Peer Review File

Manuscript Title: Mobility network models of COVID-19 explain inequities and inform reopening

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referee #1 (Remarks to the Author):

This paper fits a bipartite network mobility-informed SEIR transmission model using data on mobility from SafeGraph and Google to data on cumulative confirmed cases from 10 MSA in the US using a highly structured model, and then uses the results to consider counterfactual scenarios related to reopening and in particular which venues are important for transmission and what disproportionate impact there may be by SES.
As a first point, the authors should be commended, thanked, and somehow rewarded for producing one of the most crystal-clear explanation of a complex model that I have ever read. The order is intuitive; the jargon is minimal and correct when used; the technical aspects are explained clearly without obfuscation and with a clear intent to communicate. I wish I could say this is common in infectious disease epidemiology, but this paper really is exceptional.

My overall evaluation is that the paper really has two nearly unconnected aspects. The first is model fitting, which shows that a model with three parameters fits each of 10 time series better than a model with 1 – in fact it is less impressive still than this, because even if the number of parameters were the same, the fact that the epidemic slows down when mobility is quashed by social distancing is not a surprise, but is a further contrast between the models that is a bit more radical than just the number of parameters – one takes account of known basic mechanisms, the other leaves one (human contact) out. While not surprising, this part of the paper is encouraging, providing quantitative support for what one would expect.

The second part of the paper – the part about counterfactual interventions – uses the exquisitely strong (albeit very reasonable -- probably where I would have also started) assumption of the model – that transmission risk in a place (separate for home block group and other places, POI) is proportional to the product of time spent there and density of infected individuals per unit area – and draws conclusions about which venues are responsible for most transmission, which groups are at highest risk in various phases of the epidemic, and so on. These are perfectly reasonable conclusions to draw from a model, and impressive to draw from a simple SEIR model (I find this paper methodologically innovative for this reason – boiling down agent-based models to very simple components). But there is nothing in the paper that suggests this assumption (transmission proportional to time x density of infected) is correct. It is reasonable and may well be, but the entire fitting exercise tests only the use of mobility data at all, not the parametric form or detailed assumptions which all the conclusions are based. Therefore, in its current form, the paper goes to great effort to fit a model to data, then uses aspects of that model that are entirely untested to make its practical conclusions. To a large degree, the well-supported and innovative parts of the paper are separate.

How could this part of the model be well supported? I'm not completely sure, but here are some thoughts:

1) Try fitting the same model with the same optimization routine, but with random permutations of the block groups, so that the POIs are visited by randomized residents of the MSA, rather than their true visitors. This would be a model that incorporates trends in mobility, but scrambles the structure of home

and outside-home. If this fits much worse, then there is something meaningful in the model; if not, it may just be capturing gross patterns in mobility.

2) Try fitting the same model with (say) density-independent risk at all POIs (essentially collapsing time use into a single dichotomy of out of home vs. in home neighborhood). There may be several variants of this strategy, for example (Have not thought deeply, maybe one or both of these is not quite right, but the general idea is) removing the area term or both the area and duration terms. The goal is to figure out if the data contain any signal that the time and density metric is important.
My suspicion is that since there is no fitting to sub-MSA data, and the epidemic is not too close to saturated (I think, see below), that the first permutation will fit pretty well, undermining the evidence for the particulars of the model. The second maybe not – and if so, I think (need to think more) this would provide evidence for the fundamental assumption of the model that is used to generate the counterfactual predictions.

3) Most convincing would be to find an independent way to test the predictions of the relative importance of various types of POI to transmission. Perhaps from contact tracing data, though these are hard to come by.

Other significant concerns:

1) The data are cumulative cases, and it is bad statistical practice to fit to cumulative cases. This article, titled "Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola" https://royalsocietypublishing.org/doi/full/10.1098/rspb.2015.0347
Contains a section titled "Deterministic models fit to cumulative incidence curves: a recipe for error and overconfidence."

2) Reporting fraction fixed at 10% throughout the period is surely highly error prone given changing testing and a lot of variation across the country -- is there some way to do better?

3) The estimate of 10^-5 to 5 10^-4 for prevalence of E (if I understand p0) on March 1 seems a bit high for many cities. Given more than 1000x increase in cumulative cases in nearly all cities over that time, doesn't that imply an unrealistically high attack rate? Perhaps this is a misinterpretation.

4) The Safegraph to Google comparison is incomplete. One city is shown, for four categories (Fig. S3), but high Pearson correlation is not really relevant since it is absolute changes that matter; deviation from the 1:1 line would be more impressive, and this seems quite deviated for some categories. What was the purpose of this validation, and did the data pass the validation in all cities? What would the conclusion be if the data sets told different stories?

Minor concerns:

1) Would the authors be willing to comment more on why the model fit substantially overestimates cumulative death counts (and to some degree, cumulative cases) for Dallas, Houston, Miami, and San Francisco (Extended Data Figures 1 & 2)?

2) The figure x-axes labels are slightly cut off in Figure 3b.

3) Line 209: Additionally, it is not clear whether each Safegraph device corresponds 1:1 to an individual (e.g. an individual may have multiple devices simultaneously).

3) Line 451: The reasoning on the range of R_base refers solely to within-household transmission but the descriptions should be amended to note that R_base will include the effect of POIs not covered by

Safegraph, such as subway stations (as mentioned in Line 771), nursing homes (line 208), etc.

4) Line 683: The correction factor is estimated using an overall ratio of US population to total # of devices - is it possible to check what the correction factor is separately for each MSA and confirm they are all similar enough to 7?

5) Line 830: What was the correlation for the next highest MSA? The Philadelphia differences in risk by income (Figure 3A) and race / ethnicity still seem hard to explain - observed incidence rates disparities have not been nearly so severe. Could the fact that the income distribution for Philadelphia is particularly right-skewed versus other cities also play into any of this (similarly, median income is the lowest of all cities)? The more worrying possibility is some artifact in the Safegraph data that could lead to this.

Referee #2 (Remarks to the Author):

Chang et al. use a simple SEIR model combined with large scale human mobility data to estimate the epidemic trajectory in 10 large urban areas in the United States. They use this model, which is primarily calibrated on human movement data to extrapolate and recommend reopening strategies. Even though this work is of potential interest in its current form I have some methodological reservations and I am concerned with the interpretation of the results.

First, the major methodological issue I can see is that models are evaluated against cumulative case counts. King et al. show convincingly that fitting SEIR models to cumulative case counts can lead to overconfidence in the observed parameters and fit of the model. Please see a detailed description here: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4426634/. I would recommend to redo all analyses with daily case counts.

Second, the authors use delay distributions between infection and confirmation and that reporting rate is fixed at rc = 0.1 (time invariant). I would recommend delay distributions to be drawn from empirical line list data as described in previous work (see some work here https://www.bmj.com/content/369/bmj.m1923.long).

Reporting rates have changed dramatically through time. Especially the lack of available testing and changing testing protocols. The covid tracking website is a good resource (https://covidtracking.com/api). Detection rates used in the presented work may be relevant early in an epidemic.

Similarly, the authors use 18 days between day of infection to death. This in fact is highly variable and should be accounted for. The work cited by the authors is now outdated. Please see earlier work by Flaxman et al. Supplementary Figure 1 (https://static-content.springer.com/esm/art%3A10.1038%2Fs41586-020-2405-7/MediaObjects/41586_2020_2405_MOESM1_ESM.pdf).

The authors state: "As a sensitivity analysis, we alternatively stochastically sampled the number of confirmed cases and the confirmation delay from distributions with mean rc and c, but found that this did not change predictions noticeably." However, no Figure or formal evaluation is provided.

As far as I understand this is a highly complex and multidimensional model where each location has a SEIR model with well mixed populations and each POI has some probability of transmission when people get together. How confident are the authors that the estimated rates of infection can be attributed to these POIs? A section on potential issues of identifiability would be helpful for the reader not familiar with these high dimensional network models. Further, I am concerned with the section lines 278 - 288.

The authors arbitrarily exclude outliers which are not shown in the manuscript.

The unexplained variation in infection rates it seems is absorbed by the base rate of new exposures not captured by visiting POIs. Does this rate vary over time? If the authors model is correct the base rate should increase as more people are staying home? Is that the case?

The role of restaurants in the transmission dynamics of COVID is interesting. However, the authors model relies on human mobility and dwell times and the model fit needs further evaluation based on daily case counts rather than cumulative counts. For example, could the authors provide summary statistics of the dwell times and frequency of visits similar to Figure 2d for all the different locations? If they scaled the same way it would make clear that the findings are more related to the underlying nature of visits rather than locations itself. That would change the story of the paper.

I like Figure 2a, second / third panel. These clearly indicate that early reductions in mobility matter a lot. Please include references to earlier work from China Tian et al. 2020 (Science) & Lai et al. 2020 (Nature) showing similar results.

A few additional comments:

The authors do not account for potential re-seeding of the epidemic into major metropolitan regions. Especially for the period until lockdowns were implemented in the USA these could have played a substantial role in transmission. Please see visualisations (like nextstrain.com) of genomic data flows in the US during the COVID pandemic.

The authors could put their findings in context of detailed transmission chain evaluations from Iceland: https://www.nejm.org/doi/full/10.1056/NEJMoa2006100.

Further mention of the age groups that become infected or symptomatic in the limitations would be useful.

The authors exclude care homes, schools and hospitals from their analysis which could be mentioned in the main manuscript.

The authors state there to be potential sampling bias in their safe graph data. A quick google search showed that this is substantial at the CBG level:
https://colab.research.google.com/drive/1u15afRytJMsizySFqA2EPlXSh3KTmNTQ#sandboxMode=true&scrollTo=E-WFUxlgxNcK

The authors evaluation of safe graph vs. google mobility data was made at the country? level.

A few considerations:
1. Have the authors sought ethical approval for using such detailed mobility data?
2. There may be some concerns regarding blame of POIs for accelerating transmission. The authors should carefully balance their findings against that. A recommendation could be that the article is written in a more nuanced way and that further caution against over-interpreting these findings may be provided.

Referee #3 (Remarks to the Author):

This is a data analysis and modelling study of travel patterns in 10 large cities in the US using data from cell phone movements. To this the authors connect an epidemiological transmission model for SARS-

CoV-2 and development of COVID-19 disease. Using this model they find that some types of POIs are riskier for transmission than others and could be subject to different reopening schemes. The authors find that the identity of these POIs varies by the racial and economic makeup of the CBGs.

The paper is quite confusing throughout, since there are a lot of acronyms and methods scattered in various sections. The key thing the reader must grasp is what a POI is and why we should care about it, but there is not a clear definition early in the paper, nor information/distribution about how many of these there are, what kinds they are and aren't, where they are, and how many trips and for how long people generally spend at them. This is needed if readers are to "buy in" that the fundamental unit is something they should believe is representative.

The following comment applies to every timeseries figure: fitting to cumulative data has been shown to induce errors. See King et al. https://royalsocietypublishing.org/doi/full/10.1098/rspb.2015.0347 Although the model does not seem to be fitted directly to cumulative data (please confirm), cumulative figures cannot be the reader's indication of good/poor fit because it is subject to the same errors as in King et al. This must be corrected before I (or the readers) can actually tell if the model fits well. The authors should also switch the fitting method to likelihood-based and must understand better which of their datasets is contributing to the goodness of fit.

The transmission model itself has had a lot of epidemiological details skimmed over, and sensitivity to those assumptions is not considered. For example, schools are not included, nor the effect of age in transmission or of visiting POIs. The authors seed the cities evenly, but this may not be realistic given the characteristics of travellers. Also, there may be a change in transmission probability even if mobility stays equal because of hygienic measures. All of these (and other assumptions) should have a sensitivity analysis to determine how much they affect the findings.

I cannot recommend publication in the current form because it is not possible to judge if the model actually fits well. The section exploring the values of R within the MSAs seems to find acceptable parameter sets throughout the ranges for the transmission parameters, and therefore the parameters may not be identifiable from available data.

I have made further comments as I read through the paper, some major, some minor:

The Results section gives very skimpy methods first, instead of results. Suggest reorganising so that the results are explained here, with enough background that the readers can understand them.

Fig 1. Is the data fitted to cumulative cases? If so, this needs to be refitted to incident data. If not – only displayed as cumulative – the plots must be changed to incidence, and it will need a re-review to determine if the fit is acceptable.

Line 12-15. Awkwardly phrased.
Is the resolution at CBG level? If so, how do you know they're at POIs?

34: Suggest reordering this paragraph to say what the data are before how you make a network from them. Also, this may be better placed in methods.

37: what is a POI? This needs better description and definition at first use.

52: The out-of-sample data is not independent of the training set. It's difficult to validate infectious disease models in this way.

Line 60-62. This is a straw man argument. No serious infectious disease modeller would expect a city-level SEIR model with no interventions – when we know there have been interventions – to do a good

job.
A better validation would be to have a model which decreases transmission rates at the times of known interventions in the cities, or something that more clearly shows the advance of this particular study.

85: This is a major part of the results, but it is just stated, and it is very difficult to determine what actually backs this up.
The transmission model insists that more transmission happens at POIs than within the CBG or at other activities that don't count as POIs (which I don't find convincing), but it is quite opaque at this point what POIs we are actually talking about, and how the data have been used to come up with this result. So, when the paper says they're densely occupied it's hard to know how to interpret this – how do they know the square footage, and what durations are we talking about here?

82: CIs?

89: CIs?

90: they're not that similar across the MSAs, and the authors should give the full range. DC is the second highest, and 7 of 10 are quite a bit lower.

94: it is important to reduce transmission at the POIs, not mobility.

124: Although there are a lot of cities studied, the authors need to find a better way than just reporting Washington DC, as it is not enough information for the reader.

126: what about normalising by capacity of POIs instead of number? Or dwell time? Does it make much difference?

Extended 5: This can be much bigger. Please label ABCD. How do the MSAs vary within the ranges?

144: variable how? Across MSAs or between simulations?

148-150: That the differences in Philadelphia are due to density, income or race is not convincingly shown.

154: "were not able" is causal. They didn't, the authors don't know why.

157: visits to where? This is POIs?

162: Please make this subtitle more precise in its meaning.

205-206: The sentence of 2) would be useful right at the beginning of the paper, with a display of the data that backs it up, saying that "people from lower income CBGs tend to visit POIs which…" This would help the reader understand better where the rest of the results come from, and give more intuitive understanding of the data.
"when they do go out" implies fewer trips. Is this true?
Do workplaces count as POIs? How is work and place of work included in this study?

208: what age of children are not included? "Young children" implies under 5s to me, but the methods says under 13s. Also does not include older adults who don't have smartphones. The excluded POI list could be a bit longer, i.e. hospitals.

222: "have not been able" is used again. I don't disagree that there are likely barriers to reduction in mobility due to economic limitations. And this is likely a social determinant of increased risk of infection

for these groups. But nevertheless, the data do not back up "have not been able". The authors will need to find some evidence for this from other studies or will need to change to reflect what the data can show.

238: "must frequent". Beyond what the data can show.

252 section: what is the coverage of the population? Does it vary by characteristics of phone users?
Suggest pointing to a full list of POI types and frequency of those in the 10 population centres.
This section is a bit cryptic and could be explained more clearly.

