Implications of Heterogeneous SIR Models for Analyses of COVID-19*

Glenn Ellison†

June 2020

Abstract

This paper provides a quick survey of results on the classic SIR model and variants allowing for heterogeneity in contact rates. It notes that calibrating the classic model to data generated by a heterogeneous model can lead to forecasts that are biased in several ways and to understatement of the forecast uncertainty. Among the biases are that we may underestimate how quickly herd immunity might be reached, underestimate differences across regions, and have biased estimates of the impact of endogenous and policy-driven social distancing.

1 Introduction

The economic literature on the COVID-19 epidemic has developed at a remarkable pace. A number of recent economics papers build on the classic Susceptible-Infectious-Recovered (SIR) model to study how the epidemic may progress and how it may be affected by various policies.[1] In this note I review some results from the epidemiological literature on an SIR extension that economists have mostly not yet adopted, incorporating heterogeneity in the activity rates of different subpopulations, and note ways in which analyses based on classic SIR models can potentially yield misleading views.

The classic SIR model of **kermack1927contribution** has been a foundational model in epidemiology for nearly a century. It illustrates basic tradeoffs and provides a simple framework that can be easily built on. Subsequent work in epidemiological theory has extended the model in various ways, and modern epidemiological forecasts typically work with variants that are more flexible in a number

*I thank Daron Acemoglu, Chris Avery, Victor Chernozhukov, Adam Clark, Jonathan Dushoff, Sara Fisher Ellison, Jim Stock, and Ivan Werning for helpful conversations and comments and Chris Ackerman and Bryan Kim for research assistance.

†Department of Economics, Massachusetts Institute of Technology, Cambridge MA 02139 and NBER, e-mail: gellison@mit.edu

[1] See among others **acemoglu2020multi**; **lippi2020simple**; **baqaee2020reopening**; **eichenbaum2020macroeconomics**; **farboodi2020internal**; **fernandez2020estimating**; **jones2020optimal**; **rowthorn2012optimal**

of dimensions.[2] In this paper I focus on some theoretical extensions developed in the 1980's and 1990's that seem quite relevant to the COVID-19 epidemic. Specifically, I discuss two classic models that focus on heterogeneity in the frequency with which different individuals engage in interactions that risk spreading the disease. Given the current understanding about how COVID-19 seems to be transmitted, it is easy to think of a number of subpopulations who will have many more risky interactions than average: those living in overcrowded urban apartments, frequenting bars and nightclubs, using public transportation, attending crowded religious services, working in a nursing home, etc. Others, e.g. farmers and those who are retired or work from home, should be relatively safe.

Section 2 reviews of the classic SIR model and extensions. Each extension discussed is a multipopulation SIR model that supposes that the subpopulations differ in their "activity" levels. As with the classic SIR model, the differential equations describing the rates at which members of each subpopulation transition from the susceptible to the infectious state can be motivated by a process in which agents are randomly matched in continuous time with each interaction between susceptible and infectious agents potentially leading to a new infection. One version assumes "uniform" matching in which the probability that any two agents are randomly matched is proportional to the product of their activity levels. The other assumes "homophilic" matching in which agents are more likely to interact with others in their own subpopulation. Each model behaves much like the classic SIR model. Small infections initially grow at an exponential rate if a composite parameter analogous to R_0 is greater than one and new infections slow (and eventually die out) once the fraction with acquired immunity passes a "herd immunity" threshold. The composite R_0 and the herd immunity thresholds depend on the characteristics of the various subpopulations, and I review results from prior work to illustrate important principles about how epidemics spread in heterogeneous populations.

Sections 3 and 4 then draw out implications of these models for our analyses of COVID-19. Section 3 emphasizes that thinking about heterogeneity in contact patterns suggests that making predictions about the course of the COVID-19 epidemic and the impacts of reopening policies is inherently difficult. Heterogeneous models have more parameters that need to be calibrated. Long run outcomes can be sensitive to activity levels of the less active, and it is difficult to calibrate these parameters early in an epidemic when there are few cases in less-active communities. This is particularly true when one contemplates removing restrictions and thereby increasing activity among the currently inactive. Predictions based on classic SIR models that do not allow for heterogeneity may be overconfident.

[2] See, for example, **champredon2018two**; **viboud2018rapidd**; **unwin2020state**.

Section 4 then focuses on ways in which conclusions drawn from applying homogeneous SIR models to a world that may be like a heterogeneous SIR model can be misleading. One observation, also found in **gomes2020individual** and **britton2020disease**, is that homogeneous SIR models may substantially overstate the fraction of the population that must be infected in order to achieve herd immunity. Intuitively, if a small high-contact group plays a central role in spreading the disease, then incidence will much higher in this group, and once many in this group have acquired immunity the epidemic may die out. Less obvious but still relevant effects are present with less extreme heterogeneity. A second related observation is that (targeted) lockdown polices can also be more cost effective in heterogeneous populations. There can be substantial gains either from taking permanent measures to reduce spread among the highly active or from temporarily locking down less active groups to minimize overshooting of herd immunity thresholds. The differences in dynamics also imply that time-series estimates of policy impacts may be biased. In each case, effects depend both on the magnitude of the heterogeneity that it present and on the degree of homophily in matching. The discussions attempt to bring out comparative statics and plausible magnitudes of effects.

The final section of the paper discusses some practical implications of the results. The message that we may be missing information for assessing reopening plans is troublesome when reopening is already upon us. But the models suggest fairly easy ways in which economists could extend their models and also point to data opportunities that might reduce the critical uncertainties. The messages that controlling the epidemic may not be as hard as it appears in some models and that herd immunity might not be as far off might be naïvely predicted said also give room for optimism.

This paper is related to a number of others in epidemiology and economics. The discussion of heteogeneous SIR models is a review of a literature in epidemiology that dates back to the late 1980s and mid 1990s, with the particular formulations drawing heavily on **dushoff1995effects**. Empirical epidemiologists have also for quite some time been interested in multipopulation SIR models both examine to interactions between age groups (important for other reasons for childhood diseases) and groups, e.g. health care workers in the Ebola epidemic, who play an important role in transmission.[3] Two very recent working papers in epidemiology have made observations similar to the observation in section 4.1 that herd immunity thresholds can be substantially lower in heterogeneous SIR models than in homogeneous SIR models. **gomes2020individual** graphs the herd immunity threshold as a function of the coefficient of variatiion in contact rates in a heterogeneous SEIR model, not-

[3]See, for example, **britton1998estimation**; **demiris2005bayesian**; **lloydsmith2005superspreading**; **champredon2018two**.

ing estimates of the coefficients of variation that have been previously reported for other diseases. **britton2020disease** give herd immunity thresholds for an 18-group model calibrated to estimated interactions across 6 age groups with assumed low-activity and high-activity individuals assumed to have activity levels that are half and twice the average activity levels and discuss partial lockdown policies that hold the infection to this level. Recent papers in epidemiology are also broadly related in that their observations motivate examining heterogeneous transmission. **worobey2020emergence** concludes that early imported cases formerly thought to have triggered epidemics in Washington and Italy appear to not be related to the subsequent epidemics there, suggesting that the communities in which they occurred had low enough R_0 so that the epidemics they started died out. **miller2020full** examine transmission in Israel using full genome sequence and conclude that there are "high levels of transmission heterogeneity ... with between 1-10% of infected individuals resulting in 80% of secondary infections.

As noted earlier, the primary motivation for the paper is the large recent literature in economics that builds on SIR models. In this literature, **avery2020policy** informally discuss the potential relevance of transmission heterogeneity. Most closely related are several very recent papers, including **acemoglu2020multi**; **baqaee2020reopening**; **favero2020restarting**; **rampini2020sequential**, that use calibrated multipopulation SIR models to examine the impact of COVID-19 mitigation policies, and in the case of **acemoglu2020multi** to identify optimal policies from a broad class. These papers use age-defined group structures to illustrate the substantial gains from age-targeted policies due to how dramatically death rates vary with age. They do not focus on the impact of contact heterogeneity, nor do most of the calibrations include within-age-group heterogeneity, which is presumably much larger than cross-age group heterogeneity, but three of them do include some heterogeneity in contact rates. **baqaee2020reopening** calibrate a five by five matrix of age group to age group contact rates using both general contact survey data and a workplace proximity survey to reflect differences in occupational mixes across age groups. **acemoglu2020multi** use uniform mixing in their main analyses, but also calibrate a three by three age group contact matrix to data from another contact survey. The groups in **favero2020restarting** are age $\times$ activity based with medium- and high-activity individuals assumed to be 12% and 18% more active than the low-activity group.

