Founder effects arising from gathering dynamics systematically bias emerging pathogen surveillance

Bradford P. Taylor[1*] **and William P. Hanage**[1]

***For correspondence:**
bradfordptaylor@gmail.com (BPT)

[1]Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA

Abstract Models of infectious disease transmission have shown the importance of heterogeneous contact networks for epidemiology; the most connected individuals are most likely to be infected early. Yet it is cumbersome to parameterize and incorporate such networks into simple models. We introduce an alternative model framework that explicitly includes attendance at and disease transmission within gatherings of different sizes, which disaggregates sequential epidemics moving from the most to least social subpopulations that underly the overall, single-peaked infection curve. This can systematically bias initial estimates of the growth rate for emerging variants and their severity, if vulnerable populations avoid large gatherings. Finally, we show that how often similarly social individuals preferentially interact (i.e., homophily, or assortative mixing) tunes the magnitude and duration of these biases. Together, we provide a simple framework for incorporating socialization and behavior in epidemic models, which can help contextualize surveillance of emerging infectious agents.

Introduction

Directly transmitted infectious agents rely on the behavior of their hosts to start new infections and maintain themselves – specifically the contacts that their hosts make along which transmission may then occur. The contact networks that link individuals in a population are hence the routes along which communicable diseases spread, and considerable effort has been made to understand the nature of these networks and the implications for control of infectious agents (***Bansal et al., 2010***). The numbers of contacts individuals make can vary greatly, just as the nature of contacts can differ; some may be merely fleeting opportunities for a pathogen to initiate infection while others are sustained exposures to a large inoculum. Such heterogeneity in secondary contact rates appears to be common across epidemics, given over-dispersion of empirically observed secondary infection distributions, meaning that few individuals disproportionately contribute more towards ongoing transmission (***May and Anderson, 1987***; ***Woolhouse et al., 1997***; ***Lloyd-Smith et al., 2005***; ***Stein, 2011***). While some of this may be due to infections with unusual properties, those infectious must still make contacts in order to transmit to them (***Quinn et al., 2000***; ***Fraser et al., 2007***; ***Matthews et al., 2005***; ***Lawley et al., 2008***; ***Gopinath et al., 2012***; ***Edwards et al., 2021***).

Another feature that contact networks exhibit is assortativity: the tendency for contacts to be made between nodes with similar properties (***Newman, 2002***). In social networks this is termed homophily, and in the case of communicable disease the consequence is that most transmission events take place among groups of similar individuals, which can be explicitly incorporated with a contact matrix (e.g., for age-structured models) (***McPherson et al., 2001***; ***Rohani et al., 2010***).

Varying contact networks can be incorporated into dynamical models of transmission and have produced insights into subjects such as the optimal allocation of vaccines (those which prevent transmission are more effective when used in those with the most contacts) or the importance of core groups in maintaining sexually transmitted infections and as a focus of interventions (***Stigum et al., 1994***; ***Endo et al., 2022***).

A special case of assortativity and homophily is when individuals share increased risks of exposure, and increased numbers of contacts – for example through gathering in groups of larger size. Larger gatherings both increase the chance that an infectious person attends, and the numbers to whom they may transmit (***Altizer et al., 2003***; ***Godfrey et al., 2009***; ***Chande et al., 2020***). The resulting variations in risk of exposure may result from voluntary social interactions (and contact rates tend to be higher in younger people), but can also include sexual networks in which the majority of at-risk contacts occur between a small proportion of the population, or crowded settings such as workplaces in which employees are unable to mitigate their own risks of exposure (e.g., meat-packing plants in the early stages of the covid pandemic are a well-known case of workplaces in which large outbreaks occurred) (***Yorke et al., 1978***; ***Volz and Meyers, 2007***; ***Taylor et al., 2020***). In the context of an emerging pathogen or variant such transmission heterogeneity and homophily mean that any estimates from local surveillance efforts may not reflect the epidemic potential in other communities. The social context of disease spread can produce biases in surveillance given differences in resources between settings.

Mathematical models may be used to project the expected consequences of outbreaks. For example, in the classic SIR model for a directly transmitted pathogen the basic reproduction number R_0, determines the final size of the epidemic, but the model makes well known unrealistic assumptions about mixing patterns and must be extended to account for heterogeneity and homophily (***Kermack and McKendrick, 1927***; ***Hébert-Dufresne et al., 2020***). Network models haven proven informative means to interrogate the effect of social structure on epidemic spread, yet by their nature they are demanding to parameterize (***Keeling and Eames, 2005***; ***Keeling, 2005***). Nonetheless, capturing these heterogeneous contacts is necessary to contextualize the rapid growth of an epidemic as a result of the social context where the outbreak arises, rather than specific properties of the pathogen. This sort of "founder effect" is familiar in population genetics, but less so in epidemiology (***Mayr, 1942***; ***Templeton, 2008***).

Here we present a simple extension of the classic SIR framework that specifically includes gathering sizes and homophily among individuals on the basis of the risk of infection. The resulting case curves are related to the gathering sizes and enable us to quantify how incorporating socialization alters our expectations for early spread of emerging viruses and variants and how this is reflected in seroprevalence and genomic surveillance.

Results

Transmission at gatherings: Group-SIR model

We propose modifying SIR models to account for varying contact rates and risk by explicitly modeling gathering dynamics using sampling processes. For any SIR-like model there are two key events, transmission and recovery, yielding infection dynamics:

$$\frac{dI}{dt} = \overbrace{B(S,I,R)}^{\text{transmission}} - \overbrace{\gamma I}^{\text{recovery}}$$

We focus on how gathering dynamics alter transmission, $B(S,I,R)$ and ignore the effects of different recovery rates by normalizing time by the mean infectious period, $\frac{1}{\gamma}$. The rate gatherings occur, r_{gather}, along with the size of the gatherings determine the social component of transmission that is controllable by non-pharmaceutical interventions. At a gathering, transmission can occur between each pair of the n attendees who were randomly sampled from the overall population. Figure 1a illustrates this sampling process for the classic SIR model where n=2 and Figure 1b illustrates this for

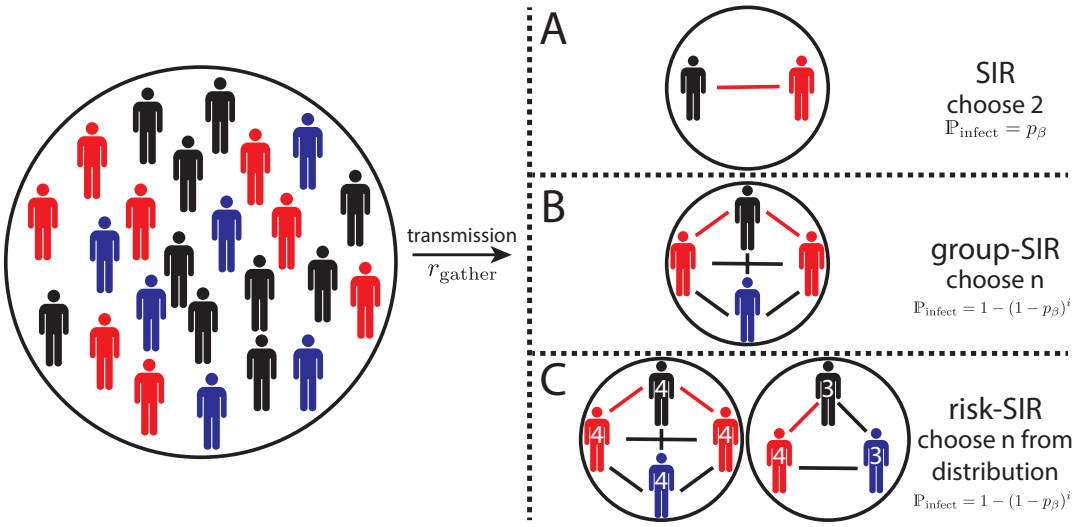


Figure 1. Schematic of sampling processes leading to different epidemic models described. A) The classic SIR model arises from sampling 2 individuals from a population of susceptible (black), infectious (red), and recovered (blue) individuals. B) Sampling n individuals yields a generalization of the SIR model, the group-SIR. Infections occur independently between each S-I pair present in a congregation. C) Denoting individuals by the maximum congregation size they are willing to attending yields the risk-SIR model. Societies can be defined by a distribution of congregation sizes which determine the epidemiological dynamics (***Boyer et al., 2022***).