260: how many is too few? Is the low number suppression each day or each hour?

273: How many POIs are excluded according to these rules? Are the excluded POIs randomly distributed geographically? And within the network?

283: this is the footprint area?

303: Is the data converted to incidence before calibration?

394: This is a strong assumption. The epidemic was seeded mostly by travellers, for which there is a strong age and sociodemographic skew. This model assumes that infection is evenly spread across the entire city. What is the effect of changing this assumption?

407: This isn't really out of sample because the epidemic is a dependent process. Suggest instead validating by simulation – simulate the epidemic using the fitted parameters and see if it looks similar to the observed epidemic.

The authors could also simulate an epidemic on the metapopulation network, and then attempt to recover the simulated parameters using their fitting method. This would give confidence that the fitting procedure is appropriate for the system.

415: why does R_base have an upper limit of 1? And what dos "approximately" mean here? What are the limits?
R_base should contain all transmission within the CBG? So, there could be transmission that occurs not at POIs but within the CBG, i.e neighbourhood, or non-recorded POIs. Does the dataset include children i.e. schools? Or public transit?
I think this is forcing the assumption that if R0 = R_base + R_poi, and R_base is household transmission, then they are saying that R0 = (R_hh + R_bh) where R_hh = R within the household, and R_bh = R between households. But R_base is probably not only R_hh because R_base is everything in the CBG, and R_bh is not R_poi because transmission does not only occurat POIs (or it is not convincing that that is the only place it is occurring). i.e. R_bh could include transmission on public transit, or in social contacts made outside POIs.

On further reflection, I think these constraints on R_base and R_poi are really strong assumptions. If R_base cannot be greater than 1, but R0 is high, then it could be forcing transmission into the POIs, even if there may be community transmission that is neither at a POI or is supercritical in R_base. The authors should determine if the proportion of POIs that are excluded or places that are somehow not included as POIs varies according to CBG characteristics, and should also try relaxing the R_base, R_poi constraints and see what difference it makes.

426: what is "roughly"? Please explicitly state the limits on the R0 values used.
What is the effect of extending these limits, i.e up to 5?

434: Why only 20 realisations? How much discrepancy in the 20? Suggest adding a table with ranges. And report the R_base and R_poi for each city.
The model is stochastic in movements or in the epi model? Or both?

451: What about if r_c is increasing? Or delta_c is decreasing? Then the test that the authors do to say the predictions weren't affected would not really account for a directional change.

Eqn 21. There's a lot of equations so it's a bit difficult to keep track of all the notation. What's m?

463: Why using RMSE? Why not likelihood based? Would better account for number of positive tests if used a Poisson likelihood.

479: 20% is arbitrary. Why was it chosen?

508: the IFR for COVID-19 is likely age-dependent and is not known. This 0.66% is given without a reference. What is the sensitivity to this assumption?

509: 18 days is a strong assumption without a reference. Please add a reference. How sensitive are the results to this assumption?

513, Extended F2: I do not agree that these "fit the deaths reasonably well". The graphs need to be bigger, have more tick marks on the y and to have more similar axes to each other. They need to be incident, not cumulative. And the colours could be more different.

Extended 1: How much of the signal in these data is coming from NYC, which has by far the highest number of cases and deaths? Since it is RMSE and not a Poisson LL, it is probably a fair bit of the signal. The authors should do a leave-one-out on each city and see how things change.

518: As previously, this is a straw man argument.
519: SLIR? Typo?

570: "clipping" term is a bit difficult to interpret and suggest they just change it - it's clearly representing a strategy, i.e. of social distancing in stores and letting few people in at a time – I haven't heard anyone calling this clipping. Therefore, suggest determining what policy you are trying to represent, and call it that. Possibly "low occupancy reopening" or "capped capacity"?

575: why do the data cut off on May 2?

601: More details on the POIs are given down here but should be presented much earlier in methods with the data, and the definition of the POIs in the first place.

603: what effect does not including schools have on the model?
604: The absence of drinking places may be a big problem, especially if newspaper reports of reopenings in these areas are to be believed. How undercounted are drinking places?
Is there a difference in missingness between CBGs?

610: R should be reserved for the reproduction number and not reused as reopening. It is confusing enough in modelling papers because of the R compartment (as used here) as well as R the reproduction number. Please change R for reopening to something else.

Extended 6. Consider the phrasing of this division. Saying the "top decile" is the one with the highest % of white people in a CBG is a little uncareful.

Extended 7: these deciles are CBGs, not people? Please add that to the caption.

Eq 31. Suggest to re-remind the reader what every symbol in this equation is. There are many. And mention the time superscript in words.

Fig 2b. Why not show them all? Or show the range? Why is Washington DC shown in main text when it is not especially representative? i.e. of the 10, 1 is more overdispersed than Washington DC. LA appears about the same, and the rest are less than 80% at 10% of POIs.

Fig 3.
a) Where are the numbers and Cis given for the values in these figures? Is this a RR in Philadephia of 30-fold? It's hard to tell from the scale what the numbers are. Are CIs from 20-100? This just isn't believable.
What is driving the difference in risk? In Washington from the Figure it might be Religious organisations, and hardware stores? There's some other dots not marked. The full-service restaurants in C don't have much of a higher transmission rate in e, so what is going on here? Are there just more of them?
f) why not show top income decile here also to show the difference?

Table S3. Gas stations 6x more transmission in Philadelphia. What are the absolute values of transmission, i.e. are gas stations generating a lot? This is a little confusing, because there is little person-to-person interaction at gas stations, and fomite transmission would have to be quite important for gas stations to be important. However, this is given as relative, so overall could be very low.
What are the CIs?
Is this a weighted median? i.e by duration of time spent, or population weighted?

Table S7. If I am interpreting this table and Fig S2 properly, then the range of R_base within 20% of the best fit implied for each city (apart from NYC) is the minimum and the maximum of the range. It appears that the lowest it can be is 0.001 and the highest is 0.012. Is this parameter actually identifiable?
It is harder to tell from psi if R_poi is also just finding the limits of the available range, because that one is different between each city.
Are these parameters correlated?
Please add to the caption the name of the parameters in words.
p0 is prevalence at time 0. How many CBGs within a city are seeded on average after Eq16 for the p0 values given here?

Extended T2: r_c is a percentage not a rate. Please add to the caption what r_c is.
How sensitive are the results to the durations?

Author Rebuttals to Initial Comments (note: the author uses bold text when quoting the reviewers' comments):

We thank the reviewers for their thoughtful and constructive comments, which have helped us significantly revise and strengthen the manuscript. While our core results remain similar, we have added many suggested sensitivity analyses and ablation studies (running a total of more than 80,000 additional models, and more than two million model realizations) to test our model assumptions more thoroughly.

To orient the reviewers, we start by summarizing the main changes we made; each of these changes are further detailed in the point-by-point replies below.

1. **An aggregate mobility baseline.** Our model uses a detailed mobility network to simulate disease spread. To test if this detailed model is necessary, or if our model is simply making use of aggregate mobility patterns, we ran a baseline SEIR model that uses the aggregate number of visits made to any POI in each

hour, but not the breakdown of visits between specific CBGs to specific POIs, as suggested by Reviewers 1 and 3 (Methods M5.1). Our model and the aggregate mobility model have the same number of free parameters (scaling transmission rates at POIs, scaling transmission rates at CBGs, and the initial fraction of infected individuals).

- We found that our network model substantially outperformed the aggregate mobility model in out-of-sample prediction (Figure ED1).
- As expected, the aggregate mobility model predicted very similar infection rates across all CBGs. This does not concord with previous work showing substantial heterogeneity in infection rates across neighborhoods. This includes higher rates of infection among disadvantaged racial and socioeconomic groups, which our network model captures, but the aggregate mobility model fails to reflect.
- Overall, these results demonstrate that our network mobility model better recapitulates observed trends than the aggregate mobility model, while also allowing us to assess fine-grained questions like the effects of POI-specific reopening policies.

2. **Modifying the parametric transmission rates.** In our model, we assume that the transmission rates at each POI are, as Reviewer 1 points out, proportional to the product of time spent there and density of infected individuals per unit area. As Reviewer 1 suggested, we tested these assumptions by computing two other variants of the transmission rate: one that removes the time spent there, and another that removes the density term. We found that the relative risks predicted by our original transmission rates concord best with the rankings of the danger of POI categories proposed by independent experts. For example, when we remove the time spent, we see unrealistic changes like limited-service restaurants being predicted to be far riskier than full-service restaurants (Figure S5).

3. **Sensitivity analyses for model fitting.** As suggested by all reviewers, we test the sensitivity of our results to several different variants of model fitting. For each variant listed below, we checked that our key results on superspreader POIs (Figure ED3), the

effects of reopening (Figure S6), and group disparities (Figure S7) were all also similar.

- We tested fitting using a Poisson error model, instead of the normal (Gaussian) error model that we originally used.
- We tested fitting on deaths instead of cases, which avoids the need to assume a constant case detection rate.
- We tested using a different threshold (10%, instead of 20%, of the best-fit model error) in our model calibration procedure.

We describe these three analyses in more detail in Methods M5.5.

Instead of assuming a constant confirmation delay (between when individuals become infectious and when their cases are confirmed), we also tested sampling the delay from two independent distributions fitted on empirical line list data, as Reviewer 2 suggested. In both cases, we found that the model's predictions were highly similar to the predictions made under the constant confirmation delay assumption (Figure S4; Methods M5.5).

4. **Identifiability of model parameters.** As suggested by Reviewer 3, we tested the identifiability of our model by simulating epidemics using the best-fit parameters for each MSA, and then checking whether our model fitting procedure correctly recovers the parameters used for simulation. For all 10 MSAs, we are able to correctly recover the true simulation parameters (Figure S8; Methods M5.3).

5. **Updated results throughout the manuscript.** Finally, we note that while our core conclusions remain unchanged, we have updated all of the figures, tables, and numbers in our revised manuscript in response to reviewer suggestions:
    - We expanded the range of our parameter grid search and increased the number of stochastic realizations used (Methods M4), as suggested by Reviewer 3.
    - We swapped out the Washington DC MSA for a more representative MSA, Chicago, in our main figures (Figures 1 to 3), as suggested by Reviewer 3. As before, we still include results for all MSAs in the supplement and describe them in the main text.
    - We worked with SafeGraph staff to further refine our data processing pipeline and have updated and clarified our data processing accordingly (Methods M1 and M7). We have also added a table of POI characteristics (Table S1) to better reflect the underlying data. These reflect questions and suggestions made by all reviewers.
    - We changed figures showing predicted daily cumulative cases to show daily incident cases instead, so that model fit can be more easily visually assessed, as suggested by all reviewers.

We have submitted a revised version of the manuscript, with our changes marked in blue, that reflect all of the changes above, as well as the other changes we made in response to the detailed reviewer feedback below. To make it easier for reviewers to cross-reference, we have

kept figure numbering in this revision as similar as possible to the original, and have added the additional figures to the supplement. We provide point-by-point replies below to the reviewers' detailed comments, with the original comments in bold. To help navigation, our response to Reviewer 1 is from pages 3 to 11; Reviewer 2, from pages 11 to 17; and Reviewer 3, from pages 18 to 35.

Reviewer 1

Comment 1: This paper fits a bipartite network mobility-informed SEIR transmission model using data on mobility from SafeGraph and Google to data on cumulative confirmed cases from 10 MSA in the US using a highly structured model, and then uses the results to consider counterfactual scenarios related to reopening and in particular which venues are important for transmission and what disproportionate impact there may be by SES.

As a first point, the authors should be commended, thanked, and somehow rewarded for producing one of the most crystal-clear explanation of a complex model that I have ever read. The order is intuitive; the jargon is minimal and correct when used; the technical aspects are explained clearly without obfuscation and with a clear intent to communicate. I wish I could say this is common in infectious disease epidemiology, but this paper really is exceptional.

We thank the reviewer for these very kind remarks and broadly agree with this framing of the paper. (We note, as a point of clarification, that we do not use Google mobility data in our model, but merely to validate the SafeGraph mobility data.)

Comment 2: My overall evaluation is that the paper really has two nearly unconnected aspects. The first is model fitting, which shows that a model with three parameters fits each of 10 time series better than a model with 1 – in fact it is less impressive still than this, because even if the number of parameters were the same, the fact that the epidemic slows down when mobility is quashed by social distancing is not a surprise, but is a further contrast between the models that is a bit more radical than just the number of parameters – one takes account of known basic mechanisms, the other leaves one (human contact) out.

We agree that it is encouraging, although not surprising, that our model fits the observed cases better than a model that leaves out human mobility. To challenge our model further, we test a stronger baseline that integrates hourly mobility patterns, but as a single aggregate measure instead of an entire POI-CBG network. In the spirit of one of the reviewer's later suggestions, this is meant to examine whether our model is simply making use of "gross patterns in mobility" or if there is something more subtle at play. We find evidence for the latter, as our model (with the full network) has substantially better out-of-sample performance than the aggregate mobility model, and also slightly better fit on the overall data. We discuss this model and our findings in

more detail in our response to Comment 4.

Comment 3: The second part of the paper – the part about counterfactual interventions – uses the exquisitely strong (albeit very reasonable -- probably where I would have also started) assumption of the model – that transmission risk in a place (separate for home block group and other places, POI) is proportional to the product of time spent there and density of infected individuals per unit area – and draws conclusions about which venues are responsible for most transmission, which groups are at highest risk in various phases of the epidemic, and so on. These are perfectly reasonable conclusions to draw from a model, and impressive to draw from a simple SEIR model (I find this paper methodologically innovative for this reason – boiling down agent-based models to very simple components). But there is nothing in the paper that suggests this assumption (transmission proportional to time x density of infected) is correct. It is reasonable and may well be, but the entire fitting exercise tests only the use of mobility data at all, not the parametric form or detailed assumptions which all the conclusions are based.
Therefore, in its current form, the paper goes to great effort to fit a model to data, then uses aspects of that model that are entirely untested to make its practical conclusions. To a large degree, the well-supported and innovative parts of the paper are separate.

We agree that the assumptions that we make on transmission rate are reasonable, but strong. We take the reviewer's helpful suggestion below to remove different terms of our transmission rate, and test our model's sensitivity to such perturbations. We find that both of the key terms that the reviewer identifies -- time spent and the density of the POI -- are valuable in helping the model make more realistic predictions of POI risk, and when either term is taken out, the predictions of the model worsen. We discuss the experiment and results in more detail in our response to Comment 5.

Comment 4: How could this part of the model be well supported? I'm not completely sure, but here are some thoughts:
1) Try fitting the same model with the same optimization routine, but with random permutations of the block groups, so that the POIs are visited by randomized residents of the MSA, rather than their true visitors. This would be a model that incorporates trends in mobility, but scrambles the structure of home and outside-home. If this fits much worse, then there is something meaningful in the model; if not, it may just be capturing gross patterns in mobility.

In the spirit of testing whether the model is "just capturing gross patterns in mobility," we introduce an alternate, simplified model that integrates hourly mobility patterns, but receives only the aggregate number of POI visits within the MSA in each hour, instead of the full
POI-CBG network. Compared to our network model, this aggregate mobility model has the same number of free parameters (scaling transmission rates at POIs, scaling transmission rates at CBGs, and the initial fraction of infected individuals).

