INFERENCE IN NETWORK-BASED EPIDEMIOLOGICAL SIMULATIONS WITH PROBABILISTIC PROGRAMMING

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate epidemiological models require parameter estimates that account for mobility patterns and social network structure. This work applies probabilistic programming to infer parameters in agent-based models. We represent mobility networks as degree-corrected stochastic block models and estimate their parameters from cell-phone co-location data. We use these networks in probabilistic programs to simulate the evolution of an epidemic, and condition on reported cases to infer disease transmission parameters. Our experiments demonstrate that the resulting models improve the accuracy-of-fit in multiple geographies relative to baselines that do not model network topology.

1 INTRODUCTION

Compartmental models of infectious diseases track the number of individuals in different stages of disease progression and with varying granularity. The least granular models use *global* compartments that track the total number of susceptible, infected, and removed individuals (Kermack & McKendrick, 1927), and rely on an assumption of uniform mixing throughout the population. This simplifies computation, but also imposes limitations on dynamics, such as the fact that there can only be a single wave of infections (Diekmann & Heesterbeek, 2000). More sophisticated simulators stratify the population according to age (Chinazzi et al., 2020) and/or geography (Chang et al., 2021) to account for variations in the frequency of interactions. The most fine-grained simulation models are agent-based (Grefenstette et al., 2013), and can account for the fact that highly-connected individuals are more likely to both contract and spread a disease (Mistry et al., 2021).

One of the challenges in developing disease models that can describe fine-grained social interactions and regional interventions is estimating parameters. In ongoing epidemics, this challenge is compounded by the fact that existing parameter estimates become invalid, since public policy responses can dramatically affect mobility and social interaction. In this setting, we would like to deploy models that can incorporate as much available data as possible, and apply techniques for approximate inference to account for uncertainty in the resulting parameter estimates and model predictions.

In this work, we present a case study in the use of probabilistic programming to infer parameters of agent-based disease simulations. We model disease spread on subsampled social networks constructed from cellphone co-location data (Sec. 2.1), where nodes transition stochastically between disease states (Alg. 1), and fit the parameters of these models to regional infection data using Black-box Variational Inference (BBVI, Wingate & Weber (2013); Ranganath et al. (2014), Sec. 2.2).

Related Work. We build on a large body of work in disease modelling, parameter inference, and probabilistic programming techniques. For a comprehensive discussion, see Appendix A.

2 METHODOLOGY

We develop a Network-SEIR model with two components. The first is a network-topology model, in the form of a degree-corrected stochastic block model (DCSBM, Karrer & Newman (2011)), that describes contact patterns in the population. The second is an agent-based compartmental model that simulates disease transmission at the level of individuals. We obtain point estimates of network-topology parameters from cell-phone co-location data (SafeGraph, 2020). To estimate transmission parameters, we incorporate the agent-based model into a probabilistic program that defines a prior over parameters and a likelihood for reported case counts. This defines a Bayesian posterior over parameters that we approximate using variational inference.

2.1 NETWORK-SEIR MODEL

Network Topology Model: We simulate a network in which vertices V^c represent individuals in communities $c = 1, \ldots, C$. Edges $E \quad V \quad V$ represent two types of social contact: (1) co-location at regional points of interest (POIs) (such as a coffee shop or a library), and (2) cohabitation. Edge weights W_{uv} describe the frequency of contact between individuals u and v. To generate the network, we fit a DCSBM to aggregated, anonymized cellphone GPS data that tracks visits to POIs in a given county (SafeGraph, 2020), and simulate from this model (See Appendix C for details).

