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

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

\textsuperscript{a}Center for Functional Nanomaterials, Brookhaven National Laboratory, Upton, NY 11973; \textsuperscript{b}Department of Physics, University of Illinois at Urbana–Champaign, Urbana, IL 61801; \textsuperscript{c}Department of Bioengineering, University of Illinois at Urbana–Champaign, Urbana, IL 61801; \textsuperscript{d}Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana–Champaign, Urbana, IL 61801; \textsuperscript{e}Department of Civil Engineering, University of Illinois at Urbana–Champaign, Urbana, IL 61801; and \textsuperscript{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, super-spreading 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 \( R_0 \) 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 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.

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.


The authors declare no competing interest.

This article is a PNAS Direct Submission.

This open access article is distributed under Creative Commons Attribution License 4.0 (CC BY).

\textsuperscript{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\textsuperscript{-1}1/DCSupplemental.

Published April 8, 2021.

https://doi.org/10.1073/pnas.2015972118
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 superspreader 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 \( \lambda \). It is related to the statistical properties of heterogeneous susceptibilities 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 \( R_e = R_0 S^{1/\lambda} \). This leads to a modified version of the result for the HIT, \( 1 - S_{HI} = 1 - R_e^{-1/\lambda} \)

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 \( \lambda \) 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 to estimate the social contribution to heterogeneous susceptibility, and the corresponding immunity factor $\lambda_s$. The biological contribution, $\lambda_b$, 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 $\lambda$. In particular, we apply our previously described epidemic model that features multichannel Bayesian calibration 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 $R_0(S)$ 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 $\lambda$ 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 susceptibility to infection $\alpha$. This heterogeneity may be biological or social in origin, and we assume these factors are independent: $\alpha = \alpha_s\alpha_b$. Effects of possible correlations between $\alpha_s$ and $\alpha_b$ have been discussed in ref. 24. The biologically driven heterogeneous susceptibility $\alpha_s$ 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 $\alpha_s$. Similarly, people with low immune response, highly social individuals (hubs in social networks), or scofflaws would all be characterized by above-average overall susceptibility $\alpha$.

In this work, we group individuals into subpopulations with similar values of $\alpha$ and describe the heterogeneity of the overall population by the probability density function (pdf) of this parameter, $\delta(\alpha)$. Since $\alpha$ is a relative measure of individual susceptibilities, without loss of generality, we set $\delta(\alpha) \equiv \int_{-\infty}^{\infty} \delta(\alpha) d\alpha = 1$. Each person is also assigned an individual reproductive number $R_\alpha$, which is an expected number of people that this person would infect in a fully susceptible population with $\delta(\alpha) = 1$. Accordingly, from each subpopulation with susceptibility $\alpha$, there is a respective mean reproductive number $R_\alpha$, 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 $R_\alpha \approx \langle \alpha \rangle$ (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 $\alpha$ and $R_\alpha$.

Let $S_\alpha(t)$ be the fraction of susceptible individuals in the subpopulation with susceptibility $\alpha$, and let $j_\alpha(t) = -S_\alpha$ 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:

$$\frac{dS_\alpha}{dt} = j_\alpha(t) = \alpha S_\alpha(t) J(t)$$

$$J(t) = \int_0^\infty \langle R_\alpha K(\tau) j_\alpha(t - \tau) \rangle \ d\tau.$$  

Here $t$ is the physical time, and $\tau$ is the time since infection for an individual; $\langle \ldots \rangle$ represents averaging over $\alpha$ with pdf $\delta(\alpha)$. $J(t)$ is the force of infection, that is, per capita incidence rate in a fully susceptible subpopulation with $\alpha = 1$. $R_\alpha$ is the previously introduced infectivity, that is, the mean reproductive number of the subpopulation with susceptibility $\alpha$. $K(\tau)$ is the pdf of the generation interval, which we assume to be independent of $\alpha$, 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 $\tau$ 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 $\alpha$ can be expressed as

$$S_\alpha(t) = \exp(-\alpha Z(t)).$$ [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 $M_\alpha$ of the distribution $f(\alpha)$ (i.e., the Laplace transform of $f(\alpha)$) according to

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

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

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

Note that, for $Z = 0$, this expression gives the basic reproduction number $R_0 = (\alpha R_e)/\alpha$. This result is reminiscent of the well-known result (23) $R_0 = (k_\text{in} k_\text{out})/(k_\text{in})$ for epidemic spread on a directed network with in and out degrees $k_\text{in}$ (analogue of our susceptibility $\alpha$) and $k_\text{out}$ (analogue of $R_e$). 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, $R_e = \kappa_0 R_0$, 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,

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

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

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

Here

$$S_e(t) = \int_0^\infty f(\alpha)e^{-\alpha Z(t)}d\alpha = -\frac{dM_\alpha(-Z(t))}{dZ}$$ [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 $1 - S(t) \approx Z(t) \ll 1$, one can linearize the effective reproduction number as

$$R_e \approx R_0(1 - \lambda(1 - S)).$$ [9]

We named the coefficient $\lambda$ 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 $\lambda$ is one, but it may be significantly larger in a heterogeneous case. By linearizing Eq. 5 in terms of $1 - S \approx Z \ll 1$ and dividing the result by $R_0 = (\alpha R_e)$, one gets

$$\lambda = \frac{\langle \alpha^2 R_e \rangle}{(\alpha R_e)}.$$ [10]

As one can see, the value of the immunity factor thus depends both on the statistics of susceptibility $\alpha$ and on its correlation with infectivity $R_e$.