We find that our network model better fits daily incident cases than the aggregate mobility model, suggesting that in addition to the information encoded in overall mobility trends, there is something meaningful in network structure itself, which highlights a contribution of our model for being able to integrate such fine-grained network information. In particular, our model substantially outperforms the aggregate mobility model on out-of-sample prediction (Figure ED1). We also find that the aggregate mobility model predicts very similar infection rates across all CBGs, which we know is unrealistic as there has been substantial heterogeneity in infection rates across neighborhoods [1] -- including higher rates of infection among disadvantaged racial and socioeconomic groups, which our network model naturally captures, but the aggregate mobility model fails to reflect.

[1] Cold Spring Harbor Laboratory. Who COVID-19 hit hardest in New York. https://www.cshl.edu/who-covid-19-hit-hardest-in-new-york/

Comment 5: 2) Try fitting the same model with (say) density-independent risk at all POIs (essentially collapsing time use into a single dichotomy of out of home vs. in home neighborhood). There may be several variants of this strategy, for example (Have not thought deeply, maybe one or both of these is not quite right, but the general idea is) removing the area term or both the area and duration terms. The goal is to figure out if the data contain any signal that the time and density metric is important.

My suspicion is that since there is no fitting to sub-MSA data, and the epidemic is not too close to saturated (I think, see below), that the first permutation will fit pretty well, undermining the evidence for the particulars of the model. The second maybe not – and if so, I think (need to think more) this would provide evidence for the fundamental assumption of the model that is used to generate the counterfactual predictions.

As the reviewer notes, we assume the transmission rate at a POI in a given hour depends on two key ingredients: how much time visitors spend there, and the density (number of visitors per sq ft) of the POI in that hour. These assumptions are based on prior expectations that a visit is more dangerous if you spend more time there and/or if the location is more crowded [1, 2], but we agree with the reviewer that it is important to test these assumptions empirically. To do this, we compare three parametric forms of transmission rate: our original formula, one that removes the density term, and one that removes the dwell time. Using each formula, we compute the risk of visiting a POI category as the average transmission rate of the category (weighting each POI by the proportion of category visits that went to that POI). Then, we evaluate whether the relative risks predicted by this setting of transmission rate concord with the rankings of POI categories proposed by independent epidemiological experts [3, 4].

Our results are summarized in Figure S5. We find that the predicted relative risks best match external sources when we use our original parametric form: restaurants, cafes, religious organizations, and gyms are among the most dangerous, while retail stores (groceries, clothing, pharmacies, etc) are less dangerous. However, when we assume only time spent matters and

we drop density, we see unrealistic changes in the ranking: e.g., restaurants drop close to grocery stores, despite multiple experts deeming them far apart in terms of risk [3,4]. When we assume only density matters and drop dwell time, we also see unrealistic changes: e.g.,
limited-service restaurants are predicted to be far riskier than full-service restaurants, and gyms and religious organizations are no longer predicted as risky, which contradicts both of our external sources deeming them among the riskiest [3, 4]. Thus, it seems that the data indeed contain signals that both of these factors are important toward faithfully modeling the transmission risk of different POIs, since the predictions become less realistic when either factor is taken out. This finding, as the reviewer says, provides evidence for fundamental assumptions of the model that are used in consequent experiments: e.g., to predict the impact of reopening different types of POIs, or to evaluate the disparities in POI riskiness between high and
low-income CBGs.

[1] Centers for Disease Control and Prevention, Considerations for Restaurants and Bars. https://www.cdc.gov/coronavirus/2019-ncov/community/organizations/business-employers/bars- restaurants.html

[2] World Health Organization, Transmission of SARS-CoV-2: implications for infection prevention precautions.
https://www.who.int/news-room/commentaries/detail/transmission-of-sars-cov-2-implications-for- infection-prevention-precautions

[3] Emanuel, E. COVID-19 Activity Risk levels.
http://www.ezekielemanuel.com/writing/all-articles/2020/06/30/covid-19-activity-risk-levels

[4] DesOrmeau, T. From hair salons to gyms, experts rank 36 activities by coronavirus risk level. https://www.mlive.com/public-interest/2020/06/from-hair-salons-to-gyms-experts-rank-36-activiti es-by-coronavirus-risk-level.html

Comment 6: 3) Most convincing would be to find an independent way to test the predictions of the relative importance of various types of POI to transmission. Perhaps from contact tracing data, though these are hard to come by.

We agree that contract tracing data would have been ideal for testing our model's predictions on the relative riskiness of different POI types; however, as the reviewer notes, this data is hard to come by. We opted instead for a second-best solution, as we searched for independent epidemiological experts who have also ranked POI types by their relative riskiness [3, 4]. As discussed above, we validated our model's predictions against these external sources, and found that the relative importance of various POI types to transmission inferred by our model concorded with that suggested by independent experts (Figure S5).

Comment 7: Other significant concerns:
1) The data are cumulative cases, and it is bad statistical practice to fit to cumulative cases. This article, titled "Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola" https://royalsocietypublishing.org/doi/full/10.1098/rspb.2015.0347

Contains a section titled "Deterministic models fit to cumulative incidence curves: a recipe for error and overconfidence."

We apologize for the lack of clarity here. In our original manuscript, we do in fact fit to daily incident cases, not cumulative cases, because we agree with all reviewers that fitting to cumulative cases is bad statistical practice. We have clarified this in Methods M4.2, including adding a citation to the above-mentioned article, and we have changed our model fit visualizations throughout the manuscript to show daily incident cases/deaths.

Comment 8: 2) Reporting fraction fixed at 10% throughout the period is surely highly error prone given changing testing and a lot of variation across the country -- is there some way to do better?

We agree: using a constant detection rate of 10% is an imperfect simplifying assumption. We have added an additional robustness check in which, rather than using the grid search models which best fit daily incident *cases*, we use the models which best fit daily incident *deaths*.
Because this does not rely on modeling the detected cases, it does not rely on the same 10% assumption. As we show in Figures ED3, S6, and S7, our primary results remain unchanged under this alternate assumption.

Comment 9: 3) The estimate of 10^-5 to 5 10^-4 for prevalence of E (if I understand p0) on March 1 seems a bit high for many cities. Given more than 1000x increase in cumulative cases in nearly all cities over that time, doesn't that imply an unrealistically high attack rate? Perhaps this is a misinterpretation.

The reviewer is correct that the parameter p0 refers to the fraction of individuals estimated to be in the E state on March 1. We believe our prevalence estimates are concordant with previous modeling work: for example, work from Vespignani et al. [1] found that there may have been thousands of infections in major US cities by March 1 (e.g., 10,700 in New York and 9,300 in San Francisco), implying that fractions of the population considerably higher than $5 * 10^{-4}$ were infected in major cities.

[1] "Hidden Outbreaks Spread Through U.S. Cities Far Earlier Than Americans Knew, Estimates Say". *The New York Times*, 2020.

Comment 10: 4) The Safegraph to Google comparison is incomplete. One city is shown, for four categories (Fig. S3), but high Pearson correlation is not really relevant since it is absolute changes that matter; deviation from the 1:1 line would be more impressive, and this seems quite deviated for some categories. What was the purpose of this validation, and did the data pass the validation in all cities? What would the conclusion be if the data sets told different stories?

These are great questions. In Figure S3 we visualize the results for New York State, but in

Table S6 we provide results for all of the states that appear in the MSAs that we model (15 states in total). The SafeGraph and Google timeseries show high Pearson correlations in all states, for all three categories of places we evaluated: Retail & Recreation, Grocery & Pharmacy, and Residential (with median correlations of 0.96, 0.76, and 0.88, respectively). This tells us that the datasets agree well on the timing and directional changes to mobility over time, which provides validation for the reliability of the SafeGraph data.

We agree it is useful to assess as well how the absolute changes in the Google and SafeGraph datasets compare in magnitude. We therefore assessed this for Retail & Recreation and Grocery & Pharmacy categories by regressing the SafeGraph time-series on the Google
time-series and computing the slope. (We did not compare absolute changes for the Residential category, since Google and SafeGraph measure different things in this category --- the number of visits to residential places versus the number of devices at home all day, respectively).
Aggregating over states, we find that the median slope for Retail & Recreation is 1.023, suggesting mobility changes are very similarly scaled in this category in both datasets. The median slope for Grocery & Pharmacy is 0.554, which means that SafeGraph's mobility changes for this category occur at around double the scale of Google's recorded mobility changes for this category. Ideally, the slope would be closer to 1, and this could reflect disagreement between the datasets about the scale of mobility reduction to Grocery & Pharmacy places. That said, we do not believe this discrepancy necessarily casts doubt on either dataset because Google is vague about how they measure visits to places and precisely which types of places are included in each category (e.g., their description of Grocery & Pharmacy is "Mobility trends for places like grocery markets, food warehouses, farmers markets, specialty food shops, drug stores, and pharmacies.") [1]. Since we cannot be sure that Google and SafeGraph are measuring visits in the same way or to the exact same types of places, we might only expect the time-series to be correlated, which we have shown, but not necessarily matched in absolute terms. Furthermore, since POIs in Grocery & Pharmacy only account for a small fraction (5%) of the overall visits in the SafeGraph dataset, we believe that even if SafeGraph data truly overestimates mobility changes to this category, it should not significantly impact our models. In contrast, Retail & Recreation constitute 35% of SafeGraph visits, so it is encouraging that SafeGraph and Google agree strongly there by both measures.

In our previously submitted draft, we also compared Park visits in Google and SafeGraph data. However, we have subsequently received information from SafeGraph that parks are sometimes inaccurately classified (e.g., other POIs are categorized as parks). For this reason, we have removed the comparison to Parks in the Google-SafeGraph analysis. (We note that Parks constitute an even smaller fraction of visits: only 2.5%.)

[1] Google. Mobility report CSV documentation.
https://www.google.com/covid19/mobility/data_documentation.html?hl=en.

Comment 11: Minor concerns:

1) Would the authors be willing to comment more on why the model fit substantially overestimates cumulative death counts (and to some degree, cumulative cases) for Dallas, Houston, Miami, and San Francisco (Extended Data Figures 1 &2)?

Following the reviewer's suggestion, Extended Data Figures 1 & 2 now show daily incident cases/deaths, instead of cumulative. Also, because we have added an additional sensitivity analysis that fits directly to deaths, Extended Data Figure 2 now uses those models for consistency (please see our response above to Comment 8), instead of models fitted on daily incident cases up to April 15, 2020. The reviewer makes a good point, though, that out-of-sample predictions tend to overestimate, and this still seems to be slightly true in Extended Data Figure 1, which shows out-of-sample case predictions. One possibility is that non-mobility-related transmission reduction measures (like mask wearing) became more widespread during the out-of-sample period, so models fit only on the train period overestimated spread during the out-of-sample period. This is consistent with data on mask wearing in the United States [1]. This is less of a concern for the full-sample fits (from March 1 - May 2) which we use to generate the main results for the paper, since by the end of that period, mask wearing was much more widespread. Nonetheless, we discuss this possibility in the Discussion section.

[1] YouGov. Personal measures taken to avoid COVID-19 (2020). Available at https://yougov.co.uk/topics/international/articles-reports/2020/03/17/personal-measures-taken-a void-covid-19.

Comment 12: 2) The figure x-axes labels are slightly cut off in Figure 3b.

Thank you! We have fixed this.

Comment 13: 3) Line 209: Additionally, it is not clear whether each Safegraph device corresponds 1:1 to an individual (e.g. an individual may have multiple devices simultaneously).

This is a good point; an individual can have multiple devices. We have edited the text to include this.

Comment 14: 3) Line 451: The reasoning on the range of R_base refers solely to within-household transmission but the descriptions should be amended to note that R_base will include the effect of POIs not covered by Safegraph, such as subway stations (as mentioned in Line 771), nursing homes (line 208), etc.

This is an important clarification and we have amended the text. Based on this reasoning, and on suggestions from Reviewer 3, we have also doubled the upper limit on R_base from 1 to 2 throughout our experiments. For more on this experiment, please see our detailed response to Comment 74.

Comment 15: 4) Line 683: The correction factor is estimated using an overall ratio of US population to total # of devices - is it possible to check what the correction factor is separately for each MSA and confirm they are all similar enough to 7?

This is a good question. We have verified that the ratio of MSA population to SafeGraph devices with home locations remains reasonably constant across all 10 MSAs: it varies from 5 to 7.6.
We note, however, that the correction factor (which follows Benzell et al. [1]) is approximate, and two additional aspects of our analysis are designed to allow for this. First, if the correction factor varies across MSAs, estimating an MSA-specific psi allows us to correct for this, since the correction factor is multiplied by psi. Second, if the correction factor varies across CBGs, we reweight CBG visits by the population of the CBG (as described in Methods M7, "To account for non-uniform sampling…"); this allows us to correct of SafeGraph's over- or under-sampling of CBGs.

[1] Benzell, S. G., Collis, A. & Nicolaides, C. Rationing social contact during the COVID-19 pandemic: Transmission risk and social benefits of US locations. Proceedings of the National Academy of Sciences (2020).

Comment 16: 5) Line 830: What was the correlation for the next highest MSA? The Philadelphia differences in risk by income (Figure 3A) and race / ethnicity still seem hard to explain - observed incidence rates disparities have not been nearly so severe. Could the fact that the income distribution for Philadelphia is particularly right-skewed versus other cities also play into any of this (similarly, median income is the lowest of all cities)? The more worrying possibility is some artifact in the Safegraph data that could lead to this.

We agree: on revisiting this line, we don't think that reporting correlations between population density and income is as persuasive as it could be. The reason is that, for the analysis in the paper, we don't necessarily care about the correlation between population density and income, but the absolute *differences* in population density between the top and bottom income deciles. Philadelphia is, indeed, an outlier in this regard: CBGs in the bottom income decile have a population density 8.2x those in the top income decile. (The overall median across MSAs is 3.3x, and the next-highest MSA is 4.5x). Overall, this substantiates the claim in the paper: that Philadelphia's unusual socioeconomic differences in density at SafeGraph POIs are consistent with its unusual socioeconomic differences in population density. We have updated the paper (Section S2) with these statistics.

In general, we agree that the numbers for Philadelphia are somewhat high and, indeed, we caution in Supplementary Methods S2 that our results can only reveal the extent of disparities under the assumption that there aren't any forces that countervail the effects of mobility on disparities -- "Since there are many other factors contributing to disparity that we do not model, we do not place too much weight on our model's prediction that Philadelphia's disparities will be larger than those of other cities". However, in general the magnitudes of the model's predicted

disparities across MSAs are plausible and consistent with observed data. For example, the overall reported black mortality rate is 2.4x higher than the white mortality rate [1], which is similar to the median racial disparity across MSAs of 3x that our model predicts.

[1] APM Research Lab. The color of coronavirus: COVID-19 deaths by race and ethnicity in the U.S. (2020). Available at https://apmresearchlab.org/covid/deaths-by-race.

Reviewer 2

Comment 17: Chang et al. use a simple SEIR model combined with large scale human mobility data to estimate the epidemic trajectory in 10 large urban areas in the United States. They use this model, which is primarily calibrated on human movement data to extrapolate and recommend reopening strategies. Even though this work is of potential interest in its current form I have some methodological reservations and I am concerned with the interpretation of the results.

