Persistent heterogeneity not short-term overdispersion determines herd immunity to COVID-19

Alexei V. Tkachenko[1,*], **Sergei Maslov**[2,4,5,*], **Ahmed Elbanna**[3], **George N. Wong**[2], **Zachary J. Weiner**[2], **and Nigel Goldenfeld**[2,5]

[1]Center for Functional Nanomaterials, Brookhaven National Laboratory, Upton, NY 11973, USA; [2]Department of Physics, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA; [3]Department of Civil Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA; [4]Department of Bioengineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA; [5]Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

It has become increasingly clear that the COVID-19 epidemic is characterized by overdispersion whereby the majority of the transmission is driven by a minority of infected individuals. Such a strong departure from the homogeneity assumptions of the traditional well-mixed compartment model is usually hypothesized to be the result of short-term super-spreader events, such as an individual's extreme rate of virus shedding at the peak of infectivity while attending a large gathering without appropriate mitigation. However, we demonstrate that the spread of epidemics is primarily sensitive to long-term, or *persistent heterogeneity* of individual susceptibility or infectivity. We demonstrate how to incorporate this heterogeneity into a wide class of epidemiological models, and derive a non-linear dependence of the effective reproduction number R_e on the susceptible population fraction S. Persistent heterogeneity has three important consequences compared to the effects of short-term overdispersion: (1) It results in a major modification of the early epidemic dynamics; (2) It significantly suppresses the herd immunity threshold; (3) It also significantly reduces the final size of the epidemic. We estimate social and biological contributions to persistent heterogeneity using data on real-life face-to-face contact networks and age variation of the incidence rate during the COVID-19 epidemic. In addition, empirical data from the COVID-19 epidemic in New York City (NYC) and Chicago, as well as 50 US states provide a consistent characterization of the level of heterogeneity. Our estimates suggest that the hardest-hit areas, such as NYC, are close to the heterogeneity-modified herd immunity threshold following the first wave of the epidemic. However, this type of immunity is fragile as it wanes over time if the pattern of social interactions changes substantially.

COVID-19 | Heterogeneity | Herd Immunity | Susceptibility

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 progression of infectious diseases (2–6). More recently, the important role played by population heterogeneity and the complex structure of social networks in spreading of epidemics has been appreciated and highlighted in multiple studies (7–22). However, an adequate 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 (i) the role played by superspreaders and superspreading events during initial outbreaks (8, 9, 14, 23–25) and (ii) substantial corrections to the herd immunity threshold (HIT) and the final size of epidemic (FSE) (10, 13, 15, 18, 22, 26). The COVID-19 pandemic has re-ignited 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 (27–31).

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 can stratify the population into several demographic groups (e.g. by age), and account for variation in susceptibility of these groups and their mutual contact probabilities (2). While 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 epidemic dynamics on real-world or artificial social networks (9, 18, 32, 33). 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 (10). In the context of the COVID-19 epidemic, this mapping suggests that the worst-case FSE may be significantly smaller than expected from classical homogeneous models (27). Such methods have so far been mostly limited to analysis of the final state of epidemics and outbreaks on a static network.

Significance Statement

This study demonstrates how a wide class of epidemiological models can be adapted for applications to heterogeneous populations in the context of the COVID-19 epidemic. It is shown that a persistent heterogeneity, rather than bursty short-term variations in infection transmission is responsible for self-limiting epidemic dynamics. Compact generalizations of the classical results for the herd immunity threshold and the final size of an epidemic are derived. The degree of persistent heterogeneity is estimated from data on real-life face-to-face contact networks, and on age variation of susceptibility to COVID-19. The estimate is further supported by the analysis of the empirical data from the epidemic in NYC and Chicago, as well as in 50 US states. The results suggest that by the end of the first wave of the epidemic, the hardest-hit areas, such as NYC, have been close to the heterogeneity-modified herd immunity, thereby limiting their vulnerability to a potential second wave of the epidemic.

*To whom correspondence should be addressed. E-mail: oleksiyt@bnl.gov or maslov@illinois.edu

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 compartment models. This third approach was developed long ago (13, 18),and has recently been applied in the context of COVID-19 (28). Here, the conclusion 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 super-spreader accidents), as opposed to *persistent heterogeneity*, which is a long-term characteristic of an individual and reflects behavioral propensity to (e.g.) socialize in large gatherings without prudent social distancing. Overdispersion is usually modeled in terms of a negative binomial branching process (8, 9, 14, 23–25), and is expected to be a much stronger source of variation compared to the longer-term characteristics that reflect persistent heterogeneity. How, then, can we bridge the gap between these model end-members in order to calculate both the herd immunity threshold and the final size of the epidemic in a unified way that treats the dynamics on long time-scales?

In this work, 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 the herd immunity threshold and the final size of the epidemic. Our starting point is a generalized version of the heterogeneous well-mixed theory in the spirit of Ref. (13), but we use the age-of-infection approach (1) rather than compartmentalized SIR/SEIR models of epidemic dynamics (see, e.g. (2)). The resulting 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 non-linear dependence of the effective reproduction number R_e on the overall population fraction S of susceptible individuals, and another non-linear function $S_e(S)$ 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 (34, 35). 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 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 herd immunity threshold , $S_0 = R_0^{-1/\lambda}$, and a corresponding result for the final size of an unmitigated epidemic.

Heterogeneity in the susceptibility of individual members of the population has several different contributions: (i) biological, which takes into account differences in factors such as strength of immune response, genetics, age, and co-morbidities; and (ii) social, reflecting differences in the number 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 (9, 15, 33, 36–39) 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. For the case of COVID 19, we determine a lower bound for it, based on the age distribution of reported cases.

To test this theory, we use the empirical data on COVID-19 epidemic to independently estimate the immunity factor λ. In particular, we apply our previously-described epidemic model that features multi-channel Bayesian calibration (34) to describe epidemic dynamics in New York City and Chicago. 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 USA, using data generated by the model in Ref. (40). Using both approaches, we find that the locations that were severely impacted by 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 4 and 5, and is in very good agreement with the value expected based on analysis of social and biological heterogeneity. This analysis shows how our model is able to make concrete and testable predictions.

Finally, we integrate the persistent heterogeneity theory into our time-of-infection epidemiological model (34), and project possible outcomes of the second wave of the COVID-19 epidemic in NYC and Chicago. 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, we expect virtually no second wave in NYC, indicating that the herd immunity has likely been achieved there. Chicago, on the other hand, has not passed the herd immunity 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 e.g. mask wearing, contact tracing, and targeted limitation of potential super-spreading events, through limitations on indoor bars, dining and other venues.