2 Heterogeneous SIR Models

In this section I'll quickly review the standard SIR model and then spend more time on two heterogeneous versions drawing on previous results.

2.1 The standard homogeneous SIR model

A number of recent economic analyses of the COVID-19 epidemic build on a standard homogeneous SIR model.

Consider a population of unit mass. Assume that at each time t each member of the population is in one of three states: Susceptible, Infectious, or Recovered. Write $S(t)$, $I(t)$, and $R(t)$ for the fractions in each state at time t. Assume that the dynamics of these fractions are:

$$\begin{aligned}\dot{I}(t) &= S(t)I(t)R_0\gamma - \gamma I(t)\\ \dot{R}(t) &= \gamma I(t)\\ \dot{S}(t) &= -S(t)I(t)R_0\gamma\end{aligned}$$

One way to motivate the model is to suppose that agents are being uniformly randomly matched in continuous time. Each agent meets another with probability $R_0\gamma dt$ in a dt time interval. A susceptible agent matched with an infectious agent becomes infectious. Agents transition from the Infectious state to the Recovered state at Poisson rate γ. These transitions reflect both true recoveries and deaths from the disease.

The parameter R_0 can be thought of as the expected number of people that a newly infected person will directly infect in a population where everyone is susceptible. It is the critical determinant of the behavior of the model. Three important facts are:

1. If $R_0 > 1$, then the equilibrium $(S, I, R) = (1, 0, 0)$ is locally unstable. Adding a small number of infected agents leads to contagious growth in I. Equilibria with $I = 0$ are locally stable if $R_0 < 1$. A small infection dies out.

2. The model has a "herd immunity" threshold of $\overline{S} \equiv 1 - 1/R_0$. Any state $(S, 0, 0)$ with $S < \overline{S}$ is a stable steady state, so a small infection introduced into such a population will not spread. This does not, however, mean that epidemics will not infect more than a fraction $1 - \overline{S}$ of the population. When the herd immunity threshold is first reached we have $\dot{I}(\overline{S}, I, R) = 0$. This

means that the infectious rate is (locally) constant with new infections occurring as fast as people are recovering. If the herd immunity threshold is reached at a point when I is large (which it typically is in models with R_0 large), then there can be substantial "overshooting" and many more than $1 - \overline{S}$ people can eventually be infected.

3. Define the growth rate of the infectious population by $g(t) = \frac{d}{dt}\log(I(t))$. Then, $g(t) = \gamma(R_0 S(t) - 1)$.

In the initial phase of an epidemic when $S(t) \approx 1$, the third fact says that the growth rate of the infectious is approximately $\gamma(R_0 - 1)$. One can think of this as a cumulative growth rate of $R_0 - 1$ over the $1/\gamma$ average duration of an infection.

Investigations of whether restrictions are "flattening the curve" often graph the log of cumulative infections, i.e. $\log(1 - S(t))$, versus time. This curve will be approximately linear with slope $\gamma(R_0 - 1)$ as long as the ever-infected fraction of the population remains small, e.g. when the US has had 10 million cases. Attempts to infer R_0 from such curves are common given the desire to assess where the herd immunity threshold might be.

One other relevant feature of SIR models is that for many values of R_0 the time-path of new infections (and of deaths) has a shape that is fairly symmetric about its peak and looks somewhat like a normal density. For example, figure **??** below reproduces Figure 1A from **ferguson2020impact** illustrating the predictions of an SIR-like model for Great Britain the US.

Epidemiologists commonly work with extensions of the SIR model. Among the standard additions are an additional state E of agents who are infected but not yet infectious, more flexible recovery processes that allow non-exponential infectious durations, an explicit death state, and population inflows/outflows. Some economic models also incorporate some of these elements. To simplify the discussion I will not incorporate any of these features here, but similar conclusions should apply.

2.2 A heterogeneous SIR model with uniform matching

In practice, some individuals are more interactive than others. For example, supermarket cashiers will be in the vicinity of many more people in a typical day than will retirees. Epidemiologists have also analyzed models that allow for such heterogeneity.[4]

[4] See **andreasen1989persistence**; **may1989transmission**; **diekmann1990definition**; **dushoff1995effects**; **jacquez1995core**; **hethcote2000mathematics**; **van2002reproduction**. The exposition below draws heavily on **dushoff1995effects**.

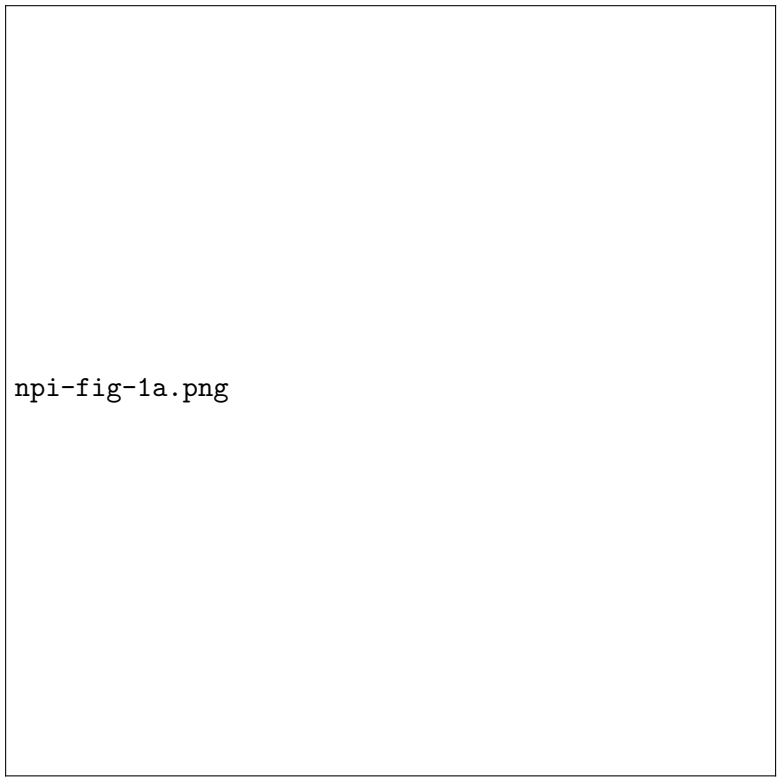

Figure 1: Figure reproduced from **ferguson2020impact** Figure 1A: "Unmitigated epidemic scenarios for GB and the US. (A) Projected deaths per day per 100,000 population in GB and US."

A tractable version is motivated by uniform matching in a population consisting of N equally sized subpopulations indexed by $i = 1, 2, \ldots, N$. Suppose that members of group i are randomly matched with probability $R_{0i}\gamma dt$ in each dt time interval. Order the populations so that $R_{01} > R_{02} > \ldots > R_{0N}$. Assume that the matchings are uniform so that the probability that a matched agent from group i meets a group j agent is $R_{0j}/\sum_k R_{0k}$. Suppose any matching between a susceptible and an infectious agent results in the susceptible agent becoming infectious. Write $S_i(t)$, $I_i(t)$, and $R_i(t)$ for the fraction of agents in group i who are susceptible, infectious, and recovered at time t, and $S(t)$, $I(t)$, and $R(t)$ for the vectors with these terms as components.

With the same recovery process as before, this matching process motivates analyzing a system of differential equations:

$$\begin{aligned}
\dot{I}_i(t) &= S_i(t)\sum_j \beta_{ij} I_j(t) - \gamma I_i(t) \\
\dot{S}_i(t) &= -S_i(t)\sum_j \beta_{ij} I_j(t) \\
\dot{R}_i(t) &= \gamma I_i(t)
\end{aligned}$$

with $\beta_{ij} \equiv \gamma R_{0i}\frac{R_{0j}}{\sum_k R_{0k}}$.