First, the major methodological issue I can see is that models are evaluated against cumulative case counts. King et al. show convincingly that fitting SEIR models to cumulative case counts can lead to overconfidence in the observed parameters and fit of the model. Please see a detailed description here: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4426634/. I would recommend to redo all analyses with daily case counts.

We apologize for the lack of clarity here. In our original manuscript, we do in fact fit to daily incident cases, not cumulative cases, because we agree with all reviewers that fitting to cumulative cases is bad statistical practice. We have clarified this in Methods M4.2, including adding a citation to the above-mentioned article, and we have changed our model fit visualizations throughout the manuscript to show daily incident cases/deaths.

Comment 18: Second, the authors use delay distributions between infection and confirmation and that reporting rate is fixed at rc = 0.1 (time invariant). I would recommend delay distributions to be drawn from empirical line list data as described in previous work (see some work here https://www.bmj.com/content/369/bmj.m1923.long).

We thank the reviewer for the reference. We have conducted a sensitivity analysis where instead of assuming a constant confirmation delay, we sample the delay from the symptom-onset-to-reporting Gamma distribution reported in Li et al. [1], which was fitted on empirical line list data. As an additional check, we also performed an experiment where delays were drawn from the exponential distribution reported in Kucharski et al. [2], which was fitted on another dataset. We find that in both cases, the model predictions barely change compared to when we assume a constant confirmation rate and delay (Figure S4). However, an advantage of our original method (using a fixed confirmation delay and reporting rate) is that it allows us to

predict confirmed cases up to 7 days after the last day of simulated infections, because the newly infectious curve is simply scaled and translated, but we cannot do the same when we sample confirmed cases and delays stochastically. Because of this advantage, we choose to stick with the fixed method, but we appreciate the reviewer's encouragement to test these assumptions.

[1] Li, R. et al. Substantial undocumented infection facilitates the rapid dissemination of novelcoronavirus (SARS-CoV2). *Science* (2020).

[2] Kucharski, A. J. et al. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *The Lancet Infectious Diseases* (2020).

Comment 19: Reporting rates have changed dramatically through time. Especially the lack of available testing and changing testing protocols. The covid tracking website is a good resource (https://covidtracking.com/api). Detection rates used in the presented work may be relevant early in an epidemic.

We agree: a constant detection rate of 10% is an imperfect simplifying assumption. We have added an additional robustness check in which, rather than using the grid search models which best fit daily incident *cases*, we use the models which best fit daily incident *deaths*. Because this does not rely on modeling the detected cases, it does not rely on the same 10% assumption. As we show in Figures ED3, S6, and S7, our primary results remain unchanged under this alternate assumption.

Comment 20: Similarly, the authors use 18 days between day of infection to death. This in fact is highly variable and should be accounted for. The work cited by the authors is now outdated. Please see earlier work by Flaxman et al. Supplementary Figure 1 (https://static-content.springer.com/esm/art%3A10.1038%2Fs41586-020-2405-7/MediaObjects/41586_2020_2405_MOESM1_ESM.pdf).

We thank the reviewer for the reference. To clarify, we use 18 days as the time period between becoming *infectious* (not infected) and death. Becoming infectious is assumed to occur around symptom onset, so we based this time period on Verity et al. [1], which found that the mean time between symptom onset and death is 17.8 days. It seems that Flaxman et al. cite Verity et al. [1] as well for the same statistic, and use it to estimate their infection-to-death distribution (bottom of page 3 of the linked supplement).

It's likely that the reviewer is also concerned about the effects of assuming a fixed delay, instead of sampling from a distribution. We thus conducted additional experiments where we sample confirmed cases and delays stochastically. As described above in the response to Comment 18, we find that this stochastic sampling does not change model predictions noticeably (Figure S4) and it prevents us from predicting cases or death trends beyond the last day of simulated infections, so we chose to remain with our fixed rate/delay method.

[1] Verity, R. et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. The Lancet 20, 669–677 (2020).

Comment 21: The authors state: "As a sensitivity analysis, we alternatively stochastically sampled the number of confirmed cases and the confirmation delay from distributions with mean rc and c, but found that this did not change predictions noticeably." However, no Figure or formal evaluation is provided.

Thank you for noting this. We have added Figure S4, which is discussed above in the response to Comment 18.

Comment 22: As far as I understand this is a highly complex and multidimensional model where each location has a SEIR model with well mixed populations and each POI has some probability of transmission when people get together. How confident are the authors that the estimated rates of infection can be attributed to these POIs? A section on potential issues of identifiability would be helpful for the reader not familiar with these high dimensional network models.

This is a good point. While the data is complex, the *model* itself has only three parameters per MSA which are fitted to the data: p0, beta, and psi. Thus, no high-dimensional parameter vectors are fitted to data, mitigating identifiability concerns. Nevertheless, identifiability may be a concern even with relatively few parameters. We perform two additional analyses to assess model identifiability:

1. We follow a useful suggestion from Reviewer 3: for each MSA, we simulate an epidemic using the best-fit parameters for that MSA. We then run our grid search fitting procedure on the simulated data. For all 10 MSAs, the parameters in our grid search that obtain the lowest RMSE on the simulated data are always the true parameters that were used to generate that data. This demonstrates that our model and fitting procedure can correctly recover the true parameters on simulated data. This analysis is shown in Figure S8.
2. We plot the RMSE on daily case count data (that is, the metric used to perform model calibration) as a function of model parameters. This is shown in Figure S9. As these plots illustrate, beta and psi are correlated, which is expected because they scale the growth of infections at CBGs and POIs respectively. We account for the uncertainty caused by this correlation in our error bars throughout the paper, by aggregating results from all parameter settings which achieve an RMSE within 20% of the best-fit model for each MSA. We also note that the parameter ranges we conduct grid search over are broad, corresponding to an overall pre-lockdown R_0 that varies between 1.1 and 5: this is a conservatively wide range, since prior work estimates a pre-lockdown R_0 of 2-3 [1]. Similarly, our grid search range allows beta to vary by a factor of 20, and psi to vary by a factor of 10. In general, then, our analysis conservatively accounts for parameter uncertainty.

We have added these analyses to the paper in Methods M4.5.

[1] Park, M., Cook, A. R., Lim, J. T., Sun, Y. & Dickens, B. L. A systematic review of COVID-19 epidemiology based on current evidence. Journal of Clinical Medicine 9, 967 (2020).

Comment 23: Further, I am concerned with the section lines 278 - 288. The authors arbitrarily exclude outliers which are not shown in the manuscript.

We would like to clarify that we do not exclude outliers from our data; rather, we truncate their values to more plausible values. The SafeGraph data is imperfect, since it tracks millions of POIs and is constantly being updated as POIs change, and we believe that preventing data outliers from having an undue influence on our results will make our analysis more robust and not potentially driven by a small number of erroneous POIs. Our specific implementation of outlier truncation is what SafeGraph staff recommended after we discussed it with them. To be transparent about the statistics of POIs that are reflected in the raw data, we summarize the total number of POIs, edges, and visits that we model (Table ED1), show transmission-related statistics for each MSA and POI category that we analyze (Figures S10-19), and provide the overall distribution in the SafeGraph dataset of POIs and visits over POI categories (Table S1).

Comment 24: The unexplained variation in infection rates it seems is absorbed by the base rate of new exposures not captured by visiting POIs. Does this rate vary over time? If the authors model is correct the base rate should increase as more people are staying home? Is that the case?

This is an interesting point! However, our model does not allow the base rate to vary over time (that is, beta remains constant). We made this decision to simplify the model and avoid the identifiability concerns that would arise from having to fit a time-varying base rate.

Comment 25: The role of restaurants in the transmission dynamics of COVID is interesting. However, the author's model relies on human mobility and dwell times and the model fit needs further evaluation based on daily case counts rather than cumulative counts.

We apologize for the lack of clarity here. In our original manuscript, we do in fact fit to daily incident cases, not cumulative cases, because we agree with all reviewers that fitting to cumulative cases is bad statistical practice. We have clarified this in Methods M4.2, and we have changed our model fit visualizations throughout the manuscript to show daily incident cases/deaths.

Comment 26: For example, could the authors provide summary statistics of the dwell times and frequency of visits similar to Figure 2d for all the different locations? If they scaled the same way it would make clear that the findings are more related to the

underlying nature of visits rather than locations itself. That would change the story of the paper.

We agree that providing the summary statistics of POIs are important: to this end, we provide the dwell times and visits per hour (per square foot) for each MSA and POI category (for each of the 10 MSAs and 20 categories that we analyze) in the supplement (Figures S10-19).

To the reviewer's question about whether POI characteristics (dwell time and physical area) are important drivers of our results, or if our findings are instead more related to the underlying nature of visits, we tested alternate parametric forms for the transmission rate at each POI: one that removes the visitor density term, and another that removes the dwell time. When we computed the risk (average transmission rate) of each POI category using these alternate forms, we found that the relative risks they predict did not concord as well with the relative ranking of POI categories proposed by independent epidemiological experts, suggesting that our findings do depend on the characteristics of the POI locations, in addition to the visit patterns (Figure S5). For more details, please see our response to Comment 5.

Comment 27: I like Figure 2a, second / third panel. These clearly indicate that early reductions in mobility matter a lot. Please include references to earlier work from China Tian et al. 2020 (Science) & Lai et al. 2020 (Nature) showing similar results.

Thank you! We have added these references.

Comment 28: A few additional comments:
The authors do not account for potential re-seeding of the epidemic into major metropolitan regions. Especially for the period until lockdowns were implemented in the USA these could have played a substantial role in transmission. Please see visualisations (like nextstrain.com) of genomic data flows in the US during the COVID pandemic.

Thank you! This visualization is beautiful (and was a nice reminder of our days writing computational genomics papers). We agree that re-seeding is a potentially important phenomenon that we do not model. In general, we do not model travel between cities because we are uncertain of SafeGraph's ability to capture travel. We have edited the Discussion section to include the limitation that our model does not capture travel or seeding between MSAs.

Comment 29: The authors could put their findings in context of detailed transmission chain evaluations from Iceland: https://www.nejm.org/doi/full/10.1056/NEJMoa2006100.

We have added this reference to Methods M4.1, where its discussion of the fraction of the infections from household transmission is very relevant when we discuss plausible ranges of beta and psi. Thank you.

Comment 30: Further mention of the age groups that become infected or symptomatic in the limitations would be useful.

We agree: age variation is an important feature we do not model, and we have added this to the limitations section in the Discussion.

Comment 31: The authors exclude care homes, schools and hospitals from their analysis which could be mentioned in the main manuscript.

We agree this is an important limitation; thanks for the suggestion. Nursing homes are already mentioned as an example in the Discussion section (as a point of clarification, we do not exclude them; they are simply undercovered in the SafeGraph dataset). We have expanded the Discussion section to include schools and hospitals more explicitly, and have also discussed them in the Results section as well.

Comment 32: The authors state there to be potential sampling bias in their safe graph data. A quick google search showed that this is substantial at the CBG level:
https://colab.research.google.com/drive/1u15afRytJMsizySFqA2EPlXSh3KTmNTQ#sandboxMode=true&scrollTo=E-WFUxIgxNcK

We agree that there is non-uniform sampling at the CBG level. We correct for this, as recommended by SafeGraph, by reweighting the sample from each CBG using the ratio of the population of the CBG to the number of SafeGraph devices in that CBG; see Methods M7.
Overall, SafeGraph data has been shown to be geographically representative (e.g., it does not oversample high-income CBGs) both by the company itself (as the reviewer's linked resource indicates) and by external academic teams [1].

[1] Athey, S., Ferguson, B., Gentzkow, M. & Schmidt, T. Experienced Segregation (2019). Available at https://gsb.stanford.edu/faculty-research/working-papers/experienced-segregation.

Comment 33: The authors evaluation of safe graph vs. google mobility data was made at the country? Level.

The evaluation was made at the state-level. We evaluate agreement between Google and SafeGraph for Washington DC and the 14 states that appear in the 10 MSAs that we model (see Table S6).

Comment 34: A few considerations:
1. Have the authors sought ethical approval for using such detailed mobility data?

Yes, definitely. Thank you for checking. We had obtained IRB exemption from the Northwestern University IRB office, and we have added this in Methods M1 where we discuss the data.

Comment 35: 2. There may be some concerns regarding blame of POIs for accelerating transmission. The authors should carefully balance their findings against that. A recommendation could be that the article is written in a more nuanced way and that further caution against over-interpreting these findings may be provided.

This is a great point and one we considered and discussed at length. Ultimately, we are not overly worried about fostering blame against particular POI categories because our analysis of which categories are most dangerous concords with prior work [1, 2, 3, 4]. We are not drawing attention to any categories of POIs which have not already been recognized as dangerous (this is also reassuring for the reliability of the results). To be safe, we have edited the Discussion to further caution against over-interpreting our findings in the context of the risks of different POIs; we are happy to make additional changes if there are specific places that cause the reviewer concern. Thank you for bringing this up.

[1] Benzell, S. G., Collis, A. & Nicolaides, C. Rationing social contact during the COVID-19 pandemic: Transmission risk and social benefits of US locations. Proceedings of the National Academy of Sciences (2020).

[2] Baicker, K., Dube, O., Mullainathan, S., Devin, P. & Wezerek, G. Is It Safer to Visit a Coffee Shop or a Gym? The New York Times (2020). Available at https://nytimes.com/interactive/2020/05/06/opinion/coronavirus-us-reopen.html.

[3] Emanuel, E. COVID-19 Activity Risk levels. http://www.ezekielemanuel.com/writing/all-articles/2020/06/30/covid-19-activity-risk-levels.

[4] DesOrmeau, T. From hair salons to gyms, experts rank 36 activities by coronavirus risk level. https://www.mlive.com/public-interest/2020/06/from-hair-salons-to-gyms-experts-rank-36-activiti es-by-coronavirus-risk-level.html.

Reviewer 3

Comment 36: This is a data analysis and modelling study of travel patterns in 10 large cities in the US using data from cell phone movements. To this the authors connect an epidemiological transmission model for SARS-CoV-2 and development of COVID-19 disease. Using this model they find that some types of POIs are riskier for transmission than others and could be subject to different reopening schemes. The authors find that the identity of these POIs varies by the racial and economic makeup of the CBGs.

The paper is quite confusing throughout, since there are a lot of acronyms and methods scattered in various sections. The key thing the reader must grasp is what a POI is and why we should care about it, but there is not a clear definition early in the paper, nor information/distribution about how many of these there are, what kinds they are and aren't, where they are, and how many trips and for how long people generally spend at them. This is needed if readers are to "buy in" that the fundamental unit is something they should believe is representative.

We agree: it's critical to explain early in the text what a POI is and give some examples. We have edited the text early in the Introduction to clarify, and also added a new Supplementary Table (S1) with the most common POI categories. We also agree that providing basic statistics about the attributes of POIs is essential, and have done so in Figures S10-S19 (see the top two plots in each figure). Thank you for the suggestions.