the generalized “group-SIR" model. The expected number of infections at a gathering determines the overall transmission rate:

$$B_{\text{group}}(n,S,I,R) = r_{\text{gather}}\mathbb{E}[\Delta I] = r_{\text{gather}} \sum_{\{s,i,r\}} s\mathbb{P}_{\text{group}}(s,i,r|S,I,R)\mathbb{P}_{\text{infect}}(i)$$

with probabilities

$$\mathbb{P}_{\text{infect}}(i) = 1-(1-p_\beta)^i,$$
$$\mathbb{P}_{\text{group}}(s,i,r|S,I,R) = \frac{n!}{s!i!r!}\left(\frac{S}{N}\right)^s\left(\frac{I}{N}\right)^i\left(\frac{R}{N}\right)^r$$

where lower-case s, i, and r denote the integer number of sampled susceptible, infectious, and recovered individuals attending the gathering and uppercase S, I, R denote the population-level densities such that $S+I+R=N$. To average, we sum over all possible combinations of n SIR attendees sampled from the population densities according to a multinomial distribution, $\mathbb{P}_{\text{group}}$. A susceptible is infected at the gathering with probability, $\mathbb{P}_{\text{infect}}$ whenever any attending infectious individual transmits according to a pairwise probability of infection, p_β. The dynamics simplify to a classic SIR model when gatherings are restricted to pairs, n=2, whereas the transmission term becomes increasingly nonlinear for larger gatherings (see Appendix for details).

Increasing gathering sizes leads to sharper epidemics for otherwise fixed parameter values (Figure 2a). This is expected since increasing group sizes implicitly increases the total number of contacts relative to the recovery rate, i.e., increasing the basic reproduction number, R_0. Rescaling the transmission rate by the total number of pairwise interactions at a gathering, $\binom{n}{2}$, collapses the group-SIR dynamics onto each other (Figure 2a inset). There are fewer cases at the epidemic peak for larger gatherings as a result of multiple infectious individuals infecting the same susceptible, effectively reducing the number of pairwise interactions. In short, congregating in larger groups increases the rate of epidemic spread, but has minor impact on the dynamics when correcting for the number of interactions except by reducing the cumulative number of infections, in line with prior observations (***Volz et al., 2011***).

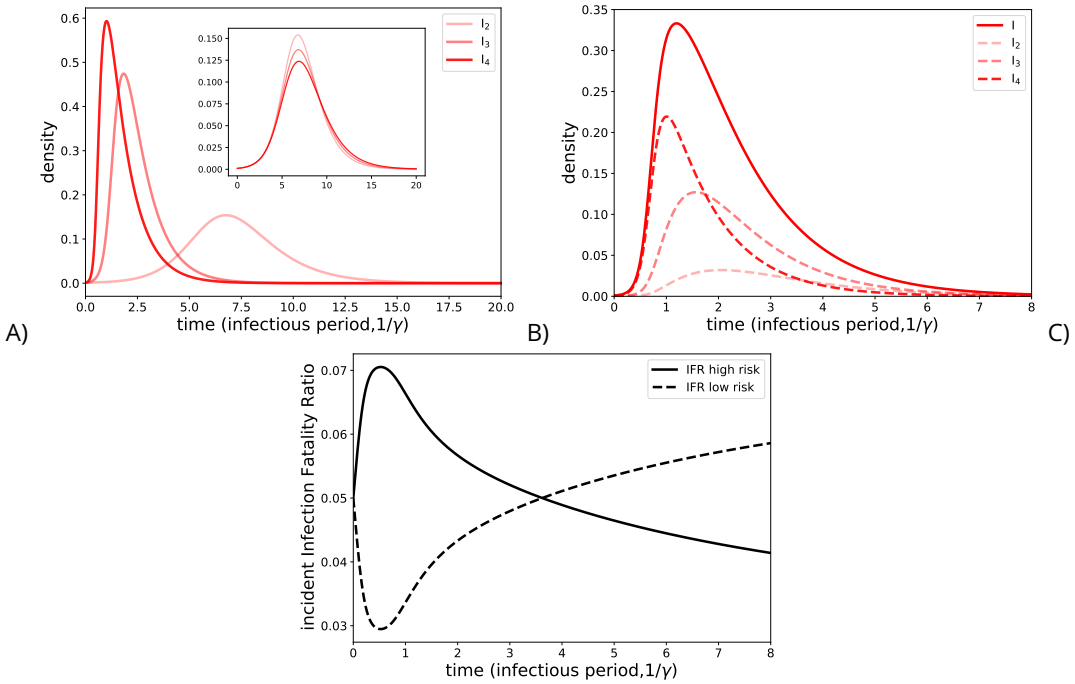


Figure 2. Generalized SIR dynamics. A) Group-SIR dynamics, where transmission occurs within groups with sizes here denoted by the subscript. B) Risk-SIR dynamics, where transmission occurs within groups according to a distribution here set with sizes n=2,3, and 4 at equal frequency. Subpopulations are denoted by their risk level shown, which specifies the maximum gathering size they are willing to attend. C) The sequential peaks of infection cause the overall infection fatality ratio (IFR) to be dynamic when it varies between subpopulations. Here we compare IFR dynamics when the IFR is positively correlated with risk level (IFR high risk) and when the IFR is negatively correlated with risk level (IFR low risk). Note, we assume that mortality dynamics are the same as recovery dynamics to demonstrate this qualitative point.

Gathering size dependent attendance: Risk-SIR model

During infectious disease outbreaks individuals may vary in how much they alter their behavior in response to perceived risk (***Perra et al., 2011***; ***Eksin et al., 2017***; ***Arthur et al., 2021***; ***Harris et al., 2023***). To capture this we can further generalize the group-SIR dynamics where individuals choose among differently sized, concurrent gatherings and avoid those larger than their "risk level":

$$\frac{dI_k}{dt} = B_{\text{risk}}(k) - \gamma I_k$$

where risk levels are denoted by subscripts. Transmission follows from repeatedly sampling from the population for a given distribution of differently sized gatherings, $\mathbb{P}_n$:

$$B_{\text{risk}}(k) = r_{\text{gather}} \sum_{n=2}^{k} \mathbb{P}_n B_{\text{group}}(n; S_k, \sum_{m \geq n} I_m, -S_k + \sum_{m \geq n} (R_m + S_m))$$

Each transmission term captures new infections within a particular focal subpopulation of individuals with the same risk level, k, such that only the respective susceptibles, S_k, contribute. These focal susceptibles can be infected when attending any gathering no larger than their risk level, $n \leq k$, by any attending infectious individual with risk levels at least as large as the respective gathering size, $m \geq n$. In contrast, infections among susceptibles from other subpopulations do not contribute to new focal infections and thus they function similarly to the recovered population. Overall, the transmission term is a weighted average over all gathering sizes according to a distribution, P_n, set by society (e.g., by lockdowns) (***Boyer et al., 2022***). The risk-SIR model includes transmission heterogeneity as subpopulations with higher risk level attend a larger fraction of gatherings, homophily as subpopulations with higher risk levels preferentially interact at larger gatherings, and super-spreading as multiple infections can occur at large gatherings.