Disease Transmission Model: Nodes transition between 4 states: susceptible, exposed, infected, and removed. We use S^t, E^t, I^t, and R^t to refer to the subset of nodes in each state at time t. We approximate the exponentially-distributed probability of exposure using a first-order Taylor expansion,

$$p\ v \in E^{t+1} \mid v \in S^t \quad = \quad 1 \quad \exp \quad E_{\text{pressure}} \quad I_{\text{pressure}} \quad ' \quad \min \quad E_{\text{pressure}} + I_{\text{pressure}}; 1 \ : \tag{1}$$

Here E_{pressure} and I_{pressure} are defined in terms the network weights for exposed neighbors $N_E^t(v)$ and infected neighbors $N_I^t(v)$, which are scaled by time-dependent parameters $\ _E^t$ and $\ _I^t$,

$$E_{\text{pressure}} = \sum_{u \in N_E^t(v)} W_{uv} \ _E^t; \qquad I_{\text{pressure}} = \sum_{u \in N_I^t(v)} W_{uv} \ _I^t: \tag{2}$$

Once individuals are exposed, they transition to infected and removed states with constant probabilities $\ $ and $\ $,

$$p\ v \in I^{t+1} \mid v \in E^t \quad = \quad ; \qquad p\ v \in R^{t+1} \mid v \in I^t \quad = \quad : \tag{3}$$

We describe the resulting disease simulator f_{SEIR} in Algorithm 1. The inputs to this model are the simulated graph G, initial rates of exposure $\ ^c$ in each community, $\ $, $\ $, and values for $\ _E$ and $\ _I$ at time points $t_1, \ldots, t_K$, from which we define parameters at time t using linear interpolation.

Algorithm 1: Stochastic Disease Simulator f_{SEIR}

Function $f_{\text{SEIR}}(G, \ ^{1:C}, f \ _E^{t_1} \ldots \ _E^{t_K} g, f \ _I^{t_1} \ldots \ _I^{t_K} g, \ , \ , T)$**:**
- **for** $c \quad 1$ **to** C **do** `// Initial Exposure`
  - **for** $v \in V^c$ **do if** $\text{Unif}(0,1) <$ $\ ^c$ **then** $v \ ! \ E^1$ **else** $v \ ! \ S^1$
- **for** $t \quad 1$ **to** $T \quad 1$ **do** `// Simulate T days`
  - $\ _E^t \quad \text{INTERPOLATE}(\ _E^{t_1}, \ldots, \ _E^{t_K})$; $\ _I^t \quad \text{INTERPOLATE}(\ _I^{t_1}, \ldots, \ _I^{t_K})$
  - **for** $v \in S_t$ **do**
    - $E_{\text{pressure}} \quad \sum_{u \in N_E^t(v)} W_{uv} \ _E^t$; $I_{\text{pressure}} \quad \sum_{u \in N_I^t(v)} W_{uv} \ _I^t$
    - **if** $\text{Unif}(0,1) < (E_{pressure} + I_{pressure})$ **then** $v \ ! \ E^{t+1}$
  - **for** $v \in E^t$ **do if** $\text{Unif}(0,1) <$ $\ $ **then** $v \ ! \ I^{t+1}$
  - **for** $v \in I^t$ **do if** $\text{Unif}(0,1) <$ $\ $ **then** $v \ ! \ R^{t+1}$
- **return** $(\sum_{t=1}^{j} I^t)_{j=1}^{T}$ `// List of Cumulative Infections`

2.2 PROBABILISTIC INFERENCE

Generative Model. We define a prior distribution over disease hyperparameters which factors into independent logistic normal distributions over the disease hyperparameters,

$$\ ^c \quad \text{LN} \quad \ ^c; \ ^2 \qquad \text{for} \quad c \in 1, \ldots, C; \tag{4}$$

$$\ _E^{t_k} \quad \text{LN} \quad \ _E^k; \ ^2_E \ ; \qquad \ _I^{t_k} \quad \text{LN} \quad \ _I^k; \ ^2_I \qquad \text{for} \quad k \in 1, \ldots, K; \tag{5}$$

$$\quad \text{LN} \quad ; \ ^2 \ ; \qquad \quad \text{LN} \quad ; \ ^2 \ : \tag{6}$$

Given these disease transmission parameters, our stochastic simulator f_{SEIR} implies a prior over latent variables $z_I^{1:T}$ representing the expected number of infected individuals on day t,

$$z_I^{1:T} = f_{\text{SEIR}}(G; \ ^{1:C}; f \ _E^{t_1} \ldots \ _E^{t_K} g; f \ _I^{t_1} \ldots \ _I^{t_K} g; \ ; \ ; T): \tag{7}$$