Comment 37: The following comment applies to every timeseries figure: fitting to cumulative data has been shown to induce errors. See King et al. https://royalsocietypublishing.org/doi/full/10.1098/rspb.2015.0347 Although the model does not seem to be fitted directly to cumulative data (please confirm), cumulative figures cannot be the reader's indication of good/poor fit because it is subject to the same errors as in King et al. This must be corrected before I (or the readers) can actually tell if the model fits well.

We apologize for the lack of clarity here. In our original manuscript, we do in fact fit to daily incident cases, not cumulative cases, because we agree with all reviewers that fitting to cumulative cases is bad statistical practice. We have clarified this in Methods M4.2, and we have changed our model fit visualizations throughout the manuscript to show daily incident cases/deaths.

Comment 38: The authors should also switch the fitting method to likelihood-based and must understand better which of their datasets is contributing to the goodness of fit.

We agree with the principle of using likelihood-based fitting methods. As the reviewer requested below, we implemented model fitting using a Poisson error model, and we show via a sensitivity analysis that its results are similar to our original fitting method, which used a normal (Gaussian) error model with constant variance. We have also implemented additional identifiability analyses requested by the reviewer. Since the reviewer provided more thorough comments below, we will respond there with experimental details.

Comment 39: The transmission model itself has had a lot of epidemiological details skimmed over, and sensitivity to those assumptions is not considered. For example, schools are not included, nor the effect of age in transmission or of visiting POIs.

Schools are included in the model. We apologize for the lack of clarity here and have clarified this in Methods M6 ("Relative risk of reopening different categories of POIs" subsection). We do not model age because the SafeGraph data contains no information about the age distribution of their POI visitors. We agree this is an important limitation, and have further clarified this in the Discussion section.

Comment 40: The authors seed the cities evenly, but this may not be realistic given the characteristics of travellers.

We allow the proportion of the population that is initially infected in each city (i.e., MSA) to vary -- this is what the parameter p_0 captures. This allows for differential travel patterns to each city. Within each city, we seed CBGs uniformly. We favored this over alternate options for three reasons:

1. There is, to our knowledge, no good data on the proportion of people in each CBG who were infected on March 1, due to large and varying under-testing across locations.
2. An alternative would be to fit the proportion of each CBG infected to the data (i.e., to make p_0 vary at the CBG rather than the MSA level). We considered and rejected this option because it would have drastically increased the dimensionality of our parameter

vector, rendering the model non-identifiable.

3. Seeding CBGs uniformly allows us to more cleanly assess the inequities in infections which are a major focus of the paper: we show that *even if the rich and poor start out uniformly infected*, inequities in infections emerge. This would be harder to assess if we chose an initialization strategy where the rich and poor were initially not infected at equal rates, since it would be unclear if the subsequent inequities were simply the product of the initialization strategy.

Comment 41: Also, there may be a change in transmission probability even if mobility stays equal because of hygienic measures. All of these (and other assumptions) should have a sensitivity analysis to determine how much they affect the findings.

We agree: this is an important limitation which we have noted and further clarified in the

Discussion section. We considered fitting an additional time-varying factor to model changes in hygienic measures, but decided not to because we believed it would create identifiability concerns.

Comment 42: I cannot recommend publication in the current form because it is not possible to judge if the model actually fits well. The section exploring the values of R within the MSAs seems to find acceptable parameter sets throughout the ranges for the transmission parameters, and therefore the parameters may not be identifiable from available data.

Assessing whether the model fits the data well is very important. We have added a number of additional analyses to address this point:

1. Regarding model identifiability, we have added two additional analyses:
    a. Following the reviewer's suggestion below, for each MSA, we simulate an epidemic using the best-fit parameters for that MSA. We then run our grid search fitting procedure on the simulated data. For all 10 MSAs, the parameters in our grid search that obtain the lowest RMSE on the simulated data are always the true parameters that were used to generate that data. This demonstrates that our model and fitting procedure can correctly recover the true parameters on simulated data. This analysis is shown in Figure S8.
    b. We plot the RMSE to daily case count data (that is, the metric used to perform model calibration) as a function of model parameters. This is shown in Figure S9. As these plots illustrate, beta and psi are correlated, which is expected because they scale the growth of infections at CBGs and POIs respectively. We account for the uncertainty caused by this correlation in our error bars throughout the paper by aggregating results from all parameter settings which achieve an RMSE within 20% of the best-fit model for each MSA; this corresponds to rejection sampling in an Approximate Bayesian Computation framework. We also note that the parameter ranges we conduct grid search over are broad (as the reviewer suggests below, we doubled the range of the base transmission parameter), corresponding to an overall pre-lockdown R_0 that varies between 1.1 and 5. This is a conservatively wide range, since prior work estimates a pre-lockdown R_0 of 2-3 [1]. Similarly, our grid search range allows beta to vary by a factor of 20, and our grid search range allows psi to vary by a factor of 10.

2. Regarding overall model fit, we have made two substantial changes in line with the reviewer's suggestions below. First, we have changed the model fit plots to show daily rather than cumulative case counts to allow reviewers and readers to better assess model fit. Second, we have fit a stronger baseline model (which incorporates aggregate mobility) and compared its fit to our model, showing that using our fine-grained mobility improves model fit.

[1] Park, M., Cook, A. R., Lim, J. T., Sun, Y. & Dickens, B. L. A systematic review of COVID-19 epidemiology based on current evidence. Journal of Clinical Medicine 9, 967 (2020).

Comment 43: I have made further comments as I read through the paper, some major, some minor:

The Results section gives very skimpy methods first, instead of results. Suggest reorganising so that the results are explained here, with enough background that the readers can understand them.

Thank you for this suggestion! We have provided more background (e.g., a better definition of a POI) at the beginning of the Results section, so the text should now be more understandable.
We have also added an additional sentence about results at the end of the Introduction.

Comment 44: Fig 1. Is the data fitted to cumulative cases? If so, this needs to be refitted to incident data. If not – only displayed as cumulative – the plots must be changed to incidence, and it will need a re-review to determine if the fit is acceptable.

We completely agree. As we described earlier in our response to Comment 37, in our original manuscript, we do in fact fit to daily incident cases, not cumulative cases. We have clarified this in Methods M4.2, and we have changed our model fit visualizations throughout the manuscript to show daily incident cases/deaths.

Comment 45: Line 12-15. Awkwardly phrased.
Is the resolution at CBG level? If so, how do you know they're at POIs?

We have edited the phrasing to emphasize that we use geo-location data that can identify both the specific POI that a person is visiting, as well as their originating CBG.

Comment 46: 34: Suggest reordering this paragraph to say what the data are before how you make a network from them. Also, this may be better placed in methods.

We have reordered this paragraph accordingly. Thank you for the suggestion.

Comment 47: 37: what is a POI? This needs better description and definition at first use.

We have more clearly defined this in the Introduction and have also added a table (Table S1) detailing the most common POI types.

Comment 48: 52: The out-of-sample data is not independent of the training set. It's difficult to validate infectious disease models in this way.

Thank you; we have edited the text to clarify that we are referring to making case/death

predictions on a held-out time period that was not used for model calibration. We believe that this remains an important validation, since the held-out time period includes model inputs (mobility data) that are new to this period and not in the training set. While a sufficiently flexible model can obtain a good fit on the training set, a good fit on the held-out period suggests that the relationship between mobility patterns and disease trajectory (as parameterized in our model) can also hold outside of the training set.

Comment 49: Line 60-62. This is a straw man argument. No serious infectious disease modeller would expect a city-level SEIR model with no interventions – when we know there have been interventions – to do a good job. A better validation would be to have a model which decreases transmission rates at the times of known interventions in the cities, or something that more clearly shows the advance of this particular study.

We agree, and have added a stronger baseline. Please see our response to Reviewer 1 (Comment 4), who makes a similar suggestion, on this point. We show that our model also better fits the data than this stronger baseline.

Comment 50: 85: This is a major part of the results, but it is just stated, and it is very difficult to determine what actually backs this up. The transmission model insists that more transmission happens at POIs than within the CBG or at other activities that don't count as POIs (which I don't find convincing), but it is quite opaque at this point what POIs we are actually talking about, and how the data have been used to come up with this result. So, when the paper says they're densely occupied it's hard to know how to interpret this – how do they know the square footage, and what durations are we talking about here?

Thank you - we agree this paragraph was unclear, and we have substantially reworded it to make it clear what POIs we are talking about, and also to clarify the basic intuition for the method. We describe the full methodology underlying this subfigure in Methods M6. We have also clarified the source of the square footage (by explaining earlier in the Results section that SafeGraph data also provides square footage data), and added the raw numbers for square footage and durations in addition to the ratios that we already report.

Comment 51: 82: CIs?

Thank you; we have added these.

Comment 52: 89: CIs?

Thank you; we have added these.

Comment 53: 90: they're not that similar across the MSAs, and the authors should give the full range. DC is the second highest, and 7 of 10 are quite a bit lower.

We have provided the full range across MSAs. In response to the reviewer's concern that DC is non-representative, we have changed all our results to feature Chicago, which is more representative in this respect and also representative in terms of its best-fit parameters (beta, p0, and psi), which lie near the middle of the pack for MSAs.

Comment 54: 94: it is important to reduce transmission at the POIs, not mobility.

We have made this edit; thank you.

Comment 55: 124: Although there are a lot of cities studied, the authors need to find a better way than just reporting Washington DC, as it is not enough information for the reader.

We agree this is unfortunate, but find it necessary due to space constraints and clarity. Wherever possible, we show the results for all MSAs in main (e.g., Figure 1d, Figure 3a, Figure 3b). When we have to only show one MSA, we show the same one, lest it seem like
cherry-picking. In response to the reviewer's concerns that Washington DC is not representative, we have switched our results to feature Chicago for the reasons stated above. However, still, we always show results for all MSAs in the supplement.

Comment 56: 126: what about normalising by capacity of POIs instead of number? Or dwell time? Does it make much difference?

While the statistics we report -- overall impact of reopening a category, and reopening impact per POI -- are highly policy-relevant, we agree other normalizations are potentially interesting as well. In particular, we believe the reviewer's suggestion to normalize by the capacity of POIs is essentially equivalent to computing the risk to each person who visits the POI, a question of clear policy interest [1]. This is captured by the transmission rate at the POI -- we report this statistic for many of the POI categories in Figure S5, and validate that our predictions of "risk of a single visit" concord with external sources.

[1] Baicker, K., Dube, O., Mullainathan, S., Devin, P. & Wezerek, G. Is It Safer to Visit a Coffee Shop or a Gym? *The New York Times* (2020). Available at https://nytimes.com/interactive/2020/05/06/opinion/coronavirus-us-reopen.html.

Comment 57: Extended 5: This can be much bigger. Please label ABCD. How do the MSAs vary within the ranges?

We have added labels to the four panels and increased the size of the plot. Regarding how the MSAs vary within the ranges, we include the same figures for each individual MSA in the Supplement (S10-19).

Comment 58: 144: variable how? Across MSAs or between simulations?

Thanks, we agree this is unclear - we meant across MSAs, and have clarified.

Comment 59: 148-150: That the differences in Philadelphia are due to density, income or race is not convincingly shown.

We have clarified this sentence to refer to how our model's predicted differences in Philadelphia are due to differences in visitor density at POIs that are frequented by different income/racial groups.

Comment 60: 154: "were not able" is causal. They didn't, the authors don't know why.

We have edited this sentence accordingly.

Comment 61: 157: visits to where? This is POIs?

Yes, visits to POIs. We have clarified this sentence accordingly.

Comment 62: 162: Please make this subtitle more precise in its meaning.

We have edited it to reflect that POIs visited by lower-income CBGs tend to have higher transmission rates. Thanks for pointing this out.

Comment 63: 205-206: The sentence of 2) would be useful right at the beginning of the paper, with a display of the data that backs it up, saying that "people from lower income CBGs tend to visit POIs which…" This would help the reader understand better where the rest of the results come from, and give more intuitive understanding of the data.
"when they do go out" implies fewer trips. Is this true?
Do workplaces count as POIs? How is work and place of work included in this study?

Thank you for the suggestion! We agree; we've added it to the introduction, and have edited the abstract and Fig 3 caption accordingly as well.

The reviewer is correct that "when they do go out" implies fewer trips, which is not true (as the preceding sentence details). We've removed "do" to make it clearer.

Workplaces do count as POIs; the SafeGraph dataset does not distinguish between an individual being at a location to work versus being at the location for leisure. Thus, individuals that work at a POI will show up in the data as being present at multiple hours over the course of a day.

Comment 64: 208: what age of children are not included? "Young children" implies under 5s to me, but the methods says under 13s. Also does not include older adults who don't have smartphones. The excluded POI list could be a bit longer, i.e. hospitals.

Thanks for pointing this out. We have edited the text to "children under 13", added adults without smartphones, and have also expanded the list of excluded POIs both here (in the Discussion section) and earlier, in the Results section.

Comment 65: 222: "have not been able" is used again. I don't disagree that there are likely barriers to reduction in mobility due to economic limitations. And this is likely a social determinant of increased risk of infection for these groups. But nevertheless, the data do not back up "have not been able". The authors will need to find some evidence for this from other studies or will need to change to reflect what the data can show.

We agree: our data do not provide the reasons that people reduce (or fail to reduce) their mobility. But, as the reviewer notes, external sources do, and in particular corroborate the fact that economic limitations have made it more difficult for lower-income groups to reduce mobility: for example, they are less likely to be able to work from home [1]. We have added this citation and further clarified.

[1] Reeves, R. V. & Rothwell, J. Class and COVID: How the less affluent face double risks. *The Brookings Institution* (2020). Available at https://www.brookings.edu/blog/up- front/2020/03/27/class-and-covid-how-the-less-affluent-face-double-risks/.

Comment 66: 238: "must frequent". Beyond what the data can show.

Thanks - we have edited to "frequent".

Comment 67: 252 section: what is the coverage of the population? Does it vary by characteristics of phone users? Suggest pointing to a full list of POI types and frequency of those in the 10 population centres. This section is a bit cryptic and could be explained more clearly.

Regarding coverage, SafeGraph data has been shown to be geographically representative (e.g., it does not oversample high-income CBGs) both by the company itself and by external academic teams [1, 2] although it does not consistently track children under 13, as we discuss. We agree a list of POI types would be helpful, and have provided it in Table S1 (providing the overall dataset statistics rather than MSA-specific statistics, and listing the 50 largest POI categories, for space reasons). We have reworded this section to explain it more clearly.

[1] Athey, S., Ferguson, B., Gentzkow, M. & Schmidt, T. Experienced Segregation (2019). Available at https://gsb.stanford.edu/faculty-research/working-papers/experienced-segregation.
[2] Squire, R. F. What about bias in the SafeGraph dataset? (2019). Available at

https://safegraph.com/blog/what-about-bias-in-the-safegraph-dataset.

Comment 68: 260: how many is too few? Is the low number suppression each day or each hour?

Good question - we have clarified in the text that the threshold is 5, computed over the course of the month.

Comment 69: 273: How many POIs are excluded according to these rules? Are the excluded POIs randomly distributed geographically? And within the network?