Theory of epidemics in heterogeneous populations

Following in the footsteps of Refs.(12, 13, 15, 18, 26, 28), we consider the spread of an epidemic in a population of individuals who exhibit significant heterogeneity in their susceptibilities to infection α. This heterogeneity may be biological or social in origin, and we assume these factors are independent: $\alpha = \alpha_b \alpha_s$. 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 towards 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 healthcare 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 sub-populations 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 sub-population with susceptibility α there is a respective mean reproductive number R_α. 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 $R_\alpha \sim \alpha$ (10, 18). 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 incident rate, i.e., the fraction of newly infected individuals per day in that sub-population. 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:

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

$$J(t) = \left\langle \int_0^\infty d\tau R_\alpha K(\tau) j_\alpha(t-\tau) \right\rangle_\alpha \qquad [2]$$

Here t is the physical time and τ is the time since infection for an individual. $\langle \ldots \rangle_\alpha$ represents averaging over α with pdf $f(\alpha)$. $J(t)$ represents the mean daily attack rate across the entire population. $K(\tau)$ is the distribution of the generation interval, which we assume is independent of α for the sake of simplicity.

According to Eq. (1), the susceptible subpopulation for any α is expressed as

$$S_\alpha(t) = \exp(-\alpha Z(t)) \qquad , \qquad [3]$$

Here $Z(t) = \int_0^t J(t')dt'$ represents the cumulative attack rate. 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:

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

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

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

Note that for $Z = 0$, this expression gives the basic reproduction number $R_0 = \langle \alpha R_\alpha \rangle$. Since both S and R_e depend on time only through $Z(t)$, Eqs. (4)–(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, i.e. $R_e = SR_0$, the relationship in the present theory is non-linear due to heterogeneity. Now one can re-write the renewal equation for the daily attack rate in the same form that it would have for a homogeneous problem:

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

Furthermore, by integrating Eqn. (1) over the whole susceptible population, we arrive at the following heterogeneity-induced modification to the relationship between the attack and the incident rates:

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

Here

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

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 non-linear function of S, defined parametrically by Eqs. (4),(8). Further generalization of this theory for the time-modulated age-of-infection model is presented in the Supplementary Information (SI). There, we also discuss the adaptation of this approach for the important special case of a compartmentalized SIR/SEIR model. Such non-linear modifications to homogeneous epidemiological models have been proposed in the past, both as plausible descriptions of heterogeneous populations and in other contexts (15, 20, 21). However, those empirical models exhibited a limited range of applicability (15) and have not had a solid mechanistic foundation, with a noticeable exception of a special case of SIR model without correlation between susceptibility and infectivity studied in Ref. (26). Our approach is more general: it provides an 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 non-linear functions $R_e(S)$ and $S_e(S)$.

We now derive a simple yet remarkably general result for the final size of an unmitigated epidemic. To do this, we integrate Eq. (6) over time t. 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 final fraction of the susceptible population, S_∞:

$$S_\infty = M_\alpha\left(-\int_{S_\infty}^1 \frac{R_e(S)dS}{S_e(S)}\right) \qquad [9]$$

This equation is valid for an arbitrary distribution of α, arbitrary correlation between susceptibility and infectivity, and for any statistics of the generation interval. It 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.(9) reproduces results of Refs. (22, 26, 30), (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 (10, 13).

One of the striking consequences of the non-linearity of $R_e(S)$ is that the effective reproduction number could be decreasing at the early stages of epidemics significantly faster than predicted by homogeneous models. Specifically, for $(1-S) \ll 1$ one obtains

$$R_e \approx R_0(1-\lambda(1-S)) \qquad [10]$$

We named the coefficient λ the *immunity factor* because it quantifies the effect that a reduction in the susceptible population due to immunity has on the spread of an epidemic. The classical value of λ is 1, but it may be significantly larger in a heterogeneous case.

$$\lambda = \frac{\langle\alpha^2 R_\alpha\rangle}{\langle\alpha R_\alpha\rangle} \qquad [11]$$

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

We previously defined the overall susceptibility α as a combination of biological and social factors: $\alpha = \alpha_s\alpha_b$ Here α_s is a measure of the overall social connectivity 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 interpersonal contacts contribution to an epidemic spread is mostly reciprocal, we assume $R_\alpha \sim \alpha_s$. On the other hand, in our analysis we neglect a correlation between the 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:

$$\lambda_b = \frac{\langle\alpha_b^2\rangle}{\langle\alpha_b\rangle^2} = 1 + CV_b^2 \qquad [12]$$

$$\lambda_s = \frac{\langle\alpha_s^3\rangle}{\langle\alpha_s\rangle\langle\alpha_s^2\rangle} = 1 + \frac{CV_s^2(2+\gamma_s CV_s)}{1+CV_s^2} \qquad [13]$$

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 both on 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: $R_\alpha \sim \alpha^\chi$ (see it 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 nature of heterogeneity, i.e., $\lambda \approx \lambda_b = 1 + CV_\alpha^2$ where CV_α is the coefficient of variation for the overall susceptibility. In the opposite limit $\chi = 1$, the variation is primarily of social origin, hence $\lambda \approx \lambda_s$ will be affected by both CV_α and the skewness γ_α of the pdf $f(\alpha)$.

Recently, real-world networks of face-to-face communications have been studied using a variety of tools, including RFID devices (36), Bluetooth and Wi-Fi wearable tags, smartphone apps (37, 38), as well as census data and personal surveys (9, 33, 39). Despite coming from a wide variety of contexts, the major features of contact networks are remarkably robust. In particular, both the degree (the number of contacts per person), and the node strength pdfs appear nearly constant when plotted in log-log coordinates, followed up by a sharp fall after a certain cut-off. This behavior is generally consistent with an exponential distribution in $f_s(\alpha_s)$ (15, 37, 39), $f_s(\alpha_s) \sim e^{-\alpha_s/\langle\alpha_s\rangle}$, leading to $\lambda_s \approx 3!/2! = 3$.

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, we estimated this parameter based on the the age distribution of cases as reported by the NYC Department of Health (41). This analysis suggests $\lambda_b = 1 + CV_b^2 \approx 1.3$. On one hand, the variation in the infection rates reported among different age groups may be exaggerated by the variation of the disease severity. On the other hand, age is not the only factor that determines biological susceptibility: it may also depend on genetics, pre-existing conditions, etc. The overall immunity factor, based on this rather conservative estimate is $\lambda \approx 4$.

