A simple model of how heterogeneous disease transmission impacts the emergence of variants and their detection

Bradford P. Taylor and William P. Hanage

Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard TH Chan School of Public Health, Boston, MA 02115

Abstract:

Understanding how epidemics spread within societies is key for establishing adequate infection control responses. Dynamical models provide a means to translate surveillance data into predictions of future disease spread, yet many epidemic models do not capture empirically observed features of socialization. Here, we utilize a connection between sampling processes and their macroscopic equations to build an epidemic model that incorporates transmission heterogeneity, homophily, and awareness of the risk of infection. Our risk-SIR model captures how diseases tend to spread initially within more social subpopulations and sequentially spread to less social subpopulations. This dynamic complicates surveillance as many metrics become dynamic when comorbidities vary with socialization. We also show fast-growing lineages that emerge within highly social subpopulations can be misidentified as potential variants of concern, and that these lineages will transiently grow in frequency as they remain distributed among more social individuals. Together, we provide a simple framework for analysis of how socialization affects disease emergence and spread, and argue that this is necessary to contextualize epidemiologic parameters like the basic and time varying reproductive numbers based on sampling and stage of the outbreak.

Introduction:

Communicable diseases transmit along networks of hosts that may vary greatly in their risks of exposure, for reasons they may or may not be able to control (Newman, 2002; Shirley & Rushton, 2005; Bansal et al., 2007). Ignoring this resulting heterogeneity can lead to biased estimates of epidemiologically relevant quantities such as the proportion of the population that must be immune to control the epidemic or the reproductive number (Kermack and McKendrick, 1927; Woolhouse et al., 1997; Hébert-Dufresne et al., 2020). The public health impact of such varying risks of transmission within the select populations are illustrated by the ongoing outbreak of monkeypox among men who have sex with men (Philpott et al., 2022), outbreaks of polio in undervaccinated communities (Link-Gelles et al., 2022), and at the largest scale, the filtering of SARS-CoV-2 through populations typically affecting younger individuals at low risk of severe illness first, before eventually moving on to older populations (Monod et al., 2021). Incorporating age structure into models can account for varying transmission risk, however at the cost of growing combinatorial complexity in parameterization and analysis (Mossong et al., 2008).

We propose a modification of the simplest SIR models to account for varying contact rates and risk. For any SIR-like model there are two key events, transmission and recovery, yielding infection dynamics:

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

For a simple memoryless recovery process, as in the classic SIR model, we can normalize time by the mean infectious period, $1/\gamma$, and thus we focus solely on how different sampling

processes change transmission dynamics alone, B(S,I,R). First, transmission events require aggregation of n individuals and allow transmission to occur between each pair of sampled individuals. Figure 1a illustrates this sampling process for the SIR model where n=2 and Figure 1b illustrates this for the general "group-SIR" model. The transmission rate is the expected number of infections whenever a group is formed:

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

with

$$\mathbb{P}_{\text{infect}}(i) = 1 - (1 - p_\beta)^i$$

and

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

where lower-case s, i, and r denote the integer number of sampled susceptibles, infected, and recovered present in a group and uppercase S, I, R denote the continuous densities such that S+I+R=N. We sum over all possible group compositions where s+i+r=n while weighting the probability of sampling, P_{group}, by a multinomial distribution parametrized by the frequency of each category in the overall population. Each susceptible in the group remains so if they avoid being infected by each infected individual in the group with a pairwise infection probability of p_β. If groups of only size 2 are permitted, then this simplifies to the classic SIR model, and as groups increase in size transmission becomes increasingly nonlinear with powers that sum to the group size. Transmission within groups can also yield multiple infections, i.e., superspreading (Lloyd-Smith et al., 2005).