The risk-SIR yields single-peaked epidemics like the classical SIR, but with underlying sequential peaks among decreasing risk level subpopulations (Figure 2b). Differences in growth rates cause these different peaks, namely that susceptibles in higher risk subpopulations attend larger gatherings on top of attending the same gatherings at the same rate as those in lower risk subpopulations. Similarly, infections in higher risk level subpopulations lead to more secondary infections. The basic reproduction number of an infectious individual with a specified risk level, k, follows a recurrence relation:

$$\begin{aligned} R_0(k) &= R_0(k-1) + \frac{r_{\text{gather}}}{\gamma} \mathbb{P}_k p_\beta (k-1) \left(1 - \left(1 - \frac{1}{\sum_{m \geq k} N_m}\right)^k\right) \\ &\approx R_0(k-1) + \frac{r_{\text{gather}}}{\gamma} \mathbb{P}_k p_\beta \frac{k(k-1)}{\sum_{m \geq k} N_m} \end{aligned}$$

such that a disproportionate number of secondary infections occur at larger gatherings as enumerated by k (see Appendix for derivation). The largest fraction of the secondary infections occur within the same risk subpopulation or in larger risk subpopulations as these susceptibles can attend all the same gathering as those infectious, whereas susceptibles in smaller risk subpopulation do not attend the larger gatherings. Together, these dynamics cause the initial growth of the epidemic to be driven by the largest risk subpopulations, in line with prior work (***May and Anderson, 1987***).

Impact on Surveillance

The sequential epidemic peaks within risk subpopulations have multiple impacts for disease surveillance. First, the cases initially accumulate within subpopulations at most risk, highlighting the importance of caution when interpreting and designing serosurveys to account for differential risk of exposure. Any underlying surveillance metrics that vary between subpopulations will now be dynamic. For example, we use infection fatality rates (IFR) to predict the threat of an emerging outbreak. If virulence depends on comorbidities that vary in frequency between each subpopulation

then the IFR will change over the successive peaks of the underlying outbreaks in different subpopulations. Figure 2c illustrates IFR dynamics for two general cases. First, when individuals with comorbidities choose to avoid larger gatherings, i.e., subpopulation IFR negatively correlates with risk level (IFR low risk), then the IFR increases over time because at risk populations are infected relatively more frequently at later times. This would also reduce case ascertainment at the outset of an epidemic, potentially biasing for lower estimates of R_0. Conversely, if individuals with comorbidities attend larger gatherings, i.e., subpopulation IFR positively correlates with risk level, then the IFR decreases over time (IFR high risk). This latter example is particular relevant for essential workers during lockdowns and communicable childhood diseases within school settings. IFR dynamics emphasize the danger in reporting aggregate statistics without accounting for disproportionate disease spread among subpopulations, such as when tracking emerging variants.

An epidemic grows at a rate determined by the socialization of all infectious individuals, i.e., it's risk distribution. Meanwhile, a variant emerges within a subset of these infectious individuals. Thus, even a neutral variant with no increased transmissibility relative to the background it emerges in is expected to increase in frequency if it is transmitted by individuals more social than average across the epidemic. Figure 3a shows the risk distribution of an epidemic and a neutral variant emerging within the most social subpopulation at a given time. The resulting variant frequency dynamics, shown in Figure 3b, mimic logistic growth–approaching an equilibrium frequency prior to sweeping. This dynamic emphasizes the role that social context plays in determining the observed growth rate of a variant. For example, a variant of concern with increased transmissibility relative to wild type will ultimately grow in frequency, but if it emerges within a less social subpopulation then the observed increase in frequency will be delayed (figure 3c) increasing its chance to stochastically become extinct. The same plot with a logarithmic y-axis shows a clear decreases in frequency of the more transmissible variants that emerge in less social subpopulations, which increases the chance for stochastic extinction (See Appendix figure 1).

Since both increased transmissibility and increased contact opportunities contribute to the apparent fitness of a variant, identifying fast growing lineages during genomic surveillance imposes a selection bias towards identifying lineages preferentially spreading among more social individuals. To see this, we simulated the individual transmission and recovery events of a risk-SIR model using the Gillespie algorithm and tracked all possible lineages (***Gillespie, 1977***). Figure 4a shows the frequency dynamics of the fastest growing lineage (having reached a given size at a given time - dotted vertical line). This lineage is neutral but appears to be growing more rapidly than its peers, and continues to increase in frequency after identification. This occurs because the risk distribution of the variant at the point of sampling, shown in Figure 4a (inset), contains relatively more social individuals relative to the rest of the epidemic and, in turn, an expected increase in frequency relative to the epidemic. Note, here and throughout we follow common usage by epidemiologists to refer to pathogen clades as lineages, i.e., all ancestors of a focal case.

Whether identifying fastest growing lineage biases for those with spreading among more risky subpopulations depends on censusing parameters. This dependence arises because selecting fast-growing lineages biases the risk distributions by two opposing processes. On one hand fast-growing lineages are more likely to spread among the riskiest subpopulations because they attend more and larger gatherings and subsequently more frequent secondary infections. On the other hand, the fastest growing lineages preferentially include recent infections, which bias towards less risky populations given the sequential spread of the epidemic from more risky to less risky subpopulations over time. Figure 4b shows the the median risk distribution bias for different censusing parameters across 250 repeated simulations of an epidemic spreading within a society with high variance in gathering rates (see caption). The x-axis specifies the minimum frequency within a 100% bandwidth that the fastest growing lineage will be tracked, e.g., 0.01 refers to the fastest growing lineage among those between .01 and .02 frequency. Fast growing lineages tend to have biased risk distribution particularly early on during epidemics, in line with greater stochastic deviations of smaller lineages for a given frequency. Figure 4c shows whether the identified lineages tend to

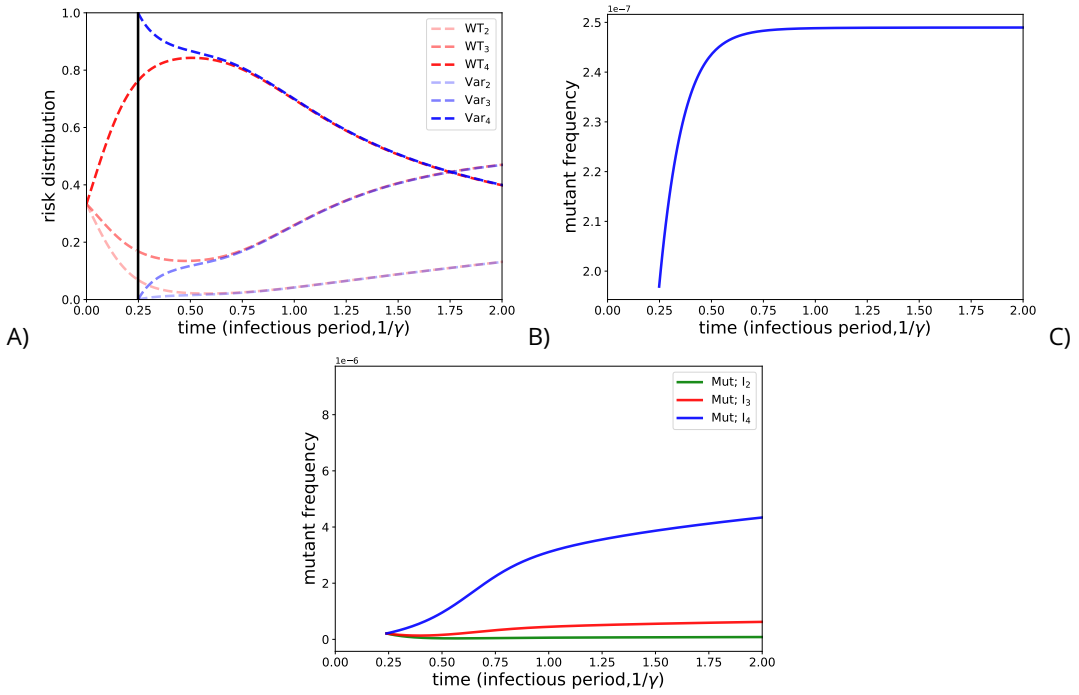


Figure 3. Variant frequency dynamics depend on which risk subpopulation it emerges within. A) Risk distribution of an epidemic (red) and a neutral variant (blue) that emerges within the highest risk subpopulation. B) The associated frequency dynamics of the emerging neutral variant. C) Frequency dynamics of a variant with 50% greater transmissibility relative to the background epidemic emerging in different risk subpopulations

grow over time due to the bias for different censusing parameters. We see that the long-term effect of the bias is larger for moderate frequencies (.04 vs .01), due to the fact that larger lineages for a given frequency take longer to deterministically relax their risk distribution to that of the entire epidemic. Note, given the impact of this bias is larger for moderately sized fast-growing lineages means we are susceptible to identifying these lineages during surveillance. For example, the CDC tracks lineages $> .01$ frequency. Our above results suggest these lineages are susceptible to being misidentified as having increased transmissibility and that their continued transient increase in frequency following censusing would incorrectly solidify these concerns. By understanding how social structure modulates these biases, we can better establish from which communities we expect to identify these spurious variants of concern.