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. (10). To describe the non-linear regime, we consider a gamma-distributed susceptibility: $f(\alpha) \sim \alpha^{1/\eta-1}\exp(-\alpha/\eta)$, 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:

$$S_e(S) = S^{1+\eta} \qquad [14]$$

and

$$R_e(S) = R_0 S^\lambda \qquad [15]$$

The exponent $\lambda = 1+(1+\chi)CV_\alpha^2 = 1+(1+\chi)\eta$ coincides with the early-epidemics immunity factor defined in Eqs. (10)–(11) 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. (26) The scaling behavior $R_e(S)$ is shown in Fig. 1(A) for $\lambda = 4 \pm 1$. This function is dramatically different from the classical linear dependence $R_e = SR_0$. To emphasize the importance of this difference, we indicate the estimated fractions of the population in New York City and Chicago susceptible to COVID-19, as of the end of May 2020. It is evident from the plot that the reduction of R_e for immunity factor between 3 and 5

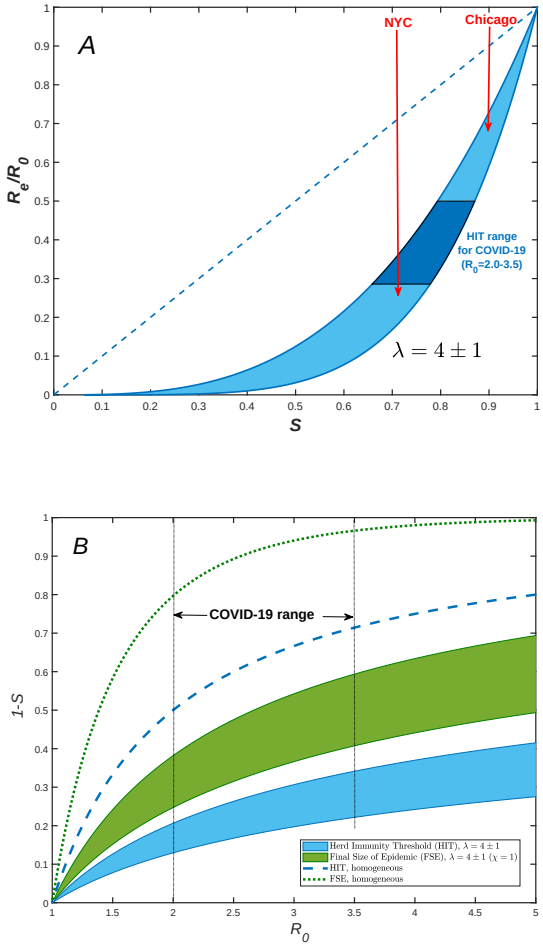

Fig. 1. A) R_e/R vs S dependence for gamma-distributed susceptibility for $\lambda = 4 \pm 1$ (blue area). The dashed line shows the classical homogeneous result, $R_e = R_0 S$. Note a substantial reduction of R_e for COVID-19 in both NYC and Chicago, compared to that value. Approximate fractions of susceptible populations, S, for both cities are estimated as of the end of May 2020, by using the model described in Ref. (34). B) Herd Immunity threshold (blue area) and final size of epidemic (FSE, green area) for gamma-distributed susceptibilities. The range of λ is the same as in (A). FSE is shown assuming maximum correlation between susceptibility and infectivity ($\chi = 1$), which corresponds to CV_α^2 ranging from 1 to 2. Notice a substantial reduction of both HIT and FSE compared to the classical results for homogeneous population which are shown as blue dashed and green dotted lines, respectively.

may substantially reduce or eliminate the chances of future outbreaks in both cities.

Eq. (15) immediately leads to a major revision of the classical result $S_0 = 1/R_0$ for the herd immunity threshold, i.e. the fraction of susceptible population below which the exponential growth stops. By setting $R_e = 1$ in Eq. (15), we obtain:

$$S_0 = \left(\frac{1}{R_0}\right)^{1/\lambda} \qquad [16]$$

An unmitigated epidemic of course does not stop once the HIT is passed, but continues until there are no more infected individuals left, a phenomenon known as overshoot. To find the FSE for the case of gamma-distributed susceptibility we apply our general result, Eq. 9, which gives:

$$S_\infty = \left(1 + \frac{R_0 \eta \left(1 - S_\infty^{\lambda-\eta}\right)}{\lambda - \eta}\right)^{-1/\eta} \qquad [17]$$

The values of HIT and FSE for various values of R_0 are presented in Figure 1B. As expected, in both cases the number of remaining susceptible individuals is substantially larger than in the homogeneous case.

Our focus on the gamma distribution is well justified by the observation that the social strength α_s is approximately exponentially distributed, i.e., it is a specific case of the gamma distribution with $\eta = CV_\alpha^2 = 1$. 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 form. In *SI Appendix*, we present analytic and numerical calculations for two other families of distributions: (i) an exponentially bounded power law $f(\alpha) \sim e^{-\alpha/\alpha_+}/\alpha^q$ ($q \geq 1$, with an additional cut-off at lower values of α) and (ii) the log-normal distribution. In addition, we give an approximate analytic result that generalizes Eq. (15) for arbitrary skewness of $f(\alpha)$. This generalization works remarkably well for all three of the families of distributions analyzed in this work. As suggested by Eqs. (12)-(13), when the distribution becomes increasingly skewed, the range between the $\chi = 0$ and $\chi = 1$ curves broadens. For instance, for distributions dominated by a power law, $f(\alpha) \sim 1/\alpha^q$, λ diverges at q slightly larger than 3 and $\chi = 1$, even if 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 epidemics effectively kills itself by immunizing the hubs on the network (13, 18, 42).

Application to COVID-19 epidemic

The COVID-19 epidemic reached the US in early 2020, and by March it was rapidly spreading across multiple states. The early dynamics was characterized by a rapid rise in the number of cases with doubling times as low as 2 days. 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 New York City, 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. (34) to New York City and Chicago. For both cities, we have access to reliable time series data on hospitalization, ICU room occupancy, and daily deaths due to COVID-19 (41, 43–45). We used these data to perform multi-channel calibration of our model (34), 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. 2A. 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. 15 with the slope corresponding to immunity factor

$\lambda = 4.5 \pm 0.05$. 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 in Fig. S2 we repeated our analysis in which $R_e(t)$ was corrected by Google's community mobility report in these two cities (46) (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 immunity factor: $\lambda = 4.1 \pm 0.1$.