Another good question - the original SafeGraph Weekly Patterns data contains 3.9 million POIs on March 1 (with similar numbers in other weeks), and after we apply all our filters, 552k POIs remain in our dataset. This considerable decrease is primarily caused by the first of our filters: i.e., filtering for POIs within MSA boundaries. This filter necessarily means that the POIs are not randomly distributed geographically, to answer the reviewer's question. Unfortunately, we cannot determine whether they are randomly distributed within the network, because many of the POIs are excluded precisely because we lack data to compute their place in the network - for example, if we lack hourly data for the POI, or if we lack information on its home visitor CBGs, we cannot compute its place in the network, which is derived using these quantities.

Comment 70: 283: this is the footprint area?

Yes, it is the area of the footprint polygon SafeGraph assigns to the POI, and this is well-correlated with external government data [1, 2]. We have clarified this when we first describe areas in Methods M1 and added the two citations below.

[1] SafeGraph. Using SafeGraph Polygons to Estimate Point-Of-Interest Square Footage (2019). Available at https://www.safegraph.com/blog/using-safegraph-polygons-to-estimate-point-of- interest-square-footage.

[2] SafeGraph. Guide to Points-of-Interest Data: POI Data FAQ (2020). Available at https://www.safegraph.com/points-of-interest-poi-data-guide.

Comment 71: 303: Is the data converted to incidence before calibration?

Yes, thanks for pointing this out - we have clarified this in the text.

Comment 72: 394: This is a strong assumption. The epidemic was seeded mostly by travellers, for which there is a strong age and sociodemographic skew. This model assumes that infection is evenly spread across the entire city. What is the effect of changing this assumption?

Please see our response to Comment 40 as to why we seed CBGs uniformly.

Comment 73: 407: This isn't really out of sample because the epidemic is a dependent process. Suggest instead validating by simulation – simulate the epidemic using the fitted parameters and see if it looks similar to the observed epidemic.

The authors could also simulate an epidemic on the metapopulation network, and then attempt to recover the simulated parameters using their fitting method. This would give confidence that the fitting procedure is appropriate for the system.

Thank you for the suggestion! We think this still represents a useful validation, for the reasons explained above in response to Comment 48. However, we agree that simulating an epidemic and attempting to recover the simulated parameters is an additional useful validation. We have now done so, and our fitting procedure does indeed recover the simulated parameters in all 10 MSAs. Please see the response above to Comment 42 where we describe this experiment in more detail.

Comment 74: 415: why does R_base have an upper limit of 1? And what dos "approximately" mean here? What are the limits?
R_base should contain all transmission within the CBG? So, there could be transmission that occurs not at POIs but within the CBG, i.e neighbourhood, or non-recorded POIs.
Does the dataset include children i.e. schools? Or public transit?
I think this is forcing the assumption that if R0 = R_base + R_poi, and R_base is household transmission, then they are saying that R0 = (R_hh + R_bh) where R_hh = R within the household, and R_bh = R between households. But R_base is probably not only R_hh because R_base is everything in the CBG, and R_bh is not R_poi because transmission does not only occur at POIs (or it is not convincing that that is the only place it is occurring). i.e. R_bh could include transmission on public transit, or in social contacts made outside POIs.

On further reflection, I think these constraints on R_base and R_poi are really strong assumptions. If R_base cannot be greater than 1, but R0 is high, then it could be forcing transmission into the POIs, even if there may be community transmission that is neither at a POI or is supercritical in R_base. The authors should determine if the proportion of POIs that are excluded or places that are somehow not included as POIs varies according to CBG characteristics, and should also try relaxing the R_base, R_poi constraints and see what difference it makes. 426: what is "roughly"? Please explicitly state the limits on the R0 values used. What is the effect of extending these limits, i.e up to 5?

Great points! We have doubled the upper limit on R_base (from 1 to 2) in all our experiments, which means that the overall range on R_base + R_poi now goes up to 5 (since R_poi goes up to 3) as the reviewer suggests. This does not significantly affect any of our conclusions, and for

9/10 MSAs, the best-fit value of the home transmission parameter (beta) lies within the original range of R_base < 1. We have updated the text to reflect this and update the logic behind the parameter ranges. We have also removed “approximately” and “roughly” in the text and added exact values.

Comment 75: 434: Why only 20 realisations? How much discrepancy in the 20? Suggest adding a table with ranges. And report the R_base and R_poi for each city.

We initially used 20 realisations in our experiments due to computational constraints: because every realisation requires multiplying matrices with millions of entries for thousands of timesteps, and we run millions of realisations across all our experiments, the computational cost becomes substantial even when using a cluster with hundreds of CPUs. However, to be safe, for this revision we reran all our experiments with 50% more stochastic realizations (30 per setting of model parameters) and found similar results. We have updated the text to reflect the new number of realizations.

For the R_base/R_POI experiments that the reviewer is referring to here, the variation across realisations was small. For R_base, the median standard deviation across realisations was 0.01 (taking the median across all R_base experiments) and for R_POI it was 0.12. The variation for R_POI is somewhat larger, but still much smaller than the variation across MSAs in R_POI (even when holding psi constant, R_POI will still vary because movement patterns vary across MSAs). For example, even when choosing psi = 1000, at the lower end of the plausible range, the range in R_POI across MSAs is 1.37, more than ten times the standard deviation across realisations within a single city. Overall, then, the variation across realisations in both R_base and R_POI, holding model parameters constant, is not a primary source of uncertainty.

As the reviewer requests, we added two additional panels to Figure S2 to report R_base and R_POI for each MSA.

Comment 76: The model is stochastic in movements or in the epi model? Or both?

The only source of stochasticity comes from the epi model. The movements themselves are not stochastic and are derived directly from SafeGraph data.

Comment 77: 451: What about if r_c is increasing? Or delta_c is decreasing? Then the test that the authors do to say the predictions weren’t affected would not really account for a directional change.

We agree this simplifying assumption does not account for a changing detection rate, and have added an additional sensitivity analysis. Please see our response to Comment 8.

Comment 78: 463: Why using RMSE? Why not likelihood based? Would better account for number of positive tests if used a Poisson likelihood.

We agree that a Poisson likelihood model, as the reviewer requested, is a natural option for model calibration, and a good idea to test. Accordingly, we implemented it as an additional sensitivity analysis (computed on daily incident cases; details in Methods M5.5). We found that ranking models via Poisson likelihood was consistent with ranking models using RMSE (also computed on daily incident cases): the median Spearman correlation over MSAs between models ranked by Poisson likelihood vs. RMSE was 0.97. Consequently, this meant that our downstream key results -- the existence of superspreader POIs (Figure ED3), the effects of reopening (Figure S6), and group disparities (Figure S7) -- were all also similar.

We note that model calibration using RMSE also has a likelihood interpretation, except using a (homoscedastic, constant variance) normal error model instead of a Poisson error model. As the reviewer notes, a constant-variance normal (Gaussian) model is likely to prioritize fitting parts of the case trajectory that have higher case counts, whereas a Poisson model will comparatively prioritize fitting parts of the case trajectory with lower case counts. In our setting, it is reassuring that both methods coincide and return a similar set of models. We have revised the text to make the likelihood interpretation clearer, in terms of how RMSE corresponds to a normal error model, and in terms of how the model calibration procedure corresponds to rejection sampling in an Approximate Bayesian Computation (ABC) framework (as in, e.g., [1]).

[1] Chinazzi, M., et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* (2020).

Comment 79: 479: 20% is arbitrary. Why was it chosen?

We agree with the reviewer that choosing the threshold (in our case, 20%) for the rejection ABC procedure is inherently arbitrary. We wished to be conservative with our predictions (selecting a larger threshold would result in accepting more models, and therefore more uncertain predictions), and selected 20% because beyond that point, model fit qualitatively deteriorated based on inspection of the case trajectories. As an additional sensitivity analysis, we selected a different threshold (10%) and verified that our key results -- the existence of superspreader POIs (Figure ED3), the effects of reopening (Figure S6), and group disparities (Figure S7) -- remained similar.

Comment 80: 508: the IFR for COVID-19 is likely age-dependent and is not known. This 0.66% is given without a reference. What is the sensitivity to this assumption?
509: 18 days is a strong assumption without a reference. Please add a reference. How sensitive are the results to this assumption?

We agree that IFR is age-dependent; unfortunately, our mobility data does not allow us to measure heterogeneity in movement patterns by age, and we have edited the Discussion to reflect this limitation. We have added references for both 0.66% and 18 days to this point in the text (references for all model parameters are listed in Table ED2). Our main results are not

sensitive to either of these parameters because we select models based on their fit to cases, not deaths.

Comment 81: 513, Extended F2: I do not agree that these "fit the deaths reasonably well". The graphs need to be bigger, have more tick marks on the y and to have more similar axes to each other. They need to be incident, not cumulative. And the colours could be more different.

Thank you for the suggestions! We have changed the colours, changed to incident as opposed to cumulative, increased the number of tick marks, changed the axes, and made the plots bigger. Also, because we have added an additional sensitivity analysis that fits directly to deaths, this figure now uses those models for consistency (please see our response above to Comment 8).

Comment 82: Extended 1: How much of the signal in these data is coming from NYC, which has by far the highest number of cases and deaths? Since it is RMSE and not a Poisson LL, it is probably a fair bit of the signal.
The authors should do a leave-one-out on each city and see how things change.

We fit model parameters (psi, beta, and p0) to each MSA individually, so the data in NYC does not influence the model parameters in other MSAs, as described in Methods M4.2.

Comment 83: 518: As previously, this is a straw man argument.

Please see our response to Comment 4 about stronger baselines ("We introduce an alternate, simplified model that integrates hourly mobility patterns…")

Comment 84: 519: SLIR? Typo?

Thank you; we have corrected this.

Comment 85: 570: "clipping" term is a bit difficult to interpret and suggest they just change it - it's clearly representing a strategy, i.e. of social distancing in stores and letting few people in at a time – I haven't heard anyone calling this clipping. Therefore, suggest determining what policy you are trying to represent, and call it that. Possibly "low occupancy reopening" or "capped capacity"?

Thank you for pointing this out. We have changed the term to "reduced occupancy reopening."

Comment 85: 575: why do the data cut off on May 2?

This was the most recent SafeGraph data available at the point when we wrote the paper. Using a time period that ended at May 2 also allowed us to more cleanly study the effects of

lockdowns and reopenings, since most locations in the United States were still locked down on May 2. Using a later time period, in which some places had reopened and had not, would have made our reopening experiments (e.g., Figures 2c and 2d) hard to interpret, since those experiments assume that the last week of the time period represents activity when the MSA has not reopened, and using a later time period would have rendered that assumption false for some MSAs.

[1] Lee, J.C. et al. See How All 50 States Are Reopening (and Closing Again). The New York Times (2020). Available at
https://www.nytimes.com/interactive/2020/us/states-reopen-map-coronavirus.html.

Comment 86: 601: More details on the POIs are given down here but should be presented much earlier in methods with the data, and the definition of the POIs in the first place.

As described above, we have added more details about POIs early in the manuscript and added a supplementary table to describe the most common types.

Comment 87: 603: what effect does not including schools have on the model?

We apologize for the lack of clarity here. We do include schools for the purposes of estimating disease spread. We have edited the text accordingly to emphasize this, and also edited Methods M6 to clarify (please see “Relative risk of reopening different categories of POIs”).

In our model, schools do not play a large role in infections, due to their relatively large physical area (and consequently their relatively low visit densities): our model attributes less than 1% of infections to Elementary and Secondary Schools (taking the median across MSAs).

We have edited our discussion to more explicitly note that one limitation of our study is our inability to fully model transmission at schools, due to limited data on children below the age of 13 (for privacy reasons) and the fact that age-dependent transmission effects are still not well understood, nor captured in our model. Thank you for bringing this up.

Comment 88: 604: The absence of drinking places may be a big problem, especially if newspaper reports of reopenings in these areas are to be believed. How undercounted are drinking places?
Is there a difference in missingness between CBGs?

We agree that drinking places can be high-risk. We exclude this category from our POI-specific analysis (but not in our model, as described in Methods M6, “Relative risk of reopening different categories of POIs”) based on [1], which reports that “SafeGraph staff suggest that part of the low count [of drinking places] is due to ambiguity in the division between restaurants and bars and pubs that serve food.” Examining the data further shows that indeed some restaurants are bars, beer gardens, breweries, cocktail lounges, Irish pubs, etc. This suggests that some

drinking places that also serve food are already accounted for in our model under restaurants, and are not entirely missing. We omit the drinking places category from the analysis to avoid giving possibly misleading conclusions based on the smaller number of drinking places that do not serve food.

More broadly, regarding data missingness between CBGs, SafeGraph released an analysis of the bias in their data along various lines [2]. Their results show that individuals are sampled consistently across counties, race, income levels, and educational levels. We correct for any residual heterogeneity in sampling across CBGs, as recommended by SafeGraph, as described in Methods M7.

We have edited the text to clarify these points. Thank you for bringing this up.

[1] Benzell, S. G., Collis, A. & Nicolaides, C. Rationing social contact during the COVID-19 pandemic: Transmission risk and social benefits of US locations. Proceedings of the National Academy of Sciences (2020).
[2] Squire, R. F. What about bias in the SafeGraph dataset? (2019). Available at https://safegraph.com/blog/what-about-bias-in-the-safegraph-dataset.

Comment 89: 610: R should be reserved for the reproduction number and not reused as reopening. It is confusing enough in modelling papers because of the R compartment (as used here) as well as R the reproduction number. Please change R for reopening to something else.

This is a good catch! We've changed R to \tau in the context of reopening.

Comment 90: Extended 6. Consider the phrasing of this division. Saying the "top decile" is the one with the highest % of white people in a CBG is a little uncareful.

Thank you, great point - we've edited this figure and Figure 3 a/b as well.

Comment 91: Extended 7: these deciles are CBGs, not people? Please add that to the caption.

Thank you, we've edited the caption.

Comment 92: Eq 31. Suggest to re-remind the reader what every symbol in this equation is. There are many. And mention the time superscript in words.

We apologize for the lack of clarity, and have elaborated on the equation accordingly.

Comment 93: Fig 2b. Why not show them all? Or show the range? Why is Washington DC shown in main text when it is not especially representative? i.e. of the 10, 1 is more

overdispersed than Washington DC. LA appears about the same, and the rest are less than 80% at 10% of POIs.

While we show one MSA in this plot for visual clarity, we agree with the reviewer that it is important to show this result for all MSAs and have done so in Figure ED3. We have also, as the reviewer suggests, switched to a more representative MSA (Chicago) throughout the main results.

Comment 94: Fig 3.
a) Where are the numbers and CIs given for the values in these figures? Is this a RR in Philadephia of 30-fold? It's hard to tell from the scale what the numbers are. Are CIs from 20-100? This just isn't believable.
What is driving the difference in risk? In Washington from the Figure it might be Religious organisations, and hardware stores? There's some other dots not marked. The full-service restaurants in C don't have much of a higher transmission rate in e, so what is going on here? Are there just more of them?

