and the daily number of deaths in and out of hospitals, $d_{\text{tot},i}(t)$. State variables are: the total number of occupied hospital beds (main floor) $H_i(t)$, and the total number of occupied ICU beds $C_i(t)$. The other parameters of the model are the fractions of infected individuals who ever become symptomatic, $p_{\sigma,i}$; the fraction of symptomatic individuals who are ever hospitalized, $p_{h,i}$; the fraction of hospital patients who ever get to ICU, $p_{c,i}$; and the fraction of ICU patients who will ultimately die $p_{d,i}$; and the multiplier, F_{tot} that converts between hospital deaths and all deaths in the state, including those outside of the hospital system. For the sake of legibility, we suppress age-group indices in the diagram.

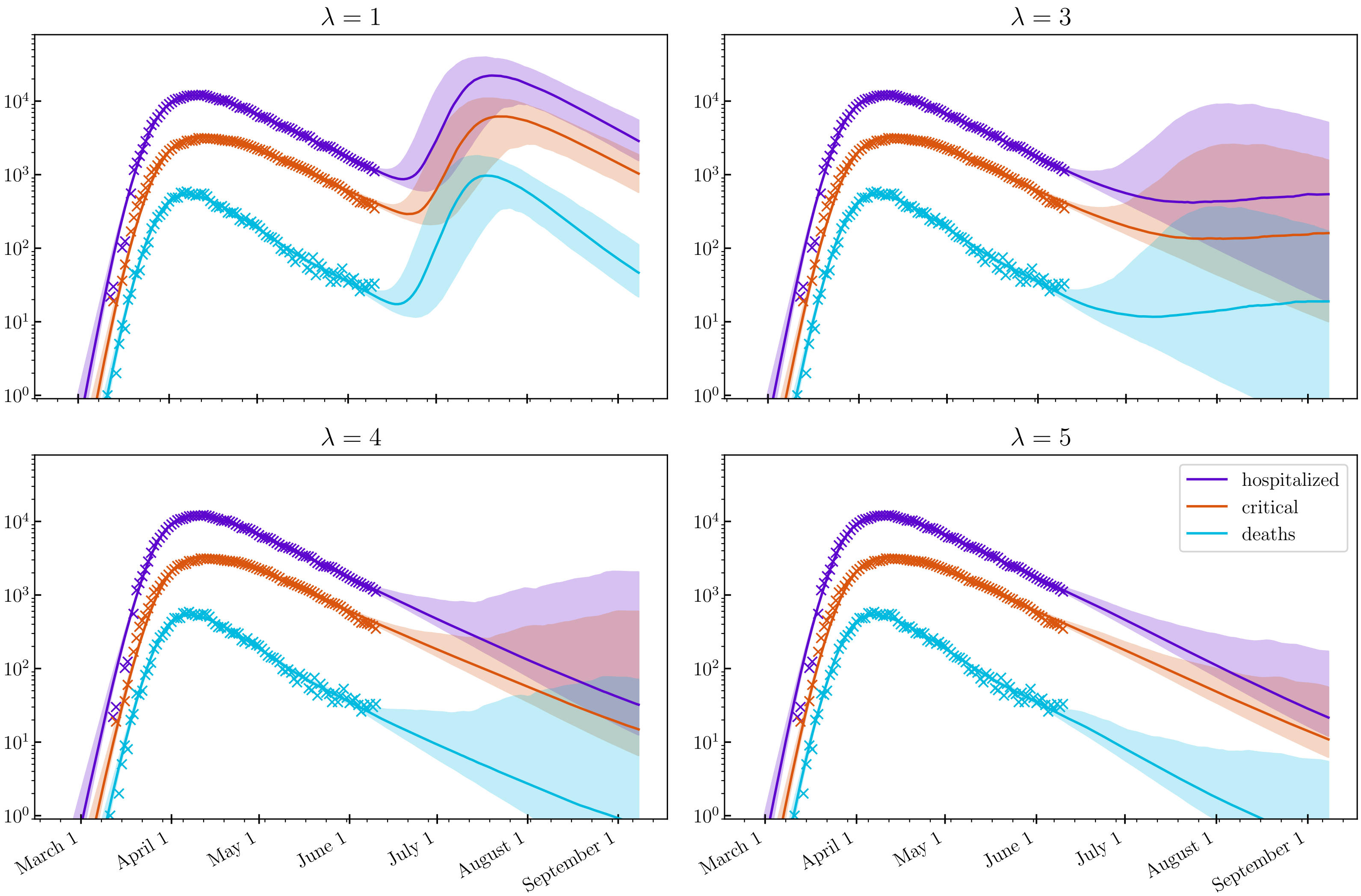

Fig. S5. Hospitalization, ICU occupancy and daily deaths in NYC modeled under hypothetical scenario when any mitigation is completely eliminated as of Jun 15 2020, for various values of λ. Model described in Ref. (8) is calibrated on data from Ref.(9), up to June 10, 2020 (shown as crosses). 95% confidence intervals are indicated.