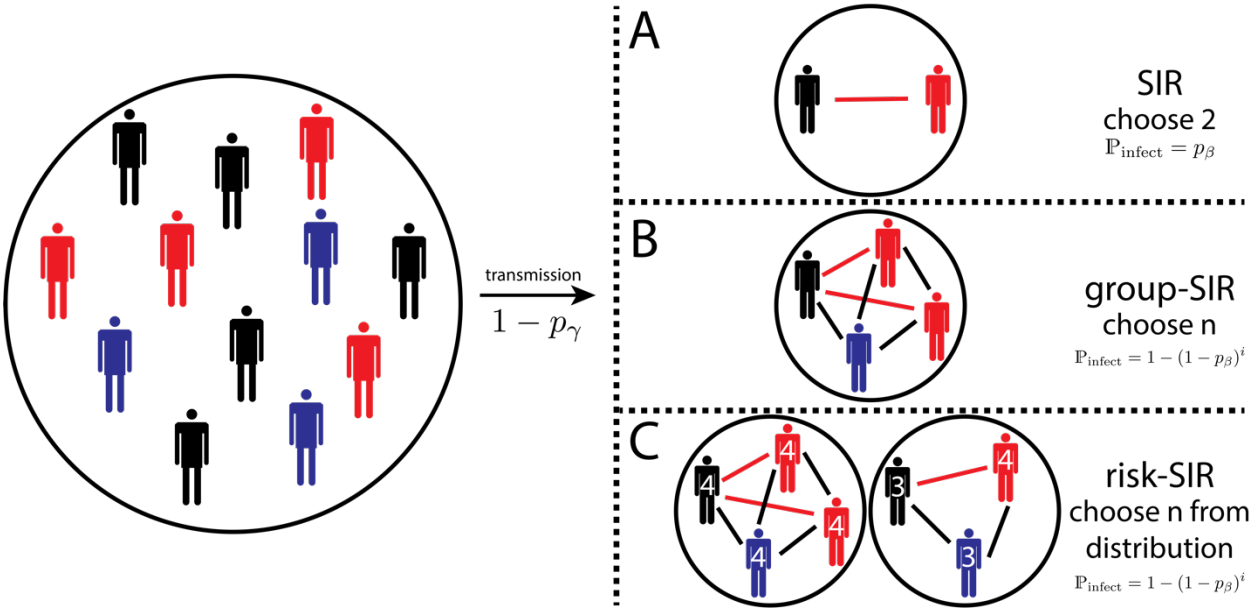


Figure 1: Schematic of sampling processes leading to different epidemic models described. A) The classic SIR model arises from sampling 2 individuals from a population of susceptible (black), infected (red), and recovered (blue). B) Sampling n individuals yields a generalization of the SIR model, the 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).

Increasing the group sizes leads to sharper epidemics for otherwise fixed parameter values (figure 2a). However, this is to be expected because increasing group sizes implicitly increases the total number of contacts relative to the recovery rate, thus increasing the reproductive number. Rescaling the transmission rate by the total number of pairwise interactions, $\binom{n}{2}$, collapses the group-SIR dynamics onto each other (Figure 2a inset). Increasing group sizes ultimately leads to fewer cases at the epidemic peak because multiple infectious individuals can infect the same susceptible within a congregation, effectively reducing the number of pairwise interactions at this point. 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).

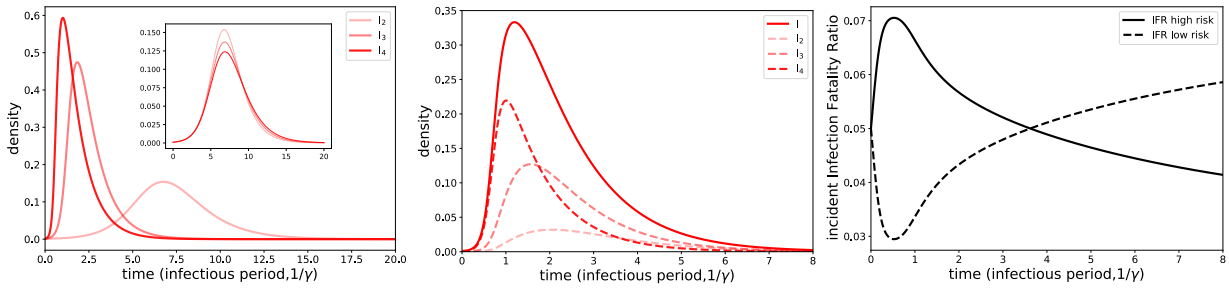


Figure 2: Generalized SIR dynamics. A) Group-SIR dynamics, where transmission occurs within groups with sizes here denoted by the subscript. B) Risk-SIR dynamics, where transmission occurs within groups according to a distribution here set with sizes n=2,3, and 4 at equal frequency. Subpopulations are denoted by their risk number 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 number (IFR high risk) and when the IFR is negatively correlated with risk number (IFR low risk).