Society modulated surveillance bias

Who interacts with whom in a society affects the magnitude of the above identified surveillance biases. For example, the amount of homophily, i.e., assortative mixing, within a community sets the time scale over which such lineages will continue to spuriously increase in frequency. Intuitively, homophily increases the probability a skewed risk distribution remains skewed as onward infections primarily occur within riskier subpopulation. As the risk-SIR model has a fixed social structure and homophily, we can better understand the relationship between homophily and skewed risk distribution retention by analyzing a simplified SIR type model that disentangles transmission heterogeneity and homophily:

$$\frac{dI_H}{dt} = S_H(\beta_{HH}I_H + \beta_{HL}I_L) - \gamma I_H$$
$$\frac{dI_L}{dt} = S_H(\beta_{HL}I_H + \beta_{LL}I_L) - \gamma I_L$$

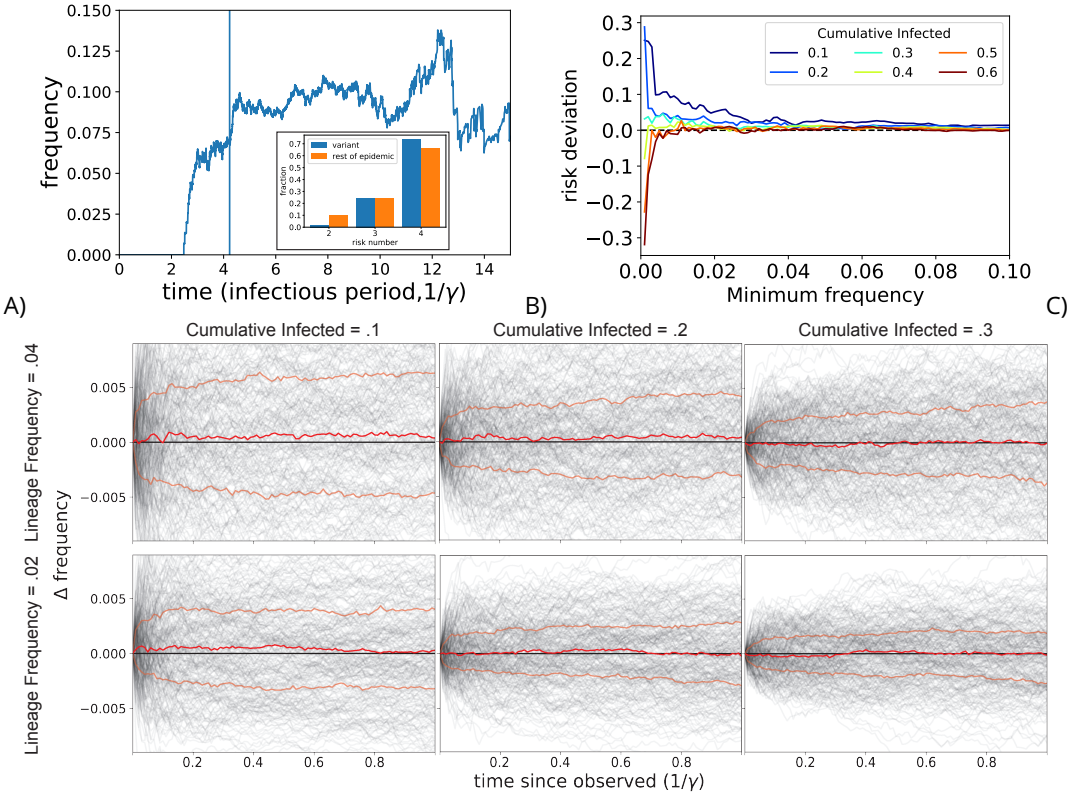


Figure 4. Selection bias of fast growing clades a) An example of a fast growing clade that continues to grow beyond the point it was identified (blue vertical line) and its risk distribution among infectious individuals as compared to all others infectious at the moment of censusing (inset). b) Median difference between the median of the fastest growing risk distribution and the risk distribution of the rest of the epidemic for given censusing frequency within a 100% bandwidth for epidemics with subpopulations $[5000, 0, 5000, 0, 5000]$ for risk numbers 2-6. c) Relative frequency dynamics of the fastest growing lineages following censusing across different censusing parameters for epidemics with subpopulations $[5000, 0, 5000, 0, 5000]$ for risk numbers 2-6. Red lines denote median and interquartile ranges across 250 simulations shown in transparent black lines.

where H denotes highly social individuals and L denotes less social individuals. This is a minimal model including transmission heterogeneity when the subpopulations differ in their basic reproduction number, i.e., $R_{0,H} = \frac{\beta_{HH}S_H+\beta_{LH}S_L}{\gamma} > \frac{\beta_{LL}S_L+\beta_{LH}S_H}{\gamma} = R_{0,L}$, while smaller values of cross-transmission, β_{HL}, correspond to societies with greater homophily. For this model, the risk distribution simplifies to the relative fraction of the more social subpopulation $f_H = \frac{I_H}{I_H+I_L}$. We can solve for the equilibrium risk distribution, f_H^*, early on in an epidemic during exponential growth by fixing the susceptibles (S_H and S_L) to constants and solving the respective dynamics, $\frac{df_H}{dt} = 0$. The equilibrium risk distribution depends on the structure of society, with increasing homophily reflected as a higher proportion of highly social individuals (see Appendix figure 2).

Randomness in individual transmission and recovery events (i.e., demographic stochasticity) causes the risk distribution to drift, even when at equilibrium. We can solve for the dynamics of the probability distribution over risk distributions as induced by demographic stochasticity via a so-called "master equation" parametrized by the average rates of the individual events (***van Kampen, 2007***). Figure 5a shows probability distributions after 500 population changes (transmission or recovery) for epidemics varying in homophily and initiated with 20 infectious individuals distributed according to their respective equilibrium risk distribution. Note, we vary homophily alone and keep the same levels of transmission heterogeneity across models by fixing the subpopulation basic reproduction numbers, $R_{0,H}$ and $R_{0,L}$. Demographic stochasticity induces a spread that eventually dissipates over time (see Appeneix figure 3). This means emerging variants are likely to deviate further from the equilibrium growth rate than the rest of the epidemic. This motivates quantifying how often a variant will grow faster versus slower than the equilibrium growth rate, $\rho = \frac{\mathbb{P}(f_H > f_H^*)}{\mathbb{P}(f_H > f_H^*)+\mathbb{P}(f_H < f_H^*)}$. Figure 5b compares the relative probability of increased sociality, ρ, dynamics after initiating infectious populations at their respective equilibrium value, f_H^*, for different amounts of homophily. Emerging variants in societies with increased homophily have larger bias towards increased growth rates relative to the overall epidemic. Figure 5c shows the bias in risk distribution for epidemics initiating either in the more social (I_H, dashed lines) or the less social (I_L, solid lines) subpopulations. The subpopulation in which a variant emerges sets an initial risk distribution bias and societies with increased homophily retain this bias longer. However, the risk distribution bias ultimately approaches the same biased quasi-equilibrium value as seen above when started at the equilibrium.