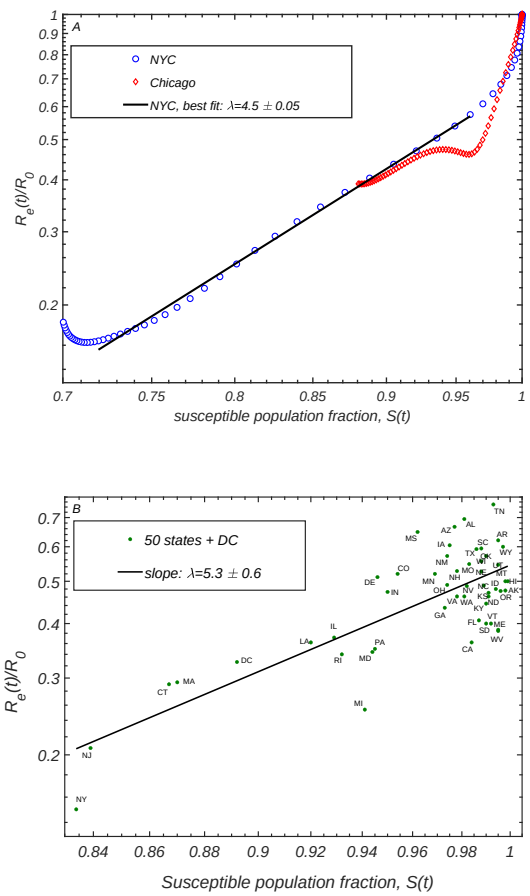

Fig. 2. Correlation between the relative reduction in the effective reproduction number $R_e(t)/R_0$ (y-axis) with the susceptible population $S(t)$. In Panel A, we present the progression of these two quantities for New York City and Chicago, as given by the epidemiological model described in Ref. (34). Panel B shows the scatter plot of $R_e(t_0)/R_0$ and $S(t_0)$ in individual states of the US, evaluated in Ref. (40) (t_0 is the latest date covered in that study).

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 (40). In that study, the COVID-19 epidemic was modelled 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 US (40). We use state level estimates of $R_e(t)$, R_0 and $S(t)$ as reported in Ref. (40) to construct the scatter plot $R_e(t_0)/R_0$ vs $S(t_0)$ shown in 2, 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 $\lambda = 5.3 \pm 0.6$ and for $S = 1$ intercept around 0.54. In Fig. S3 (see *SI Appendix*), 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 $\lambda = 4.7 \pm 1.5$. Both estimates of the immunity factor based on the state data are consistent with our earlier analysis of NYC and Chicago. Furthermore, the range of λ between 4 and 5 extracted from these COVID-19 data sources, is in a very good agreement with the value $\lambda = \lambda_s \lambda_b \approx 4$ that we obtained above, based on the statistics of interpersonal contacts and the age variation of biological susceptibility to COVID-19 infection.

We can now incorporate heterogeneity into our epidemiological model, and examine how future projections change as a result of this modification. This is done by plugging the scaling relationships, Eq. (14)-(15) into the attack rate and incident rate equations of the original model. These equations are similar to Eqs. (6)-(7), but also include time modulation due to 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 1, 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 herd immunity, $R_e = R_0 S^\lambda$. The projected daily deaths for each of the two cities under this (unrealistically harsh) scenario are presented in Fig. 3 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).

Discussion

We have demonstrated that population heterogeneity due to biological and/or social susceptibility to infection may lead to dramatic changes affecting the early dynamics, the herd immunity threshold and the final size of an epidemic. Heterogeneity can be easily integrated into a wide class of traditional epidemiological models in the form of two non-linear functions

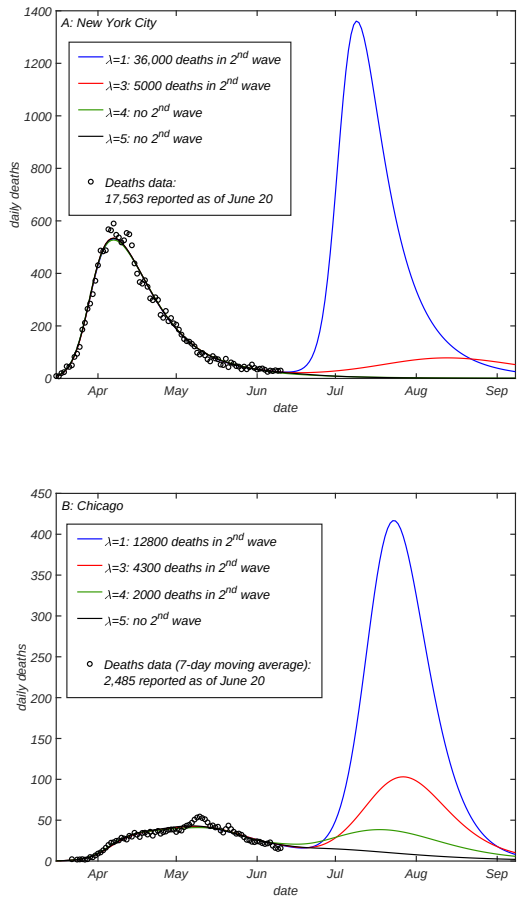

Fig. 3. Projections of daily deaths under the hypothetical scenario in which any mitigation is completely eliminated as of June 15 2020, for (A) NYC and (B) Chicago. Different curves correspond to different values of the immunity factor $\lambda = 1$ (blue), 3 (red), 4 (green) and 5 (black lines). The model described in Ref. (34) was fully calibrated on daily deaths (circles), ICU occupancy and hospitalization data up to the end of May. See Figs. S4-S5 in SI appendix for additional details, including confidence intervals.

$R_e(S)$ 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 (see Eqs. (12-13)). By applying our theory to the COVID-19 epidemic we found evidence that the hardest hit areas such as New York City, have likely passed the heterogeneity-modified herd immunity threshold. Other places that had intermediate exposure, such as e.g. Chicago, while still above the HIT, 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, e.g. by contact tracing, control of potential super-spreading events, etc.

According to our results, a significant suppression of the second wave in both cities is expected even for a rather moderate value of the immunity factor. This is because, in the limit of a strong correlation between susceptibility and infectivity $\lambda = 1 + 2CV_\alpha^2$, yielding $CV_\alpha = 1$ and thus a moderate value of $\lambda = 3$. A conceptually similar analysis from Ref. (28) for Italy and Austria suggested that a much greater level of heterogeneity would be needed to suppress second waves in those countries. Specifically, $CV_\alpha = 1$ did not give a noticeable effect in that study, so $CV_\alpha = 3$ had to be assumed, which corresponds to λ as high as 19. While the fraction of susceptible population in Austria is indeed very low, Italy was among the the hardest-hit countries during the COVID-19 epidemic, with some areas affected as strongly as NYC, and the national average number of deaths per capita is comparable to that in Chicago. We therefore expect the hardest-hit regions of Italy, such as Lombardy, to be close to herd immunity, despite our more conservative estimate of the level of heterogeneity. The quantitative difference between our conclusions and the results of Ref. (28) is likely to have methodological origin. First, we used daily deaths, hospitalization and ICU occupancy, rather than case statistics for calibration of our model. Second, we focused on city rather than country level, which certainly enhances the overall effect of the herd immunity. In Table 1 we show how R_e is suppressed as a result of depletion of susceptible population in selected locations in the world, as of early June 2020.