Last, our likelihood is a Gaussian whose noise scales with time, graph size, and hyperparameter $\ $,

$$x^{1:T} \quad N \quad z_I^{1:T}; \ _{obs}(G; \ ; t)^2 \ : \tag{8}$$

Variational Distribution. To approximate the model posterior over latent variables, we define a variational distribution $q(\ ;\ _E;\ _I;\ ;\ )$ which mirrors the prior of the generative model, with parameters $= (\ ^{1:C};\ \ ;\ f\ ^{t_1}_E \ldots\ ^{t_K}_E g;\ f\ ^{t_1}_E \ldots\ ^{t_K}_E g;\ f\ ^{t_1}_I \ldots\ ^{t_K}_I g;\ f\ ^{t_1}_I \ldots\ ^{t_K}_I g;\ ;\ ;\ ;\ )$ for the individual logistic-normal distribution. The variational distribution factors as

$$q\ = \left(\prod_{c=1}^{\ } q_{\ c;\ }(\ ^c)\right)\left(\prod_{k=1}^{\ } q_{\ ^{t_k}_E;\ ^{t_k}_E}(\ ^{t_k}_E)\right)\left(\prod_{k=1}^{\ } q_{\ ^{t_k}_I;\ ^{t_k}_I}(\ ^{t_k}_I)\right) q_{\ ;\ }(\ )\, q_{\ ;\ }(\ ): \quad (9)$$

Note that we share the variance parameter for all communities c. We estimate using a BBVI implementation in Gen (Cusumano-Towner et al., 2019).

3 Experiments

We grid search over hyperparameters for our probabilistic model, our prior over disease parameters, and the BBVI algorithm. A table of the explored parameter ranges can be found in Appendix E.1. Note that we scale our data to heuristically account for under-reporting of infections (See E.3)

3.1 Validation on Simulated Data

To confirm our inference procedure is well-calibrated, we perform inference using simulated infection counts. We perform this experiment with 6 different time-varying patterns for ^t_E such as low-high-low. For each case, we generate synthetic data by running our generative model with fixed disease parameters. Then, we run our inference procedure and compare our learned variational distribution to the ground truth disease hyperparameters. For each setting, we generate inferred cumulative infection curves and compute mean daily absolute error (MDAE) as defined in Eq. (10). Note that our inference model is conditioned only on infection counts, and for our overparametrized model, there exist multiple ambiguous solutions whose output infection counts would be of similar quality. Thus, we compare our learned model to the ground truth in the space of infection counts.

$$\text{MDAE} \quad \mathop{\mathbb{E}}_{q\ (z)} \frac{k f_{\text{SEIR}}(z) \quad xk_1}{TN} \qquad \frac{1}{N}\frac{1}{ST} \sum_s \sum_t f_{\text{SEIR}}(z_s)^t \quad x^t \qquad (10)$$

Figure 1: Validation on a simulated model on Miami-Dade topology. Generated disease trajectory using "high-low-high" $_E = 0{:}45; 0{:}1; 0{:}45$ (left) and "low-high-low" $_E = 0{:}1; 0{:}45; 0{:}1$ (right).

Table 1: MDAE for different counties and disease dynamics

County	low	high	low-high	high-low	low-high-low	high-low-high
Miami-Dade	0.0052	0.0046	0.0042	0.0051	0.0043	0.0050
Los Angeles	0.0037	0.0046	0.0050	0.0044	0.0048	0.0047

3.2 Fitting Parameters in Different Regions

Next, we apply our method as described above to networks constructed using regional data from Los Angeles, CA, and Miami-Dade, FL. We fit our guide distribution over the parameters for Network-SEIR by conditioning on the cumulative infection counts in each county, as reported on the Johns Hopkins University Center for Systems Science and Engineering dashboard (Dong et al., 2020).

We find that our network model with parameters learned through probabilistic inference fits observed data better than two alternative baselines: (1) Compartmental CE-EM: a compartmental SEIR model with parameters fit using CE-EM (Menda et al., 2021), and (2) Network R_t-Analytic: a simplified analytic solution for f_{SEIR} parameters (see Appendix D for details). Quantitative and qualitative results are shown in Table 2 and Figure 2 respectively.