Preferential stochastic extinction of slower growing epidemics exacerbates the growth bias induced by homophily. An epidemic goes stochastically extinct when its last remaining infectious individual recovers prior to onwards transmission. Intuitively, epidemics are more likely to go stochastically extinct when the last remaining infectious individual is among a less social subpopulation. This further increases the growth bias beyond that induced by homophily alone (see Appendix Figure 4). Hence, we expect emerging variants to increase in frequency by preferentially spreading among more social individuals, particularly when they emerge in societies with high levels of homophily.

Discussion

The dynamics of communicable diseases are intimately linked to the contacts that are made between their hosts (***Buckee et al., 2021***). These are the opportunities to transmit and initiate new infections, and must be accounted for if we wish to interpret recent data during the course of an outbreak or forecast its course. Here we have presented a simple and flexible means to incorporate network structure implicitly into the classic SIR-framework and relate it to varying risks of exposure. By allowing interactions within groups of size > 2, representing varying risk-groups as a result of choice or necessity, the dynamics can no longer be described simply as the product of the proportions of the population that are susceptible or infectious, allowing us to examine the dynamics in separate risk groups and the impact on individual chains of transmission within them.

Prior studies on how contact heterogeneity impacts disease spread has been a major contrib-

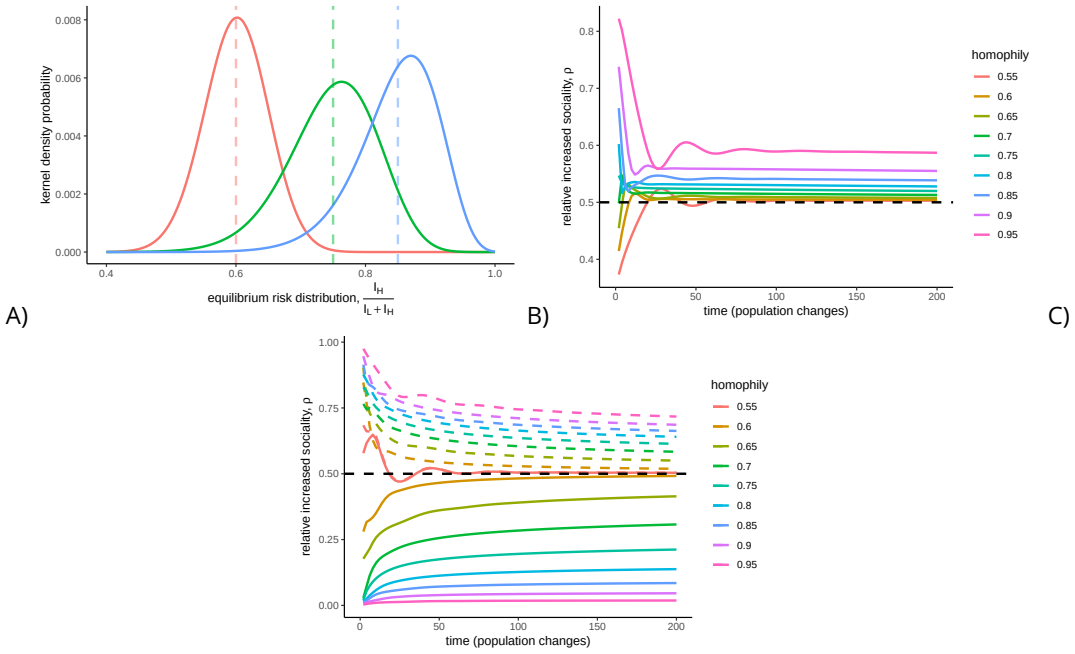


Figure 5. Homophily biases risk distributions a) Smoothed probability distributions of the risk distribution of epidemics initiated with 20 individuals at the respective equilibrium distribution after 500 population changes (total recovery or transmission events). Vertical dashed lines specify respective equilibrium risk distributions. B) Relative probability of a more social risk distribution ($f_H > f_H^*$) than a less social risk distribution ($f_H < f_H^*$) for epidemics initiated with 20 infectious individuals distributed according to the equilibrium risk distribution (f_H^*) denoted as homophily. C) Relative probability of a more social risk distribution ($f_H > f_H^*$) than a less social risk distribution ($f_H < f_H^*$) for epidemics initiated with 1 infectious individuals either in the more social subpopulation (dashed lines) or the less social subpopulation (solid lines).

utor to the epidemiology of HIV-1 among at-risk populations, and more recently Mpox transmission along dense contact networks among men who have sex with men (***May and Anderson, 1987***; ***Endo et al., 2022***). This includes sequential spread to populations with low contact rates from 'core groups' characterized by higher numbers of contacts (***Yorke et al., 1978***; ***Colgate et al., 1989***). Our work explicitly includes gatherings as a mechanism for the separation of and homophily between high and low contact subpopulations. This context is more relevant for pathogens with shorter latency time than HIV including respiratory viruses like SARS-CoV-2 in which super-spreading (i.e., overdispersed transmission) plays a large role in the observed dynamics – especially at the outset as the pathogen is introduced and becomes established. By emphasizing the importance of gatherings as the locus of spread, our model is aligned with interventions that utilize retrospective contact tracing to enable epidemiologists to monitor key locations where transmission occurs and link these data to population-level predictions (***Boyer et al., 2022***).

In an unmitigated outbreak, the only factor reducing the rate of new infections is convalescent immunity, and it has been recognized for some time that variable contact rates can have a large impact on the point at which this begins to take effect (***Hébert-Dufresne et al., 2020***). This is important for vaccination strategies aimed at those most likely to become infected and transmit, as it can markedly reduce the 'herd immunity threshold' necessary to achieve control and prevent large outbreaks (***Woolhouse et al., 1997***). Our model readily captures this and shows moreover that a single outbreak is itself made up of multiple ones with different dynamics in distinct but overlapping networks with distinct but overlapping epidemic curves. Those with the most contacts are rapidly infected, leading to the counterintuitive result that those who make the fewest contacts and/or have the fewest risks of exposure are more likely to be infected at times when the force of infection is lower, particularly late in the epidemic. This finding is important when interpreting the relationship between crude case counts and severe outcomes, where the latter vary markedly among individuals. It is illustrated by the use of hospitalizations and deaths as a lagging indicator of cases early in the COVID-19 pandemic, and likely contributes to the much smaller variance in numbers of severe outcomes over time as compared with estimates of case counts or wastewater data (***Khan et al., 2023***). Intuitively, those who are most vulnerable are least likely to be infected early on, but they and their networks will remain capable of supporting a sustained outbreak over a longer period of the epidemic.

This can also produce substantial bias in the apparent value of key parameters that are important for forecasting and estimating the impacts of an outbreak. If there is any correlation between the risk groups and the probability of severe illness, quantities such as the infection fatality rate or infection hospitalization rate will be biased. And this bias is in addition to those already known to arise through under-ascertainment of mild infections (***Accorsi et al., 2021***).

The model also suggests caution when interpreting and estimating the fitness of emerging lineages such as putative 'variants of concern'. The founder effect in population genetics refers to a spurious appearance of fitness as a result of stochastic factors, as when a small number of migrating individuals colonizes a deserted island (***Wright, 1942***). Here, it indicates that early on in an epidemic the initial growth rate of any lineage that becomes common enough to be noticed is expected to be biased upward in relation to its true fitness, simply because it is spreading by definition in the networks which make the most contacts. Conversely, as the overall epidemic is declining those lineages that are persisting are infecting a higher proportion of fewer contacts, and so past the peak increases in proportion of a sample becomes a more reliable indicator of fitness.

Because we have developed an extension of the SIR model, we have not explicitly considered immune evasion. However it can readily be seen that the host population in which a pathogen with immune evasion is most likely to experience both high fitness and start spreading is that which makes the most contacts and within which seroprevalence is highest (***Bushman et al., 2021***). Although this may be complicated by the timescale of waning immunity this is beyond the scope of the present work.