In another recent study (31), the reduction of HIT due to heterogeneity has been illustrated on 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. 13, the immunity factor in that model is $\lambda = 1.54$. For this immunity factor, Eq. (16) 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. (31): $S_0 = 62.5\%$, 53.5%, and 47.5%, respectively.

Table 1. Effect of heterogeneity-modified herd immunity on R_e in selected locations. The fraction of susceptible population as of early June 2020 is estimated from reported death count, assuming infection fatality rate of 0.7% (47). The range in the last column corresponds to $\lambda = 4 \pm 1$.

Location	Deaths per 1000	$1 - S$	$R_e/R_0 = S^\lambda$
New York City, USA (41)	2.1	30%	$25\% \pm 9\%$
Lombardy, Italy (48)	1.7	24%	$35\% \pm 9\%$
London,UK (49)	0.9	13%	$58\% \pm 8\%$
Chicago, USA (43)	0.9	13%	$58\% \pm 8\%$
Stockholm, Sweden (50)	0.9	13%	$58\% \pm 8\%$

Finally, we summarise the assumptions and limitations of our study. First, we assume a long-lasting immunity of recovered individuals. Second, we present a slightly different perspective on heterogeneity than that used in other recent papers. The susceptibility used in our model is defined as a *persistent* (or time-averaged) property of each individual. Thus there is a crucial distinction between the heterogeneity relevant for our study, which is long-term, and overdispersion in transmission statistics associated with short-term super-spreading events (8, 9, 14, 23–25). In our theory, a personal decision to attend a large party or a meeting would only be significant to the epidemic dynamics 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 (8, 24) would be significantly exaggerated compared to time-averaged quantities relevant for self-limiting epidemic dynamics and herd immunity. Specifically, the statistics of superspreading events is commonly described by the negative binomial distribution with dispersion parameter k estimated to be about 0.1 for the COVID-19 (25). According to Ref. (8), this is consistent with the expected value of the individual-level reproduction number R_i drawn from a gamma distribution with the shape parameter $k \simeq 0.1$. This distribution has a very high coefficient of variation, $CV^2 = 1/k \simeq 10$. 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: $\lambda = 1 + 2CV^2 = 1 + 2/k \simeq 20$. For this reason, according to our perspective and calculation, the final size of the COVID-19 epidemic may have been substantially underestimated in Ref. ((27)).

One of the consequences of the persistent nature of α_s is that the heterogeneity-modified herd immunity might wane after some time as individuals change their social interaction patterns. In particular, 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, the social susceptibility α_s inevitably changes. The impact of these changes on the herd immunity depends on whether each person's α_s during and after the mitigation are sufficiently correlated. For example, herd immunity would be compromised if people who practiced strict self-isolation would compensate for it by an above-average social activity after the first wave of the epidemic has passed.

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 New York City during the COVID-19 epidemic would gain full or partial herd immunity thereby limiting the spread of infection to other regions during a potential second wave 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.

As we were finalizing this paper for submission, a preprint by Aguas et al. appeared online (51). They independently obtained the analytic expression for the HIT in case of a Gamma-distributed susceptibility, a special case of our analysis. However our estimates for the coefficient of variation of heterogeneity and therefore the immunity factor are significantly lower than the estimates reported in Ref. (51), reflecting methodological differences.

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 United States Department of Energy Computational Science Graduate Fellowship, provided under Grant No. DE-FG02-97ER25308. A.E. acknowledges partial support by NSF CAREER Award No. 1753249. This work made use of the Illinois Campus Cluster, a computing resource that is operated by the Illinois Campus Cluster Program (ICCP) in conjunction with the National Center for Supercomputing Applications (NCSA) 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 U.S. DOE Office of Science Facility, at Brookhaven National Laboratory under Contract No. DE-SC0012704.

