medRxiv preprint doi: https://doi.org/10.1101/2020.07.26.20162420; this version posted January 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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

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

Epidemics generally spread through a succession of waves that reflect factors on multiple time-scales. On short time-scales, super-spreading events lead to burstiness and overdispersion, while long-term persistent heterogeneity in susceptibility is expected to lead to a reduction in the infection peak and the herd immunity threshold (HIT). Here, we develop a general approach to encompass both time-scales, including time variations in individual social activity, and demonstrate how to incorporate them phenomenologically into a wide class of epidemiological models through parameterization. We derive a non-linear dependence of the effective reproduction number R_e on the susceptible population fraction S. We show that a state of transient collective immunity (TCI) emerges well below the HIT during early, high-paced stages of the epidemic. However, this is a fragile state that wanes over time due to changing levels of social activity, and so the infection peak is not an indication of herd immunity: subsequent waves can and will emerge due to behavioral changes in the population, driven (e.g.) by seasonal factors. Transient and long-term levels of heterogeneity are estimated by using empirical data from the COVID-19 epidemic as well as from real-life face-to-face contact networks. These results suggest that the hardest-hit areas, such as NYC, 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 reqions can still experience subsequent waves.

COVID-19 | Heterogeneity | Overdispersion | Epidemic theory

The COVID-19 pandemic is nearly unprecedented in the level of disruption it has caused globally, but also, potentially, in the degree to which it will change our understanding of epidemic dynamics and the efficacy of various mitigation strategies. Ever since the pioneering works of Kermack and McKendrick (1), epidemiological models have been widely and successfully used to quantify and predict the progression of infectious diseases (2–6). More recently, the important role in epidemic spreading played by population heterogeneity and the complex structure of social networks has been appreciated and highlighted in multiple studies (7–25). However, integration of this conceptual progress into reliable, predictive epidemiological models remains a formidable task. Among the key effects of heterogeneity and social network structure are (i) the role played by superspreaders and superspreading events during initial outbreaks (12, 13, 16, 26–28) and (ii) a substantial reduction of the final size of epidemic (FSE) as well as the herd immunity threshold (HIT) (9, 10, 19–21, 23, 29). 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 (30–34). In studying epidemics in heterogeneous populations it is important to emphasize the qualitative nature of two important time-scales. First, overdispersion reflects short-term patterns of behavior and even accidental events, and not the degree of persistent population behavioral heterogeneity. Second, short-term overdispersion is generally supposed to have a limited impact on the long term epidemic dynamics, being important primarily in early outbreaks which are dominated by super-spreading events. In other words, epidemics are generally assumed to have distinct descriptions on widely-differing time-scales. In this paper, we attempt to provide a multi-scale theory for epidemic progression and show that both overdispersion and persistent heterogeneity affect the overall progression of the COVID-19 epidemic. The significance of this multi-scale perspective is that it provides a natural formalism to predict the occurrence and nature of successive epidemic waves, even when it might seem that a first wave has attained a state which could be mistaken for herd immunity.

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

Significance Statement

Epidemics generally spread through a succession of waves that reflect factors on multiple time-scales. Here, we develop a general approach to encompass super-spreading and population heterogeneity, and demonstrate that a fragile state of transient collective immunity (TCI) emerges well below the HIT during early, high-paced stages of the epidemic. However, this is not an indication of herd immunity: subsequent waves can and will emerge due to behavioral changes in the population, driven (e.g.) by seasonal factors. Analysis of empirical data suggests that even in locations with strong first waves of COVID-19, subsequent waves will still emerge.

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

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.26.20162420; this version posted January 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

on real-world or artificial social networks (8, 13, 23, 35, 36). 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). As demonstrated in Refs. (21) the FSE is a very robust property of the epidemic, insensitive to fine details of its dynamics but defined by (i) distribution of susceptibilities in the population (20, 29, 37); (ii) 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 have so far 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, 29). The overall impact of heterogeneity occurs because as the disease spreads, it preferentially "vaccinates" 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 (30, 31, 34, 38, 39). Here, the conclusion was that the herd immunity threshold 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 (12, 13, 16, 26–28), and is expected to be a much stronger source of variation compared to the longer-term characteristics that reflect persistent heterogeneity. It is generally presumed that this short-term overdispersion has no effect on the epidemic dynamics outside the early outbreaks. 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 days 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 both to 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 SIR/SEIR models of epidemic dynamics (see, e.g. (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, e.g. for the super-spreading phenomenon) is 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 days), this will strongly affect the overall epidemic dynamics. Therefore, our initial analysis is eventually expanded to a more general theory accounting for the non-negligible 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 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 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 (40, 41). We find that with our approach, the entire effect of heterogeneity is in many cases well-characterized by a single parameter which we call the *immunity factor* λ. It is related to the statistical properties of heterogeneous susceptibility across the population and to its correlation with individual infectivity. The immunity factor determines the rate at which R_e drops during the early stages of the epidemic as the pool of susceptibles is being depleted: $R_e \approx R_0(1-\lambda(1-S))$. Beyond this early linear regime, for an important case of gamma-distributed individual susceptibilities, we show that the classical proportionality, $R_e = R_0 S$,

medRxiv preprint doi: https://doi.org/10.1101/2020.07.26.20162420; this version posted January 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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, $1 - S_0 = 1 - R_0^{-1/\lambda}$.

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 comorbidities; and (ii) social, reflecting differences in the number and frequency of close contacts of different people. The immunity factor λ in our model combines these sources of heterogeneous susceptibility as well as its correlation with individual infectivity. As we demonstrate, under certain assumptions the immunity factor is simply a product of social and biological contributions: $\lambda = \lambda_s \lambda_b$. In our study, we leverage existing studies of real-life face-to-face contact networks (13, 19, 36, 42–45) 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, there is little indication that biological variations in susceptibility, unrelated to one's social activity, play a significant role in the epidemic dynamics.

To test this theory, we use empirical data for the COVID-19 epidemic to independently estimate the immunity factor λ. In particular, we apply our previously-described epidemic model that features multi-channel Bayesian calibration (40) 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. (46). 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 4 and 5. Importantly, this represents a transient value of the parameter λ observed on intermediate timescales and dependent on both persistent and short-term heterogeneity. Our estimates of the long-term, value of the immunity factor, defined by persistent heterogeneity only, is considerably lower: $\lambda_\infty \simeq 2$. This difference explain why achieving the state of transient collective immunity 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 (40), and project possible outcomes of the second wave of the COVID-19 epidemic in during the summer months in NYC and Chicago, using data up to the end of May 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, we expect virtually no second wave in NYC in the immediate future, 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 e.g. mask wearing, contact tracing, and targeted limitation of potential super-spreading 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, 29, 31), we consider the spread of an epidemic in a population of individuals who exhibit significant persistent heterogeneity in their susceptibilities to infection α. This heterogeneity may be biological or social in origin, and we assume these factors are independent: $\alpha = \alpha_b \alpha_s$. Effects of possible correlations between α_b and α_s have been discussed in Ref. (24). The biologically-driven heterogeneous susceptibility α_b is shaped by variations of several intrinsic factors such as the strength of individuals' immune responses, age, or genetics. In contrast, the socially-driven heterogeneous susceptibility α_s is shaped by extrinsic factors, such as differences in individuals' social interaction patterns (their degree in the network of social interactions). Furthermore, individuals' different risk perceptions and attitudes 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 infectivity or 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 $R_\alpha \sim \alpha$ (9, 10, 23). In reality, not all transmissions involve face-to-face contacts, and biological susceptibility need not be strongly correlated with infectivity. Therefore, it is reasonable to expect only a partial correlation between α and R_α.

Let $S_\alpha(t)$ be the fraction of susceptible individuals in the subpopulation with susceptibility α, and let $j_\alpha(t) = -\dot{S}_\alpha$ be

medRxiv preprint doi: https://doi.org/10.1101/2020.07.26.20162420; this version posted January 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

the corresponding daily incidence 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:

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

$$J(t) = \int_0^\infty \langle R_\alpha K(\tau) j_\alpha(t-\tau)\rangle \, d\tau \quad [2]$$

Here t is the physical time and τ is the time since infection for an individual. $\langle \ldots \rangle$ represents averaging over α with pdf $f(\alpha)$. $J(t)$ represents the mean daily attack rate, i.e. hypothetical incidence rate in a fully susceptible homogeneous population with $\alpha = 1$. $K(\tau)$ is the distribution of the generation interval, which we assume to be independent of α for the sake of simplicity.