We previously defined the overall susceptibility as a combination of biological and social factors: $\alpha = \alpha_b \alpha_s$. Here $\alpha_b$ 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 $\alpha_b \approx \alpha_s$. On the other hand, in our analysis, we neglect a correlation between biological susceptibility and infectivity, as well as between $\alpha_b$ and $\alpha_s$. 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 $\alpha_b$ and $\alpha_s$, respectively,

$$\lambda_b = \frac{\langle \alpha_b^2 \rangle}{(\alpha_b)^2} = 1 + CV_b^2,$$ [11]

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

These equations follow from Eq. 10 in the limit $R_e = \text{const}$ and $R_0 \approx \alpha$, 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 $\lambda_e$ in Eq. 12 is proportional to the third moment of $\alpha_s$ instead of the second moment in the case of $R_0 \approx \langle \alpha R_0 \rangle \approx \alpha^2$. Note that the biological contribution to the immunity factor depends only on the coefficient of variation $CV_b$ of $\alpha_b$. On the other hand, the social factor $\lambda_s$ depends on both the coefficient of variation $CV_s$ and the skewness $\gamma_s$ of the distribution of $\alpha_s$. Due to our normalization, $\langle \alpha_s \rangle = 1$.

The relative importance of biological and social contributions to the overall heterogeneity of $\alpha$ may be characterized by a single parameter $\chi$. For a log-normal distribution of $\alpha_s$, $\chi$ appears as a scaling exponent between infectivity and susceptibility: $R_e \approx \alpha^\chi$ (see SI Appendix for details). The corresponding expression for the overall immunity factor is $\lambda \approx (\alpha^{\chi+2})/(\alpha^{1+\chi})$.

The limit $\chi = 0$ corresponds to a predominantly biological source of heterogeneity, that is, $\lambda \approx \lambda_b = 1 + CV_b^2$, where $CV_b$ 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 $CV_s$ and the skewness $\gamma_s$ of the pdf $f(\alpha)$. The biological contribution $\lambda_b$ 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.

https://doi.org/10.1073/pnas.2015972118

Tkauchenko et al. Time-dependent heterogeneity leads to transient suppression of the COVID-19 epidemic, not herd immunity
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: \( f(\alpha) \approx \alpha^{2/\eta - 1} \exp(-\alpha/\eta) \), where \( \eta = CV^2_\alpha \). In this case, according to Eqs. 4 and 5, \( R_e, S_e, \) and \( S \) are related by scaling relationships (see SI Appendix),

\[
S_e(S) = S^{1+\eta}.
\]

and

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

The exponent \( \lambda = 1 + (1 + \chi)CV^2_\alpha = 1 + (1 + \chi)\eta \) 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 \( \lambda \) 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, \( 1 - S_0 \) is the relative size of the epidemic at that time. By setting \( R_e = 1 \) in Eq. 14, we obtain

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

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 \( R_e \) 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 \( \alpha \) and \( R_0 \), 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 \( (\alpha) \) distribution.

Our focus on the gamma distribution is well justified by the observation that the social strength \( \alpha_s \) is approximately exponentially distributed, that is, it is a specific case of the gamma distribution with \( \eta = CV^2_\alpha = 1 \) (see more discussion of this in the next section). A moderate biological heterogeneity would lead to an increase of the overall \( CV_\alpha \), 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 \( f(\alpha) \approx e^{-\alpha/\alpha_c}/\alpha^q (q \geq 1, \) with an additional cutoff at lower values of \( \alpha \)), 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. \( R_e/R_0 \) 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.
Time-dependent heterogeneity leads to transient suppression of the COVID-19 epidemic, not herd immunity.

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 used to study 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 $f_s(a)$ (19, 46, 48), $f(a) \propto e^{-a/\langle a\rangle}$. That sets the value of $\eta = CV_\text{ISO}^2 = 1$. If not for short-term overdispersion, that would yield $\lambda = (\langle a^3 \rangle / \langle a^2 \rangle) = 3! / 2! = 3$ according to Eq. 12. However, with temporal effects taken into account, the buildup of long-term collective immunity is determined by $\lambda_{eff} = \lambda - \lambda_{SO}$. 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: $\delta a^2 = \alpha_i$. This leads to

$$\lambda_{SO} = 1 + \eta(1 + \chi^\ast).$$

Here $\chi^\ast = (\langle a^2 \rangle / \langle a_i \rangle)^2$ 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 $\lambda$ in the purely persistent case, with $\chi^\ast$ replacing the parameter $\chi$ that originally quantified the correlation between infectivity and susceptibility. By assuming the limit of strong short-term overdispersion ($\chi^\ast \ll 1$), we obtain $\lambda_{SO} \approx 2$. 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 $\lambda_{SO}$ should be used as a scaling exponent for long-term behavior of $R_e(S)$. Therefore, HIT is set by Eq. 15 with $\lambda = \lambda_{SO} \approx 2$. 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 $\lambda_{eff}$, 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 curves broadens. For instance, for distributions dominated by a power law, $f(\alpha) \approx 1/\alpha^\gamma$, with $3 < \gamma < 4$ and $\chi = 1$, $\lambda$ diverges even though $CV_\text{ISO}$ remains finite. This represents a crossover to the regime of so-called scale-free networks ($2 \leq \gamma < 3$), 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 overdispersion transmision 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 short-term overdispersion in transmission is commonly pre-