The SIR model has been a valuable source of intuition for generations of epidemiologists but is

known to be limited. Here we have shown the value of a small change to incorporate larger group sizes than pairwise interactions. The results help interpret the expected course of epidemics where socialization and awareness of the epidemic drive transmission, such as during the recent Mpox outbreak (***Endo et al., 2022***). Furthermore, our model highlights how varying exposure leads to under-appreciated sources of bias when tracking putative variants of concern. Overall, the simplicity of the model may allow it to be readily applied to more complicated scenarios in which risk perception varies over time and the relative compositions of risk groups change accordingly.

Code availability

All code used to generate the figures is available at: https://github.com/bradfordptaylor/risk_sir.

Acknowledgements

This project has been funded (in part) by contract 200-2016-91779 with the Centers for Disease Control and Prevention and by contract U01 CA261277 through SeroNet. Disclaimer: The findings, conclusions, and views expressed are those of the author(s) and do not necessarily represent the official position of the Centers for Disease Control and Prevention (CDC).

References

Accorsi EK, Qiu X, Rumpler E, Kennedy-Shaffer L, Kahn R, Joshi K, Goldstein E, Stensrud MJ, Niehus R, Cevik M, Lipsitch M. How to detect and reduce potential sources of biases in studies of SARS-CoV-2 and COVID-19. European Journal of Epidemiology. 2021 2; 36(2):179–196.

Altizer S, Nunn CL, Thrall PH, Gittleman JL, Antonovics J, Cunningham AA, Dobson AP, Ezenwa V, Jones KE, Pedersen AB, Poss M, Pulliam JRC. Social Organization and Parasite Risk in Mammals: Integrating Theory and Empirical Studies. Annual Review of Ecology, Evolution, and Systematics. 2003 11; 34(1):517–547.

Arthur RF, Jones JH, Bonds MH, Ram Y, Feldman MW. Adaptive social contact rates induce complex dynamics during epidemics. PLOS Computational Biology. 2021 feb 10; 17(2):e1008639.

Bansal S, Read J, Pourbohloul B, Meyers LA. The dynamic nature of contact networks in infectious disease epidemiology. Journal of Biological Dynamics. 2010 9; 4(5):478–489.

Boyer CB, Rumpler E, Kissler SM, Lipsitch M. Infectious disease dynamics and restrictions on social gathering size. Epidemics. 2022 9; 40:100620.

Buckee C, Noor A, Sattenspiel L. Thinking clearly about social aspects of infectious disease transmission. Nature. 2021 jun 30; 595(7866):205–213.

Bushman M, Kahn R, Taylor BP, Lipsitch M, Hanage WP. Population impact of SARS-CoV-2 variants with enhanced transmissibility and/or partial immune escape. Cell. 2021 12; 184(26):6229–6242.e18.

Chande A, Lee S, Harris M, Nguyen Q, Beckett SJ, Hilley T, Andris C, Weitz JS. Real-time, interactive website for US-county-level COVID-19 event risk assessment. Nature Human Behaviour. 2020 nov 9; 4(12):1313–1319.

Colgate SA, Stanley EA, Hyman JM, Layne SP, Qualls C. Risk behavior-based model of the cubic growth of acquired immunodeficiency syndrome in the United States. Proceedings of the National Academy of Sciences. 1989; 86(12):4793–4797.

Edwards DA, Ausiello D, Salzman J, Devlin T, Langer R, Beddingfield BJ, Fears AC, Doyle-Meyers LA, Redmann RK, Killeen SZ, Maness NJ, Roy CJ. Exhaled aerosol increases with COVID-19 infection, age, and obesity. Proceedings of the National Academy of Sciences. 2021 feb 9; 118(8).

Eksin C, Shamma JS, Weitz JS. Disease dynamics in a stochastic network game: a little empathy goes a long way in averting outbreaks. Scientific Reports. 2017 mar 14; 7(1).

Endo A, Murayama H, Abbott S, Ratnayake R, Pearson CAB, Edmunds WJ, Fearon E, Funk S. Heavy-tailed sexual contact networks and monkeypox epidemiology in the global outbreak, 2022. Science. 2022 oct 7; 378(6615):90–94.

Fraser C, Hollingsworth TD, Chapman R, de Wolf F, Hanage WP. Variation in HIV-1 set-point viral load: Epidemiological analysis and an evolutionary hypothesis. Proceedings of the National Academy of Sciences. 2007 oct 30; 104(44):17441–17446.

Gillespie DT. Exact stochastic simulation of coupled chemical reactions. The Journal of Physical Chemistry. 1977 12; 81(25):2340–2361.

Godfrey SS, Bull CM, James R, Murray K. Network structure and parasite transmission in a group living lizard, the gidgee skink, Egernia stokesii. Behavioral Ecology and Sociobiology. 2009 apr 2; 63(7):1045–1056.

Gopinath S, Carden S, Monack D. Shedding light on Salmonella carriers. Trends in Microbiology. 2012 7; 20(7):320–327.

Harris MJ, Cardenas KJ, Mordecai EA. Social divisions and risk perception drive divergent epidemics and large later waves. Evolutionary Human Sciences. 2023; 5.

Hébert-Dufresne L, Althouse BM, Scarpino SV, Allard A. Beyond R_0: heterogeneity in secondary infections and probabilistic epidemic forecasting. Journal of The Royal Society Interface. 2020 11; 17(172):20200393.

van Kampen NG. Stochastic Processes in Physics and Chemistry (Third Edition). Elsevier; 2007.

Keeling M. The implications of network structure for epidemic dynamics. Theoretical Population Biology. 2005 2; 67(1):1–8.

Keeling MJ, Eames KTD. Networks and epidemic models. Journal of The Royal Society Interface. 2005 jun 20; 2(4):295–307.

Kermack WO, McKendrick AG. 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. 1927 8; 115(772):700–721.

Khan D, Park M, Burkholder J, Dumbuya S, Ritchey MD, Yoon P, Galante A, Duva JL, Freeman J, Duck W, Soroka S, Bottichio L, Wellman M, Lerma S, Lyons BC, Dee D, Haile S, Gaughan DM, Langer A, Gundlapalli AV, et al. Tracking COVID-19 in the United States With Surveillance of Aggregate Cases and Deaths. Public Health Reports. 2023 mar 24; 138(3):428–437.

Lawley TD, Bouley DM, Hoy YE, Gerke C, Relman DA, Monack DM. Host Transmission of *Salmonella enterica* Serovar Typhimurium Is Controlled by Virulence Factors and Indigenous Intestinal Microbiota. Infection and Immunity. 2008 1; 76(1):403–416.

Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature. 2005; 438(7066):355–359.

Matthews L, McKendrick IJ, Ternent H, Gunn GJ, Synge B, Woolhouse MEJ. Super-shedding cattle and the transmission dynamics of*Escherichia coli*O157. Epidemiology and Infection. 2005 jun 3; 134(1):131–142.

May RM, Anderson RM. Transmission dynamics of HIV infection. Nature. 1987 mar 1; 326(6109):137–142.

Mayr E. Systematics and the Origin of Species. New York: Columbia University Press.; 1942.

McPherson M, Smith-Lovin L, Cook JM. Birds of a Feather: Homophily in Social Networks. Annual Review of Sociology. 2001 8; 27(1):415–444.

Newman MEJ. Assortative Mixing in Networks. Physical Review Letters. 2002 oct 28; 89(20).

Perra N, Balcan D, Gonçalves B, Vespignani A. Towards a Characterization of Behavior-Disease Models. PLoS ONE. 2011 aug 3; 6(8):e23084.

Quinn TC, Wawer MJ, Sewankambo N, Serwadda D, Li C, Wabwire-Mangen F, Meehan MO, Lutalo T, Gray RH. Viral Load and Heterosexual Transmission of Human Immunodeficiency Virus Type 1. New England Journal of Medicine. 2000 mar 30; 342(13):921–929.