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; Chande et al., 2020; Arthur et al., 2021; Harris & Mordecai, 2022). This may be voluntary or involuntary, reflecting personal choice and agency; for example many low paid workers were unable to avoid gatherings in early 2020, and the result was large SARS-CoV-2 superspreading events and substantial mortality in settings like meatpacking plants (Waltenburg et al., 2020). To capture this we can further generalize the group-SIR dynamics where there is a distribution of group sizes that occur and individuals only attend gatherings no larger than their "risk level".

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

with

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

where risk levels are denoted by subscripts. Here, each transmission term captures new infections within a particular focal subpopulation, k, such that only the respective susceptibles, S_k, contribute. These focal susceptibles can be infected by any attending infected regardless of subpopulation, here denoted by the dummy variable m. Meanwhile, 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 congregation sizes according to a distribution, P_n, set by society (e.g., by lockdowns) (Boyer et al., 2022). The risk-SIR captures transmission heterogeneity as subpopulations with higher risk numbers attend a larger fraction of gatherings, homophily as subpopulations with higher risk numbers 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 number 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 numbered subpopulations lead to more secondary infections. This leads to a recurrence relation for the basic reproduction number, R_0, of an infected individual with a specified risk number, k:

$$R0(k) = R0(k-1) + \mathbb{P}_k p_\beta \left(1 - \left(1 - \frac{1}{N}\right)^k\right)(k-1) \approx R0(k-1) + P(k)p_\beta \frac{k(k-1)}{N}$$

such that a disproportionate number of secondary infections occur at larger gatherings as enumerated by k. 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 the infected, 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).

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. Additionally, any underlying surveillance metrics that vary between subpopulations will now be dynamic. For example, infection fatality rates (IFR) for an emerging outbreak are required to assess the overall threat it poses to society. If virulence depends on comorbidities that vary in frequency between each subpopulation then the IFR will change over successive peaks driven by outbreaks in different subpopulations. Figure 2c illustrates IFR dynamics for two general cases. When individuals with comorbidities choose to avoid larger gatherings, i.e., subpopulation IFR negatively correlates with risk number (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 number, 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 infected individuals, i.e., it's risk distribution. Meanwhile, a variant emerges within a subset of these infected individuals. Thus, even a neutral variant with no increased transmissiblity relative to the background it emerges is expected to increase in frequency if it is transmitted primarily 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 (supplementary figure 1).

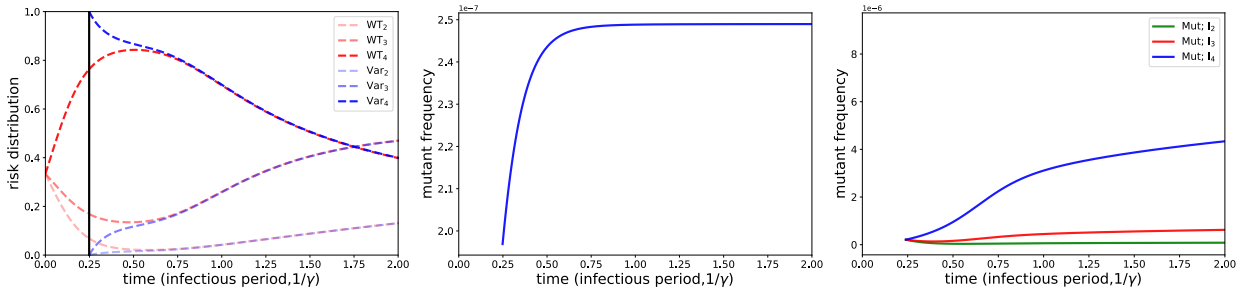


Figure 3: Variant frequency dynamics depend on which risk subpopulation it emerges within. A) Risk distribution of an epidemic (red) and a neutral variant (blue) that emerges within the highest risk subpopulation. B) The associated frequency dynamics of the emerging neutral variant. C) Frequency dynamics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

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. 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. This occurs because the risk distribution of the variant at the point of sampling, shown in Figure 4b, contains relatively more social individuals relative to the rest of the epidemic and, in turn, an expected increase in frequency relative to the epidemic. Indeed, repeating this simulation shows that the fastest growing lineages are on average biased towards an increased growth after censusing, as shown in Figure 4c. This means a fast-growing lineage that initially spreads among more social individuals may continue to appear to be more transmissible.