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

$$S_\alpha(t) = \exp(-\alpha Z(t)). \quad [3]$$

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

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

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

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

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 force of infection in the same form as if this were a homogeneous problem:

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

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

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

Here

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

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 *Supporting Information Appendix (SI Appendix)*. 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 (17–19). However, those empirical models exhibited a limited range of applicability (19) 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. (29). Our approach is more general: it provides the exact mapping of a wide class of heterogeneous well-mixed models onto homogeneous ones, and provides a specific relationship between the underlying statistics of α and R_α and the non-linear functions $R_e(S)$ and $S_e(S)$.

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 an epidemic significantly faster than predicted by homogeneous models. Specifically, for $1-S(t) \simeq Z(t) \ll 1$ one can linearize the effective reproduction number as

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

We named the coefficient λ the *immunity factor* because it quantifies the effect that a gradual build-up of population 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. By linearizing Eq. (5) in terms of $1 - S \simeq Z \ll 1$ and dividing the result by $R_0 = \langle \alpha R_\alpha \rangle$ one gets:

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

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

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

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

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

These equations follow from Eq. 10 in the limit $R_\alpha = \text{const}$ and $R_\alpha \sim \alpha$ 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

medRxiv preprint doi: https://doi.org/10.1101/2020.07.26.20162420; this version posted January 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

R_e in response to depletion of susceptible population S. That is why λ_s in Eq. 12 is proportional to the third moment of α_s instead of the second moment in the case of $R_0 = \langle \alpha R_\alpha \rangle \sim \alpha_s^2$. Note that the biological contribution to the immunity factor depends only on the coefficient of variation CV_b of α_b. On the other hand, the social factor λ_s depends 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 *SI Appendix* for details). The corresponding expression for the overall immunity factor is $\lambda = \langle \alpha^{2+\chi} \rangle / \langle \alpha^{1+\chi} \rangle$. The limit $\chi = 0$ corresponds to a predominantly biological source of heterogeneity, 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 $\chi = 1$, population heterogeneity is primarily of social origin, hence $\lambda \approx \lambda_s$ is affected by both CV_α and the skewness γ_α of the pdf $f(\alpha)$. The biological contribution λ_b depends on specific biological details of the disease and thus is unlikely to be as universal and robust as the social one. For the COVID-19 epidemic, there is no strong evidence of a wide variation in attack rates unrelated to social activity, geographic location, or socioeconomic status. For instance, there is very little age variability in COVID-19 prevalence as reported by the NYC Department of Health (47) 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.

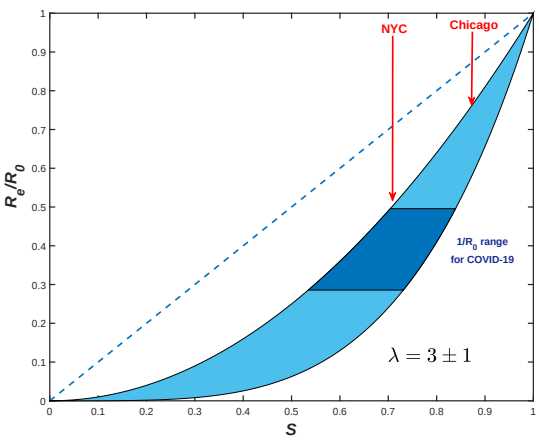

Fig. 1. R_e/R vs S dependence for gamma-distributed susceptibility with $\lambda = 3 \pm 1$ (blue area). The dashed line shows the classical homogeneous result, $R_e = R_0 S$. Note a substantial reduction of R_e for COVID-19 epidemic expected in both NYC and Chicago, compared to that value. Approximate fractions of susceptible populations, S, for both cities are estimated as of the end of May 2020, by using the model described in Ref. (40).

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 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 (see *SI Appendix*):

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

and

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

The exponent $\lambda = 1 + (1+\chi) CV_\alpha^2 = 1 + (1+\chi)\eta$ coincides with the early-epidemics immunity factor defined in Eqs. (9)–(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. (29) and more recently reproduced in Ref. (38) in the context of COVID-19. The scaling behavior $R_e(S)$ is shown in Fig. 1 for $\lambda = 3 \pm 1$. 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.

Eq. (14) immediately leads to a major revision of the classical result for the herd immunity threshold $1 - S_0 = 1 - 1/R_0$. S_0 is the fraction of susceptible population at which the exponential growth stops, while $1 - S_0$ is the relative size of the epidemic at that time. By setting $R_e = 1$ in Eq. (14), we obtain:

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

As we were finalizing this paper for public release, a preprint by Aguas et al. appeared online (48) that independently obtained our Eqs. (14, 15) for gamma-distributed susceptibilities. The same result has also been recently obtained in Ref. (39). Note however that the full quasi-homogeneous description of the epidemic dynamics requires both $R_e(S)$ and $S_e(S)$, that in the general case is characterized by different scaling exponents.

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 $\eta = CV_\alpha^2 = 1$ (see more discussion of this in the next section). A moderate biological heterogeneity would lead to an increase of the overall CV_α, but the pdf $f(\alpha)$ will still be close to the gamma distribution family. From the conceptual point of view, it is nevertheless important to understand how the function $R_e(S)$ would change if $f(\alpha)$ had a different functional form. In *SI Appendix*, we present analytic and numerical calculations for two other families of distributions: (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. (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. (11-12), as the distribution becomes more skewed, the range between the $\chi = 0$ and $\chi = 1$ curves broadens. For instance, for distributions dominated by a power law, $f(\alpha) \sim 1/\alpha^q$, with $3 < q < 4$ and $\chi = 1$, λ diverges even though CV_α remains finite. This represents a crossover to the regime of so-called scale-free networks ($2 \leq q \leq 3$), which are characterized by zero epidemic threshold yet strongly self-limited dynamics: the epidemic effectively kills itself by immunizing the hubs on the network (10, 23, 49).

medRxiv preprint doi: https://doi.org/10.1101/2020.07.26.20162420; this version posted January 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Role of short-term variations in social activity

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

Since fast variations are primarily caused by bursty dynamics of social interactions (50–53), and since heterogeneous biological susceptibility appears subdominant in the context of COVID-19, we set $\alpha_b = 1$ for all individuals. So α has purely social origin. Let $a_i(t) = \alpha_i + \delta a_i(t)$ be the time-dependent susceptibility of a person, which we associate with variable level of social activity. Hence, the same function determines also individual infectivity $a(t)R$ around time t. Interestingly, even the classical result for basic reproduction number in a heterogeneous system, $R_0 = R\langle\alpha^2\rangle$, needs to be modified due to correlated short-term variations in social activity:

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

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

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

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

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

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

Note that according to Eq. (17), $\delta\lambda(t')$ is being averaged with the weight proportional to the force of infection $J(t-t')$ since $1-S(t) \approx \int_0^\infty J(y-t')dt'$. Since $\delta\lambda(t,t')$ decreases with time difference t', its effect on λ_{eff} should be the strongest during the initial period of fast exponential growth. The initial suppression of the epidemic is caused by the combined effect of mitigation measures and both terms in λ_{eff}. Since $\lambda_{\text{eff}} > \lambda_\infty$, the population may reach the state of Transient Collective Immunity (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, bring $\lambda_{\text{eff}}(t)$ closer to λ_∞. According to Eq.(19), it is the correlation time of bursty social activity $\delta a(t)$ that sets the timescale over which this TCI state deteriorates, and the new epidemic wave may get ignited. The relationship between this relaxation time and the duration of a single epidemic wave also determines the typical value of λ_{eff} during that wave.