Rohani P, Zhong X, King AA. Contact Network Structure Explains the Changing Epidemiology of Pertussis. Science. 2010 nov 12; 330(6006):982–985.

Stein RA. Super-spreaders in infectious diseases. International Journal of Infectious Diseases. 2011 8; 15(8):e510–e513.

Stigum H, Falck W, Magnus P. The core group revisited: The effect of partner mixing and migration on the spread of gonorrhea, chlamydia, and HIV. Mathematical Biosciences. 1994 3; 120(1):1–23.

Taylor CA, Boulos C, Almond D. Livestock plants and COVID-19 transmission. Proceedings of the National Academy of Sciences. 2020 nov 19; 117(50):31706–31715.

Templeton AR. The reality and importance of founder speciation in evolution. BioEssays. 2008; 30(5):470–479.

Volz E, Meyers LA. Susceptible–infected–recovered epidemics in dynamic contact networks. Proceedings of the Royal Society B: Biological Sciences. 2007 sep 18; 274(1628):2925–2934.

Volz EM, Miller JC, Galvani A, Ancel Meyers L. Effects of Heterogeneous and Clustered Contact Patterns on Infectious Disease Dynamics. PLoS Computational Biology. 2011 jun 2; 7(6):e1002042.

Woolhouse ME, Dye C, Etard JF, Smith T, Charlwood J, Garnett G, Hagan P, Hii J, Ndhlovu P, Quinnell R, et al. Heterogeneities in the transmission of infectious agents: implications for the design of control programs. Proceedings of the National Academy of Sciences. 1997; 94(1):338–342.

Wright S. Statistical genetics and evolution. Bulletin of the American Mathematical Society. 1942 April; 48(4):223–246.

Yorke JA, Hethcote HW, Nold A. Dynamics and Control of the Transmission of Gonorrhea. Sexually Transmitted Diseases. 1978 4; 5(2):51–56.

Appendix 1

$n = 2$ and $n = 3$ group-SIR transmission terms

Here, we expand the group-SIR transmission term to connect it to the classic SIR model and to show how transmission grows nonlinearly with increasing group size. The sampling process yields transmission rates proportional to the expected number of infections at gathering with a given size n:

$$B_{\text{group}}(n, S, I, R) \propto \mathbb{E}_{\Delta I}(n) = \sum_{\{s,i,r\}} s\mathbb{P}_{\text{group}}(s, i, r|S, I, R)\mathbb{P}_{\text{infect}}(i)$$

with probabilities

$$\mathbb{P}_{\text{infect}}(i) = 1 - (1 - p_\beta)^i,$$
$$\mathbb{P}_{\text{group}}(s, i, r|S, I, R) = \frac{n!}{s!i!r!}\left(\frac{S}{N}\right)^s\left(\frac{I}{N}\right)^i\left(\frac{R}{N}\right)^r$$

Transmission only occurs at gatherings when at least one infected and at least one susceptible attend. Hence, only one combination of attendees contributes in the $n = 2$ case:

$$B_{\text{group}}(2, S, I, R) \propto 1 \times \mathbb{P}(s = 1, i = 1, r = 0|S, I, R)\left(1 - (1 - p_\beta)^1\right) = 2p_\beta\frac{S}{N}\frac{I}{N}$$

which is proportional to the bilinear classic SIR model. Note, the emphasis on proportionality, as this from differs from standard presentation of the SIR transmission by a constant factor of $\frac{1}{N}$ which can be accounted for by the rate of gathering r_{gather}. The $n = 3$ case includes transmission with more combinations of s,i, and r attendees:

$$\begin{aligned}B_{\text{group}}(3, S, I, R) \propto \left(1 - (1 - p_\beta)^1\right)&[\mathbb{P}(s = 1, i = 1, r = 1|S, I, R) + 2\mathbb{P}(s = 2, i = 1, r = 0|S, I, R)] + \\ &\left(1 - (1 - p_\beta)^2\right)\mathbb{P}(s = 1, i = 2, r = 0|S, I, R) = \\ &p_\beta\left[6\left(\frac{S}{N}\right)\left(\frac{I}{N}\right)\left(\frac{R}{N}\right) + 2 \times 3\left(\frac{S}{N}\right)^2\left(\frac{I}{N}\right)\right] + \\ &p_\beta(2 + p_\beta)\left[3\left(\frac{S}{N}\right)\left(\frac{I}{N}\right)^2\right]\end{aligned}$$

which deviates from a classical SIR model due to higher order nonlinearity.

risk-SIR R_0 derivation

The basic reproduction number, R_0, is defined as the average number of secondary infections transmitted by an index case in an otherwise uninfected population. This can be calculated from the dynamics of gathering and the expected number of infections given the combinatorics of attendance. Because individuals with different risk levels attend different gatherings, we calculate the basic reproduction number for each risk subpopulation. Three terms contribute to $R_0(k)$: (1) the total number of gatherings at given sizes that an individual with risk level k can attend while infectious, (2) the probability that the individual attends each gathering, and (3) the average number of infections at each gathering given the infectious individual attends. The average number of gatherings for a given size $n \leq k$ that occur during the average infectious period is:

$$N_{\text{gather}}(n) = \frac{r_{\text{gather}}}{\gamma}\mathbb{P}_n$$

where r_{gather} is the rate all gatherings occur, γ is the recovery rate, and $\mathbb{P}_n$ is the probability a gathering is size n. The probability an individual with risk level $\geq n$ attends a gathering of size n is:

$$\mathbb{P}_{\text{attend}}(n) = 1 - \left(1 - \prod_{j=1}^{n} \frac{\sum_{m \geq n} N_m - j - 1 + 1}{\sum_{m \geq n} N_m - j + 1}\right)$$

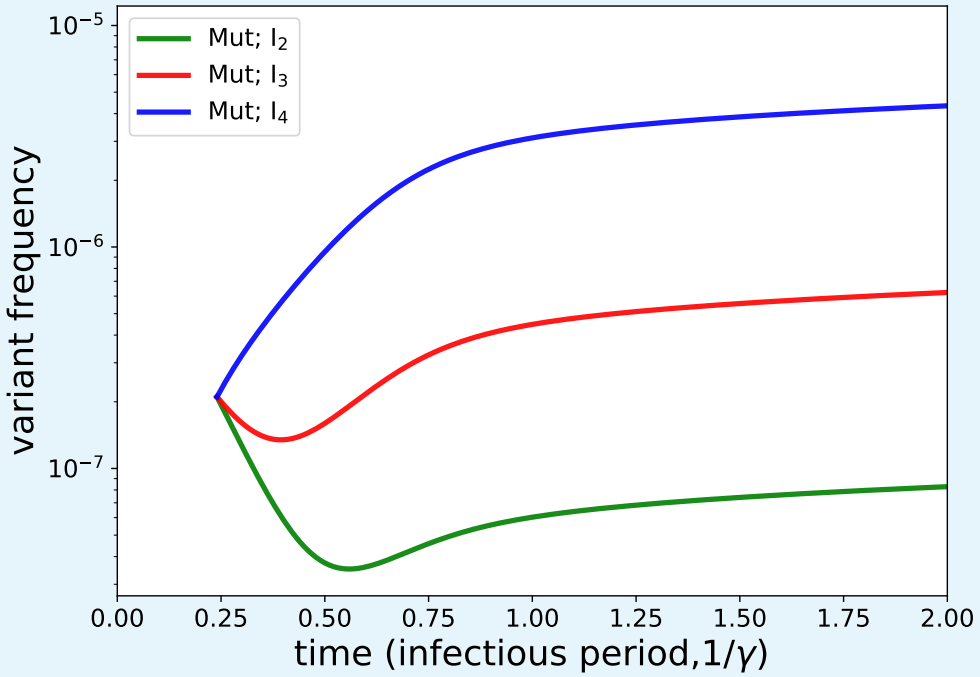
which is the complement of the probability an infectious individual not being sampled as any of the n attendees from the total individuals with sufficiently high risk levels, N_m. The expected number of infections at a gathering of size n when the infectious individual attends is simply:

$$\mathbb{E}_{\Delta I}(n) = (n-1)p_\beta$$

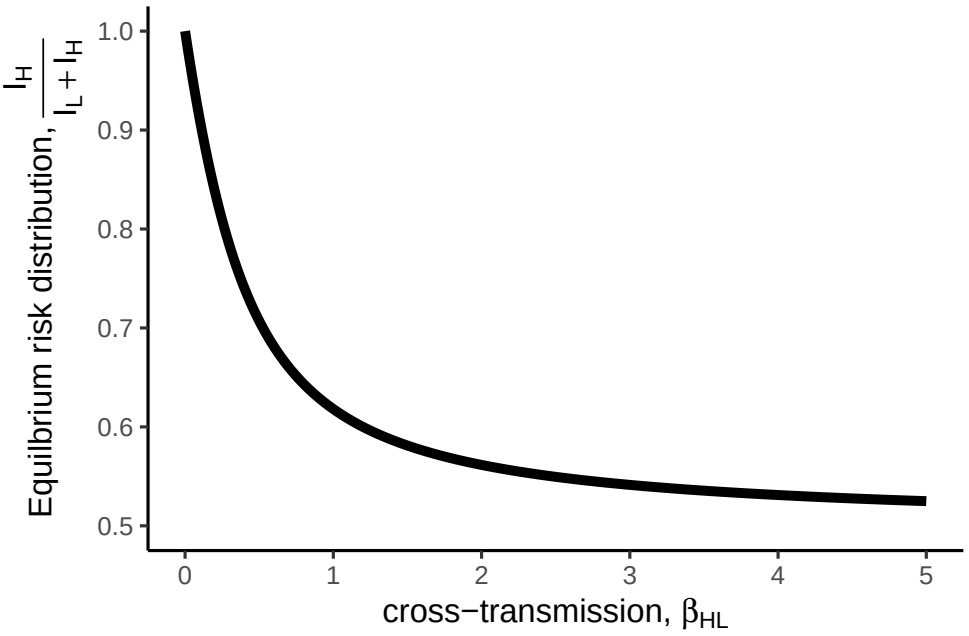
where p_beta is the pairwise infection probability. We combine these terms, sum over all possible gathering sizes that the infectious individual with risk level k is willing to attend, and simplify $\mathbb{P}_{\text{attend}}$ in the limit of infinite population sizes, $N \to \infty$, to get the final result:

$$\text{risk: } R_0(k) = \sum_{n=2}^{k} N_{\text{gather}}(n)\mathbb{E}_{\Delta I}(n)\mathbb{P}_{\text{attend}}(n) = \frac{r_{\text{gather}}}{\gamma} \sum_{n=2}^{k} \mathbb{P}_n \frac{p_\beta (n-1) n}{\sum_{m \geq n} N_m}$$

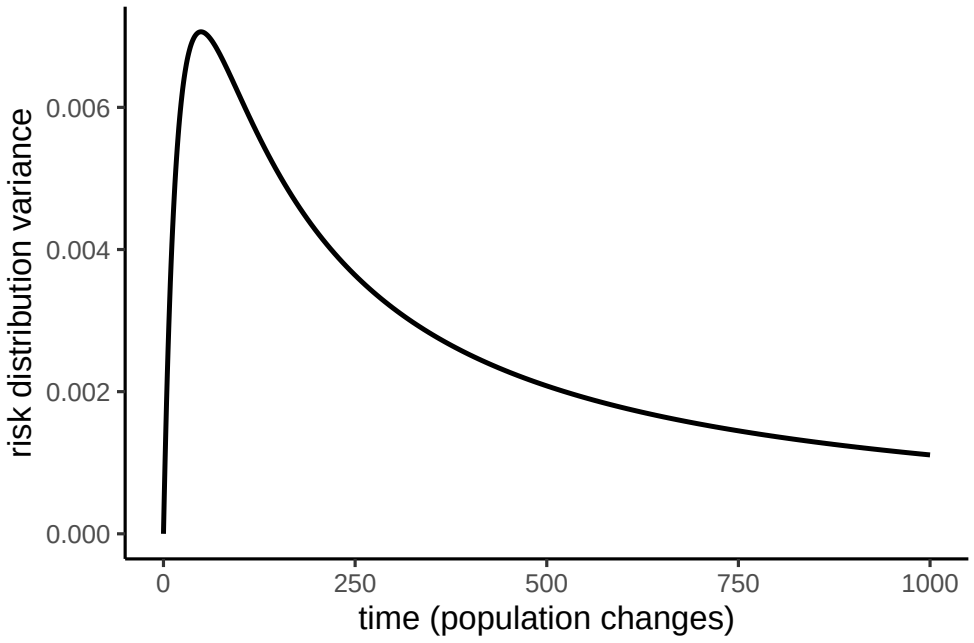
We show this result as a recurrence relation in the text to highlight how it scales with k.



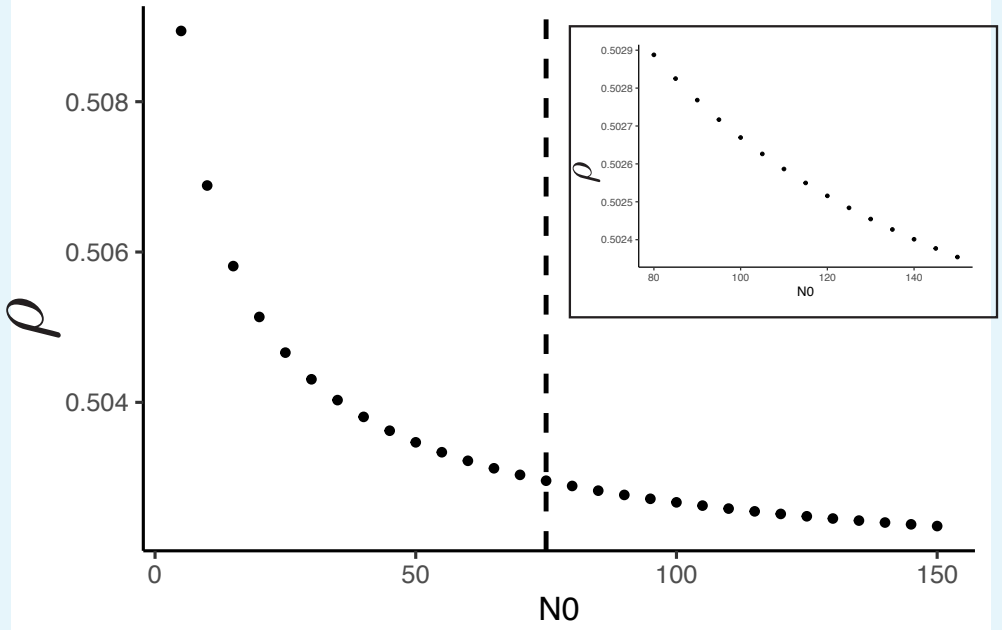
Appendix 1—figure 1. Dynamics of a variant with increased transmissibility relative to wildtype emerging in different risk subpopulations.



Appendix 1—figure 2. Equilibrium risk distribution for exponentially growing epidemics in societies with transmission heterogeneity as the amount of cross transmission between two subpopulation varies.



Appendix 1—figure 3. Dynamics of the variance of the probability density function over risk distributions for a growing epidemic with two subpopulations. It initially increases because we initialize the population at a specific risk distribution (delta distribution) and it eventually decreases as the risk distribution drifts less for larger populations.



Appendix 1—figure 4. Growth rate bias, ρ, for epidemics initialized with different population sizes, N_0, of infected distributed by the equilibrium risk distribution (here $f_H = .6$. Epidemics were run for 75 populations changes and the vertical line delineates epidemics that stochastic extinction can affect (smaller initial populations). (inset) Focusing on epidemics where not enough population changes have occurred for stochastic extinction to occur.