With the assumption that the population size remains constant, the state is fully described by $S(t)$ and $I(t)$ and we will usually omit $R(t)$ from the state vector. For any vector S^0 giving the fraction of susceptibles in each group, the disease free state $(S, I) = (S^0, 0)$ is a steady state. To analyze the stability of such a steady state and the behavior of the system in a neighborhood thereof, we linearize the system around the steady state. Note also that all derivatives $\frac{\partial \dot{I}_i}{\partial S_j}$ are equal to zero when evaluated at a state with $I = 0$. Hence, the behavior of I in a neighborhood of $(S^0, 0)$ in the full 2N-dimensional system has the same first-order approximation as that of I in the N-dimensional system

$$\dot{I} = A^{S^0} I,$$

where A^{S^0} is the partial derivative matrix with ijth element

$$a_{ij} = \left.\frac{\partial \dot{I}_i}{\partial I_j}\right|_{(S^0,0)} = \begin{cases} S_i^0\beta_{ij} - \gamma & \text{if } j = i \\ S_i^0\beta_{ij} & \text{if } j \neq i. \end{cases}$$

In particular, the equilibrium is locally stable if all eigenvalues of this matrix have negative real parts, and unstable if any eigenvalue has a positive real part.

The A^{S^0} matrix has positive off-diagonal elements, so the eigenvalue with the largest real part is real, and corresponds to a strictly positive eigenvector. This eigenvector gives the relative prevalence of the infected across groups for which the total number infected grows most rapidly. The special structure of this matrix allows one to easily find this eigenvector. It is $v_1 = (S_1^0 R_{01}, \ldots, S_N^0 R_{0N})$, i.e. prevalence is proportional to the product of the susceptible fraction and the contact rate. The eigenvalue corresponding to this eigenvector is

$$\lambda_1 = \gamma \left(\frac{\sum_i S_i^0 R_{0i}^2}{\sum_i R_{0i}} - 1 \right).$$

Two important implications of this are:

1. The equilibrium $(S^0, 0)$ is locally stable if $\frac{\sum_i S_i^0 R_{0i}^2}{\sum_i R_{0i}} < 1$ and locally unstable if $\frac{\sum_i S_i^0 R_{0i}^2}{\sum_i R_{0i}} > 1$.
2. For small δ, the growth rate of the log of the total infected population at the state $(S^0, \delta v_1)$ is approximately $\gamma \left(\frac{\sum_i S_i^0 R_{0i}^2}{\sum_i R_{0i}} - 1 \right)$.

Note that if we start from any state with a very small fraction δ infected, then the initial cases will initially grow at different rates in the different groups in a way that makes the distribution of cases across groups aligned with the principal eigenvector v_1.[5] Hence, provided that this alignment has already occurred by the time the epidemic starts to be measured, the early growth of a heterogeneous-SIR epidemic with activity vector R_0 will resemble the early growth of a homogeneous-SIR epidemic with parameter $\overline{R}_0 \equiv \frac{\sum_i R_{0i}^2}{\sum_i R_{0i}}$.

Two ways of rewriting this expression are informative. First,

$$\overline{R}_0 = \frac{\sum_i R_{0i}^2}{\sum_i R_{0i}} = \sum_i \frac{R_{0i}}{\sum_k R_{0k}} R_{0i}.$$

This formula makes clear that growth rates depend on a weighted average of group-level R_{0i}'s, with the weights being proportional to the activity level in each group. This weighted average can be substantially higher than the unweighted mean. The relation to the unweighted average is made clearer by a second rewriting:

$$\overline{R}_0 = \frac{\sum_i R_{0i}^2}{\sum_i R_{0i}} = \frac{N E(R_{0i}^2)}{N E(R_{0i})} = E(R_{0i}) + \frac{\mathrm{Var}(R_{0i})}{E(R_{0i})}.$$

[5] Suppose the initial population infected is δv with $v = \sum a_i v_i$ where the v_i are the eigenvectors of A. In a neighborhood of this point we will have $I(t) \approx \sum a_i e^{\lambda_i t} v_i$, which becomes aligned with v_i.

This equality indicates that growth rate is the sum of the unweighted average of the R_{0i} and the ratio of the variance of the R_{0i} across groups to the mean. The latter can easily be quite important quantitatively.

2.3 A heterogeneous SIR model with homophily

While supermarket cashiers may interact with a fairly representative sample of the population, some other highly active groups disproportionately interact with others in their group. For example, those who frequent nightclubs, take public transportation, attend crowded religious services, or live in a working class neighborhood with overcrowded housing disproportionately interact with others who do the same things. Those who live in rural areas will disproportionately interact with others who live in the same rural area.

Heterogeneous SIR models with homophilic matching are more difficult to analyze, but epidemiologists have also derived insightful characterizations of some such models, referred to sometimes as models with "preferred mixing" or "like-with-like preference". To motivate one such model, consider an N group model as in the previous subsection, but suppose that when an agent from group i is randomly matched the probability that the person with whom they are matched is in group j is

$$p_{ij} = \begin{cases} h + (1-h)\frac{R_{0j}}{\sum_k R_{0k}} & \text{if } j = i \\ (1-h)\frac{R_{0j}}{\sum_k R_{0k}} & \text{if } j \neq i. \end{cases}$$

Such a matching process would lead to an SIR model nearly identical to that in the previous subsection with

$$\dot{I}_i(t) = S_i(t) \sum_j \beta_{ij}^h I_j(t) - \gamma I_i(t),$$

where

$$\beta_{ij}^h = \begin{cases} \gamma R_{0i}(h + (1-h)\frac{R_{0j}}{\sum_k R_{0k}}) & \text{if } j = i \\ \gamma R_{0i}(1-h)\frac{R_{0j}}{\sum_k R_{0k}} & \text{if } j \neq i. \end{cases}$$

Once again, any $(S^0, 0)$ is a steady state of the system and we can analyze the stability of this steady state by looking at a linearized N-dimensional system:

$$\dot{I} = A^{S^0 h} I,$$

where $A^{S^0 h}$ is the partial derivative matrix with ijth element

$$a_{ij}^h = \left.\frac{\partial \dot{I}_i}{\partial I_j}\right|_{(S^0,0)} = \begin{cases} S_i^0 \beta_{ij}^h - \gamma & \text{if } j = i \\ S_i^0 \beta_{ij}^h & \text{if } j \neq i. \end{cases}$$

The off-diagonal elements of this matrix are again positive, so the eigenvector with the largest real part is again unique and corresponds to a positive eigenvector. It is no longer easy to give an explicit formula for the eigenvalue, but as noted by **diekmann1990definition**; **dushoff1995effects** we can give explicit necessary and sufficient conditions for the equilibrium to be stable.

1. If $S_i^0 h R_{0i} > 1$ for any i, then $(S^0, 0)$ is unstable. This is obvious: the number of infected in population i will increase solely from within-group contacts, and cross-group contacts only add to the growth.

2. If $S_i^0 h R_{0i} < 1$ for all i, then $(S^0, 0)$ is unstable if $\sum_i \frac{R_{0i}}{\sum_k R_{0k}} \frac{1}{1-hS_i^0 R_{0i}}(S_i^0 R_{0i} - 1) > 0$ and stable if $\sum_i \frac{R_{0i}}{\sum_k R_{0k}} \frac{1}{1-hS_i^0 R_{0i}}(S_i^0 R_{0i} - 1) < 0$.