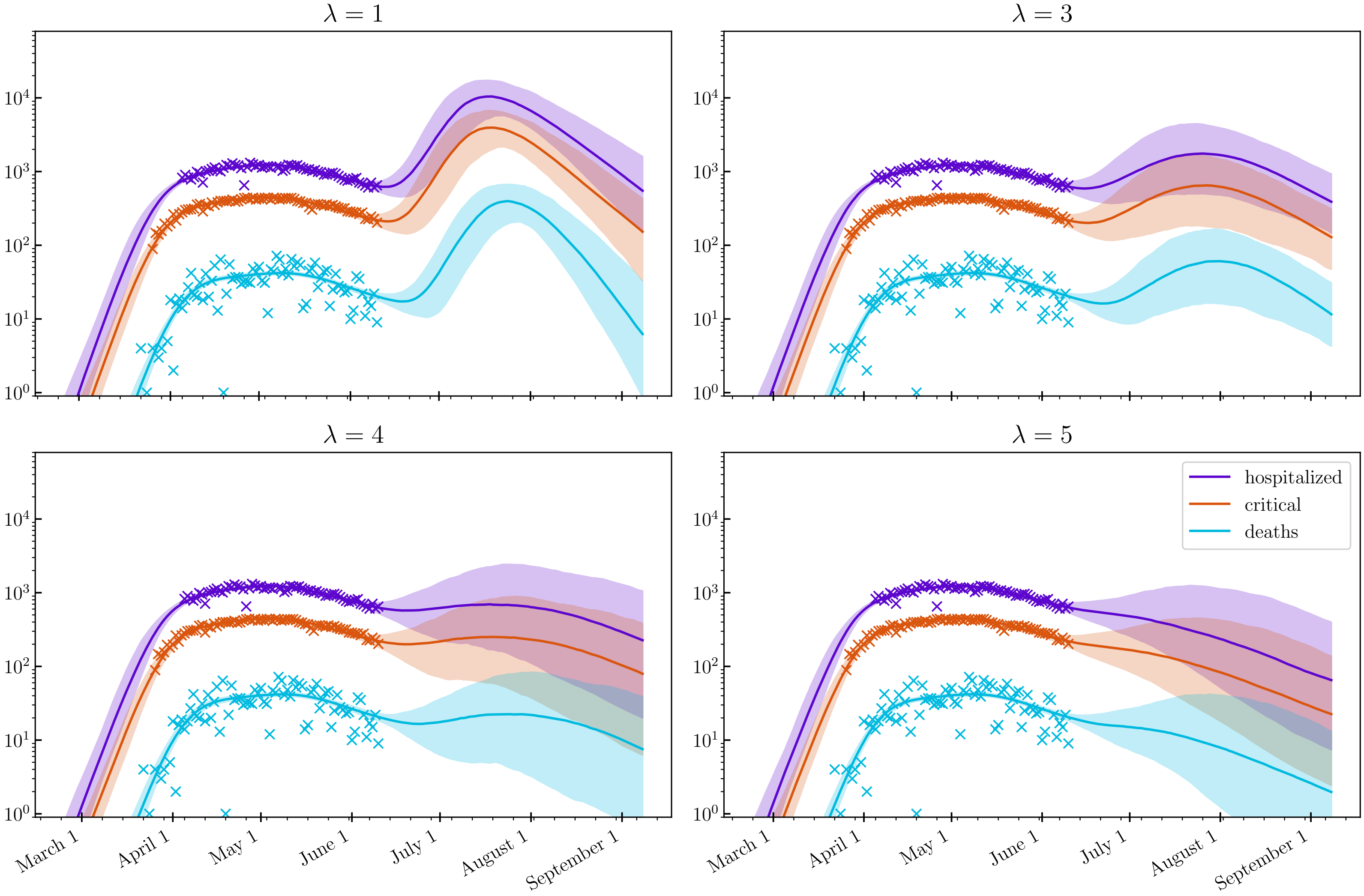

Fig. S6. Hospitalization, ICU occupancy and daily deaths in Chicago modeled under hypothetical scenario when any mitigation is completely eliminated as of Jun 15 2020, for various values of λ. Model described in Ref. (8) is calibrated on data from Ref.(9), up to June 10, 2020 (shown as crosses). 95% confidence intervals are indicated.