Despite a large number of empirical studies of contact networks (50–53), information about the temporal correlations in $\alpha(t)$ or its proxies remains limited. On the other hand, much more is known about parameters of persistent heterogeneity. Recently, real-world networks of face-to-face communications have been studied using a variety of tools, including RFID devices (42), Bluetooth and Wi-Fi wearable tags, smartphone apps (43, 44), as well as census data and personal surveys (13, 36, 45). 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 cut-off. This behavior is generally consistent with an exponential distribution in $f_s(\alpha_s)$ (19, 43, 45), $f(\alpha) \sim e^{-\alpha/\langle\alpha\rangle}$. That sets the value of $\eta = CV_\alpha^2 \approx 1$. If not for short-term overdispersion, that would yield $\lambda = \langle\alpha_s^3\rangle/\langle\alpha_s^2\rangle = 3!/2! = 3$ according to Eq. (12). However, with temporal effects taken into account, the build-up of long-term collective immunity is determined by $\lambda_{\text{eff}}(\infty) = \lambda_\infty$. In order to estimate it, we make a simple model assumption that the short-term overdispersion for a particular individual is proportional to the persistent value of that person's social activity: $\overline{\delta a_i^2} \sim \alpha_i$. This leads to

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

Here $\chi^* = \langle\alpha^2\rangle/\overline{a_i(t)^2}$ is a parameter that measures the relative strength of persistent heterogeneity and the overdispersion on the timescale of a single generation interval. Note that formally we recover our original result for λ in the purely persistent case, with χ^* replacing the parameter χ that originally quantified the correlation between infectivity and susceptibility. By assuming the limit of strong short-term overdispersion ($\chi^* \ll 1$), we obtain $\lambda_\infty \approx 2$. As shown in *SI Appendix*, the very same value of λ_∞ should be used as a scaling exponent for long-term behavior of $R_e(S)$. Therefore, HIT is set by Eq.(15) with $\lambda = \lambda_\infty \approx 2$. Its value is plotted vs. R_0 in Fig. 2, along with the homogeneous result, and the estimated threshold of TCI. To estimate the corresponding transient immunity factor λ_{eff}, we analyze the empirical data for the first wave of the COVID-19 epidemic below.

Application to the COVID-19 epidemic

The COVID-19 epidemic reached the 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

medRxiv preprint doi: https://doi.org/10.1101/2020.07.26.20162420; this version posted January 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

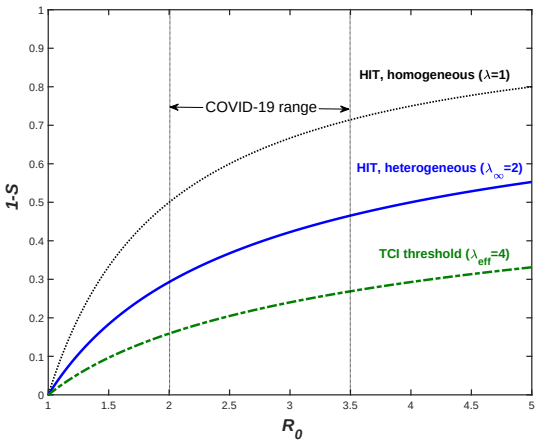

Fig. 2. TCI threshold (dot-dashed), long-term heterogeneous HIT (solid), and homogeneous HIT (dotted) for various values of R_0 (x-axis). HIT (solid line) is determined by persistent heterogeneity. The corresponding immunity factor $\lambda_\infty \approx 2$ is estimated from empirical face-to-face contact network. For transient behavior $\lambda_{\rm eff} \approx 4$ is assumed based on analysis of empirical data for COVID-19 epidemic in select locations.

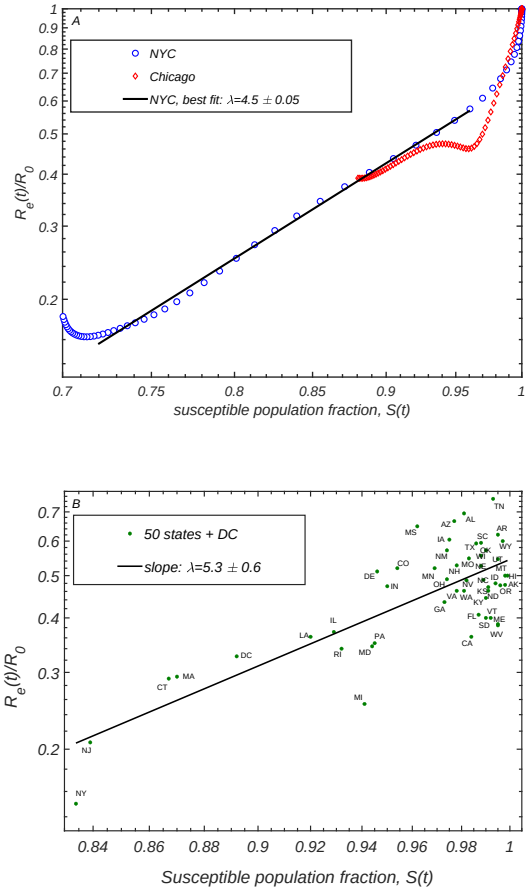

Fig. 3. Correlation between the relative reduction in the effective reproduction number $R_e(t)/R_0$ (y-axis) with the susceptible population $S(t)$. In Panel A, we present the progression of these two quantities for New York City and Chicago, as given by the epidemiological model described in Ref. (40). 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. (46) (t_0 is the latest date covered in that study).

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. (40) 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 (47, 54–56). We used these data to perform multi-channel calibration of our model (40), 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 $\lambda_{\rm eff} = 4.5 \pm 0.05$. 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 $\lambda_{\rm eff}$ in Fig. S2 we repeated our analysis in which $R_e(t)$ was corrected by Google's community mobility report in these two cities (57) (see *SI Appendix*). Although the range of data consistent with the constant slope shrank somewhat, our main conclusion remains unchanged. This provided us with a lower bound estimate for the transient immunity factor: $\lambda_{\rm eff} = 4.1 \pm 0.1$.

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 (46). 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 (46). We use state level estimates of $R_e(t)$, R_0 and $S(t)$ as reported in Ref. (46) to construct the scatter plot $R_e(t_0)/R_0$ vs $S(t_0)$ shown in 3, 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 $\lambda_{\rm eff} = 5.3 \pm 0.6$ 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 $\lambda_{\rm eff} = 4.7 \pm 1.5$. Both estimates of the immunity factor

medRxiv preprint doi: https://doi.org/10.1101/2020.07.26.20162420; this version posted January 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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, $\lambda_{\rm eff} \simeq 4$, is set by the the pace of the first epidemic wave in the US. As expected, it exceeds our estimate of $\lambda_\infty \approx 2$ associated with persistent heterogeneity and responsible for the long-term herd immunity.

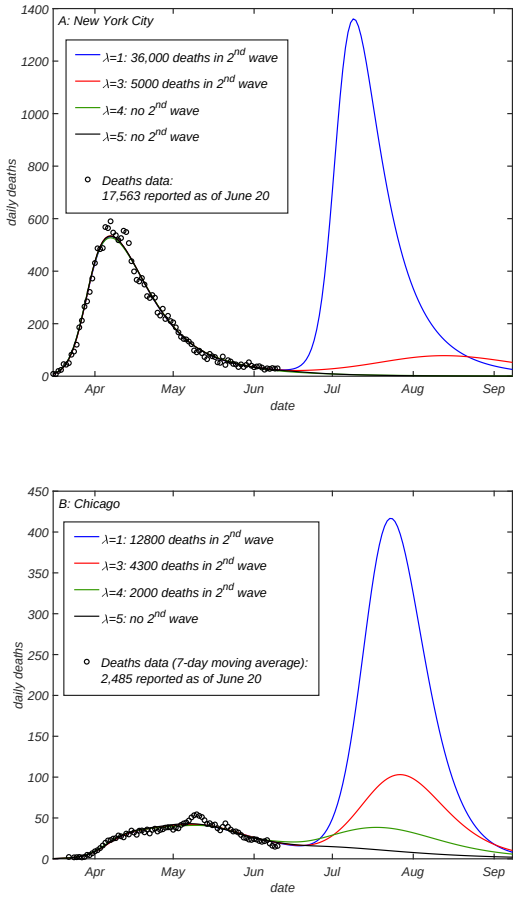

Fig. 4. Projections of daily deaths under the hypothetical scenario in which any mitigation is completely eliminated as of June 15 2020, for (A) NYC and (B) Chicago. Different curves correspond to different values of the transient immunity factor $\lambda_{\rm eff} = 1$ (blue), 3 (red), 4 (green) and 5 (black lines). The model described in Ref. (40) was fully calibrated on daily deaths (circles), ICU occupancy and hospitalization data up to the end of May. See *SI Appendix*, Figs. S4-S5 in for additional details, including confidence intervals.

We can now incorporate this transient level of heterogeneity into our epidemiological model, and examine how future projections change as a result of this modification. This is done by plugging scaling relationships given by Eqs. (13)-(14) into the force of infection and incidence rate equations of the original model. These equations are similar to Eqs. (6)-(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 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 TCI, $R_e = R_0 S^\lambda$. The projected daily deaths for each of the two cities under this (unrealistically harsh) scenario are presented in Fig. 4 for various values of λ. For both cities, the homogeneous model ($\lambda = 1$, blue lines) predicts a second wave which is larger than the first one with an additional death toll of around $35,000$ in NYC and $12,800$ in Chicago. The magnitude of the second wave is greatly reduced by heterogeneity, resulting in no second wave in either of the two cities for $\lambda = 5$ (black lines). Even for a modest value $\lambda = 3$ (red lines), which is less than our estimate, the second wave is dramatically reduced in both NYC and Chicago (by about 90% and 70%, respectively).