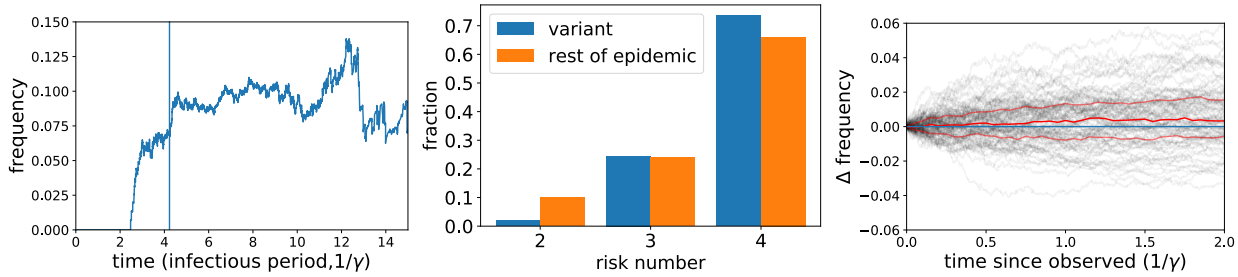


Figure 4: Identifying fast growing lineages during genomic surveillance preferentially identifies those spreading among more social individuals. A) Example frequency dynamics of the fastest growing neutral lineage at a censusing time (vertical line). B) Risk distribution of this fast growing neutral lineage compared to the rest of the epidemic C) Aggregate statistics of growth in frequency of the fastest growing lineage. The darker red line denotes the median across 100 replicate simulations and lighter red lines demarcate the interquartile range.

The amount of assortative mixing, also termed homophily, within a community sets the time scale over which such lineages will continue to spuriously increase in frequency. Intuitively, the risk distribution will remain 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 disaggregates transmission heterogeneity and homophily:

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

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, $ß_{HL}$, 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 supplementary 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 probability distribution of risk distributions over time induced by demographic stochasticity by numerically integrating a so-called "master equation" describing the probabilities associated with individual events as determined by their average rates (Van Kampen, 1992). Figure 5a shows probability

distributions after 500 population changes (transmission or recovery) for epidemics varying in homophily and initiated with 20 infected individuals distributed according to their respective equilibrium risk distribution. Note, to facilitate comparison across models, we rescale within-subpopulation transmission terms, $ß_{HH}$ and $ß_{LL}$ such that we vary homophily alone and not transmission heterogeneity by maintaining the same subpopulation basic reproduction numbers, $R_{0,H}$ and $R_{0,L}$. The demographic stochasticity induces a spread which eventually dissipates over time (supplementary figure 3)). The peak of the probability distribution is biased towards increased risk distributions, with an increased bias as homophily increases (i.e., lower f_2). This further motivates quantifying contact networks and risk distribution over the course of an epidemic, $\rho = \frac{\mathbb{P}(f_2 > f_2^*)}{\mathbb{P}(f_2 > f_2^*) + \mathbb{P}(f_2 < f_2^*)}$. Figure 5b compares the relative probability of increased sociality, ρ , dynamics after initiating the populations at their respective equilibrium value, f_2^*, for different amounts of homophily. Increased homophily leads to an increased quasi-equilibrium bias arising from demographic stochasticity. Figure 5c shows the bias in risk distribution for epidemics initiating either in the more social (dashed lines) or the less social (solid lines) subpopulations. Despite starting in just one subpopulation, the risk distribution bias ultimately approaches the same biased quasi-equilibrium value as seen above when started at the equilibrium, doing so more quickly with lower homophily.