Note that when $h = 0$ the stability condition in the part 2. simplifies to a version of expression we gave earlier for the uniform model: $\sum_i \frac{R_{0i}}{\sum_k R_{0k}}(S_i^0 R_{0i} - 1) < 0$. For a disease-free equilibrium to be stable in a model with $h > 0$ it must satisfy the additional constraints in 1. that $S_i^0 h R_{0i} < 1$ for all i as well as the modified inequality given in 2. Note that the summation in this inequality differs from the summation for $h = 0$ in that we multiply the ith term by $\frac{1}{1-hS^0 R_{0i}}$. These multiplicative factors are positive for all terms, and they are larger for the terms with $S_i^0 R_{0i}$ larger. Hence, we can think of the sum as proportional to a reweighting of the $h = 0$ sum that puts greater weight on the terms with $S_i^0 R_{0i}$ large and less weight on the terms with $S_i^0 R_{0i}$ small. As a result, if the model with $h = 0$ is unstable, then the model with $h > 0$ is unstable as well.[6]

The argument above extends easily to a full monotonicity theorem: if a disease-free equilibrium is unstable for some value of h, then it is unstable for any value $h' > h$. Hence, a clear intuition one can take away from this model is that homophilic matching is an obstacle to the stability of disease free states.

Note that homophily on its own does not affect the dynamics of an epidemic. If we consider a model in which the R_{0i} are identical across groups, then as long as h is not extremely close to one,

[6]Mathematically, this is a classic Chebyshev inequality argument: if a_i and b_i are monotone increasing, then $\sum_i a_i b_i > \frac{1}{N} \sum_i a_i \sum_i b_i$.

a small infection introduced into any one population will soon equalize across populations. With equal fractions infected in each population, the dynamics of the homophilic multipopulation model are identical to the $h = 0$ model. Hence, all of the effects of homophily discussed above should be understood as the effects of the combination of homophily and contact heterogeneity.

3 Challenges Inherent in Analyzing Heterogeneous Population Epidemics

In this section and the one that follow I turn to the task of drawing out implications of the above models for analyses of COVID-19. This section stresses a cautionary implication: it can be difficult to provide policy advice in epidemics that are well described by heterogeneous population SIR models. In particular, with currently available data it is challenging to estimate activity rates in less active populations, and important outcomes can be sensitive to these hard-to-estimate parameters.

3.1 Difficulty in calibrating models

Early in the COVID-19 epidemic several authors noted that it is difficult to calibrate critical parameters of homogeneous SIR model in the initial phase of an epidemic.[7] In the initial phase we may not have reliable data on anything but deaths. The fact that deaths in the model increase at an exponential rate makes it fairly easy to estimate R_0. But when many cases go unreported it is hard to calibrate the death rate. Different death rates would lead to dramatically different future paths of the epidemic. This weak identification problem goes away when we get some other piece of information that lets us estimate the death rate. Some potential sources for this are random serology tests to estimate the fraction that have ever been infected, fatality data from locations, e.g. South Korea or the Diamond Princess, where we think almost all cases have been identified, or seeing an epidemic peak, which is informative about S.[8]

As more information has become availble many economists have analyzed SIR models with calibrated R_0 and death rate parameters. The state of the art, in fact, has already moved well beyond this with a few recent working papers analyzing calibrated multipopulation SIR models that allow death rates and contacts to vary by age group.[9] The contact rate calibrations in these papers

[7] See **atkeson2020deadly**; **fernandez2020estimating**; **korolev2020identification**; **stock2020coronavirus**.

[8] **fernandez2020estimating** note that more complex SIR models fit under a variety of assumptions about accessory parameters make very similar predictions about the future course of epidemics in locations where epidemics have peaked.

[9] See **acemoglu2020multi**, **baqaee2020reopening**, and **favero2020restarting**.

rely on three survey datasets. POLYMOD (**mossong2008social**) and the BBC Pandemic Project (**klepac2020contacts**) are survey datasets which asked respondents to list those with which they had contact in the previous 24 hours. And employment website O*Net asked workers in a large number of occupations to report how physically close to others they worked on a 5 point scale. One can also now capture some changes in activity over time using movement data available from firms with phone-tracking capabilities.

Heterogenous SIR models that allow for idiosyncratic variation in contact rates by breaking a population (or each age group or other cell) into subpopulations that differ in activity levels have more parameters than do SIR models that do not consider such divisions. The fact that predictions can be dramatically affected by heterogeneity in R_0 suggests that it is important to try to capture some of the heterogeneity that surely exist within the cells that economists have been using with these extra parameters.

One approach to calibrating the extra parameters might be to use data on the variance of reported contacts in contact surveys. Although the surveys mentioned above have been used to estimate the relative prevalence of different age-group to age-group contacts, they seem less compelling as a source for estimating contact heterogeneity. For one thing, the way that contacts were defined, e.g. in the BBC survey contacts were defined as those whom one had physically touched or had a face-to-face conversation of at least three words with, leaves out many contacts that may be important in spreading COVID-19: singing near someone in a choir practice, standing near someone in a crowded bar, riding on the same subway train, being served by a cruise ship waiter, etc. The obvious heterogeneities in the frequency of such unrecorded contacts may mostly cancel out when one computes means for a large group, but we would definitely want to capture them to calibrate a model of contact heterogeneity. Another limitation of the main contact surveys is that they record contacts on a single day. Hence, recorded cross-subject variation confounds differences is cross-sectional means and time-series variation.

Another approach might be to calibrate the model to the path of the epidemic to-date. The initial growth rate of an epidemic should let us estimate the parameter composite $\overline{R}_0$ mentioned earlier. However, in a heterogeneous population SIR model, there is a weak identification problem when one tries to get more than this: it can be very difficult to obtain estimates of the activity rates in the less-active populations even after there has been substantial spread of the infection. Intuitively, when there is substantial heterogeneity in the R_{0i}, there will be a substantial number of infections when the epidemic surges in the highest R_{0i} subpopulations. At that point, there may still be few infections in

many of the less-active groups, particularly if matching is homophilic. This can make it very difficult to estimate activity parameters for the low infection groups from aggregate infection data.

While it is hard to have any confidence in a calibration, I know that many economist readers will want to know if the differences between heterogeneous and homogeneous SIR models are salient for plausible parameters. Accordingly, I will at times discuss a simple numeric example in which the mean and variance of activity levels across groups has been chosen to be in the plausible range. Specifically, I will sometimes discuss a population with five equally-sized subpopulations having activity rates 3.5, 1.5, 1, 0.5, 0.5. With uniform matching the model has $\overline{R_0} \approx 2.3$ which roughly matches the the growth rates assumed by **ferguson2020impact** and **acemoglu2020multi**.[10] The coefficient of variation of the cross-group differences, 0.8, roughly matches the variation in reported contacts in the BBC Pandemic Project data.

3.2 Difficulty in predicting future epidemic paths

The fact that some parameters of the heterogenous SIR model are difficult to calibrate would not be troubling if the hard-to-estimate parameters of the model did not affect model predictions that we care about. Unfortunately, this is not true for the heterogeneous SIR model. One reason is that activity levels in the relatively low activity groups can have a substantial impact on the long run course of the epidemic. As an illustration, Figure **??** graphs new daily cases for two heterogeneous SIR models.[11] The parameters of the two models were chosen so that new cases take off at about the same time, rise to a peak at about the same rate, and peak at about the same level. Despite the nearly identical behavior up to the point when the epidemics peak, however, the epidemics proceed very differently on the way back down. In the end, one epidemic eventually infects more than twice as many people as the other, 58% vs. 28% of the population. The fraction who will eventually be infected under a given constant policy is obviously highly policy-relevant, and this example indicates that it will sometimes be very difficult to predict even when an epidemic is sufficiently far along as to have already reached its peak.

Intuitively, the way in which the example was constructed is that the two models each have fairly homophilic matching ($h = 0.7$) and feature a highly-active subpopulation in which the epidemic peaks

[10] This is also consistent with some of the more sophisticated recent estimates growth rates such as that of **miller2020full**.