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 $\alpha_i$. 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 case for the basic reproduction number in a homogeneous system, $R_0 = R(\alpha^2)$, needs to be modified due to correlated short-term variations in social activity,

$$R_0 = R \left( \langle a^2 \rangle + \delta a_0^2 \right).$$

[16]

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

In the time-dependent generalization of our theory, $R_0$ 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 $\lambda$ becomes time dependent, and Eqs. 9 and 10 become

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

[17]

$$\lambda_{SO} = \langle a^2 \rangle + \frac{1}{2} \delta a_0^2$$

[18]

$$\delta \lambda(t, t') = \frac{\delta a_0^2(t) \delta a_i(t - t')}{\langle a^2 \rangle + \delta a_0^2}.$$  

[19]

Constant $\lambda_{SO}$ 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 $\delta \lambda(t, t')$ decreases with time difference $t'$, its effect on $\lambda_{eff}$ 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 $\lambda_{eff}$. Since $\lambda_{eff} \gg \lambda_{SO}$, 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 $\lambda_{eff}$ closer to $\lambda_{SO}$. 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 $\lambda_{eff}$ during that wave.
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 $\alpha$ ($\eta = 1$). For transient behavior, $\lambda_{\text{eff}} \approx 4$ is assumed based on analysis of empirical data for COVID-19 epidemic in select locations.

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. 3A. 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 $\lambda_{\text{eff}}$, 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 shrunk 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. 3B, 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, $\lambda_{\text{eff}} \approx 4$, 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
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 $R_t(t)/R_0$ and $S(t)$ 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 $\lambda$. 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 dynamics 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 $\lambda_{\text{eff}}$, 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 in 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 $\alpha_s$ inevitably changes. The impact of these changes on collective immunity depends on whether each person’s $\alpha_s$ 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 $\alpha_i(t)$. This corresponds to $\lambda_{\text{eff}}(0) = 3$ and $\lambda_{\text{eff}}(\infty) = \lambda_{\infty} = 1$. In this model, each individual completely changes the set of his/her social connections at some timescale $\tau_a$. 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,

$$S_a = S_a J - \frac{1}{\tau_a} (S_a - S) \quad [21]$$

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 $\alpha_s$ 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 $\tau_a = 150$ days 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.
Since constant rewiring eliminates correlations in individual social activity on scales longer than $\tau_s$, 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...
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 $\lambda$, introduced in our study. Like susceptibility itself, $\lambda$ 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 $\lambda_{\text{eff}}$ 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...
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 $\alpha$. 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 $k \approx 0.1 \ldots 0.3$. This distribution has a very high coefficient of variation, $CV^2 = 1/k \approx 3 \ldots 10$. In the case of a perfect correlation between individual infectivity and susceptibility $\alpha$, this would result in an unrealistically high estimate of the immunity factor: $\lambda = 1 + 2CV^2 = 1 + 2/k \approx 7 \ldots 20$. 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 $\lambda$ 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 $R_i$ 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.

<table>
<thead>
<tr>
<th>Location</th>
<th>Deaths, per 1,000 (refs.)</th>
<th>Estimated Seroprevalence (refs.)</th>
<th>Transient</th>
<th>Long-term</th>
</tr>
</thead>
<tbody>
<tr>
<td>NYC</td>
<td>2.1 (51)</td>
<td>30</td>
<td>23 (68)</td>
<td>0.24 to 0.33</td>
</tr>
<tr>
<td>Lombardy</td>
<td>1.7 (69)</td>
<td>24</td>
<td>23 (70, 71)</td>
<td>0.33</td>
</tr>
<tr>
<td>London</td>
<td>0.9 (72)</td>
<td>13</td>
<td>13 (73)</td>
<td>0.58</td>
</tr>
<tr>
<td>Chicago</td>
<td>0.9 (58)</td>
<td>13</td>
<td>20 (74)</td>
<td>0.41 to 0.58</td>
</tr>
<tr>
<td>Stockholm</td>
<td>0.9 (75, 76)</td>
<td>13</td>
<td>12 (76)</td>
<td>0.58</td>
</tr>
</tbody>
</table>

The transient and long-term suppression coefficients $R_i/R_0$ 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.

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 $\lambda_{\text{eff}}$, 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.I.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 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-SC0017204.


Kachenko et al.

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

https://doi.org/10.1073/pnas.2015972118