1. WO Kermack, AG McKendrick, A contribution to the mathematical theory of epidemics. *Proc. Royal Soc. London. Ser. A, Containing papers a mathematical physical character* **115**, 700–721 (1927).
2. MJ Keeling, P Rohani, *Modeling infectious diseases in humans and animals.* (Princeton University Press), (2011).
3. K Rock, S Brand, J Moir, MJ Keeling, Dynamics of infectious diseases. *Reports on Prog. Phys.* **77**, 026602 (2014).
4. J Ma, Estimating epidemic exponential growth rate and basic reproduction number. *Infect. Dis. 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. AL Lloyd, RM May, Epidemiology - how viruses spread among computers and people. *Science* **292**, 1316–1317 (2001).
8. JO Lloyd-Smith, SJ Schreiber, PE Kopp, WM Getz, Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, 355–9 (2005).
9. LA Meyers, B Pourbohloul, MEJ Newman, DM Skowronski, RC Brunham, Network theory and sars: predicting outbreak diversity. *J. Theor. Biol.* **232**, 71–81 (2005).
10. MEJ Newman, Spread of epidemic disease on networks. *Phys. Rev. E* **66** (2002).
11. MJ Ferrari, S Bansal, LA Meyers, ON Bjornstad, Network frailty and the geometry of herd immunity. *Proc. Royal Soc. B-Biological Sci.* **273**, 2743–2748 (2006).
12. S Bansal, LA Meyers, The impact of past epidemics on future disease dynamics. *J. Theor. Biol.* **309**, 176–184 (2012).
13. Y Moreno, R Pastor-Satorras, A Vespignani, Epidemic outbreaks in complex heterogeneous networks. *Eur. Phys. J. B* **26**, 521–529 (2002).
14. M Small, C Tse, DM Walker, Super-spreaders and the rate of transmission of the sars virus. *Phys. D: Nonlinear Phenom.* **215**, 146–158 (2006).
15. S Bansal, BT Grenfell, LA Meyers, When individual behaviour matters: homogeneous and network models in epidemiology. *J. Royal Soc. Interface* **4**, 879–891 (2007).
16. 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. Public Heal.* **15**, 2369 (2018).
17. Z Dezső, AL Barabási, Halting viruses in scale-free networks. *Phys. Rev. E* **65**, 055103 (2002).
18. R Pastor-Satorras, C Castellano, P Van Mieghem, A Vespignani, Epidemic processes in complex networks. *Rev. Mod. Phys.* **87**, 925 (2015).
19. MJ Keeling, KTD Eames, Networks and epidemic models. *J. Royal Soc. Interface* **2**, 295–307 (2005).
20. M Roy, M Pascual, On representing network heterogeneities in the incidence rate of simple epidemic models. *Ecol. Complex.* **3**, 80–90 (2006).
21. PD 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).
22. G Katriel, The size of epidemics in populations with heterogeneous susceptibility. *J. Math. Biol.* **65**, 237–262 (2012).
23. AP Galvani, RM May, Epidemiology - dimensions of superspreading. *Nature* **438**, 293–295 (2005).
24. Z Shen, et al., Superspreading sars events, beijing, 2003. *Emerg. Infect. Dis.* **10**, 256–260 (2004).
25. 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).
26. AS Novozhilov, On the spread of epidemics in a closed heterogeneous population. *Math. Biosci.* **215**, 177–185 (2008).
27. L Hébert-Dufresne, BM Althouse, SV Scarpino, A Allard, Beyond r0: Heterogeneity in secondary infections and probabilistic epidemic forecasting. *medRxiv* **2020.02.10.20021725** (2020).
28. MGM Gomes, et al., Individual variation in susceptibility or exposure to sars-cov-2 lowers the herd immunity threshold. *medRxiv* **2020.04.27.20081893** (2020).
29. PV Brennan, LP Brennan, Susceptibility-adjusted herd immunity threshold model and potential r0 distribution fitting the observed covid-19 data in stockholm. *medRxiv* **2020.05.19.20104596** (2020).
30. F Ball, Deterministic and stochastic epidemics with several kinds of susceptibles. *Adv. Appl. Probab.* **17**, 1–22 (1985).
31. T Britton, F Ball, P Trapman, A mathematical model reveals the influence of population heterogeneity on herd immunity to sars-cov-2. *Science* **eabc6810** (2020).
32. JC Stack, S Bansal, VSA Kumar, B Grenfell, Inferring population-level contact heterogeneity from common epidemic data. *J. Royal Soc. Interface* **10** (2013).
33. S Eubank, et al., Modelling disease outbreaks in realistic urban social networks. *Nature* **429**, 180–184 (2004).
34. GN Wong, et al., Modeling covid-19 dynamics in illinois under non-pharmaceutical interventions. *medRxiv* **2020.06.03.20120691** (2020).
35. M Castro, S Ares, JA Cuesta, S Manrubia, Predictability: Can the turning point and end of an expanding epidemic be precisely forecast? *arXiv:* **2004.08842** (2020).
36. L Isella, et al., What's in a crowd? analysis of face-to-face behavioral networks. *J. Theor. Biol.* **271**, 166–180 (2011).
37. BF Nielsen, K Sneppen, L Simonsen, J Mathiesen, Heterogeneity is essential for contact tracing. *medRxiv* **2020.06.05.20123141** (2020).
38. M Starnini, et al., Robust modeling of human contact networks across different scales and proximity-sensing techniques in *Social Informatics*, eds. GL Ciampaglia, A Mashhadi, T Yasseri. (Springer International Publishing, Cham), pp. 536–551 (2017).
39. L Danon, JM Read, TA House, MC Vernon, MJ Keeling, Social encounter networks: characterizing great britain. *Proc. Royal Soc. B-Biological Sci.* **280** (2013).
40. HJT Unwin, et al., 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 (2020).
41. Data w from https://github.com/nychealth/coronavirus-data (2020).
42. R Pastor-Satorras, A Vespignani, Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* **86**, 3200–3203 (2001).
43. Data were downloaded from https://www.dph.illinois.gov/covid19/covid19-statistics. (2020).
44. Data from https://github.com/thecityny/covid-19-nyc-data (2020).
45. Data originally due to https://www.thecity.nyc/ (2020).
46. https://www.blog.google/technology/health/covid-19-community-mobility-reports?hl=en (2020).
47. G Meyerowitz-Katz, L Merone, A systematic review and meta-analysis of published research data on COVID-19 infection-fatality rates. *medRxiv* **2020.05.03.20089854** (2020).
48. Data from https://github.com/pcm-dpc/COVID-19 (2020).
49. Data from https://data.london.gov.uk/dataset/coronavirus--covid-19--deaths (2020).
50. Data from https://github.com/mrunesson/covid-19 (2020).
51. R Aguas, et al., Herd immunity thresholds for sars-cov-2 estimated from unfolding epidemics. *medRxiv* **2020.07.23.20160762** (2020).

Supporting Information

Derivation of quasi-homogeneous model.

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

$$J(t) = \mu(t) \left\langle \int_0^\infty d\tau R_\alpha K(\tau) j_\alpha(t-\tau) \right\rangle_\alpha \quad [S1]$$

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

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

Here $R_0 = \int_0^\infty \alpha R_\alpha f(\alpha) d\alpha$ is basic reproduction number. Now one can write integral equation for the attack rate which is formally identical to the one for a homogeneous case:

$$J(t) = \mu(t) R_0 \int_0^\infty d\tau K(\tau) j^*(t-\tau) \quad [S3]$$

Here introducing infectivity-weighted incident rate, $j^* = S_R J$. Eq. (7) completes the set of our quasi-homogeneous equations:

$$\frac{dS}{dt} = -S_e J \quad [S4]$$

As discussed in the main text, the inhomogeneity is fully accounted for by non-liner function $S_R(S)$, and variable effective susceptibility $\alpha_e(S)$.

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:

$$\dot{S}_\alpha = -\alpha S_\alpha J \quad [S5]$$

$$\dot{E}_\alpha = \alpha S_\alpha J - \gamma_E E_\alpha \quad [S6]$$

$$\dot{I}_\alpha = \gamma_E E_\alpha - \gamma_I I_\alpha \quad [S7]$$

Here attack rate is $J(t) = \mu(t)\gamma_I \int_0^\infty R_\alpha I(\alpha) f(\alpha) d\alpha$. We define infectivity-weighted "Exposed" and "Infectious" fractions as

$$E = \int_0^\infty R_\alpha E(\alpha) f(\alpha) d\alpha \quad [S8]$$

$$I = \frac{J}{\gamma_I \mu(t)} = \int_0^\infty R_\alpha I(\alpha) f(\alpha) d\alpha \quad [S9]$$

[S10]

This leads to a complete description of epidemic dynamics with three ODEs formally equivalent to those for the homogeneous case. The difference are, once again, functions $R_e = \mu(t) S_R(S) R_0$ and $S_e(S)$:

$$\dot{S} = -\mu(t)\gamma_I S_e I \quad [S11]$$

$$\dot{E} = R_e(t)\gamma_I I - \gamma_E E \quad [S12]$$

$$\dot{I} = \gamma_E E - \gamma_I I \quad [S13]$$

Correlation parameter and scaling relationship between infectivity and susceptibility. . Below we consider a model in which biological susceptibility α_b is not correlated either 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 pdfs of variables $x \equiv \ln \alpha_s$) and $y \equiv \ln \alpha_b$. It is reasonable to assume log-normal distribution for α_b, since biological susceptibility can be modeled as a product several random factors (due to age, gender, genetics, pre-existent conditions, etc). This corresponds to Gaussian 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:

$$R_\alpha \sim \langle \alpha_s \rangle \sim \frac{\int f_x(x) \exp\left(x - \frac{(x - \ln\alpha - \sigma^2/2)^2}{2\sigma^2}\right) dx}{\int f_x(x) \exp\left(-\frac{(x - \ln\alpha - \sigma^2/2)^2}{2\sigma^2}\right) dx} \quad [S14]$$

