Supplementary information for

Quarantine fatigue thins fat-tailed coronavirus impacts in U.S. cities by making epidemics inevitable

Place-based measures of contact rate distribution

To generate our contact rate distributions, we rely on cell phone data from the Safegraph weekly patterns dataset, which aggregates data from 45 million mobile devices in the United States, and visits to 6 million 'Points-of-Interest' (POIs). We build contact rate distributions based on visits to each POI in the dataset as follows:

Contact Rate = ADV*CR*DF*s

ADV is the average daily visitors to the point of interest during the week. CR is the contact radius, a proxy for the crowdedness of a POI, defined as some radius within which transmission could occur, expressed as a fraction of the POI's square footage. In practice we use 10 and 20 feet and find similar results. DF is the median dwell time in a given POI expressed as a fraction of the hours a POI is open, and s is a scaling factor that scales up to the population size given the fraction of devices observed by Safegraph. This effectively makes our contact rate equivalent to the expected number of people that would come within a certain radius of a visitor to a given POI during the week in question, assuming visitors are equally likely to go anywhere within the POI at any time that it is open. Thus, the uniform mixing assumption of SEIR models holds within POIs in our model, but our approach allows for more complex contact patterns across POIs within a CBSA.

To generate a complete 'place-based contact rate distribution', we also need to account for contacts within the home, which have shown to be an important source of transmission (Lee et al. 2020). Using the Safegraph social distancing dataset, we calculate the average fraction of devices that remain completely at home for a day during the week. We then make that fraction of the POIs in our distribution 'homes' and assign them a contact rate of the average household size in the CBSA. We then calculate the sample mean and variance of these distributions, and estimate the shape parameter of the tail of the distribution using maximum likelihood on the upper 50th percentile of the POI distribution (Grimshaw, 1993). We confirm these results using the mean-excess plots to estimate the shape parameter (Ghosh and Resnick 2010). We also do not estimate shape parameters for CBSAs with very few POIs (less than 21). We do this for 34 weeks for nearly 2,300 CBSAs resulting in parameters for more than 75,000 distributions.

Correlations between contact rate parameters and epidemiological dynamics

We run regressions using the following specification:

$$Y_{i,t+1} = \beta_1 Y_{i,t} + \beta_2 \Omega_{i,t} + \beta_3 Y_{i,t} * \Omega_{i,t} + \beta_4 X_{it} + \mu_i + \gamma_t$$

Y is the log of the outcome of interest, either Covid-19 cases or deaths per 100,000 population, in CBSA i during week t. For Covid-19 deaths and case data, we use the Johns Hopkins Covid 19 data repository from https://github.com/CSSEGISandData/COVID-19 (Dong, Du, and Gardner 2020). Given the large number of zeroes, we use the inverse hyperbolic sine transformation. Ω is a vector of variables describing the distribution of the log of the contact rate, including the mean, variance, and shape parameters. We also include these contact rate variables interacted with the lagged case or death rate, which tells us how their marginal effects change as cases/deaths increase. $\mu_i$and γ_t are CBSA and week fixed effects, and X_{it} is a vector of control variables, including the fraction of

devices observed in a CBSA relative to the months before the epidemic. Regressions are also weighted by CBSA population.

All regressions have high goodness-of-fit as measured by R squared, especially the regressions on case rates, and the coefficients on our contact rate distribution variables are highly significant, showing that changes in the mean, variance, and tail behavior of our constructed contact rate all have predictive power.

These results are robust to changing the size of the lead on the outcome variable, alternative definitions of the contact radius, as well as breaking out stay at home behavior and household size separately, and constructing a contact rate distribution based solely on visits to POIs outside the home.

Stochastic SIR model with fat-tailed contact rates

To determine if the impacts of the outbreak are fat-tailed, we use a stochastic variant of the standard SIR framework, which tracks the numbers of susceptible, infected, and recovered individuals over the course of an infectious disease outbreak (Kermack and McKendrick 1927, Hethcote 2000):

$$dS(t) = -\beta C(t) I(t) S(t) dt \tag{1}$$

$$dI(t) = \left[\beta C(t) I(t) S(t) - \frac{\gamma}{1-m} I(t)\right] dt \tag{2}$$

$$dR(t) = \gamma I(t) dt \tag{3}$$

where $I(t)$, $S(t)$, and $R(t)$ are the proportion of the total population N that is infected, susceptible, and recovered respectively and $\frac{\gamma m}{1-m} I(t)$ is the daily change in the percent of the population that dies due to COVID-19. Assuming frequency dependent transmission, the per capita infectiveness of COVID-19 is $\beta C(t)$ which captures both the natural infectiveness of the disease, β, and the rate of contact between individuals in a population $C(t)$.