Fragility of Transient Collective Immunity

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

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

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

Epidemiological models with rewiring of underlying social networks have been studied before (58) (see (59) for a review), but under a constraint that the individual level of social activity quantified by network degree is preserved. In contrast, the dynamics described above stems from the individual level of social activity α_s changing in time.

Fig. 5 shows the simulation of SIR model, where the first wave of the epidemic is mitigated, thereby reducing the effective reproduction number $R_0 = 2.5$. During the course of the mitigation R_0 is multiplied by $\mu = 0.7$. After the mitigation measures have been lifted at the end of the first wave, the population is positioned slightly below the TCI threshold preventing the immediate start of the second wave. However, gradual rewiring of the social network with time constant $\tau_s = 150$ days ultimately results in the second and even the third wave of the epidemic (see Fig. 5). The inset of this figure shows $R_e(t)$ plotted as a function of $S(t)$ in this epidemic.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.26.20162420; this version posted January 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Note that each of the waves follows the power law relationship between $R_e(t) \sim S(t)^\lambda$ predicted by Eq. 14. Since constant rewiring eliminates correlations in individual social activity on scales longer than τ_s, the epidemic stops after multiple waves bring the total fraction of infected individuals close to the unmodified (homogeneous) herd immunity threshold $1/R_0$. 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.

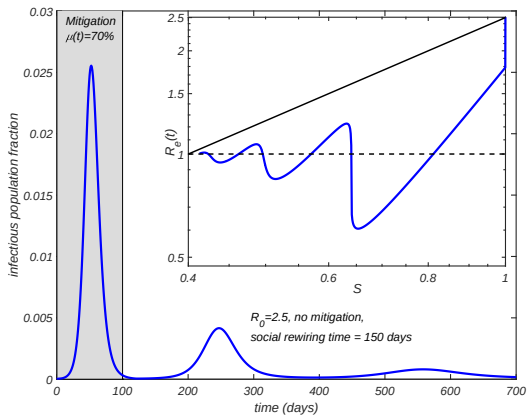

Fig. 5. Effect of social rewiring on the epidemic dynamics. The time course of an epidemic in a heterogeneous SIR model with $R_0 = 2.5$ and $\lambda = 3$. During the first 100 days a mitigation factor $\mu = 0.7$ is applied. Social networks gradually rewire with a time constant $\tau_s = 150$ days.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

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 time scales, 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 non-linear functions $R_e(S)$ and $S_e(S)$, both of which are fully determined by the statistics of individual susceptibilities and infectivities. Furthermore, $R_e(S)$ is largely defined by a single parameter, the immunity factor λ, introduced in our study. Like susceptibility itself, λ has two contributions: biological and social (see Eqs. (11-12)).

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 slow down of a fast-paced epidemic, and to its medium-term stabilization. However, this state of Transient Collective Immunity (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, 29, 33) 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 (see Fig. 5). The FSE result may still be used, but primarily as an estimate for the size of the first wave of an (unmitigated) epidemic. In that case, the transient value of immunity factor λ_{eff} should be assumed.

By applying our theory to the COVID-19 epidemic we found evidence that the hardest-hit areas such as New York City, have likely passed TCI threshold by the end of the first wave, but are less likely to have achieved a real long-term herd immunity. Other places that had intermediate exposure, such as e.g. 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, e.g. 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 (34), 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. (34): $S_0 = 62.5\%$, 53.5%, and 47.5%, respectively.

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

medRxiv preprint doi: https://doi.org/10.1101/2020.07.26.20162420; this version posted January 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

behavioral pattern. On the other hand, superspreading events are shaped by short-time variations in individual infectivity (e.g. a person during 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 about 0.1 for COVID-19 (28). According to Ref. (12), this is consistent with the expected value of the individual-level reproduction number R_i drawn from a gamma distribution with the shape parameter $k \simeq 0.1$. 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. (30). Similarly, the degree of heterogeneity assumed in other recent studies (31, 48) is considerably larger than our estimates. Based on our analysis, the value of the immunity factor λ depends on the pace of the epidemic and on the timescale under consideration. We estimated its long-term value (responsible for the permanent HIT) as $\lambda_\infty \approx 2$. However, the transient values are expected to be higher, especially during the first several waves of COVID-19 in select locations, characterized by large growth rates. Our analysis of the empirical data in NYC and Chicago indicate that the slowdown of the epidemic dynamics in those locations was consistent with $\lambda \approx 4$. In Table 1 we present our estimates of the factor by which R_e is transiently suppressed as a result of depletion of susceptible population in selected locations in the world, as of early June 2020, as well as the predicted long-term suppression related to acquisition of a partial herd immunity.

Table 1. Effects of heterogeneity on suppression of the effective reproduction number R_e in selected locations. The transient and long-term suppression coefficients R_e/R_0 are calculated using $\lambda = 4$ and $\lambda = 2$, 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 (IFR) of 0.7% (60).

Location	Deaths, per 1000	Attack Rate (1-S)	R_e suppression transient	R_e suppression long-term
New York City, USA (47)	2.1	30%	0.24	0.50
Lombardy, Italy (61)	1.7	24%	0.33	0.58
London,UK (62)	0.9	13%	0.58	0.75
Chicago, USA (54)	0.9	13%	0.58	0.75
Stockholm, Sweden (63)	0.9	13%	0.58	0.75

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 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.

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. Kermack WO, McKendrick AG (1927) A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing papers of a mathematical and physical character* 115(772):700–721.
2. Keeling MJ, Rohani P (2011) *Modeling infectious diseases in humans and animals.* (Princeton University Press).
3. Rock K, Brand S, Moir J, Keeling MJ (2014) Dynamics of infectious diseases. *Reports on Progress in Physics* 77(2):026602.
4. Ma J (2020) Estimating epidemic exponential growth rate and basic reproduction number. *Infectious Disease Modelling* 5:129 – 141.
5. Fraser C (2007) Estimating individual and household reproduction numbers in an emerging epidemic. *PLoS One* 2(8):e758.
6. Chowell G (2017) Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A Primer for parameter uncertainty, identifiability, and forecasts. *Infect Dis Model* 2(3):379–398.
7. Lloyd AL, May RM (2001) Epidemiology - how viruses spread among computers and people. *Science* 292(5520):1316–1317.
8. May RM, Lloyd AL (2001) Infection dynamics on scale-free networks. *Physical Review E* 64(6):066112.
9. Newman MEJ (2002) Spread of epidemic disease on networks. *Physical Review E* 66(1).
10. Moreno Y, Pastor-Satorras R, Vespignani A (2002) Epidemic outbreaks in complex heterogeneous networks. *European Physical Journal B* 26(4):521–529.
11. Dezső Z, Barabási AL (2002) Halting viruses in scale-free networks. *Physical Review E* 65(5):055103.
12. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM (2005) Superspreading and the effect of individual variation on disease emergence. *Nature* 438(7066):355–9.
13. Meyers LA, Pourbohloul B, Newman MEJ, Skowronski DM, Brunham RC (2005) Network theory and sars: predicting outbreak diversity. *Journal of Theoretical Biology* 232(1):71–81.
14. Keeling MJ, Eames KTD (2005) Networks and epidemic models. *Journal of the Royal Society Interface* 2(4):295–307.
15. Ferrari MJ, Bansal S, Meyers LA, Bjornstad ON (2006) Network frailty and the geometry of herd immunity. *Proceedings of the Royal Society B-Biological Sciences* 273(1602):2743–2748.
16. Small M, Tse C, Walker DM (2006) Super-spreaders and the rate of transmission of the sars virus. *Physica D: Nonlinear Phenomena* 215(2):146–158.
17. Roy M, Pascual M (2006) On representing network heterogeneities in the incidence rate of simple epidemic models. *Ecological Complexity* 3(1):80–90.
18. Stroud PD, et al. (2006) Semi-empirical power-law scaling of new infection rate to model epidemic dynamics with inhomogeneous mixing. *Mathematical Biosciences* 203(2):301–318.
19. Bansal S, Grenfell BT, Meyers LA (2007) When individual behaviour matters: homogeneous and network models in epidemiology. *Journal of the Royal Society Interface* 4(16):879–891.
20. Katriel G (2012) The size of epidemics in populations with heterogeneous susceptibility. *Journal of Mathematical Biology* 65(2):237–262.
21. Miller JC (2012) A note on the derivation of epidemic final sizes. *Bulletin of mathematical biology* 74(9):2125–2141.
22. Bansal S, Meyers LA (2012) The impact of past epidemics on future disease dynamics. *Journal of Theoretical Biology* 309:176–184.
23. Pastor-Satorras R, Castellano C, Van Mieghem P, Vespignani A (2015) Epidemic processes in complex networks. *Reviews of Modern Physics* 87(3):925.
24. Gou W, Jin Z (2017) How heterogeneous susceptibility and recovery rates affect the spread of epidemics on networks. *Infectious Disease Modelling* 2(3):353–367.
25. Kim Y, Ryu H, Lee S (2018) Agent-based modeling for super-spreading events: A case study of mers-cov transmission dynamics in the republic of korea. *International Journal of Environmental Research and Public Health* 15(11):2369.
26. Galvani AP, May RM (2005) Epidemiology - dimensions of superspreading. *Nature* 438(7066):293–295.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.26.20162420; this version posted January 30, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