Disease Model	Fitting Method	LA-MDAE	Miami-Dade-MDAE
Compartmental	CE-EM	0:0251	0:0161
Network	R_t-Analytic	0:0075	0:0086
Network	BBVI	**0:0029**	**0:0053**

Table 2: Comparison of the MDAE of different disease models and fitting methods. Our method can fit different regions with different multi-peaked dynamics as in Figure 4.

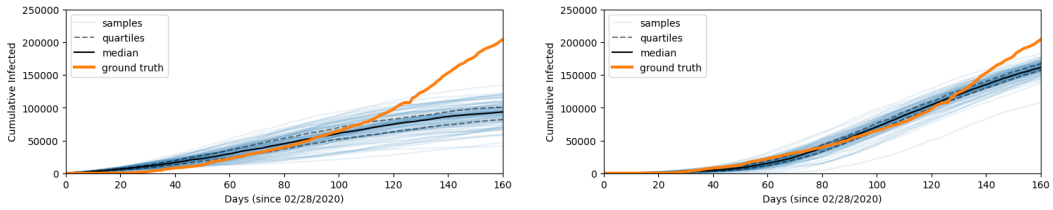

Figure 2: Inference for Los Angeles using R_t-Analytic fitting (left) and our method (right).

3.3 Inferring Starting Communities

Since our variational distribution includes means for the proportion of initial exposure in each community c, we can interpret learned values for the parameters ^{c} as indicative of which communities were likely to have had higher initial exposure given the observed disease data. Note this is not the same as inferring the actual precise location of the initial exposure within a region since our cumulative global infection data is too coarse to deduce this. There are many possible initial exposure scenarios which may result in similar global infection data. Rather, we conclude the location of certain communities in the network topology is more consistent with observed disease dynamics.

In Figure 3, we see that for high observational noise the inferred parameters are close to the uniform prior ^{c} = :05, whereas for low observation noise, we find an initial exposure in communities 1; 9; 10; 13 is more consistent with the observed daily cumulative infection data.

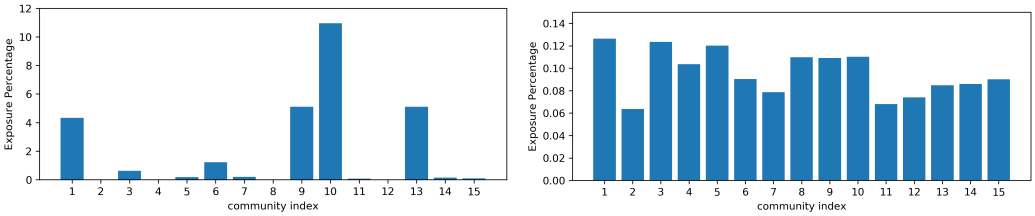

Figure 3: The network topology of Miami-Dade county is modeled using 15 communities which correspond to actual geographic areas. We plot ^{c} for 1 c 15: In the left plot, we use = 0:00025, a tighter observational noise than the right plot where = 0:0005.

4 Conclusion

We use probabilistic programming to allow sophisticated agent-based, topological disease models to be fit to real data. This enables improved accuracy in downstream simulations of disease control interventions, reducing the need for real-world experiments. Future work includes better guide distributions, time-varying network topologies, control variates, and amortized multi-region inference.

REFERENCES

David Adam. A guide to R—the pandemic's misunderstood metric. *Nature*, 583(7816):346–348, 2020.

Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan : A Probabilistic Programming Language. *Journal of Statistical Software*, 76(1), January 2017. doi: 10/b2pm.

Serina Chang, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David Grusky, and Jure Leskovec. Mobility network models of covid-19 explain inequities and inform reopening. *Nature*, 589(7840):82–87, 2021.

Matteo Chinazzi, Jessica T Davis, Marco Ajelli, Corrado Gioannini, Maria Litvinova, Stefano Merler, Ana Pastore y Piontti, Kunpeng Mu, Luca Rossi, Kaiyuan Sun, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (covid-19) outbreak. *Science*, 368(6489): 395–400, 2020.