The contact rate is a stochastic variable reflecting the unpredictability of contacts between individuals captured in our contact rate distributions. While human behavior makes contact rates inherently unpredictable (Sims et al. 2013), our place-based contact rate measures also highlight how a CBSA's mix of businesses and building stock also enable or discourage variability in contact rates. For example, a CBSA with more superspreader points of interest such as full-service restaurants, fitness centers, and cafes (Chang et al. 2020) will also exhibit greater variability in contacts between individuals. To convert the place-based measure of contact rate to a contact rate compatible with our compartmental model, we assume the contact rate in CBSA i,$C_i(t)$, is the expected contacts per person in CBSA i: $C_i(t) = \frac{1}{2}\left[\frac{\mu_{i,t}^2 + v_{i,t}}{\mu_{i,t}} - 1\right]$ where $\mu_{i,t}$ and $v_{i,t}$ are the mean and variance of the place-based contact rate distribution respectively.

We use this time-series of contact rates to estimate a stochastic volatility model for contact rates for the four most populous U.S. cities. To inform the specification of our model, we first performed an augmented Dickey-Fuller unit root test to determine whether the time series of contact rates is a random walk or trend stationary. For all 4 CBSAs, we fail to reject the null hypothesis of a unit root thus ruling out a variety of mean-reverting processes. Based on this finding:

$$dC_i = \alpha_i C_i dt + \sigma_i(t) C_i dz_{1i} \tag{4}$$

where α_i is a drift term, $\sigma_i(t)$ is a diffusion or variance term, and dz_{i1} is the increment of a standard Wiener process which are independent across CBSA. $\alpha_i > 0$ indicates that, on average, the contact rate is rising over time in CBSA *i*. More unpredictable human behavior would manifest as a more variable contact rate and a larger value for $\sigma_i(t)$. If $\sigma_i(t)$ were a fixed parameter, contact rates would be log-normally distributed leading to thick-tailed contact rates, $\xi = 0$. However, all 4 CBSAs exhibited fat tails ($\xi > 0$) throughout the 34 weeks of our study. To allow for fat tails, the diffusion or volatility parameter is also stochastic and follows an arithmetic Ornstein-Uhlenbeck (or AR(1)) process:

$$d\sigma_i = \theta_i(\bar{\sigma}_i - \sigma_i)dt + \kappa_i dz_{i2} \tag{5}$$

where $\bar{\sigma}_i$ is the long-run average level of percent volatility in contact rates, θ_i is the speed of mean reversion, $\kappa_i$is the volatility in the percent volatility, and dz_{i2} is a second independent Wiener process. The model in equations (4) and (5) ensures $C_i(t)$ will be fat-tailed with finite variance and can exhibit infinite variance with certain parameter combinations (Stein and Stein 1991). Unlike stochastic SIR models with thin tails, the stochastic SIR model in equations (1)-(5) may not converge to its deterministic counterpart, because the Central Limit Theorem fails with infinite moments (Fukui and Fukuwara 2020).

We estimate the drift term and the long-run average level of percent volatility as:

$$\widehat{\alpha_i} = \frac{1}{N}\sum_{t=1}^{N} log\left(\frac{C_{ti}}{C_{t-1i}}\right) + \frac{\hat{\sigma}_i^{\,2}}{2}$$

$$\hat{\sigma}_i = \sqrt{\frac{1}{N-1}\sum_{t=1}^{N}\left(log\left(\frac{C_{ti}}{C_{t-1i}}\right) - \frac{1}{N}\sum_{t=1}^{N} log\left(\frac{C_{ti}}{C_{t-1i}}\right)\right)^2}$$

Our estimates for the drift term range from 0.019 in Chicago to 0.002 in New York City. Our estimates for the long-run average level of percent volatility range from 0.21 in New York City to 0.04 in Houston. Together, these parameter values provide an indication of the degree of quarantine fatigue and unpredictability in each city. For example, contact rates in New York City increased 1.9% each week with a 21% volatility around this trend.

Our weekly contact rate data do not provide enough observations to confidently estimate an AR(1) model needed to recover estimates of θ_i and κ_i. However, the relationship between these parameters and results from the augmented Dickey-Fuller test give us several clues about their relative magnitude. The unconditional standard deviation of volatility in our model is given by $\kappa_i/(2\theta_i)^2$. Thus, small values of κ_i can have a large effect on the data generating process if the speed of mean reversion is low. Conversely, large values of $\kappa_i$can have a small effect on the data generating process if the speed of mean reversion is fast. In each CBSA, we set $\kappa_i = \frac{1}{2}\bar{\sigma}_i$. The results of our augmented Dickey-Fuller test suggest $\theta_i$is relatively large to ensure the random walk properties of the data are preserved. Given $\kappa_i = \frac{1}{2}\bar{\sigma}_i$, we select values for θ_i in each CBSA that ensures consistent rejection of the null hypothesis of a unit root. These values imply that the half-life of a volatility shock ranges from 2 weeks (θ_i=19.22) in New York City to 6 weeks (θ_i=3.82) in Houston. Our parameter values imply that the unconditional standard deviation of volatility ranges from 0.02 in New York City (8% of its mean) to 0.008 in Houston (18% of its mean). Our results do not qualitatively change under various combinations of θ_i and $\kappa_i$values that results in rejection of the unit root null hypothesis.