For Philadelphia, the model predicts that the median risk ratio between the top and bottom income deciles is 30; Figure 3a shows the lower and upper quartiles, which are 23 and 44, respectively. For race (Figure 3b), the median is 20 and the lower and upper quartiles are 18 and 24. (We are happy to provide a table with all the numbers in 3a/b if the reviewer believes it would clarify.) We share the reviewer's instinct that these ratios are large, which is why we analyze how they arise in detail in Supplementary Methods S2. In general, we agree that the numbers for Philadelphia are somewhat high and, indeed, and, indeed, we caution in this section that our results can only reveal the extent of disparities under the assumption that there aren't any forces that countervail the effects of mobility on disparities -- "Since there are many other factors contributing to disparity that we do not model, we do not place too much weight on our model's prediction that Philadelphia's disparities will be larger than those of other cities".

However, in general the magnitudes of the model's predicted disparities across MSAs are plausible and consistent with observed data. For example, the overall reported black mortality rate is 2.4x higher than the white mortality rate [1], which is similar to the median racial disparity across MSAs of 3x that our model predicts.

[1] APM Research Lab. The color of coronavirus: COVID-19 deaths by race and ethnicity in the U.S. (2020). Available at https://apmresearchlab.org/covid/deaths-by-race.

Comment 95: f) why not show top income decile here also to show the difference?

Good question, and we discussed it. We decided against it because the point of this subfigure is that it's important to evaluate the effect of a policy on disadvantaged groups specifically, rather than just evaluating overall impact, so we wanted to compare the bottom decile to the population as a whole.

Comment 96: Table S3. Gas stations 6x more transmission in Philadelphia. What are the absolute values of transmission, i.e. are gas stations generating a lot? This is a little confusing, because there is little person-to-person interaction at gas stations, and fomite transmission would have to be quite important for gas stations to be important. However, this is given as relative, so overall could be very low.
What are the CIs?
Is this a weighted median? i.e by duration of time spent, or population weighted?

Gas stations account for only a small proportion of infections: 3.5% in Philadelphia (overall median across MSAs, 3.5%). There are no CIs on the transmission rates in Table S3 because they are the result of our transmission rate equation (eq 8) applied directly to the data, and we do not incorporate uncertainty in the data measurements or in the parametric form. The median is not weighted.

Comment 97: Table S7. If I am interpreting this table and Fig S2 properly, then the range of R_base within 20% of the best fit implied for each city (apart from NYC) is the minimum and the maximum of the range. It appears that the lowest it can be is 0.001 and the highest is 0.012. Is this parameter actually identifiable?
It is harder to tell from psi if R_poi is also just finding the limits of the available range, because that one is different between each city.
Are these parameters correlated?
Please add to the caption the name of the parameters in words.

Regarding identifiability, please see our response to Comment 42; we also provide a figure showing the correlation between parameters. As a point of clarification, the range we search over for psi is the same for all MSAs, as explained in Methods M4.1. We have updated the caption.

Comment 98: p0 is prevalence at time 0. How many CBGs within a city are seeded on average after Eq16 for the p0 values given here?

The median number of CBGs seeded, using the best-fit values of p0, is 1795 (median across all MSAs), or 31% of CBGs. As discussed above, our estimates of p0 are consistent with those in other work; for example, work from Vespignani [1] finds that there may have been thousands of infections in major US cities by March 1 (e.g., 10,700 in New York and 9,300 in San Francisco).

[1] “Hidden Outbreaks Spread Through U.S. Cities Far Earlier Than Americans Knew, Estimates Say”. *The New York Times*, 2020.

Comment 99: Extended T2: r_c is a percentage not a rate. Please add to the caption what r_c is.
How sensitive are the results to the durations?

Thank you; we have corrected this and better explained what r_c is in the table ("percentage of cases which are detected"). Regarding sensitivity of the results to durations, we assume, because the comment is about r_c, that this refers to the sensitivity of the results to delta_c (which refers to the duration of time between when someone becomes infectious and when their case is confirmed). We conducted additional sensitivity analyses where, instead of assuming a fixed confirmation delay, we sampled delays stochastically from two independent distributions that were fitted on empirical line list data [1, 2]. For both distributions, we found that this stochastic sampling did not change model predictions noticeably (Figure S4), which suggests that our results also are not sensitive to this parameter. Please see our response to Comment 18 for more details.

[1] Li, R. et al. Substantial undocumented infection facilitates the rapid dissemination of novelcoronavirus (SARS-CoV2). *Science* (2020).

[2] Kucharski, A. J. et al. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *The Lancet Infectious Diseases* (2020).

Reviewer Reports on the First Revision:

Referee #1 (Remarks to the Author):

This paper remains well-written, and it addresses a number of the key criticisms of the earlier draft.

In its current form, it shows

1) A very parsimonious model with SEIR dynamics in CBGs and POIs (roughly homes and businesses resp.), detailed travel data, and three free parameters provides good out of sample aggregate predictive power for 6 or 7 of 10 MSAs considered, among which the details in the model seem visually to make a substantial difference in most or all of the well-fitting MSAs. This is remarkable and surprising, as well as being methodologically innovative.

2) This model also replicates ethnic and income disparities in infection rates, based purely on how many trips to which POI individuals from different CBG make. An interesting corollary of this finding is that within categories of POI, individuals from lower-SES CBG have a higher predicted risk of infection because the (say) grocery stores they go to are smaller and more crowded. This is interesting, but requires caution, because the model assumes across CBG, transmission at home is the same, that duration of stay in a given POI is the same, and biological susceptibility is the same, as well as assuming that modes of transport have no impact on risk – making the specific POIs people visit (and how many and how often) the only source of disparities.In this sense, the finding is also remarkable, but in a way explains too much, because low SES is also associated with crowded housing, longer duration of stay at some POI (say full-service restaurants, where presumably a less wealthy person is more likely to work a many-hour shift while a richer one might be a customer who spends an hour), use of public transport, and many other risk factors. The spotlighting of particular categories of POIs and their differential effect by SES is a fascinating and novel hypothesis, but is not completely compelling because all of these other issues are assumed not to contribute.

3) By considering different reopening strategies the model finds that certain categories of reopening will make a bigger difference overall (and more disparate to low SES and persons of color) than others. This is not implausible, but the inference is limited by the fact that the density of visitors (visit rate/sqft*average dwell time) is forced to "absorb" all the variability in these other parameters. Thus the conclusions about reopening strategies are quite uncertain. Somewhat more robust are the conclusions

about the two kinds of reopening strategies, where capping the occupancy is better than uniform reductions in visits. This is more robust because it relies on the notion that transmission opportunities under the model are proportional to density squared, a reasonable assumption.

Overall, this paper is innovative in its methods and in its predictions that 1) transmission outside the home is fully responsible for racial and SES disparities; 2) The same category of POI may be more hazardous to low-income people than high-income because it is smaller and more crowded; 3) reducing density in high-density establishments is better than reducing overall visits to the same category. On the other hand, these should be seen as hypotheses generated by the model rather than as descriptions of COVID-19 transmission, and that needs to be clarified throughout.

Specific comments:
Line 95: this comparison between timing and magnitude is arbitrary. The timing effect is related to the doubling time of the infection and the amount of time delay; the mobility reduction effect is related to the amount of mobility reduction – If instead of comparing one week to 25% magnitude, the comparison had been 2 weeks vs 75% magnitude, the results would have gone the other way, presumably – this is a quantitative not a qualitative result and should be stated as such.

Line 165: One might hypothesize that many of the low-income individuals in the full-service restaurants were working there, not clients – suggests the wrong inference if reducing occupancy for example, as kitchens maybe different from dining rooms

Throughout, should not talk about infections or people being infected—this is all in a heavily constrained model

Line 192: The claims about different rates at POIs frequented by low income people may not be correct – how do we know it isn't the duration of stay?

Referee #2 (Remarks to the Author):

I thank the authors for their revised manuscript which has improved since the first submission.

A few suggestions below:

The concern regarding fitting the model to cumulative case counts has been addressed. The fit however does not look as good anymore in Figure 1d. Do the authors have particular insights as to why Atlanta, San Francisco do not seem to have such a good fit? I was wondering whether it may be due to large numbers of continued introductions rather than a self sustained epidemics in the city. This may be added to the discussion of the manuscript.

The authors perform sensitivity analysis with respect to a comment considering the delay between date of onset and confirmation and date of infection to confirmation. The authors results do not change which is surprising given that other simulation studies have shown that this delay is crucial in determining the epidemic shape (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7097845/). This makes me wonder whether the model is overparameterised?

The authors state that the base rate cannot change over time in the model. However, it is clear that those rates change over time due to different behaviours and interventions put in place during the epidemic. For example, the transmission dynamics have shifted from household, hospital and care home

transmission in the beginning of the outbreak to community transmission later. See detailed investigations of that for example from Germany:
https://www.rki.de/DE/Content/Infekt/EpidBull/Archiv/2020/Ausgaben/38_20.pdf?__blob=publicationFile

Would the authors be able to look at time dependent variation of where transmission may have occurred (Figure 2d at different time slices of the epidemic?)

Referee #3 (Remarks to the Author):

The authors have engaged with the comments in the main and this has improved the manuscript.

Putting Chicago in the main text is ok but it would be good to exactly replicate the Chicago-specific figures (i.e. Fig 2 and most of Fig 3) for each city in the supplement, so that readers can quickly go to them and look for the same patterns. This would mean keeping axes limits the same, so that readers could make a straightforward comparison. This is analogous to what they did in S10-19.

Can the model be updated to include later data?

Could consider dropping the MSA acronym – define the area at first, and then call them "cities" to reduce the acronyms – CBGs and POIs in MSAs is quite a bit to handle.
New line 42, MSA is defined for a second time, as is CBG and POI.

The Model paragraph doesn't feel like results.

101. this isn't "earlier findings" that showed this, this is epidemiological theory and doesn't need to be re-proven at every epidemic. Please put a reference to a classical text. Recent ones are useful for saying how much and how quickly for COVID.

175: Are these visits by people from CBGS which have a different socioeconomic and racial makeup, not individuals in different socioeconomic and racial groups? It's not clear that the demographics and the movement are linked at an individual level.

*187-205: Interesting paragraph. The table that backs it up is S4, but the ratios across cities vary by city ranging from no effect in Atlanta and Dallas to 10x in Philadelphia (which still seems pathological). The results paragraph does not convey this nuance. Suggest to unpack this result a little more, e.g. a set of plots, with a panel for each city, x axis is the decile categories, and y is the transmission rate, with a dot for the rate in each grocery store POIs as dots (with boxplot overlay). This will help readers understand the pattern in the increasing income deciles, the differences between cities, and convey more information than a ratio. Same could be done for the racial groups in S5.
And to add nuance to the paragraph reporting this result.

221 add "in 10 cities".
226: "clipped" was going to be replaced in this version?

Extended 6: Green/red comparisons are difficult for colorblind people. Suggest to change colours.
696: Clipping is still here.

Suggest to add to the Discussion that the transmission rate at POIs is constant in the model, which may

not be true anymore - important if people use it for prediction/policy.

Comments:
I agree with comment 35 for caution. This paper might accidentally suggest that closing restaurants and other POIs in poor and black neighbourhoods may be a transmission-reducing intervention. So, suggest to be quite careful with it.

*39 and 87: Line 143 explicitly says it excludes schools. As does 730-31. This is very confusing, and it needs to be much clearer to readers how schools are in the model but not in the POI analysis.

*42: This seems like a very confusing way of presenting the data. It's not intuitive what an "acceptable" loss value is, whether it should be the same from city to city, and what are the model ranks? Can you put x = simulated parameter, y = value of the estimate?
I don't understand S9 very well. If it's RMSE I think converting it to a ratio makes interpretation much harder, so suggest to make this a heatmap of the RMSE or Poisson likelihood instead. The white area indeed appears to show that these parameters are not identifiable in some cities – Atlanta, LA, DC . Also, suggest to add what every Greek letter is in words in the supplementary figures – readers will likely have forgotten.

94: Philadelphia's big values do affect the medians reported, though, so may be having an outsize effect on the overall message.

98: suggest to add the number to methods or a caption somewhere – it's interesting to readers and will allow other people to more easily use the results of the model (especially if you give a city-specific number).

Author Rebuttals to First Revision (note: the author uses bold text when quoting the reviewers' comments):

We thank the reviewers for their thoughtful and constructive comments, and provide a point-by-point response to reviewer comments below. To help navigation, our response to Reviewer 1 is from pages 1 to 3; Reviewer 2, from pages 3 to 5; and Reviewer 3, from pages 5 to 10.

Reviewer 1

Comment 1: This paper remains well-written, and it addresses a number of the key criticisms of the earlier draft.

In its current form, it shows

1) A very parsimonious model with SEIR dynamics in CBGs and POIs (roughly homes and businesses resp.), detailed travel data, and three free parameters provides good out of sample aggregate predictive power for 6 or 7 of 10 MSAs considered, among which the details in the model seem visually to make a substantial difference in most or all of the well-fitting MSAs. This is remarkable and surprising, as well as being methodologically innovative.

2) This model also replicates ethnic and income disparities in infection rates, based purely on how many trips to which POI individuals from different CBG make. An interesting corollary of this finding is that within categories of POI, individuals from lower-SES CBG have a higher predicted risk of infection because the (say) grocery stores they go to are smaller and more crowded. This is interesting, but requires caution, because the model assumes across

CBG, transmission at home is the same, that duration of stay in a given POI is the same, and biological susceptibility is the same, as well as assuming that modes of transport have no impact on risk – making the specific POIs people visit (and how many and how often) the only source of disparities.In this sense, the finding is also remarkable, but in a way explains too much, because low SES is also associated with crowded housing, longer duration of stay at some POI(say
full-service restaurants, where presumably a less wealthy person is more likely to work a many-hour shift while a richer one might be a customer who spends an hour), use of public transport, and many other risk factors. The spotlighting of particular categories of POIs and their differential effect by SES is a fascinating and novel hypothesis, but is not completely compelling because all of these other issues are assumed not to contribute.

3) By considering different reopening strategies the model finds that certain categories of reopening will make a bigger difference overall (and more disparate to low SES and persons of color) than others. This is not implausible, but the inference is limited by the fact that the density of visitors (visit rate/sqft*average dwell time) is forced to "absorb" all the variability in these other parameters. Thus the conclusions about reopening strategies are quite uncertain. Somewhat more robust are the conclusions about the two kinds of reopening strategies, where capping the occupancy is better thanuniform

reductions in visits. This is more robust because it relies on the notion that transmission opportunities under the model are proportional to density squared, a reasonable assumption.

Overall, this paper is innovative in its methods and in its predictions that 1) transmission outside the home is fully responsible for racial and SES disparities; 2) The same category of POI may be more hazardous to low-income people than high-income because it is smaller and more crowded; 3) reducing density in high-density establishments is better than reducing overall visits to the same category. On the other hand, these should be seen as hypotheses generated by the model rather than as descriptions of COVID-19 transmission, and that needs to be clarified throughout.

We thank the reviewer for these comments, and broadly agree with this framing of the paper. We have added the points raised in 2) to the Discussion, and more broadly clarified in the discussion that we do not believe that "transmission outside the home is fully responsible for racial and SES disparities" because many other factors including differences in household size, access to care, and comorbidities likely contribute to disparities as well. ("Beyond these mechanisms, racial and socioeconomic disparities in infection rates may also be driven by mobility differences our dataset cannot capture…") Finally, as the reviewer suggests, we have clarified throughout the paper which findings are predicted outputs of the model as opposed to descriptions of COVID transmission.