Marco F. Cusumano-Towner, Feras A. Saad, Alexander K. Lew, and Vikash K. Mansinghka. Gen: A general-purpose probabilistic programming system with programmable inference. In *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*, PLDI 2019, pp. 221–236, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-6712-7. doi: 10.1145/3314221.3314642. URL `http://doi.acm.org/10.1145/3314221.3314642`.

Christian Schroeder de Witt, Bradley Gram-Hansen, Nantas Nardelli, Andrew Gambardella, Rob Zinkov, Puneet Dokania, N Siddharth, Ana Belen Espinosa-Gonzalez, Ara Darzi, Philip Torr, et al. Simulation-based inference for global health decisions. *arXiv preprint arXiv:2005.07062*, 2020.

Jonas Dehning, Johannes Zierenberg, F Paul Spitzner, Michael Wibral, Joao Pinheiro Neto, Michael Wilczek, and Viola Priesemann. Inferring change points in the spread of covid-19 reveals the effectiveness of interventions. *Science*, 369(6500), 2020.

Odo Diekmann and Johan Andre Peter Heesterbeek. *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*, volume 5. John Wiley & Sons, 2000.

Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534, 2020.

Nathan Eagle and Alex Pentland. Reality mining: Sensing complex social systems, 2005.

Seth Flaxman, Swapnil Mishra, Axel Gandy, H Juliette T Unwin, Thomas A Mellan, Helen Coupland, Charles Whittaker, Harrison Zhu, Tresnia Berah, Jeffrey W Eaton, et al. Estimating the effects of non-pharmaceutical interventions on covid-19 in Europe. *Nature*, 584(7820):257–261, 2020.

John J Grefenstette, Shawn T Brown, Roni Rosenfeld, Jay DePasse, Nathan TB Stone, Phillip C Cooley, William D Wheaton, Alona Fyshe, David D Galloway, Anuroop Sriram, et al. Fred (a framework for reconstructing epidemic dynamics): an open-source software system for modeling infectious diseases and control strategies using census-based populations. *BMC public health*, 13 (1):1–14, 2013.

Mathieu Génois and Alain Barrat. Sociopatterns datasets, December 2017. URL `https://doi.org/10.5281/zenodo.2540795`.

Shaobo He, Yuexi Peng, and Kehui Sun. Seir modeling of the covid-19 and its dynamics. *Nonlinear Dynamics*, 101(3):1667–1680, 2020.

David Holtz, Michael Zhao, Seth G Benzell, Cathy Y Cao, Mohammad Amin Rahimian, Jeremy Yang, Jennifer Allen, Avinash Collis, Alex Moehring, Tara Sowrirajan, et al. Interdependence and the cost of uncoordinated responses to covid-19. *Proceedings of the National Academy of Sciences*, 117(33):19837–19843, 2020.

Brian Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Phys. Rev. E*, 83:016107, Jan 2011. doi: 10.1103/PhysRevE.83.016107.

William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M Blei. Automatic differentiation variational inference. *arXiv preprint arXiv:1603.00788*, 2016.

Qun Li, Xuhua Guan, Peng Wu, Xiaoye Wang, Lei Zhou, Yeqing Tong, Ruiqi Ren, Kathy S.M. Leung, Eric H.Y. Lau, Jessica Y. Wong, Xuesen Xing, Nijuan Xiang, Yang Wu, Chao Li, Qi Chen, Dan Li, Tian Liu, Jing Zhao, Man Liu, Wenxiao Tu, Chuding Chen, Lianmei Jin, Rui Yang, Qi Wang, Suhua Zhou, Rui Wang, Hui Liu, Yinbo Luo, Yuan Liu, Ge Shao, Huan Li, Zhongfa Tao, Yang Yang, Zhiqiang Deng, Boxi Liu, Zhitao Ma, Yanping Zhang, Guoqing Shi, Tommy T.Y. Lam, Joseph T. Wu, George F. Gao, Benjamin J. Cowling, Bo Yang, Gabriel M. Leung, and Zijian Feng. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *New England Journal of Medicine*, 382(13), March 2020. doi: 10.1056/NEJMoa2001316.