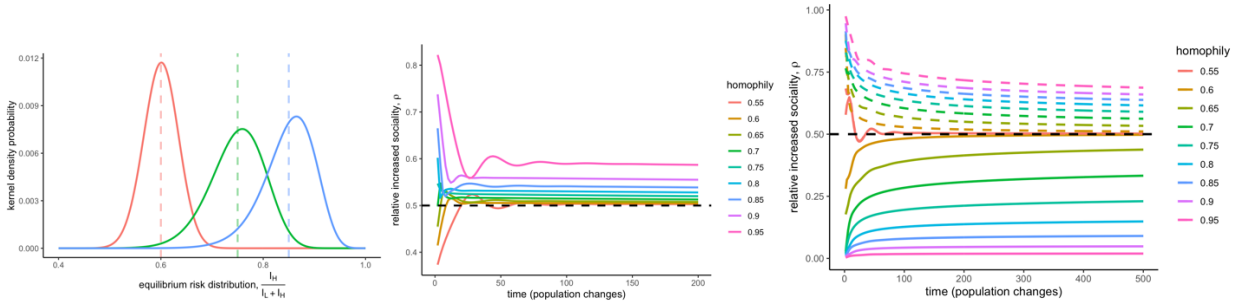


Figure 5: Homophily biases risk distributions a) Smoothed probability distributions of the risk distribution of epidemics initiated with 20 individuals at the respective equilibrium distribution after 500 population changes (total recovery or transmission events). Vertical dashed lines specify respective equilibrium risk distribution. B) Relative probability of a more social risk distribution ($f_2>f_2^*$) than a less social risk distribution ($f_2<f_2^*$) for epidemics initiated with 20 infected individuals distributed according to the equilibrium risk distribution (f_2^*) denoted as homophily. C) Relative probability of a more social risk distribution ($f_2> f_2^*$) than a less social risk distribution ($f_2<f_2^*$) for epidemics initiated with 1 infected individuals either in the more social subpopulation (dashed lines) or the less social subpopulation (solid lines).

This homophily bias arises due to stochastic extinction. Small populations go extinct by chance when all infected individuals recover prior to transmitting. All else equal, epidemics with a lower risk distribution are more likely to go extinct than those with higher risk distributions because more social individuals are more likely to transmit before recovery. Increasing homophily increases how long an epidemic remains deviated from the equilibrium risk distribution, thus increasing the likelihood for stochastic extinction. The homophily bias decreases to 0 as the epidemic gets larger meaning that emerging variants, which are smaller populations, are more likely than not to have an apparent increased relative fitness compared to the population as a whole.

Discussion:

These models offer a simple and flexible approach to capture the independent role socialization plays in epidemic spread relative to the inherent epidemiologic properties of the pathogen. In particular, the distribution of congregation sizes determine rates of epidemic spread with the most social subpopulations disproportionately driving growth. Differences in degree of socialization, which leads to transmission heterogeneity and may vary over the course of an epidemic, bias surveillance in multiple ways that can influence early estimates of key epidemiologic parameters and the detection of emerging variants. The extent of these biases are also modulated by social structure, with the degree of homophily setting the time scale and magnitude over which they change. While we analyze dynamics where socialization rates are static, we expect these to change in response to both exogenous factors like non-pharmaceutical interventions and endogenous factors like individual avoidance when cases are high, both of which have been evident over the course of the covid-19 pandemic. Indeed, it is all but certain that avoidance leads to a reduction in case counts prior to that expected from depletion of susceptibles alone as predicted from a classic SIR model.

Exactly how long apparent fitness biasing effects persist depends on both epidemiological and social factors. We numerically showed that the bias can be retained for 100s of infectious generations which for a pathogen with a generation interval of 3-4 days and R0=3 (as for emerging SARS-CoV-2) gives retention times in the order of months (Ganyani et al., 2020). This time frame is consistent with the incomplete, logistic growth of certain lineages that was observed in populations subjected to intense genomic surveillance (Hodcroft et al 2020), although we caution against over-interpreting as other mechanisms can lead to similar dynamics.

Repeat introductions during a pandemic complicate variant dynamics. Each introduction can be interpreted as a single trajectory sampled from the probability distribution of risk distributions with the total variant transmissibility as a weighted average of these distributions over time, as repeat introductions happen at different times. Furthermore, any variant fitness bias depends strongly on the subpopulation in which it emerges. If individuals willing to travel during a pandemic are more likely to engage in behavior that risks exposure in general then we expect introductions to preferentially occur in such subpopulations further contributing to an increased apparent fitness. Properly interpreting emerging epidemic and variant growth rates rely on establishing such social drivers. Indeed, informing epidemiology with sociology is imperative for the future of public health, as sustained transmission increasingly becomes more likely within unvaccinated communities.