Specific comments:

Comment 2: Line 95: this comparison between timing and magnitude is arbitrary. The timing effect is related to the doubling time of the infection and the amount of time delay; the mobility reduction effect is related to the amount of mobility reduction – If instead of comparing one week to 25% magnitude, the comparison had been 2 weeks vs 75% magnitude, the results would have gone the other way, presumably – this is a quantitative not a qualitative result and should be stated as such.

Thank you, this is a good point - we agree, and have corrected the wording here to be more consistent with the figure caption.

Comment 3: Line 165: One might hypothesize that many of the low-income individuals in the full-service restaurants were working there, not clients – suggests the wrong inference if reducing occupancy for example, as kitchens maybe different from dining rooms

We agree, and have added this point to the Discussion. ("Beyond these mechanisms, racial and socioeconomic disparities in infection rates may also be driven by mobility differences our dataset cannot capture…")

Comment 4: Throughout, should not talk about infections or people being infected—this is all in a heavily constrained model

We absolutely agree. Thanks for bringing this up. We have clarified this language throughout the entire draft.

Comment 5: Line 192: The claims about different rates at POIs frequented by low income people may not be correct – how do we know it isn't the duration of stay?

To clarify, the equation for transmission rate does take into account the duration of the stay (see equation 8 in the Methods). Indeed, one reason that our model predicted that POIs frequented by people from low-income CBGs had higher transmission rates was that, based on the mobility data, visitors to these POIs tended to stay longer on average. We have clarified this in the manuscript ("We use the inferred density of infectious individuals at each POI to determine its transmission rate…"). We note that this accounts for heterogeneity in the duration of stay between POIs, but as the reviewer pointed out in their general comments, not for heterogeneity in the duration of stay within a single POI; as we mentioned above, we have added this limitation into the Discussion.

Reviewer 2

Comment 6: I thank the authors for their revised manuscript which has improved since the first submission.

Thank you!

A few suggestions below:

Comment 7: The concern regarding fitting the model to cumulative case counts has been addressed. The fit however does not look as good anymore in Figure 1d. Do the authors have particular insights as to why Atlanta, San Francisco do not seem to have such a good fit? I was wondering whether it may be due to large numbers of continued introductions rather than a self sustained epidemics in the city. This may be added to the discussion of the manuscript.

We agree that daily case counts in Atlanta and San Francisco are noisier -- i.e., there is more variability in day-to-day numbers than in other cities, possibly caused by differences in reporting practices across cities that, for example, increase variance between days of the week. However, our model does fit the smoothed weekly average trend well, and we have added a line to the figure to showcase this. The point about how the model does not include cases introduced from other cities is a good one, and is included in the Discussion ("travel and seeding between MSAs").

Comment 8: The authors perform sensitivity analysis with respect to a comment considering the delay between date of onset and confirmation and date of infection to confirmation. The authors results do not change which is surprising given that other simulation studies have shown that this delay is crucial in determining the epidemic shape (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7097845/). This makes me wonder whether the model is overparameterised?

Good question! We do not think the similarity in curves is due to overparameterisation of the model. In Figure S4, which we assume the reviewer is referring to, we show that our assumption of a deterministic infectious-to-confirmation delay of seven days is consistent with two stochastic delay distributions used in previous work based on previous line list data. We believe the similarity occurs because the two stochastic distributions --- Gamma(1.85, 3.57) in Li et al. and Exp(6.1) in Kucharski et al. --- both have means which are very similar to our deterministic delay of 7 days.

Comment 9: The authors state that the base rate cannot change over time in the model. However, it is clear that those rates change over time due to different behaviours and interventions put in place during the epidemic. For example, the transmission dynamics have shifted from household, hospital and care home transmission in the beginning of the outbreak to community transmission later. See detailed investigations of that for example from Germany: https://www.rki.de/DE/Content/Infekt/EpidBull/Archiv/2020/Ausgaben/38_20.pdf? blob=p ublicationFile

Thanks for the very helpful link. We have added a sensitivity analysis (Methods M5.2) where we incorporated a time-varying base transmission rate into the model. Instead of using a fixed β_{base} to represent the base transmission rate (Equation 11), we modified the base transmission rate to equal β_{base} * $proportion\_at\_home_{i}^{(t)}$, i.e., the proportion of CBG c_i that stayed at home at time t, a measure included in SafeGraph Social Distancing dataset [1]. This captures the reviewer's earlier suggestion that "If the authors' model is correct the base rate should increase as more people are staying home". We ran the same grid-search procedure (Methods M4) to calibrate this modified model, and evaluated its ability to fit incident daily cases. Fit quality was very similar for the original model and the modified model: taking the median across MSAs, the RMSEs are within 2% of each other. Fits are qualitatively similar as well, as we show in Figure S23. Since there was only minimal impact of adding this feature, we chose to keep the simpler original model in the main text but have added the results of this sensitivity analysis to the supplement.

[1] We also considered an alternative method of allowing the base transmission rate to vary over time: fitting a separate base transmission rate for every week (or every two weeks). We decided against this for two reasons: first, introducing new free parameters would result in identifiability issues, and second, grid searching over all parameter configurations would be computationally infeasible due to the exponential blowup of the search space. Instead, we opted

for the approach described in the above paragraph, where we parametrically vary the base transmission rate as a function of the proportion-at-home and keep the number of free parameters at 3, just as in the original model.

Comment 10: Would the authors be able to look at time dependent variation of where transmission may have occurred (Figure 2d at different time slices of the epidemic?)

Thanks for this interesting idea! We implemented this analysis with a plot analogous to Figure 3c, which looks at where transmissions occurred at each POI category retrospectively. (Figure 2d predicts how many new infections would occur one month into the "future" if each category were reopened, so we opted for the retrospective analysis instead, which would allow us to use the actual recorded mobility data.)

Figure S21 shows the results of our analysis: we selected the POI categories that our model predicted contributed the most to infections, and plotted the predicted *proportion* of POI infections each category accounted for over time. Our model predicts that there was indeed time-dependent variation of where transmissions may have occurred, and furthermore that this variation aligns with known real-world events. For example, Full-Service Restaurants and Fitness Centers contributed to a smaller proportion of predicted infections over time, likely due to lockdown orders closing these types of POIs, while Grocery Stores remained steady or even grew in their contribution, which may be because they are essential businesses that remained open. Hotels & Motels also feature in these plots; most notably, the model predicts a peak in their contributed infections in Miami around mid-March --- this would align with college spring break, with Miami as a popular vacation spot for students. We believe these results provide additional evidence of the strength of our model, and we are grateful for the suggestion to include this analysis.

Reviewer 3

Comment 11: The authors have engaged with the comments in the main and this has improved the manuscript.

Thank you!

Comment 12: Putting Chicago in the main text is ok but it would be good to exactly replicate the Chicago-specific figures (i.e. Fig 2 and most of Fig 3) for each city in the supplement, so that readers can quickly go to them and look for the same patterns. This would mean keeping axes limits the same, so that readers could make a straightforward comparison. This is analogous to what they did in S10-19.

We agree it would be helpful to provide the information in Figure 2 and 3 for each MSA in the Supplement in a way which facilitates comparisons across MSAs. (Using the same axes for all MSAs for all plots unfortunately renders some of the plots very difficult to read, due to the very

different sizes of the MSAs, although we are sometimes able to do it.) The corresponding Supplementary Figures and Tables are as follows:

Figure 2a: Tables S2 and S3. Figure 2b:
Extended Data Figure 3.
Figure 2c: Extended Data Figure 4, Figure S20 (our previously submitted manuscript had no Supplementary counterpart for the right side of this subfigure; thank you for drawing our attention to this).
Figure 2d: Extended Data Figure 5, Figures S10-S19. Figure 3a, b:
These already contain results for all MSAs. Figure 3c: Figure S1.
Figure 3d: Extended Data Figure 6. Figure 3e:
Tables S4 and S5.
Figure 3f: Extended Data Figure 4.

We have edited the captions for Figures 2 and 3 to contain references for each subfigure to its corresponding Supplementary Figures/Tables.

Comment 13: Can the model be updated to include later data?

We considered and discussed this at length, but decided ultimately that extending the data would unfortunately present too much of a computational burden. This would require reprocessing data and rerunning all of our experiments, of which we would have to run at least 100,000 models and three million stochastic realizations. Overall, this update would take weeks, if not months, to execute carefully; this could lead to a long delay in sharing our work, likely reducing its impact on the scientific community and responses to the ongoing pandemic.

Comment 14: Could consider dropping the MSA acronym – define the area at first, and then call them "cities" to reduce the acronyms – CBGs and POIs in MSAs is quite a bit to handle.

New line 42, MSA is defined for a second time, as is CBG and POI.

Thanks for the suggestion, we agree that reducing the acronyms could help readability. We prefer not to use the term "city" since it is a bit imprecise here; for example, New York City has a population of only around 8 million, but the New York MSA has a population of around 20 million. However, we've replaced all usage of "MSA" with "metro area", which should be easier to understand.

Comment 15: The Model paragraph doesn't feel like results.

This is a fair point; thanks for raising it. We were attempting to follow the style of other *Nature* papers we referenced.

Comment 16: 101. this isn't "earlier findings" that showed this, this is epidemiological theory and doesn't need to be re-proven at every epidemic. Please put a reference to a classical text. Recent ones are useful for saying how much and how quickly for COVID.

We agree, and have added a citation!

Comment 17: 175: Are these visits by people from CBGS which have a different socioeconomic and racial makeup, not individuals in different socioeconomic and racial groups? It's not clear that the demographics and the movement are linked at an individual level.

Yes, these are visits from people with CBGs that have different socioeconomic and racial makeups. We have clarified the language here.

Comment 18: *187-205: Interesting paragraph. The table that backs it up is S4, but the ratios across cities vary by city ranging from no effect in Atlanta and Dallas to 10x in Philadelphia (which still seems pathological). The results paragraph does not convey this nuance. Suggest to unpack this result a little more, e.g. a set of plots, with a panel for each city, x axis is the decile categories, and y is the transmission rate, with a dot for the rate in each grocery store POIs as dots (with boxplot overlay). This will help readers understand the pattern in the increasing income deciles, the differences between cities, and convey more information than a ratio. Same could be done for the racial groups in S5.
And to add nuance to the paragraph reporting this result.

We agree that nuance in this section would be useful, and have added some discussion at the end of the paragraph highlighting the heterogeneity across MSAs. We have also added a reference to Supplementary Section 2, which further unpacks why the ratios are especially high in Philadelphia. For example, the model's prediction of a 10x ratio for grocery stores in Philadelphia, which the reviewer points out, is the result of two underlying factors: the average grocery store visited by lower-income CBGs has 5x the number of hourly visitors per square foot and visitors tend to stay 86% longer. We appreciate the suggestion for an additional plot, but we hesitate to add a figure that looks at every decile, since this is not something we analyze for any of the other plots (we just compare the top and bottom income deciles) and we worry it would confuse the reader. However, we hope that the edits to this section and reference to the supplemental analysis will add nuance for the reader, highlighting heterogeneity across MSAs and explaining where Philadelphia's higher ratios may come from.

Comment 19: 221 add "in 10 cities".

Thanks, we have clarified here.

Comment 20: 226: "clipped" was going to be replaced in this version?

Thanks for catching this! We have changed this here and throughout the text.

Comment 21: Extended 6: Green/red comparisons are difficult for colorblind people. Suggest to change colours.

Thank you for the suggestion! We have changed the colors to dark green and light tan.

Comment 22: 696: Clipping is still here.

Thanks for catching this! We have changed this here and throughout the text.

Comment 23: Suggest to add to the Discussion that the transmission rate at POIs is constant in the model, which may not be true anymore - important if people use it for prediction/policy.

To clarify, the transmission at POIs (as given by equation 8) is not constant over time in the model, since it depends on visit counts which change over time. We agree, however, that there may be time-varying factors (like mask-wearing or outdoor dining) which affect POI transmission and are not captured in the model, and we have updated the discussion to include more of those. ("various time-varying transmission-reducing behaviors").

Comments

Comment 24: I agree with comment 35 for caution. This paper might accidentally suggest that closing restaurants and other POIs in poor and black neighbourhoods may be a transmission-reducing intervention. So, suggest to be quite careful with it.

We agree that this would be an unfortunate interpretation of our results, and have added language to the Discussion to emphasize caution and encourage weighing the potential public health benefits of shutdowns against the economic harms they may cause.

Comment 25: *39 and 87: Line 143 explicitly says it excludes schools. As does 730-31. This is very confusing, and it needs to be much clearer to readers how schools are in the model but not in the POI analysis.

Thank you for pointing this out - we have clarified the text at both these points. To summarize: we include schools in our model (i.e., we assume that people visit these POIs, and that transmissions occur there); we simply do not perform an analysis focusing specifically on schools, because we are not confident that all people who visit schools are included in the data.

Comment 26: *42: This seems like a very confusing way of presenting the data. It's not intuitive what an "acceptable" loss value is, whether it should be the same from city to

city, and what are the model ranks? Can you put x = simulated parameter, y = value of the estimate?

We apologize for the confusion. In these simulations, for all of the cities (MSAs), we exactly recover the simulated parameters (which correspond to the best-fit parameters that we find for each city). Thus, a figure that plots the value of the estimate against the simulated parameter would show all points lying on the y=x line. Figure S8 unpacks this in a bit more detail by showing the loss that each parameter set obtains on the simulated data. The main takeaway is that the leftmost point (which corresponds to the parameters that best fit the real data and that we use as ground truth for the simulated data) also obtains the lowest loss on the simulated data. We have clarified the text, the figure caption, and the axis labels. Thank you for pointing this out.

Comment 27: I don't understand S9 very well. If it's RMSE I think converting it to a ratio makes interpretation much harder, so suggest to make this a heatmap of the RMSE or Poisson likelihood instead.

Thank you for the good point. We've edited the heatmap in S9 so that the legend now shows the actual RMSE and not the ratio.

Comment 28: The white area indeed appears to show that these parameters are not identifiable in some cities – Atlanta, LA, DC .

We agree that there is uncertainty in the parameter estimates for the cities that the reviewer mentions. We reflect this uncertainty in our errorbars, which contain the predictions of all of the parameter estimates within the white area. We have edited the text in M5.3 ("Parameter identifiability") to clarify this. Thank you for pointing it out.

Comment 29: Also, suggest to add what every Greek letter is in words in the supplementary figures – readers will likely have forgotten.

Thank you for the suggestion. We have edited the axis labels and the captions as appropriate in the supplementary figures.

Comment 30: 94: Philadelphia's big values do affect the medians reported, though, so may be having an outsize effect on the overall message.

We share the reviewer's concern about outliers affecting our overall results, which is why we report the median across MSAs throughout. The median is specifically chosen as a statistic which is more robust to outliers than, e.g., the mean. For example, regardless of whether Philadelphia has a value 10x that of the next-largest MSA, or 1x that of the next-largest MSA, the median across MSAs will be the same.

Comment 31: 98: suggest to add the number to methods or a caption somewhere – it's interesting to readers and will allow other people to more easily use the results of the model (especially if you give a city-specific number).

Good idea; we have added this number to the text.