27. Shen Z, et al. (2004) Superspreading sars events, beijing, 2003. *Emerging Infectious Diseases* 10(2):256–260.
28. Endo A, Abbott S, Kucharski A, Funk S (2020) Estimating the overdispersion in covid-19 transmission using outbreak sizes outside china. *Wellcome Open Research* 5:67.
29. Novozhilov AS (2008) On the spread of epidemics in a closed heterogeneous population. *Mathematical Biosciences* 215(2):177–185.
30. Hébert-Dufresne L, Althouse BM, Scarpino SV, Allard A (2020) Beyond r0: Heterogeneity in secondary infections and probabilistic epidemic forecasting. *medRxiv* 2020.02.10.20021725.
31. Gomes MGM, et al. (2020) Individual variation in susceptibility or exposure to sars-cov-2 lowers the herd immunity threshold. *medRxiv* 2020.04.27.20081893.
32. Brennan PV, Brennan LP (2020) Susceptibility-adjusted herd immunity threshold model and potential r0 distribution fitting the observed covid-19 data in stockholm. *medRxiv* 2020.05.19.20104596.
33. Ball F (1985) Deterministic and stochastic epidemics with several kinds of susceptibles. *Advances in Applied Probability* 17(1):1–22.
34. Britton T, Ball F, Trapman P (2020) A mathematical model reveals the influence of population heterogeneity on herd immunity to sars-cov-2. *Science* 369(6505):846–849.
35. Stack JC, Bansal S, Kumar VSA, Grenfell B (2013) Inferring population-level contact heterogeneity from common epidemic data. *Journal of the Royal Society Interface* 10(78).
36. Eubank S, et al. (2004) Modelling disease outbreaks in realistic urban social networks. *Nature* 429(6988):180–184.
37. Hickson R, Roberts M (2014) How population heterogeneity in susceptibility and infectivity influences epidemic dynamics. *Journal of Theoretical Biology* 350:70–80.
38. Rose C, et al. (2020) Heterogeneity in susceptibility dictates the order of epidemiological models. *arXiv:* 2005.04704.
39. Neipel J, Bauermann J, Bo S, Harmon T, Jülicher F (2020) Power-law population heterogeneity governs epidemic waves. *PLoS one* 15(10):e0239678.
40. Wong GN, et al. (2020) Modeling covid-19 dynamics in illinois under non-pharmaceutical interventions. *Physical Review X* 10.
41. Castro M, Ares S, Cuesta JA, Manrubia S (2020) The turning point and end of an expanding epidemic cannot be precisely forecast. *Proceedings of the National Academy of Sciences* 117(42):26190–26196.
42. Isella L, et al. (2011) What's in a crowd? analysis of face-to-face behavioral networks. *Journal of Theoretical Biology* 271(1):166–180.
43. Nielsen BF, Sneppen K, Simonsen L, Mathiesen J (2020) Heterogeneity is essential for contact tracing. *medRxiv* 2020.06.05.20123141.
44. Starnini M, et al. (2017) Robust modeling of human contact networks across different scales and proximity-sensing techniques in *Social Informatics*, eds. Ciampaglia GL, Mashhadi A, Yasseri T. (Springer International Publishing, Cham), pp. 536–551.
45. Danon L, Read JM, House TA, Vernon MC, Keeling MJ (2013) Social encounter networks: characterizing great britain. *Proceedings of the Royal Society B-Biological Sciences* 280(1765).
46. Unwin HJT, et al. (2020) Report 23: State-level tracking of COVID-19 in the United States WHO Collaborating Centre for Infectious Disease Modelling MRC Centre for Global Infectious Disease Analytics.
47. (2020) Data w from https://github.com/nychealth/coronavirus-data.
48. Aguas R, et al. (2020) Herd immunity thresholds for sars-cov-2 estimated from unfolding epidemics. *medRxiv* 2020.07.23.20160762.
49. Pastor-Satorras R, Vespignani A (2001) Epidemic spreading in scale-free networks. *Physical Review Letters* 86(14):3200–3203.
50. Barabasi AL (2005) The origin of bursts and heavy tails in human dynamics. *Nature* 435(7039):207–211.
51. Kossinets G, Watts DJ (2006) Empirical analysis of an evolving social network. *science* 311(5757):88–90.
52. Rybski D, Buldyrev SV, Havlin S, Liljeros F, Makse HA (2009) Scaling laws of human interaction activity. *Proceedings of the National Academy of Sciences* 106(31):12640–12645.
53. Saramäki J, Moro E (2015) From seconds to months: an overview of multi-scale dynamics of mobile telephone calls. *The European Physical Journal B* 88(6):164.
54. (2020) Data were downloaded from https://www.dph.illinois.gov/covid19/covid19-statistics.
55. (2020) Data from https://github.com/thecityny/covid-19-nyc-data.
56. (2020) Data originally due to https://www.thecity.nyc/.
57. (2020) https://www.blog.google/technology/health/covid-19-community-mobility-reports?hl=en.
58. Volz E, Meyers LA (2007) Susceptible–infected–recovered epidemics in dynamic contact networks. *Proceedings of the Royal Society B: Biological Sciences* 274(1628):2925–2934.
59. Bansal S, Read J, Pourbohloul B, Meyers LA (2010) The dynamic nature of contact networks in infectious disease epidemiology. *Journal of biological dynamics* 4(5):478–489.
60. Meyerowitz-Katz G, Merone L (2020) A systematic review and meta-analysis of published research data on COVID-19 infection-fatality rates. *medRxiv* 2020.05.03.20089854.
61. (2020) Data from https://github.com/pcm-dpc/COVID-19.
62. (2020) Data from https://data.london.gov.uk/dataset/coronavirus--covid-19--deaths.
63. (2020) Data from https://github.com/mrunesson/covid-19.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.26.20162420; this version posted January 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Supporting Information Appendix

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 force of infection :

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

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

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

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

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

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

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

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

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

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

medRxiv preprint doi: https://doi.org/10.1101/2020.07.26.20162420; this version posted January 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

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 $f(\alpha)$

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

This function also determines the 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 the 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 the 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. (14).

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

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

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

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

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

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

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

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

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.

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

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

medRxiv preprint doi: https://doi.org/10.1101/2020.07.26.20162420; this version posted January 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Final Size of Epidemic. Here we derive a simple result for FSE in a population with a persistent heterogeneity. 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 fraction of the susceptible population, S_∞:

$$S_\infty = M_\alpha\left(-\int_{S_\infty}^1 \frac{R_e(S)dS}{S_e(S)}\right) \tag{S30}$$

This equation is valid for an arbitrary distribution of α, arbitrary correlation between susceptibility and infectivity, and for any statistics of the generation interval. This result can be also obtained as a solution to a general integral equation derived in Ref. (21) for the well-mixed case. Eq. S30 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.(S30) reproduces results of Refs. (20, 21, 29, 33), (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 (9, 10, 21).

For the case of gamma-distributed persistent susceptibility Eq. (S30), gives:

$$S_\infty = \left(1 + \frac{R_0\eta\left(1-S_\infty^{\lambda-\eta}\right)}{\lambda-\eta}\right)^{-1/\eta} \tag{S31}$$

It should be emphasized however that this result is of limited relevance to more realistic situations. Even if one assumes no government-imposed mitigation or societal response to the epidemic, the case of fully persistent heterogeneity is just an approximation. As we demonstrate in our paper, short-term correlations of time-dependent individual susceptibilities and infectivities lead to transient stabilization of a fast-pacing epidemic. Because of this effect, Eq.(S30)-Eq.(S31) should be interpreted as an estimate of the size of the first wave rather than the actual FSE.