Understanding the dynamics of socialization rates themselves is the next step to better model real epidemics and the risk-SIR model provides a framework for these future studies. However, identifying empirical proxies for socialization to inform these models is challenging due to the limited scope of measurements. Mobility data from tracked cell phones offer one possibility, with prior work showing lockdowns reducing socialization during the initial stages of the SARS-CoV-2 pandemic (Wesolowski et al., 2015; Kishore et al., 2021). However, such data ignore individuals without cell phones, which may include key social subpopulations like young schoolchildren. Additionally, even inferring interactions from mobility data alone may be

fraught as it is agnostic to epidemiological details such as where and when masking may be in use, or spurious connections between phones when proximate but between barriers. One possible route to estimate varying contact patterns over time is methods from population genetics, leveraging the extraordinary amounts of genetic data that are increasingly routinely collected (Scarpino et al., 2015).

This work, along with others, illustrates how simple models can capture the way social context determines and complicates epidemic dynamics, without explicit consideration of host networks (Weitz et al., 2020). To begin to quantify the emergent complexity parameters such as the reproduction number R_0 can be compared among relevant contexts. Prior work has shown the value of these key parameters depends on social context through the distribution of secondary contacts (Meyers et al., 2005; Hébert-Dufresne et al., 2020). Nevertheless, secondary infection rates and case count growth rates are easily measurable and can be incorporated into this modeling framework, allowing us to contextualize discussions within specific social contexts instead of a single mean value which can understate the capacity for transmission in certain populations. Understanding social contexts associated with rapid growth may allow early warning of for future spread in different, neighboring social contexts (Petros et al., 2022), and will be a key component of improved epidemic forecasting that accounts for changing behavior. Determining these social "routes of transmission" could be key for broad infection control, where resources can be prioritized towards mitigating epidemics within and between communities with increased growth.

References:

Arthur, R. F., Jones, J. H., Bonds, M. H., Ram, Y., & Feldman, M. W. (2021). Adaptive social contact rates induce complex dynamics during epidemics. PLoS computational biology, 17(2), e1008639.

Bansal, S., Grenfell, B. T., & Meyers, L. A. (2007). When individual behaviour matters: homogeneous and network models in epidemiology. Journal of the Royal Society Interface, 4(16), 879-891

Boyer CB, Rumpler E, Kissler SM, Lipsitch M. Infectious disease dynamics and restrictions on social gathering size. Epidemics. 2022 Sep;40:100620. doi: 10.1016/j.epidem.2022.100620. Epub 2022 Aug 17. PMID: 36058184; PMCID: PMC9384337.

Chande, A., Lee, S., Harris, M., Nguyen, Q., Beckett, S. J., Hilley, T., ... & Weitz, J. S. (2020). Real-time, interactive website for US-county-level COVID-19 event risk assessment. Nature Human Behaviour, 4(12), 1313-1319.

Eksin, C., Shamma, J. S., & Weitz, J. S. (2017). Disease dynamics in a stochastic network game: a little empathy goes a long way in averting outbreaks. Scientific reports, 7(1), 1-13.

Ganyani, T., Kremer, C., Chen, D., Torneri, A., Faes, C., Wallinga, J., & Hens, N. (2020). Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020. *Eurosurveillance*, *25*(17), 2000257.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. The journal of physical chemistry, 81(25), 2340-2361.

Harris, M. J., & Mordecai, E. A. (2022). Social divisions and risk perception can drive divergent epidemic dynamics and large second and third waves. medRxiv.

Hébert-Dufresne, L., Althouse, B. M., Scarpino, S. V., & Allard, A. (2020). Beyond R 0: heterogeneity in secondary infections and probabilistic epidemic forecasting. Journal of the Royal Society Interface, 17(172), 20200393.

Hodcroft EB, Zuber M, Nadeau S, Vaughan TG, Crawford KHD, Althaus CL, Reichmuth ML, Bowen JE, Walls AC, Corti D, Bloom JD, Veesler D, Mateo D, Hernando A, Comas I, González-Candelas F; SeqCOVID-SPAIN consortium, Stadler T, Neher RA. Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. Nature. 2021 Jul;595(7869):707-712. doi: 10.1038/s41586-021-03677-y. Epub 2021 Jun 7. PMID: 34098568.

Kermack, W. O., & McKendrick, A. G. (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.