This integral, for most pdfs f_x and f_y, will be dominated by the vicinity of point x_0 defined by 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

$$R_\alpha \sim \exp\left(\frac{(\sigma^2 + \ln\alpha)^2 - (\ln\alpha)^2}{2\sigma^2(1+\kappa\sigma^2)}\right) \sim \alpha^\chi \quad [S15]$$

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

Functions $S_R(S)$ and $S_e(S)$.. According to Eq.(4), function $S(Z)$ is directly related to the moment generating function M_α for pdf $f(\alpha)$

$$S = \langle e^{-\alpha Z} \rangle_\alpha = M_\alpha(-Z) = 1 - Z + \frac{\langle \alpha^2 \rangle Z^2}{2} - \frac{\langle \alpha^3 \rangle Z^3}{6} + ... \quad [S16]$$

This function also determines effective fraction of susceptible population S_e:

$$S_e = \langle \alpha e^{-\alpha Z} \rangle_\alpha = -\frac{d\ln S}{dZ} \quad [S17]$$

Remarkably, once function $S_e(S)$ is found, it completely determines how S_R, and hence R_e, behaves in the limiting cases of both the strong and weak correlations:

$$S_R = \begin{cases} \langle \alpha e^{-\alpha Z} \rangle_\alpha = -dS/dZ = S_e, & \chi = 0 \\ \frac{1}{\langle \alpha^2 \rangle}\frac{dS^2}{dZ^2} = \frac{S_e}{\langle \alpha^2 \rangle}\frac{dS_e}{dS}, & \chi = 1 \end{cases} \quad [S18]$$

Application to specific distributions of susceptibility..

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

$$f(\alpha) \sim \alpha^{1/\eta - 1} \exp(-\alpha/\eta) \quad [S19]$$

By using Eqs. (4)-(5), we obtain:

$$S = (1 + \eta Z)^{-1/\eta} \quad [S20]$$

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

$$S_R = (1 + \eta Z)^{-(1+\chi+1/\eta)} = S^\lambda \quad [S22]$$

This leads to the scaling relationship $R_e = R_0 S^\lambda$, Eq. (15).

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 exponential factor $\exp(-\alpha/\alpha_+)$, we recover the functional form identical to the gamma distribution, Eq. (S19) discussed above, but with negative values of the shape factor:

$$f(\alpha) = \frac{\alpha_+^{q-1}\exp(-\alpha/\alpha_+)}{\alpha^q \Gamma(1-q,\epsilon)} \quad [S23]$$

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

$$\alpha_+ = \frac{\Gamma(1-q,\epsilon)}{\Gamma(2-q,\epsilon)} \quad [S24]$$

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 $1 \le q \le 3$, the value of $\eta = CV^2$ sets the dynamic range the between upper and lower cut-offs, i.e. parameter ϵ:

$$1 + \eta = \langle \alpha^2 \rangle = \frac{\Gamma(1-q,\epsilon)\Gamma(3-q,\epsilon)}{\Gamma(2-q,\epsilon)^2} \quad [S25]$$

By using Eq. (4)-(5), we can obtain exact results for S S_R in terms of Z:

$$S = \frac{\Gamma(1-q,\epsilon(1+\alpha_+ Z))}{\Gamma(1-q,\epsilon)(1+\alpha_+ Z)^{1-q}} \quad [S26]$$

$$S_R = \frac{\Gamma(\nu, \epsilon(1+\alpha_+ Z))}{\Gamma(\nu,\epsilon)(1+\alpha_+ Z)^\nu} \qquad [S27]$$

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

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

$$S_e(S) \approx \frac{S}{(1+\Delta_\lambda(1-S))^{(\lambda_0-1)/\Delta_\lambda}} \qquad [S28]$$

According to Eq. (S18), this function completely defines behavior of S_R in both limits of the weak and strong correlation regimes :

$$S_R \approx \frac{(1+(\Delta_\chi-1)(1-S))\,S}{(1+\Delta_\lambda(1-S))^{(\lambda-\Delta_\chi)/\Delta_\lambda}} \qquad [S29]$$

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.

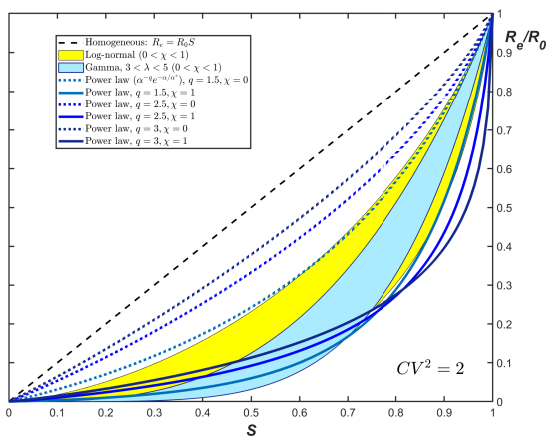

Fig. S1. R_e/R_0 vs S dependence for three different families of probability distribution $f(\alpha)$: Gamma (light blue), truncated power law (dashed lines), and log-normal (yellow). Different curves correspond to the same value of the coefficient of variation $CV_\alpha^2 = 2$, and two limiting values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

Log-normal distribution. 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 log-normal distribution, the initial drop in R_e according to Eq. (11), is noticeably faster than for gamma: $\lambda = (1+CV_\alpha^2)(1+\chi CV_\alpha^2)$. However, the initial linear regime is also much narrower. Figure S1 shows dependence $R_e(S)$ both for log-normal alongside with the above results for gamma and scaling distributions, for the same values of CV (specifically, $CV_\alpha^2 = 2$). As one can see from the plots, despite the stronger effect of heterogeneuity at the early stage, the curves generated by log-normal distribution approach $R_e = 0$ significantly slower than those corresponding to gamma. Note that the overall $R_e(S)$ behavior generated by log-normal distribution closely matches the one obtain for power law with certain scaling exponent q. That exponent would depend on CV and should approach 1 in the limit of sufficiently wide distribution when log-normal pdf asymptotically approaches a power law $1/\alpha$ with upper and lower cut-offs.