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

$$\mathrm{E}\left[\dot{s}_i(t)\right] = -a_i(t)s_i(t-0)J(t) \tag{S32}$$

$$J(t) = -\int_0^\infty R\cdot K(\tau)\overline{a_i(t)\dot{s}_i(t-\tau)}d\tau \tag{S33}$$

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

The overall quasi-homogeneous description given by Eqs.(7)-(6), remains valid. It is obtained by averaging Eqs. (S32)-(S33) over the entire population. However, in contrast to the case of persistent heterogeneity, variables $S(t)$, $S_e(t)$ and R_e are no longer connected to each other by a simple functional relationship. To relate them we first note that the average probability that an individual is still susceptible at time t is given by $E[s_i(t)] = \exp\left(-\int_{-\infty}^t J(t')a_i(t')dt'\right)$. Therefore,

$$S(t,[J(t')]) \equiv \overline{s_i(t)} = \overline{\exp\left[-\int_{-\infty}^t J(t')a_i(t')dt'\right]} \tag{S34}$$

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

$$S_e(t) = \overline{a_i(t)s_i(t)} = -\frac{\delta S(t,[J(t')])}{\delta J(t)} \tag{S35}$$

$$R_e(t) = R\overline{a_i^2(t)s_i(t)} = R\frac{\delta^2 S(t,[J(t')])}{\delta J(t)^2} \tag{S36}$$

For the sake of simplicity, in deriving Eq.(S36) we assumed $\alpha_i(t)$ to be smoothed over the timescale of a single generation interval. As a result, $\int_0^\infty \overline{a_i(t)a_i(t-\tau)}K(\tau)d\tau \approx \overline{a_i^2(t)}$. By applying Eq.(S36) to the initial state of fully susceptible population we obtain the result for R_0, Eq. (16):

$$R_0 = R\left(\langle\alpha^2\rangle + \overline{\delta a_i^2}\right) \tag{S37}$$

At the early stages of epidemic $E[s_i(t)] \approx 1 - \int_{-\infty}^t a_i(t')J(t')dt'$ for the entire population. After substituting this expression for $s_i(t)$ to Eq.(S36) one obtains a generalization of our previous result for the initial suppression of R_e, Eqs.(9)-(10):

$$R_e(t) \approx R_0\left(1 - \int_0^\infty \Lambda(t,t')J(t-t')dt'\right) \tag{S38}$$

$$\Lambda(t,t') = \frac{\overline{\tilde{\alpha}_i^2(t)a_i(t-t')}}{\langle\alpha^2\rangle + \overline{\delta a_i^2}} = \lambda_\infty + \delta\lambda(t,t') \tag{S39}$$

medRxiv preprint doi: https://doi.org/10.1101/2020.07.26.20162420; this version posted January 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Here $\lambda_\infty = \Lambda(\infty)$ and $\delta\lambda(t,t')$ are the constant and time-dependent contributions to "immunity kernel" $\Lambda(t,t')$ which are discussed in the main text.

To obtain a corrected result for HIT, we assume a very slow progression of the epidemic (e.g. due to a gradual relaxation of the level of mitigation). In this case, any intermediate-term correlations between time dependent variations $\delta\alpha_i(t)$ become negligible, and we largely recover the formalism developed for pure persistent heterogeneity. Under the same assumption that was used for the estimate of λ_∞ in the main text (i.e. $\overline{\delta a_i^2} \sim \alpha_i$), the only modification that needs to be done is to replace the term αR_α in Eq. (5) with $(\chi^*\alpha^2 + (1-\chi^*)\alpha)R_0$ (where $\chi^* = \langle\alpha^2\rangle/\overline{a_i^2}$). This gives an expression for $R_e(S)$ in terms of two earlier results, $R_e^{(\chi)}(S)$ (for $\chi = 0$ and 1, respectively):

$$\frac{R_e}{R_0} = \chi^* R_e^{(1)} + (1-\chi^*)R_e^{(0)} \approx S^{\lambda_\infty} \qquad \text{[S40]}$$

Here $\lambda_\infty = 1 + (1+\chi^*)\eta$. To obtain this result, we substituted the scaling functions for gamma-distributed persistent heterogeneity, $R_e(S) = R_0 S^\lambda$, with $\lambda = 1 + (1+\chi)\eta$, as given by Eq.(14).

medRxiv preprint doi: https://doi.org/10.1101/2020.07.26.20162420; this version posted January 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

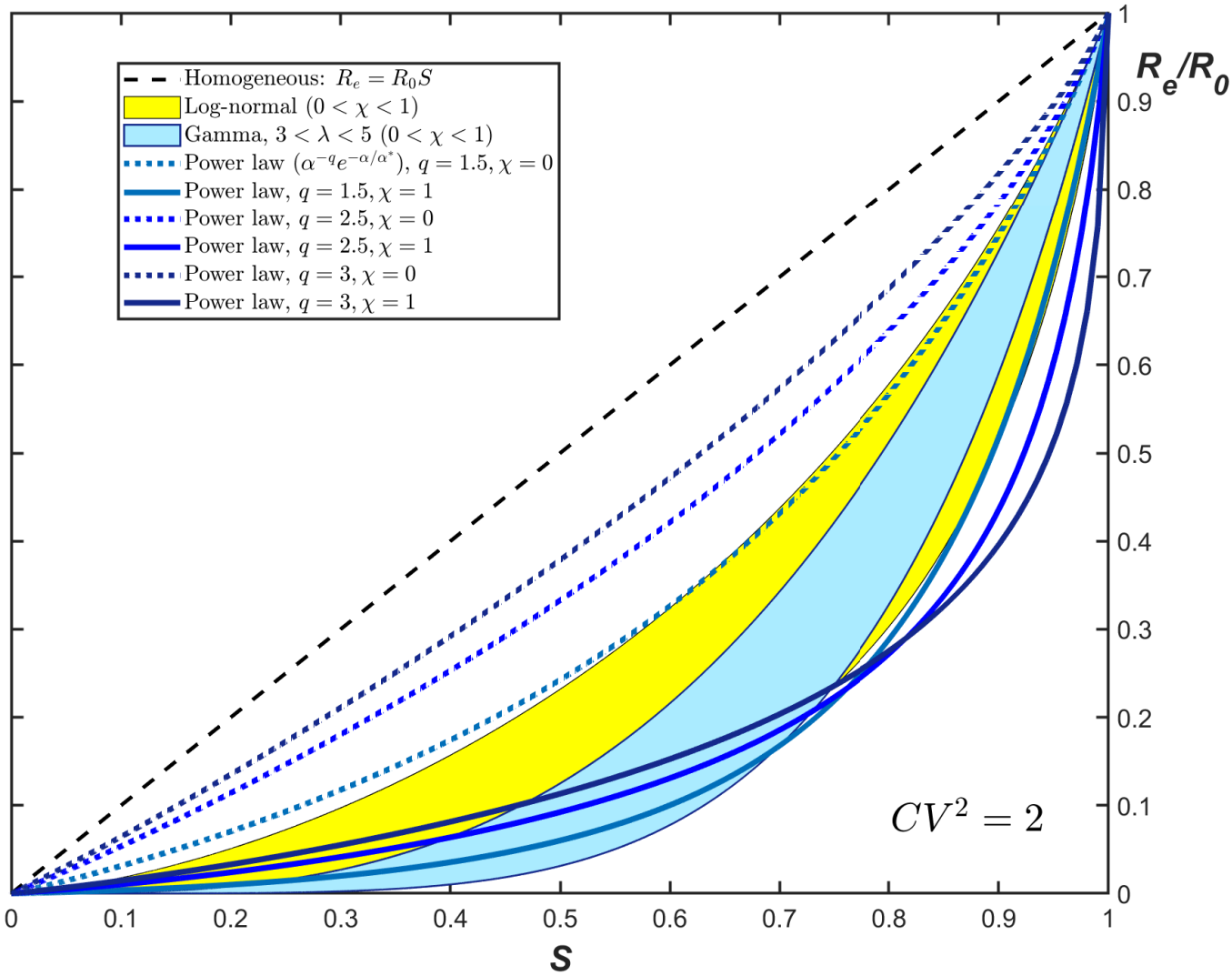

Fig. S1. R_e/R_0 vs S dependence for three different families of probability distribution $f(\alpha)$: Gamma (light blue), truncated power law (dashed lines), and log-normal (yellow). Different curves correspond to the same value of the coefficient of variation $CV_\alpha^2 = 2$, and two limiting values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

medRxiv preprint doi: https://doi.org/10.1101/2020.07.26.20162420; this version posted January 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