[11] Both models have ten equally-sized subpopulations with $h = 0.7$. The population with the long-lasting epidemic has $R_0 = (5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1, 1, 1)$. The population with the shorter-lived epidemic has $R_0 = (5.4, 2.6, 0.6, 0.4, \ldots, 0.4)$ and a lower fraction initially infected.

before many in the less active subpopulations have been extensively infected. The models differ in the activity levels of the less-active. In one population, corresponding the dashed red line, seven of the ten subpopulations have $R_{0i} = 0.4$. The epidemic never really takes off in these groups and this results in the fairly rapid decline in infection rates once the epidemic has burned through the highly active groups. In the other population, corresponding to the solid blue simulation, nine of the ten groups have R_{0i} equal to 1.5 or 1.0. The infections coming out of the most active group set off a spread in these groups that goes on for quite some time. This produces an asymmetric peak with a decline that is much more gradual than the run up. Most of the total infections occur post-peak.

Figure 2: Example illustrating the difficulty in predicting the long-run course of a heterogeneous-population epidemic given the path of infection rates up to the point when the infection peaks. New daily cases are graphed for two ten population heterogeneous SIR models with $h = 0.7$. Model 1 has $R_0 = (5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1, 1, 1)$. Model 2 has $R_0 = (5.4, 2.6, 0.6, 0.4, \ldots, 0.4)$

Analyses that do not consider the possibility that there may be heterogeneity in contact rates can report confidence intervals that are too narrow for this reason. For example, while the US remains quite far from the no-social-distancing herd immunity threshold, many states (and countries) are well beyond their peaks, which means that SIR-based models will make highly confident predictions about the future course of the epidemic presuming that individual behaviors and government policies remain fixed. For example, the April 29th update of the widely discussed IHME model gave its confidence interval for August 1st Massachusetts COVID-19 deaths as just 0 to 2.

3.3 Difficulty in predicting policy impacts

As reopening has become salient, a number of economic analyses have modeled the effects that relaxations of restrictions may have.[12] Thinking about heterogeneous models, however, suggests that it will be challenging to confidently make such predictions. Uncertainty about activity levels in low-activity population can become even more important when considering policies relax social distancing. Intuitively, if we are far from the herd-immunity region (given the new policy), then the relaxation will

[12] See, for example, **baqaee2020reopening**.

set off a substantial second wave. If we are already close to or in the herd-immunity region, then the second wave will be smaller or nonexistent. Where we are relative to the herd immunity threshold depends on the full set of R_{0i}, including the hard-to-estimate activity levels in populations that have seen few infections while activity is tightly restricted.

Figure **??** provides a numerical illustration. It shows the time paths that an epidemic would follow under the same nonconstant policy path in two heterogeneous SIR populations. The policy involves a severe lockdown, reducing activity levels by 65%, imposed gradually over a two-week period just as the epidemic is taking off, and a partial relaxation about a month later that allows activity levels to return to 70% of their pre-lockdown values. The left panel plots new daily cases. The right panel plots cumulative cases to date. The vertical lines mark the dates when the initial lockdown starts its phase in and the date on which it is relaxed. The epidemics rise at very similar rates in the two populations prior to the lockdown. They have similar declines once the initial severe lockdown is imposed. Indeed, in the right panel it is very hard to see any difference in the courses of the two epidemics up through the date at which the relaxation occurs.

Despite this similarity in the initial run up and through the lockdown, the two epidemics follow very different paths following the relaxation. As in the previous example, this reflects that the parameters were chosen so that activity levels in the less active subpopulations differ. In one population, whose outcomes correspond to the solid blue line, the relatively low activity populations have $R_{0i} = 1.5$. When we relax distancing rules, a large second wave takes off in these groups, infecting nearly three times as many people as had the first wave. In the other population, corresponding to the dotted red line, the low activity populations have $R_{0i} = 0.7$ and this makes the second wave much smaller. Difficulty in distinguishing the blue from the red population at the point when the relaxation is occurring will make it difficult to predict which future course we should anticipate.

Figure 3: Example of epidemics that diverge after a policy relaxation. The figure graphs new daily cases and cumulative cases for heterogeneous SIR models with $h = 0.7$ under a policy intervention involving a severe lockdown and a partial relaxation. Model 3 has $R_0 = (3.63, 3.63, 1.5, \ldots, 1.5)$. Model 4 has $R_0 = (3.55, 3.55, 0.7, \ldots, 0.7)$.

4 Potential Biases From Ignoring Heterogeneity

Many economic analyses of the COVID-19 epidemic build on the simpler homogeneous SIR model, even though heterogeneous models seem more natural. This section notes several ways in which ignoring or understating heterogeneity in contact rates may bias the conclusions these analyses reach.

4.1 Overstatement of the damage incurred in reaching herd immunity

COVID-19 may remain widespread until after we pass a herd immunity threshold. Understanding how many cases must occur before herd immunity is reached is critical to assessing policies that lead to herd immunity.[13] An influential paper by **ferguson2020impact** suggested that deaths with uncontained spread could be very high and **greenstone2020does** note that they correspond to extremely high economic costs under standard value-of-life assumptions. Two critiques of these calculations are that deaths may not be as high as the models suggest even without a government response due to endogenous social distancing, and that fatality rates could be lower due to asymptomatic cases. We note here another reason: models with heterogeneous activity suggest that herd immunity thresholds may be lower than naïve calculations based on homogeneous SIR models suggest.

In the homogeneous SIR model herd immunity is reached when $S = 1/R_0$, implying that the fraction of the population infected on the path to herd immunity must be at least $1 - 1/R_0$. If the system is instead described by the heterogeneous SIR model with uniform matching, then the naïve estimation of an R_0 parameter may lead us to misestimate the herd immunity threshold as

$$\hat{S} = \frac{1}{\overline{R}_0} = \frac{\sum_i R_{0i}}{\sum_i R_{0i}^2}.$$

This is indeed the threshold at which herd immunity is reached **if** the susceptible fraction is equal in all groups, but we can reach herd immunity with fewer infected by concentrating infections in the more active populations, and infections will naturally concentrate in the more active populations.

If $R_{0i} > 1$ for all i, then one state that obviously achieves herd immunity is to set $S_i = 1/R_{0i}$ for all i. The fraction susceptible is $\frac{1}{N}\sum_i 1/R_{0i}$. That this is always greater than $\hat{S}$ can be seen via an elegant two-step argument comparing both expressions to the reciprocal of the arithmetic mean of the

[13] Also critical to such calculations are an assessment of the extent to which we will overshoot herd immunity and the excess deaths that may occur due to exceeding hospital capacity.

R_{0i},

$$\frac{1}{N}\sum_i 1/R_{0i} = \frac{1}{1/\left(\frac{1}{N}\sum_i \frac{1}{R_{0i}}\right)} \geq \frac{1}{\frac{1}{N}\sum_i R_{0i}} \geq \frac{\frac{1}{N}\sum_i R_{0i}}{\frac{1}{N}\sum_i R_{0i}^2} = \hat{S},$$

with the two inequalities coming from the two parts of the root mean square-arithmetic-harmonic mean inequality. More important than the elegance is that the difference can be quite large in practical terms. For example, in our loosely calibrated five-population example with $R_0 = (3.5, 1.5, 1, 0.5, 0.5)$, the naive homogeneous SIR calculation gives $\hat{S} = 7/16 \approx 0.44$, suggesting that 56% of the population must be infected before herd immunity is reached. However, the $S_i = 1/R_{0i}$ state is in the herd immunity region and has just 21% of the population infected.

More generally, the maximum fraction that can remain uninfected at the herd immunity point is calculated by solving

$$\max_{S_1,\ldots,S_N} \sum_i S_i$$

s.t.

$$\sum_i \frac{R_{0i}}{\sum_k R_{0k}} S_i R_{0i} \leq 1.$$

The linearity of the objective function and constraint make clear that the optimal solution involves concentrating the infections in the highest activity groups, i.e. S_i equal to zero in the highest-activity groups, S_i equal to one in the lowest activity groups, with S_i perhaps at an intermediate level in some marginal group to make the constraint hold with equality.[14] This will require even fewer infections than the $S_i = 1/R_{0i}$ state. In the five-population example above, we can achieve herd immunity with just 14% of the population infected by fully concentrating infections in the highest-activity subpopulation. While the example is clearly very loosely calibrated, the fact that the true level of infection needed to reach herd immunity is just one-fourth of what a naïve homogeneous-SIR based calculation indicates that contact heterogeneity is potentially a very important consideration.