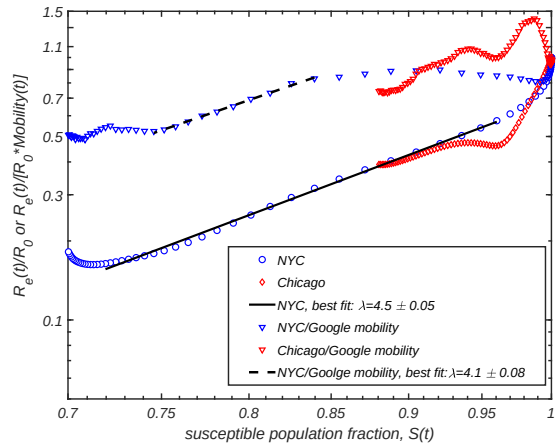

Fig. S2. Exploration of effect of mobility on data presented in Figure 2(A). Triangles represent data points for NYC and Chicago with $R_e(t)/R_0$ corrected by a mobility factor calculated from Google community mobility report, Ref. (46). We compute the mobility for NYC using average mobility of its five counties: New York county, Bronx county, Kings county, Richmond county, and Queens county, weighted by their population fraction.

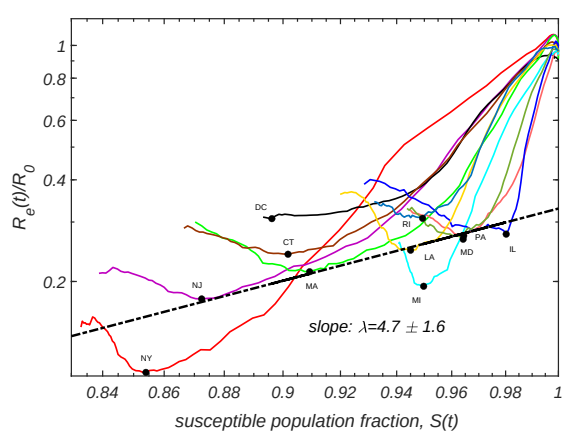

Fig. S3. Time progressions of $Re(t)/R_0$ and $S(t)$ for the hardest-hit US states and DC, as reported in Ref. (40). Black dots correspond to absolute minima of transmission and population susceptible fractions. The dashed line with slope $\lambda = 4.7 \pm 1.6$ is the best power law fit through these black dots.

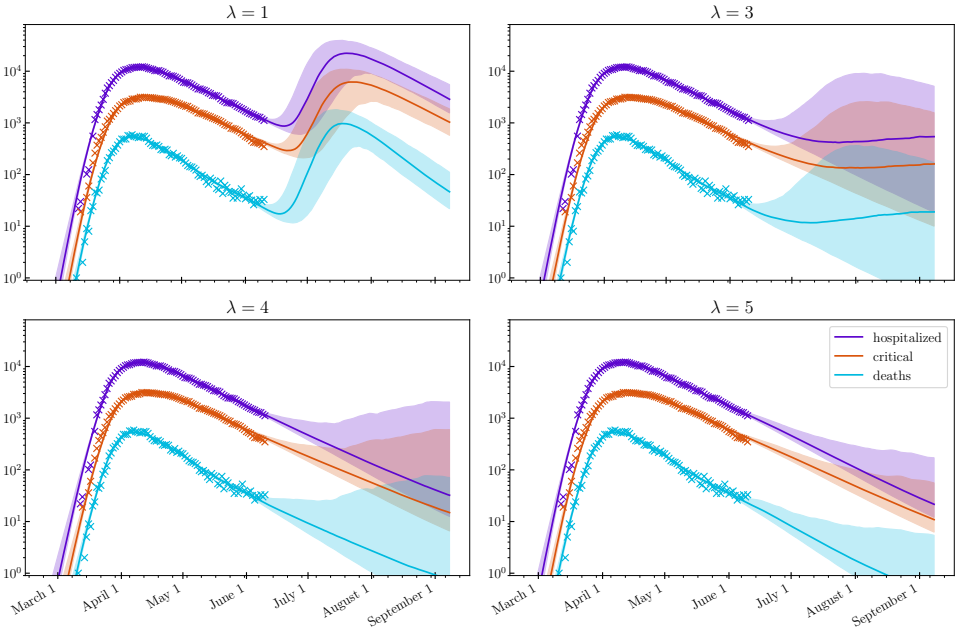

Fig. S4. Hospitalization, ICU occupancy and daily deaths in NYC modeled under hypothetical scenario when any mitigation is completely eliminated as of Jun 15 2020, for various values of λ. Model described in Ref. (34) is calibrated on data from Ref.(43), up to June 10, 2020 (shown as crosses). 95% confidence intervals are indicated.

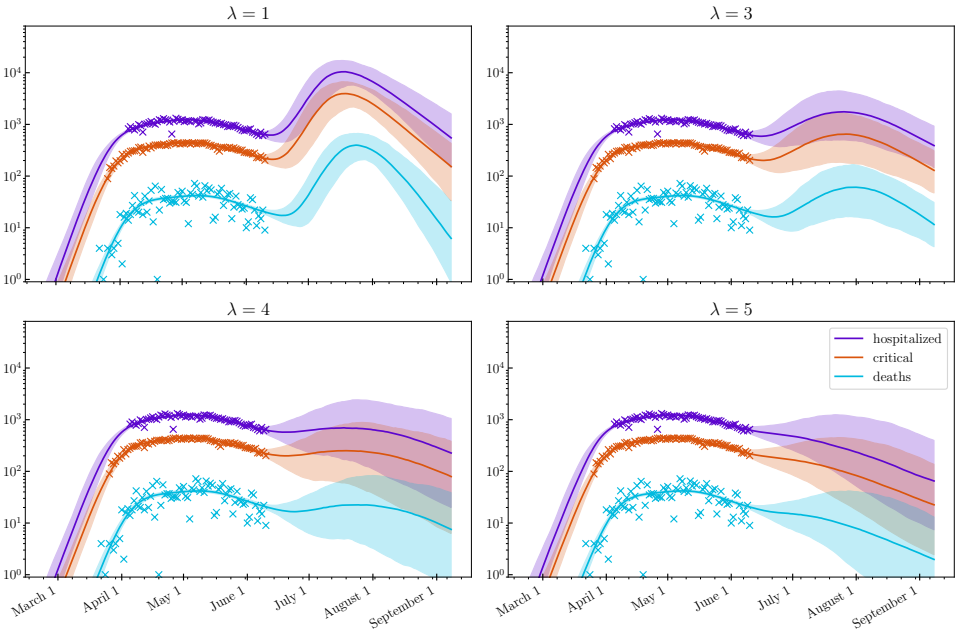

Fig. S5. Hospitalization, ICU occupancy and daily deaths in Chicago modeled under hypothetical scenario when any mitigation is completely eliminated as of Jun 15 2020, for various values of λ. Model described in Ref. (34) is calibrated on data from Ref.(43), up to June 10, 2020 (shown as crosses). 95% confidence intervals are indicated.