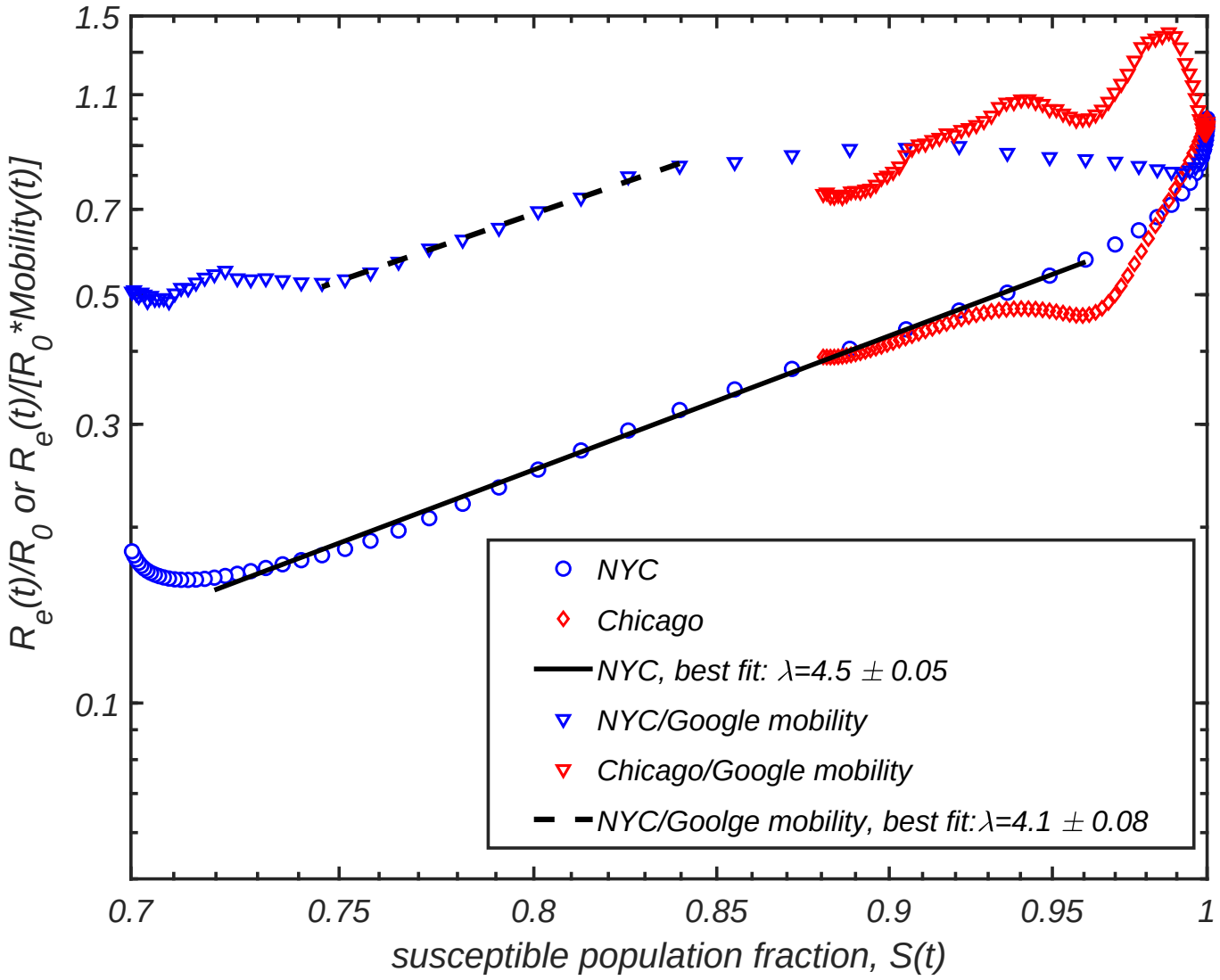

Fig. S2. Exploration of effect of mobility on data presented in Figure 3(A). Triangles represent data points for NYC and Chicago with $R_e(t)/R_0$ corrected by a mobility factor calculated from Google community mobility report, Ref. (57). We compute the mobility for NYC using average mobility of its five counties: New York county, Bronx county, Kings county, Richmond county, and Queens county, weighted by their population fraction.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.26.20162420; this version posted January 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

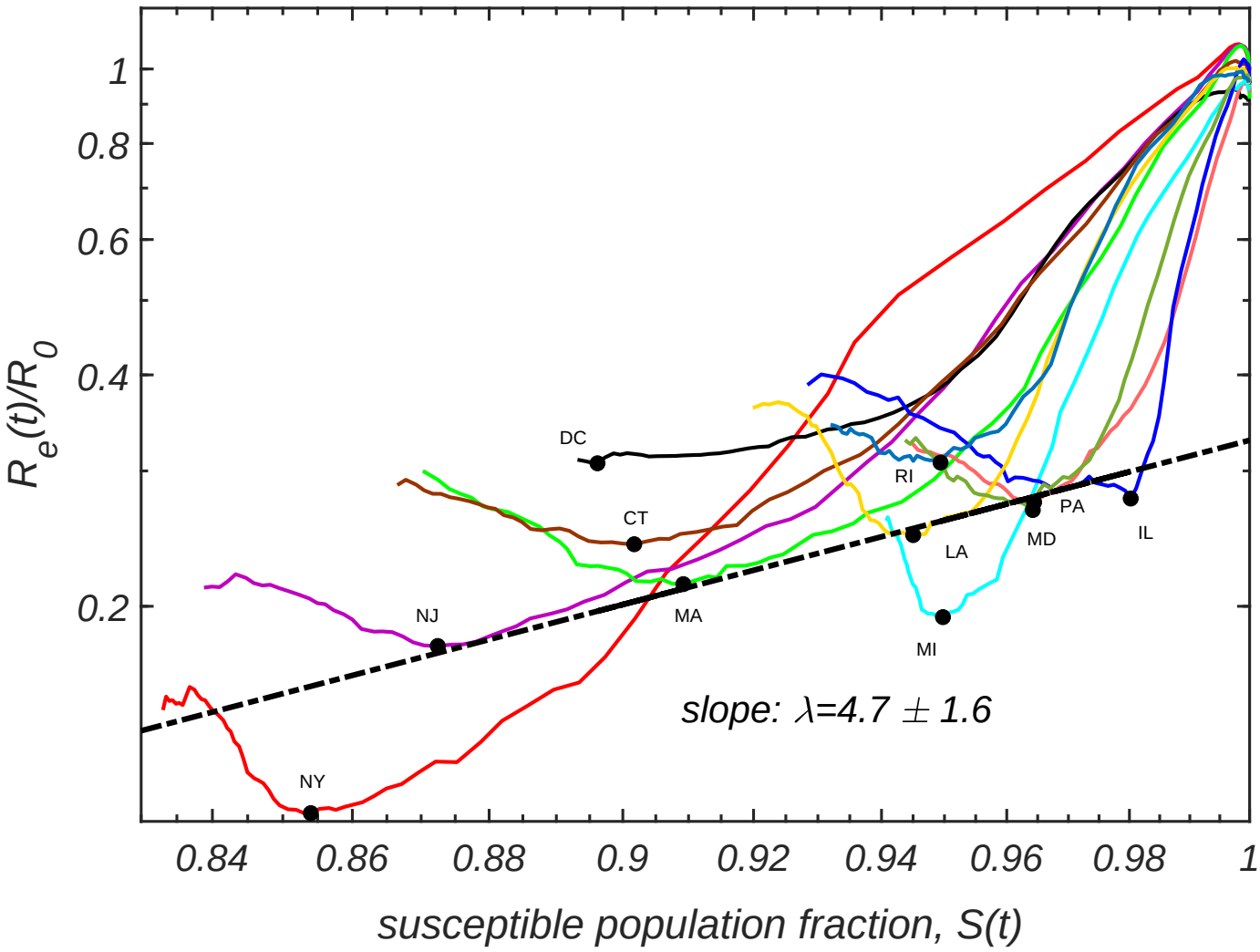

Fig. S3. Time progressions of $Re(t)/R_0$ and $S(t)$ for the hardest-hit US states and DC, as reported in Ref. (46). Black dots correspond to absolute minima of transmission and population susceptible fractions. The dashed line with slope $\lambda = 4.7 \pm 1.6$ is the best power law fit through these black dots.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.26.20162420; this version posted January 30, 2021. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