Kishore, N., Kahn, R., Martinez, P. P., De Salazar, P. M., Mahmud, A. S., & Buckee, C. O. (2021). Lockdowns result in changes in human mobility which may impact the epidemiologic dynamics of SARS-CoV-2. Scientific reports, 11(1), 1-12

Link-Gelles, R. (2022). Public health response to a case of paralytic poliomyelitis in an unvaccinated person and detection of poliovirus in wastewater—New York, June–August 2022. MMWR. Morbidity and Mortality Weekly Report, 71.

Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E., & Getz, W. M. (2005). Superspreading and the effect of individual variation on disease emergence. Nature, 438(7066), 355-359.

Mobility

May, R. M., & Anderson, R. M. (1987). Commentary transmission dynamics of HIV infection. Nature, 326(137), 10-1038.

Meyers, L. A., Pourbohloul, B., Newman, M. E., Skowronski, D. M., & Brunham, R. C. (2005). Network theory and SARS: predicting outbreak diversity. Journal of theoretical biology, 232(1), 71-81

Monod, M., Blenkinsop, A., Xi, X., Hebert, D., Bershan, S., Tietze, S., ... & Imperial College COVID-19 Response Team. (2021). Age groups that sustain resurging COVID-19 epidemics in the United States. Science, 371(6536), eabe8372.

Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., ... & Edmunds, W. J. (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS medicine, 5(3), e74.

Newman, M. E. (2002). Spread of epidemic disease on networks. Physical review E, 66(1), 016128.

Perra, N., Balcan, D., Gonçalves, B., & Vespignani, A. (2011). Towards a characterization of behavior-disease models. PloS one, 6(8), e23084.

Petros, B. A., Turcinovic, J., Welch, N., White, L., Kolaczyk, E., Bauer, M., ... & Springer, M. (2022). Early introduction and rise of the Omicron SARS-CoV-2 variant in highly vaccinated university populations. medRxiv.

Philpott, D. (2022). Epidemiologic and clinical characteristics of monkeypox cases—United States, May 17–July 22, 2022. MMWR. Morbidity and Mortality Weekly Report, 71.

Scarpino, S. V., Iamarino, A., Wells, C., Yamin, D., Ndeffo-Mbah, M., Wenzel, N. S., ... & Townsend, J. P. (2015). Epidemiological and viral genomic sequence analysis of the 2014 Ebola outbreak reveals clustered transmission. *Clinical Infectious Diseases*, *60*(7), 1079-1082.

Shirley, M. D., & Rushton, S. P. (2005). The impacts of network topology on disease spread. *Ecological Complexity*, *2*(3), 287-299.

Van Kampen, N. G. (1992). Stochastic processes in physics and chemistry (Vol. 1). Elsevier.

Volz, E. M., Miller, J. C., Galvani, A., & Ancel Meyers, L. (2011). Effects of heterogeneous and clustered contact patterns on infectious disease dynamics. *PLoS computational biology*, *7*(6), e1002042.

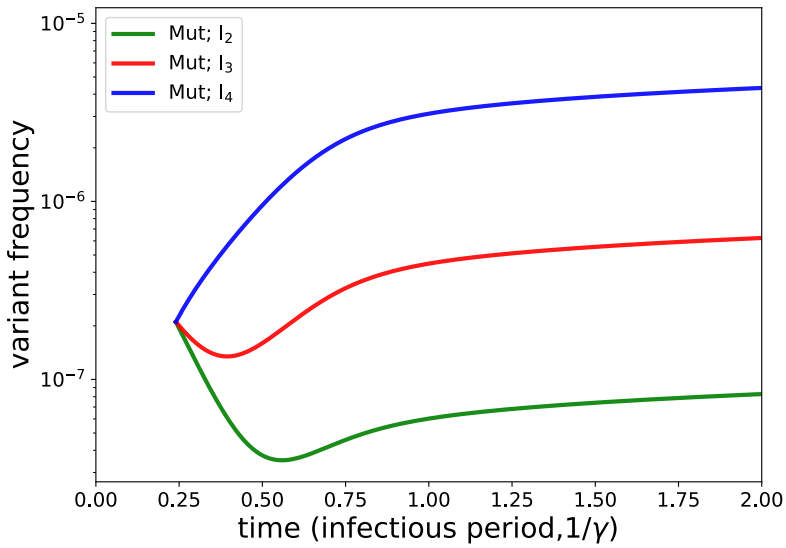
Waltenburg, M. A., Victoroff, T., Rose, C. E., Butterfield, M., Jervis, R. H., Fedak, K. M., ... & Zarate-Bermudez, M. (2020). Update: COVID-19 among workers in meat and poultry processing facilities—United States, April–May 2020. Morbidity and Mortality Weekly Report, 69(27), 887.