For another example that may provide additional intuition, consider the spread of an epidemic in a less-developed country that lacks adequate personal protective equipment for its health care workers. In such an environment, transmissions from COVID-infected patients to health care workers to patients who are in hospitals for other reasons could play a major role in disease transmission. Suppose that this transmission resembled that in a ten-group uniform matching model with $R_0 = (6, 1, \ldots, 1)$. The most-active group in this model could represent the health care workers. In the early stages of the

[14] **acemoglu2020multi** also include a discussion of targeting which point in the herd immunity region the system reaches.

epidemic any non-health care worker who is infected will infect on average one other, 0.4 health care workers and 0.6 non-health care workers. An infected health care worker will in turn infect six others, again with 40%-60% split between health care workers and others. If a homogeneous SIR model is fit to early growth of such an epidemic one would estimate $\hat{R}_0 = \overline{R}_0 = (6^2 + 1^2 + \ldots + 1^2)/(6 + 1 + \ldots + 1) = 45/15 = 3$ and infer that herd immunity will not be reached until two-thirds of the population is infected. In fact, herd immunity can be reached much more easily. The key is to stop the within-hospital transmission. If five-sixths of the health care workers are immune, then each new infection will lead to just one other. The health care workers are just one-tenth of the population, so we can reach herd immunity with just 8.3% of the population having been infected.

In the model with homophilic matching we achieve herd immunity by choosing $S_1, \ldots, S_N$ such that $hS_iR_{0i} < 1$ for all i and so that

$$\sum_i \frac{R_{0i}}{\sum_k R_{0k}} \frac{S_i R_{0i} - 1}{1 - hS_i R_{0i}} \leq 0.$$

Here, the herd-immunity point with the lowest total number infected again involves having a lower fraction susceptible in the more active groups, but the solution will typically not be to fully concentrate the infected. Although the initial change in the constraint from reducing S_i away from one is largest in the most active group, the marginal benefit of reducing the fraction susceptible decreases as the fraction susceptible in a group is reduced, which may make the solution interior in multiple populations.

Achieving herd immunity with homophilic matching is more difficult than achieving herd immunity with uniform matching. This follows directly from the contrapositive of the result noted at the end of section 2.3: if a disease free state is stable for any h', then it must also be stable for all $h < h'$. This implies that the herd immunity region for a model with homophily parameter h', i.e. the set of S^0 for which $(S^0, 0)$ is stable, is a subset of the herd immunity region for a model with parameter h. The minimum fraction of the population that must have been infected to achieve herd immunity is therefore monotonically increasing in h. Finding the minimum threshold is very easy in the $h = 1$ case: the model is essentially a set of separate homogeneous SIR models so the solution is simply to set $S_i = \text{Min}(1, 1/R_{0i})$ in each subpopulation. In our five population example with $R_0 = (3.5, 1.5, 1, 0.5, 0.5)$ this involves infecting 2/7 of those in subpopulation 1, 1/3 of those in subpopulation 2, and no others, which is the 21% of the total population mentioned earlier. For intermediate h one needs to solve the maximization problem described above, but we know the threshold increases continuously from 14% to 21% as h goes from 0 to 1. For $h = 0.5$ it is 15.5%.

An important factor to keep in mind when thinking about implications of results on herd immunity is that heterogeneous SIR models, like homogeneous SIR models, always "overshoot" their herd immunity thresholds in an uncontrolled epidemic. In the homogeneous SIR model, overshooting occurs because many are infected when herd immunity is reached and the infection is then reproducing at an approximately contant rate. For example, a homogenous SIR model with $R_0 = 16/7 \approx 2.3$ reaches its herd immunity threshold when just $1 - 7/16 \approx 56\%$ of the population has been infected but the infection will eventually hit about 87% of the population in an uncontrolled epidemic. In heterogeneous SIR models, uncontrolled spread infects more people than are infected in the minimal herd immunity state for two reasons: the same overshooting effect as before, and because the path of the infection does not concentrate infections in the high-activity population to as great a degree as does a path that enters the herd immunity region at the minimal-infection point. This additional source of excess infections can be extremely potent. For example, whereas the five population uniform-matching SIR model with $R_0 = (3.5, 1.5, 1, 0.5, 0.5)$ (which has $\overline{R}_0 = 16/7$) can be in the herd-immunity region with as little as 14% of the population infected, an infection starting from a small evenly distributed mass of infected will not reach the herd immunity region until 33% of the population is infected, and overshooting will result in 54% eventually being infected.

While I noted above that the minimal-infection herd immunity point entails more infections when matching is more homophilic, overshooting can be less extreme in homophilic models and when epidemics spread in an uncontrolled manner this can more than offset the difference in the herd immunity thresholds. For example, with the same R_0 vector as above, the fraction eventually infected is 46% with $h = 0.5$, 42% with $h = 0.75$, and just 31% with $h = 1$.

4.2 Overestimation of the difficulty of controlling an epidemic

While heterogeneous population models suggest that reaching herd immunity need not involve nearly as many infections as homogeneous SIR models suggest, they also suggest that avoiding herd immunity via selective lockdown policies may not be as difficult as homogeneous SIR models suggest.

Several recent papers have discussed optimal policies using frameworks in which transmission rates constant at time t can be reduced to $R_0(1 - x_t)$ by "locking down" a fraction x_t of the population.[15] This can reduce the fraction infected before a vaccine is developed, and reduce excess deaths from exceeding hospital capacity. In a homogeneous SIR model, lockdown policies that keep the population

[15] See **acemoglu2020multi**; **lippi2020simple**; **rowthorn2012optimal**.

from reaching herd immunity incur large economic costs because the fraction infected will grow unless we keep the initial x_0 large enough so that $1 - x_0$ is below the herd immunity threshold. As a result, some optimal-policy simulations suggest that we may mostly want to use lockdowns just to temporarily slow the epidemic when hospitals would otherwise be overwhelmed.

The lower herd immunity thresholds of homogeneous models imply that targeted permanent lockdowns could keep the fraction infected from ever expanding by locking down a smaller fraction of the population. For example, in the $R_0 = (3, 1.5, 1.0, 0.5, 0.5)$ example discussed in the herd immunity section, the problem of determining the minimum fraction the population that must be permanently locked down to keep the epidemic from ever expanding is mathematically equivalent to earlier calculation of the minimal herd-immunity threshold. Hence, the epidemic can be stopped by permanently locking down 14% of the population in the uniform matching case or at most 21% of the population in the homophilic model.

Temporary lockdowns can also be appealing in heterogeneous population models because they can serve as a means to guide the system toward a more desirable part of the herd-immunity region and/or reduce overshooting For example, to prevent the dramatic overshooting noted in the previous section, one could lock down all members of the lowest-activity populations once prevalence there reached a fraction of a percent, keep them locked down as the infection spreads through the most active populations, and then release them from lockdown once the population is close to herd immunity.

Figure **??** provides a numerical illustration. The solid blue series is the time path of new daily cases in a homogeneous population with $R_0 = 16/7 \approx 2.3$. The dashed red line is the time path of new daily cases in a heterogeneous uniform matching population with $R_0 = (3.5, 1.5, 1.0, 0.5, 0.5)$. Recall that the heterogeneous population has $\hat{R}_0 = 16/7$ and indeed the two series initially look identical. The infection in the heterogeneous population reaches herd immunity sooner and many fewer people are eventually infected than in the homogeneous population. But the lower damage absent a lockdown does not mean that the incremental benefit from a temporary lockdown is lower. The gray dashed line shows the path of the infection under a temporary targeted lockdown: we reduce activity by 20% in the highest activity populations and by 60% in the lower activity populations for a 60 day period. This reduces overshooting in the highest-activity population and reduces the number of low-activity people who are infected by members of the high-activity population as it is going through its peak. In the numeric example, it reduces the fraction who are ever infected from 54% to 38%.[16]

[16]In the homogeneous model, implementing the same policy would make cases decline during the lockdown period,

Figure 4: Effect of a temporary targeted shutdown in a heterogeneous population. The figure graphs new daily cases in three models. The dashed red line is a heterogeneous SIR model uniform matching and $R_0 = (3.5, 1.5, 1.0, 0.5, 0.5)$. The dashed gray line graphs cases for the same populaton assuming a temporary 60-day lockdown is imposed during the peak. The solid blue line is a homogenous SIR model with $R_0 = 16/7$ with no lockdown.