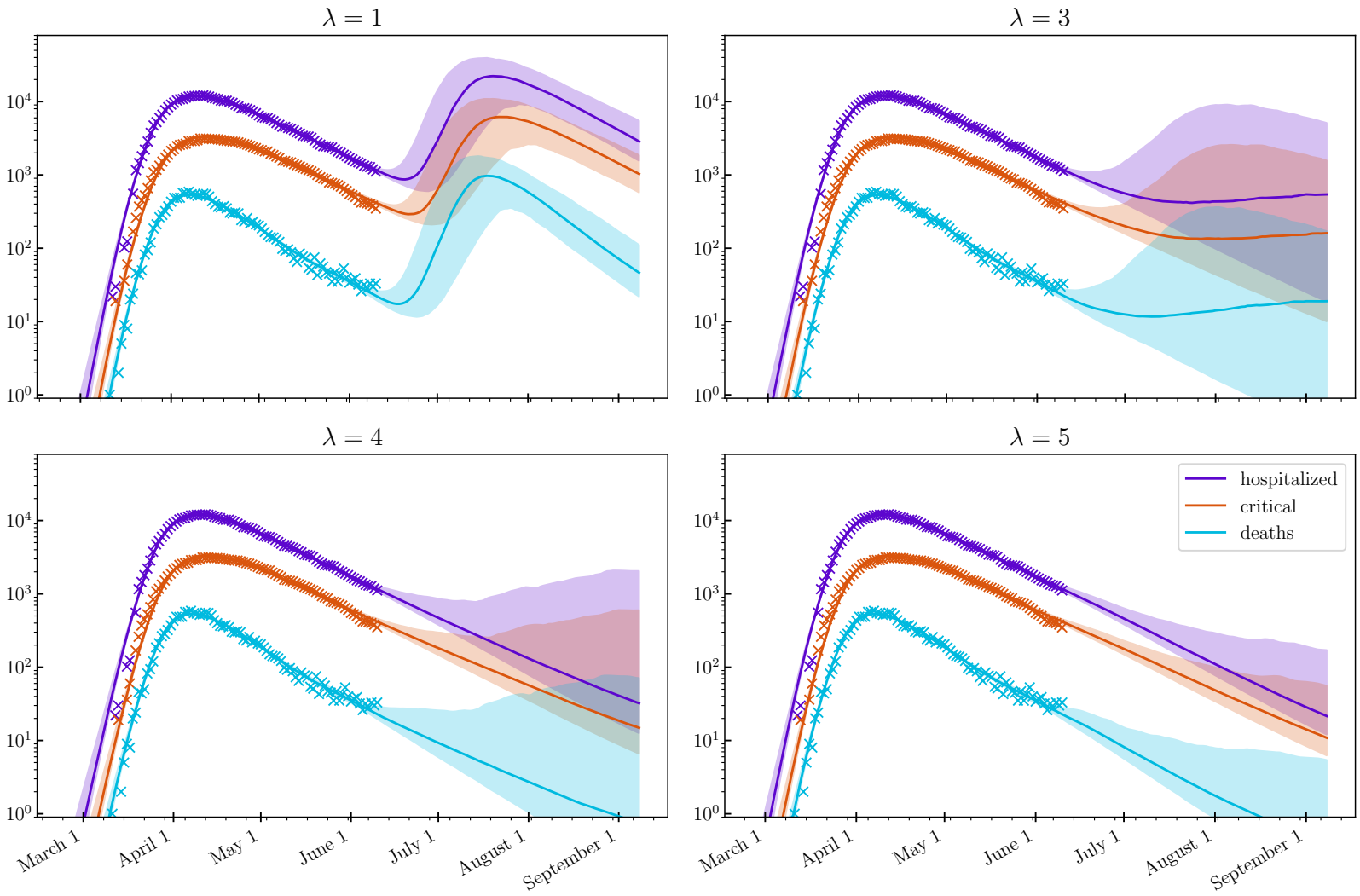

Fig. S4. Hospitalization, ICU occupancy and daily deaths in NYC modeled under hypothetical scenario when any mitigation is completely eliminated as of Jun 15 2020, for various values of λ. Model described in Ref. (40) is calibrated on data from Ref.(54), up to June 10, 2020 (shown as crosses). 95% confidence intervals are indicated.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.26.20162420; this version posted January 30, 2021. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

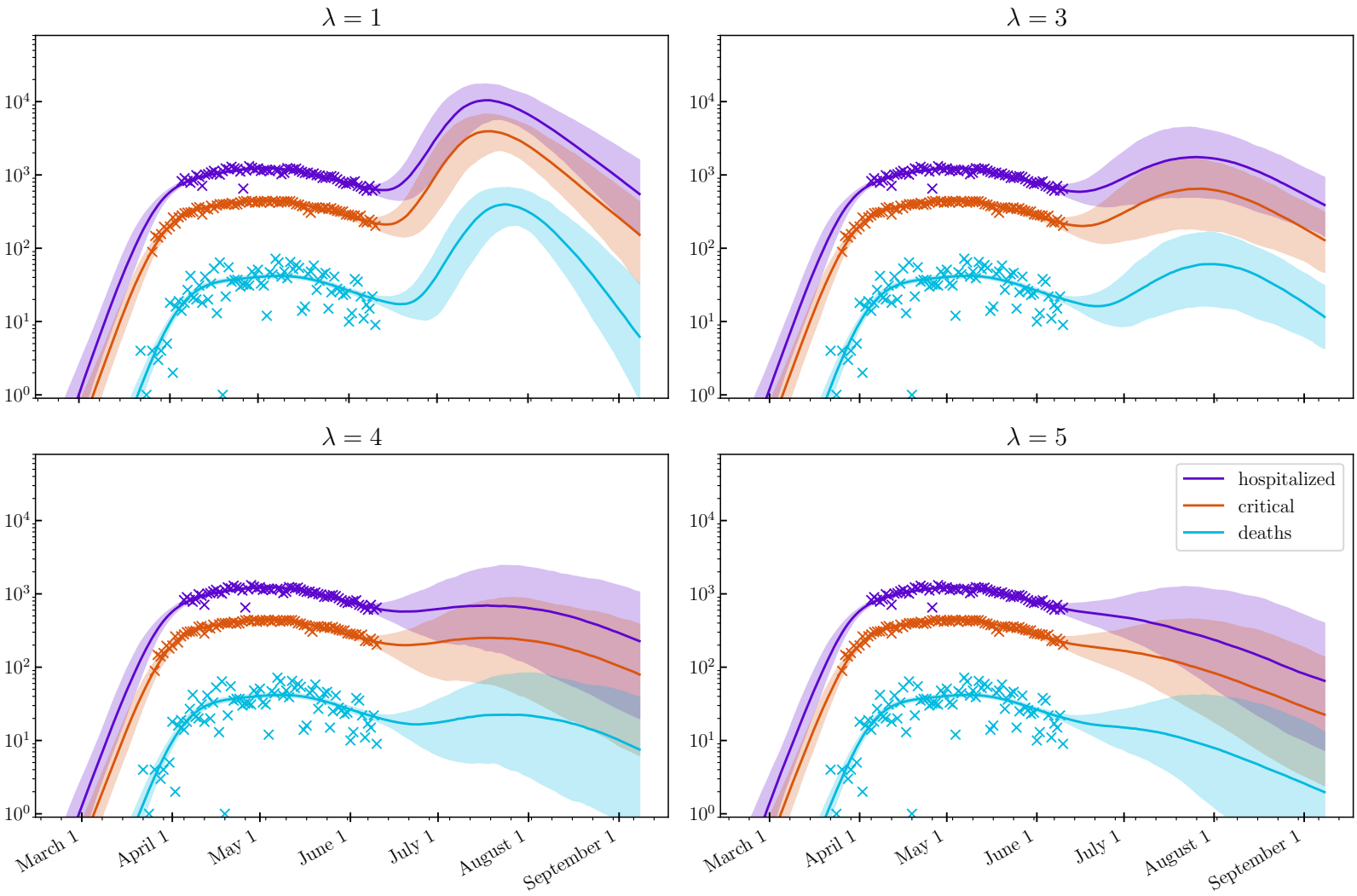

Fig. S5. Hospitalization, ICU occupancy and daily deaths in Chicago modeled under hypothetical scenario when any mitigation is completely eliminated as of Jun 15 2020, for various values of λ. Model described in Ref. (40) is calibrated on data from Ref.(54), up to June 10, 2020 (shown as crosses). 95% confidence intervals are indicated.