Weitz, J. S., Park, S. W., Eksin, C., & Dushoff, J. (2020). Awareness-driven behavior changes can shift the shape of epidemics away from peaks and toward plateaus, shoulders, and oscillations. Proceedings of the National Academy of Sciences, 117(51), 32764-32771.

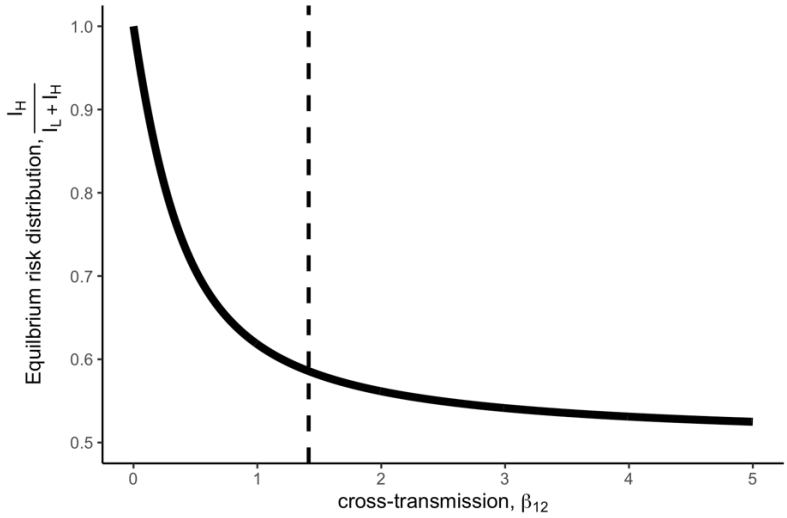
Wesolowski, A., Qureshi, T., Boni, M. F., Sundsøy, P. R., Johansson, M. A., Rasheed, S. B., ... & Buckee, C. O. (2015). Impact of human mobility on the emergence of dengue epidemics in Pakistan. Proceedings of the National Academy of Sciences, 112(38), 11887-11892.

Woolhouse, M. E., Dye, C., Etard, J. F., Smith, T., Charlwood, J. D., Garnett, G. P., ... & Anderson, R. M. (1997). Heterogeneities in the transmission of infectious agents: implications for the design of control programs. Proceedings of the National Academy of Sciences, 94(1), 338-342.

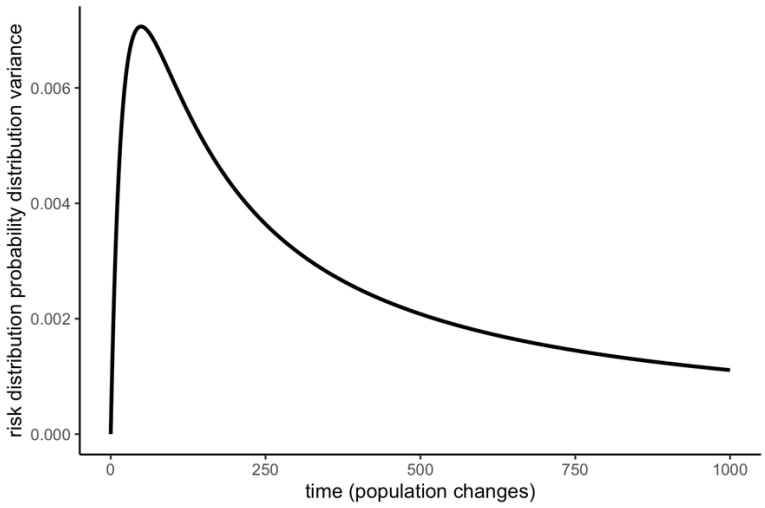
Supplementary Information:



Supplementary Figure 1: Dynamics of a variant with increased transmissibility relative to wildtype emerging in different risk subpopulations.



Supplementary Figure 2: Equilibrium risk distribution for exponentially growing epidemics in societies with transmission heterogeneity as the amount of cross transmission between two subpopulation varies.



Supplementary Figure 3: Dynamics of the variance of the probability distribution 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.