4.3 Overestimation of the impact of social distancing policies and endogenous behavioral responses

Cumulative US COVID-19 deaths grew roughly exponentially throughout March, passing 5000 on April 1st. Growth subsequently slowed dramatically. Many have noted that both government-mandated policies and endogenous individual reactions would be expected to contribute to the change.[17] Understanding the the causal impact of each factor is critical to forecasting the impact of reopenings. A third effect also contributes to slowing growth in SIR models—the growth rate $\gamma(R_0 S(t) - 1)$ decreases as $S(t)$ declines—but this effect will be small in homogeneous SIR models calibrated to current conditions because $S(t)$ remains close to one.[18] In heterogeneous SIR models, however, this third effect can be nontrivial even in the early stages of an epidemic, particularly if matching is homophilic.

The effect is easiest to quantify in the uniform matching model. If the susceptible fraction has been reduced to S and the infection is still small, the growth of the infection will resemble that of a homogeneous SIR model with parameter $\overline{R}_0(S) = \sum_i \frac{R_{0i}}{\sum_k R_{0k}} S_i R_{0i}$. Writing $\overline{S}$ for the average fraction susceptible, the dominant eigenvector implies that the relative frequencies in the early infected population will be roughly proportional to their activity levels so $S_i \approx 1 - N\frac{R_{0i}}{\sum_k R_{0k}}(1 - \overline{S})$. Differentiating with respect to $\overline{S}$ we find $\frac{d\overline{R}_0(\overline{S})}{d\overline{S}} = N \sum_i w_i^2 R_{0i}$, where we have written $w_i \equiv \frac{R_{0i}}{\sum_k R_{0k}}$ for the fraction

but there would be a massive second wave after the policy is lifted and the eventual total infected would only be reduced by about ten percentage points.

[17] See **baqaee2020reopening**; **farboodi2020internal**; **fernandez2020estimating**; **jones2020optimal**; **kudlyak2020for**. Epidemiological estimates of changes in growth rates include **miller2020full**; **unwin2020state**.

[18] A recent study in Sweden indicated that despite their embrace of herd immunity the fraction with antibodies is just 7% in Stockholm.

of early infections which are in population i. Focusing just on the effect due to reductions in the most active group we have

$$\frac{d\log \overline{R}_0(\overline{S})}{d\overline{S}} = \frac{N\sum_i w_i^2 R_{0i}}{\sum_i w_i R_{0i}} \geq \frac{N w_1^2 R_{01}}{\sum_i w_i R_{0i}} = w_1 \cdot \frac{w_1 R_{01}/\sum_i w_i R_{0i}}{1/N}.$$

Note that the first term in the product is population 1's share of early infections and the second is the ratio of population 1's contribution to $\overline{R}_0$ to its share of the total population. In extreme examples where almost all early infections are in one small subpopulation this effect can be very large. For example, $w_1 \approx 1$ in a model in which population 1 is just $1/N$ of the total population we have $\frac{d\log \overline{R}_0(\overline{S})}{d\overline{S}} \approx N$, i.e. the apparent $\overline{R}$ will have been reduced by about $N\%$ by the time 1% of the population has been infected. (In a homogeneous SIR model, the reduction in the apparent R_0 would be 1%.) The effect is smaller in uniform-matching models with less extreme heterogeneity in R_{0i}. For example, in the $R_0 = (3.5, 1.5, 1.0, 0.5, 0.5)$ example I have used frequently, $w_1 = \frac{1}{2}$, $\frac{w_1 R_{01}}{\sum_i w_i R_{0i}} \approx \frac{3}{4}$, and $1/N = 0.2$, so $\frac{d\log \overline{R}_0(\overline{S})}{d\overline{S}} \approx 2$. This suggests that the apparent $\overline{R}_0$ will have been reduced by about 10% when cumulative infections have reached 5%. This is larger than the 5% prediction of a homogenous SIR model, but is not a dramatic difference.

The reduction in the apparent $\overline{R}_0$ can be much larger in models with homophilic matching because infections and loss of susceptibility are both more concentrated in the highest activity groups. For a simple illustration, think of a model with $h \approx 1$. Here, the power of exponential growth means that if we start from a tiny fraction infected in each group we will soon have almost all of the infected in the highest-activity group. As a result, we can perceive the growth process early in the epidemic to be close to R_{01} growth. If the infection peaks and declines in population 1 before it reaches a substantial size in population 2, the apparent growth rate can temporarily fall to well below one even though the epidemic is still in its early stages. Growth will then rise back to look like R_{02} growth in a second wave, and so on. Such nonmonotonic growth rates only occur when h is very close to one, but the fairly rapid early decline in the apparent growth rate as the epidemic burns itself out in the highest-activity population is a feature that persists well away from the $h = 1$ limit. If we take $h = 0.7$ in the $R_0 = (3.5, 1.5, 1.0, 0.5, 0.5)$ example, growth that looks like $\overline{R}_0 \approx 3$ growth early in the epidemic will slow to what looks like $\overline{R}_0 \approx 2.5$ growth by the time 5% of the population has been infected. Almost 20% of the highest-activity population is no longer susceptible at this point, and this substantially reduces the epidemic growth rate.

The slowdown of an epidemic continues as it approaches and passes the herd immunity threshold.

Hence, viewing an epidemic in light of a homogeneous SIR model can both lead one to to mistakenly conclude that initial behavior changes were more effective than they were at slowing the epidemic and that later reopenings caused less acceleration than they did.

4.4 Underestimation of heterogeneity in R_0 across regions

The SIR parameter R_0 reflects both the contagiousness of a disease and the frequency and closeness of interactions in a population. It seems natural that R_0 should be larger in some countries or states than in others. For example, we might expect it to be larger in more densely populated and highly urbanized Belgium than in Sweden. But few economic analyses incorporate heterogeneity in R_0 across regions. This presumably reflects at least in part that the early epidemiological literature did not provide clear evidence of cross-country or cross-region differences. For example, **flaxman2020report** provided estimates for 11 European countries from the period before lockdowns went into effect, and the 50% credible interval for Sweden (roughly 3.7–4.3) overlaps with the 50% credible intervals for 9 of the other 10 countries including Belgium.[19]

The limited heterogeneity in reported R_0 parameters could reflect what is estimated when one applies a homogeneous SIR model to a heterogeneous world with homphilic matching. As an illustration, suppose that the differences between two countries lie not in differences between activity vectors $(R_{01}, R_{02}, \ldots, R_{0N})$, but in the fact that country a has a higher fraction of its population in the high activity groups than does country b. For example, it may be that both countries have working class subpopulations living in crowded urban housing and riding public transportation to jobs where they work in close proximity to others and rural populations with much lower contact rates, with the primary cross-country difference being in the relative fractions in each group. In an extreme homophilic model with $h \approx 1$, an estimation of R_0 would yield identical estimates of $\hat{R}_0 = R_{01}$ in both countries, regardless of whether important differences in the population compositions were present.

Again, these differences persist well away from the $h = 1$ limit. For example, with $h = 0.5$ a model with five equal sized populations with $R_0 = (3.5, 1.5, 1.0, 0.5, 0.5)$ will resemble $\overline{R}_0 \approx 2.75$ growth early in the epidemic, whereas a model with he same R_0 vector, but in which the three most active populations are each 10% of the population rather than 20% will resemble $\hat{R}_0 = 2.5$ growth. Homogeneous SIR epidemics with $R_0 = 2.75$ and $R_0 = 2.5$ follow similar paths – herd immunity is

[19] **unwin2020state** estimates a more flexible model with more recent data and reports much more substantial heterogeneity across US states, as do **fernandez2020estimating**. There is also a substantial range in early estimates of the rate at which COVID-19 spread in China.

reached when 64% are infected in one model vs. 60% in the other and with overshooting the epidemics eventually infect 93% and 90%. It would be natural to not bother to incorporate such differences in an economic analysis of homogeneous SIR-based models. But the two heterogeneous models follow quite different paths with one eventually infecting 49% of the population and the other eventually infecting 29%. Accounting for the potential impacts of such differences seems much more important.