We use weekly case counts for each of the 4 CBSAs to estimate the epidemiological parameters β_i, γ_i, and m_i that minimize the sum of squared errors between the weekly observed case counts in a CBSA and the expected path of $I_i(t)$ for that CBSA over a period of 34 weeks. For Covid-19 deaths and case data, we use the Johns Hopkins Covid 19 data repository from https://github.com/CSSEGISandData/COVID-19 (Dong, Du, and Gardner 2020).

We then perform 100,000 simulations of the fitted stochastic SIR model using simbyEuler in Matlab. To alleviate concerns about under reporting of cases early in the outbreak, we set the initial condition for the simulation equal to the week where the percent infected in a CBSA first exceeds 0.0005. This initial condition is as early as March 23, 2020 (New York City) and as late as April 13, 2020 (Houston). We then calculate the cumulative sum of cases and deaths for each simulation as a proxies for the damages incurred by COVID-19. In each week, we fit a GPD to the 100,000 simulated cumulative cases and deaths yielding a weekly estimate of the thickness of the tails (shape parameter, ξ_i) for the cumulative impacts of COVID-19 in each CBSA. We use the gpfit.m routine in Matlab to fit the GPD to the cumulative cases.

The stochastic epidemiological system in (1)-(5) has several attractive features. First, it assumes the current proportion of the population infected, susceptible, and recovered is known but the future course of the outbreak is unknown due to the inability to predict future behaviors that determine the contact rate of a population. This approach ensures that the fat tails we find are due to extreme draws in individual behavior due to factors such as superspreaders. While uncertainty in current cases and deaths (i.e., state uncertainty) are important sources of uncertainty to consider when developing testing protocols, and other public health responses, they can lead to fat tails in cumulative cases due to intermittent changes in testing efforts and individual willingness to volunteer for testing. For example, an "extreme draw" for COVID-19 cases would be expected prior to Thanksgiving and Christmas as individuals seek out testing as part of quarantine procedures prior to visiting family.

Second, fat tails in the contact rate distribution also correspond to fat tails in the average number of secondary cases per infectious case which has been shown to exhibit fat tails (Wong and Collins 2020). Our stochastic SIR model implies the effective reproductive number is stochastic since $C_i(t)$ is stochastic: $R_i^e(t) = \frac{\beta_i C_i(t)(1-m_i)}{\gamma_i}$. Ito's Lemma ensures that if $C_i(t)$ is fat-tailed, $R_i^e(t)$ will also be fat-tailed.

Third, unlike most stochastic epidemiological models that result in a single noisy wave of COVID-19 cases (Allen 2008), the stochasticity in $C(t)$ leads to multiple waves of cases whose timing and magnitude are unpredictable. A single peak in cases is not consistent with the vast majority of CBSAs in our study.

SI References

Allen L.J.S. An introduction to stochastic epidemic models. In: Brauer F., van den Driessche P., Wu J., editors. Mathematical epidemiology. Vol. 1945. Springer; Berlin: 2008. pp. 81–130. (Lecture notes in mathematics). Ch. 3.

Chang, Serina, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David Grusky, and Jure Leskovec. "Mobility network models of COVID-19 explain inequities and inform reopening." *Nature* (2020): 1-6.

Dong, Ensheng, Hongru Du, and Lauren Gardner. "An interactive web-based dashboard to track COVID-19 in real time." *The Lancet infectious diseases* 20, no. 5 (2020): 533-534.

Ghosh S, Resnick S. 2010. A discussion on mean excess plots. *Stochastic Processes and their Applications*120:1492–1517.

Grimshaw SD. 1993. Computing maximum likelihood estimates for the generalized pareto distribution. *Technometrics* 35:185–191.

Hethcote, H.W. 2000. "The mathematics of infectious diseases." *SIAM review* 42:599-653.

Hosking JR, Wallis JR. 1987. Parameter and quantile estimation for the generalized pareto distribution. *Technometrics* 29:339–349.

Kermack, W.O., and A.G. McKendrick. 1927. "A contribution to the mathematical theory of epidemics." *Proceedings of the royal society of london. Series A, Containing papers of a mathematical physical character* 115:700-721.

Lee, Elizabeth C., Nikolas I. Wada, M. Kate Grabowski, Emily S. Gurley, and Justin Lessler. "The engines of SARS-CoV-2 spread." *Science* 370, no. 6515 (2020): 406-407.

Sims, Charles, David Finnoff, and Suzanne M. O'Regan. "Public control of rational and unpredictable epidemics." *Journal of Economic Behavior & Organization* 132 (2016): 161-176.