4.5 Misestimation of when epidemics start

A number of early papers fitting SIR models produced estimates of when epidemics started. In addition to satisfying intellectual curiosity, one motivation for such an exercise is that it may provide evidence on the size of the asympotomatic population. Features of the heterogeneous SIR model suggest that it will be very difficult to produce reliable estimates via such a method. Specifically, while I emphasized earlier that a heterogeneous SIR model will appear to grow at a rapid pace $\overline{R}_0$ from quite early on, this is not true at the very, very beginning. The infection only starts to grow at a rate related to the largest eigenvalue once the pattern of the distribution of infections across the populations is aligned with the principal eigenvector. Before this occurs, the growth rate can be very different depending on whether the initial infections are in a low- or high-activity population. This makes early growth rates unpredictable, and makes inferences about when an epidemic started very imprecise.

One of the most influential and inaccurate early papers on the COVID-19 epidemic may have misled in part this reason. **lourencco2020fundamental** calibrated an SIR model to estimate the fraction of the UK and Italy populations who were already infected as of March 19. In the three primary scenarios included in Figure 1, they estimate that the start of the UK epidemic occurred about 30 days before the first reported death, and then project forward to estimate that between 36% and 67% of the UK population was already infected as of March 19. One reason for the inaccuracy is that the analysis assumed that the death rate was much lower than now appears to be the case. Another source of the inaccuracy, however, may be another pair of assumptions—that deaths do not occur until well after infection and that the time series of infections followed an SIR path with R_0 equal to 2.25 or 2.75 from the very beginning.

In addition to being imprecise, homogeneous SIR-based inferences about epidemic origins may be biased. Growth rates were probably lower in the very early days than they were by the time the epidemic grew to the size where estimates of R_0 were first made. This difference may help reconcile why the fraction that antibody tests indicate have ever been infected is not larger, despite revelations

that there was a case in France in late December and a death in California on February 6.[20] It may also help account for why some models, e.g. that shown in Figure 3 of **baqaee2020reopening**, find it difficult to match data on deaths from very early in the epidemic.[21]

5 Implications and Conclusions

The most basic message of this paper is that thinking about an epidemic in terms of homogeneous SIR models can lead to mistaken conclusions if the interactions are better described by a model with heterogeneous contact rates. Incorporating at least some heterogeneity need not be so difficult—in many cases what is being done with a single population model could be done quite similarly in a multipopulation model. But the remarkable pace at which the economics literature on COVID-19 has been progressing makes keeping up with the state of the art sufficiently difficult that my primary hope is that others will take the "heterogeneity matters" message to heart and incorporate it in their work.

Early in any epidemic, there is a great deal of scientific uncertainty about the disease transmission process. This paper's most important message about the COVID-19 epidemic itself is that to the extent that the epidemiological literature is suggesting that heterogeneity in transmission may be important, economists should be cognizant that we may still not understand its dynamics very well. Estimates of R_0 derived from observations in the early days of the epidemic may not reflect how the epidemic would spread now absent restrictive policies, and it is particularly difficult to estimate the parameters describing how the virus is spreading in less active communities. These parameters are critical to understanding how the epidemic may progress as restrictions are loosened. As questions about the impact of reopening policies become most salient, recognizing our limitations and doing our best to estimate the hard-to-estimate parameters is important. The greater speed with which the apparent R_0 can decline in heterogeneous models, particularly when matching is homophilic, also suggests that there may be more uncertainty than has been assumed in estimates of the impact both of distancing policies and of reopenings. The natural directions of bias are that we may overstate the impact that initial shutdown policies had in slowing the spread of COVID-19 and underestimate the extent to which the partial relaxations have accelerated the spread. It is particularly important to keep these biases in mind when estimates obtained in some region are used to provide advice to

[20] **worobey2020emergence** provide genome-based evidence that later early cases were not part of the main epidemics in Washington and Italy.

[21] Data inaccuracies may, of course, also be relevant here, so it is possible that the model predictions are closer to the truth than are the data.

others.

A more optimistic implication of heterogeneous SIR models is that the COVID-19 epidemic may not be as bad as some models suggest. Models using growth rates estimated in the early days of the epidemic may overstate how rapidly the epidemic would have spread absent government intervention even if people had not taken it upon themselves to socially distance. And it is possible that epidemic growth can be slowed by herd immunity effects at prevalence levels substantially lower than naïve models suggest. If so, the option of reaching herd immunity, becomes less unattractive, particularly if the herd immunity level being contemplated is that which applies when cost-effective mitigation measures, such as universal mask wearing, are maintained, and if extensive efforts are made to keep infections out of vulnerable populations along the path. The possibility that the impact of restrictive policies may have been overestimated also suggests that some partial reopenings may be less damaging than anticipated.

Another important conclusion, however, is that the optimistic message that reaching herd immunity may not be as damaging as feared should not be taken to imply that trying to reach herd immunity is more advisable than earlier analyses suggest. Models with heterogeneity also suggest that controlling the spread of COVID-19 may be easier than thought. For one thing, benefits similar to those which herd immunity provides can be obtained by implementing targeted measures to prevent high-contact people, e.g. health care and nursing home workers, those riding public transportation, etc., from ever being infected. This makes measures such as ensuring nursing home workers have adequate personal protective equipment even more powerful and cost-effective. And heterogeneous models also suggest that both permanent and temporary targeted lockdowns may be more effective as a means to limit the spread of the epidemic than homogeneous SIR models suggest. The good news on both sides of the equation makes it possible that correctly accounting for heterogeneity in the epidemic process could bolster the case for keeping in place policies that shut down high spread activities. Obviously, it would be valuable to know more about the nature of contact heterogeneity (and about the long-run health consequences of COVID-19 for survivors) to make this assessment.

I also noted that estimating SIR models on data early in an epidemic may lead one to underestimate the extent to which critical parameters of the epidemic process differ across regions. The changes in the course of epidemics in the aftermath of severe lockdown policies indicate that in aggregate policies and behavioral changes had a very large impact on R_0. It seems likely that to avoid a resurgence during the reopening process, we will need to retain some restrictions. The limits to what is safe, however, could be very different in different locations and at different times. It would be valuable to

have tailored guidance so that we do not simply have to rely on trying to infer the effect of each set of incremental changes.

The recent economics literature includes several papers examining multipopulation SIR models including **acemoglu2020multi**; **baqaee2020reopening**; **favero2020restarting**. While the analyses in these papers have not been calibrated to fully capture within age-group heterogeneity in contact rates, they certainly could move in this direction. **baqaee2020reopening**, for example, could in theory have "simply" used age $\times$ occupation groups instead of age groups as the basis of their model, replacing 5×5 matrices with 330×330 matrices, to capture contact heterogeneity across those working in each of the sectors they consider. This still, however, would not have captured within-occupation heterogeneity.

In addition to the computational challenges, a factor that will limit our ability to calibrate more complex models is the limited data that is available on heterogeneity in contact rates. Just as serology data has helped compensate for the weak identification of asymptomatic cases in regular SIR models, more data may also provide the solution to the weak identification problem noted here. Although several firms have already made location tracking data available to researchers, privacy concerns have limited public releases to means within various cells. One simple step that could potentially greatly enhance the value of this data is to also release within-cell variances and within-individual time series correlations. While those developing apps that use Bluetooth interactions to track phone-to-phone proximity are rightly being careful with privacy, they could potentially provide an even more valuable source of information on contact distributions. Epidemiologists are also able to exploit variation in virus genomes to provide more micro-based estimates of disease-transmission.[22] While economists are unlikely to have the expertise to take advantage of genomic data, keeping current on insights coming out of these analyses will be important.

[22] See, for example, **miller2020full** and **worobey2020emergence**.