Kunal R Menda, Lucas Laird, Mykel J Kochenderfer, and Rajmonda S Caceres. Explaining covid-19 outbreaks with reactive seird models. *medRxiv*, 2021.

Dina Mistry, Maria Litvinova, Ana Pastore y Piontti, Matteo Chinazzi, Laura Fumanelli, Marcelo FC Gomes, Syed A Haque, Quan-Hui Liu, Kunpeng Mu, Xinyue Xiong, et al. Inferring high-resolution human mixing patterns for disease modeling. *Nature communications*, 12(1):1–12, 2021. doi: https://doi.org/10.1038/s41467-020-20544-y.

Gaurav Pandey, Poonam Chaudhary, Rajan Gupta, and Saibal Pal. Seir and regression model based covid-19 outbreak predictions in India. *arXiv preprint arXiv:2004.00958*, 2020.

Nicola Perra. Non-pharmaceutical interventions during the covid-19 pandemic: A review. *Physics Reports*, 2021. doi: https://doi.org/10.1016/j.physrep.2021.02.001.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial intelligence and statistics*, pp. 814–822. PMLR, 2014.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pp. 1278–1286. PMLR, 2014.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pp. 400–407, 1951.

SafeGraph, 2020. URL `https://docs.safegraph.com/docs/weeklypatterns`.

Marcel Salathé, Maria Kazandjieva, Jung Woo Lee, Philip Levis, Marcus W Feldman, and James H Jones. A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences*, 107(51):22020–22025, 2010.

John Salvatier, Thomas V. Wiecki, and Christopher Fonnesbeck. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55, April 2016. doi: 10.7717/peerj-cs.55.

Jan-Willem van de Meent, Brooks Paige, David Tolpin, and Frank Wood. Black-Box Policy Search with Probabilistic Programs. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pp. 1195–1204, 2016.

David Wingate and Theophane Weber. Automated variational inference in probabilistic programming. *arXiv preprint arXiv:1301.1299*, 2013.

Frank Wood, Andrew Warrington, Saeid Naderiparizi, Christian Weilbach, Vaden Masrani, William Harvey, Adam Scibior, Boyan Beronov, and Ali Nasseri. Planning as inference in epidemiological models. *arXiv preprint arXiv:2003.13221*, 2020.

A Related Work

Compartmental and Agent-based Disease Models. A large amount of work has been done using variants of the global SEIR model. For example, Pandey et al. (2020) fit parameters for a standard SEIR model; He et al. (2020) propose an extended model with additional compartments for Quarantined and Hospitalized individuals. More sophisticated disease models incorporate social and geographic network structure. Holtz et al. (2020) develop a mobility network based on SafeGraph GPS data and Facebook social connectivity data in order to examine how social distancing policies affect regional mobility patterns, especially between neighboring municipalities. Chang et al. (2021) construct a mobility network using SafeGraph data, with time-varying edge weights, and overlay this network with a simple compartmental SEIR model containing 3 free parameters. Chinazzi et al. (2020) use an epidemic simulator where many small regions, each undergoing a simple compartmental SEIR process, are connected by a stochastic transportation grid. They use Monte Carlo likelihood analysis to fit the epidemic parameters. For a more comprehensive list of disease models, both compartmental and network-based, see a recent survey by Perra (2021).

Probabilistic Programming for Disease Models. Methods from probabilistic programming have been applied to epidemiological models in a number of contexts. Flaxman et al. (2020) investigate the effectiveness of non-pharmaceutical interventions by modeling the disease with a discrete stochastic SIR model implemented in Stan (Carpenter et al., 2017) and estimating the daily reproduction number using Hamiltonian Monte Carlo. Dehning et al. (2020) infer the parameters of a differentiable SIR model implemented in PYMC3 (Salvatier et al., 2016) using a combination of automatic differentiation variational inference (Kucukelbir et al., 2016) and MCMC. These approaches differ from ours by considering models that do not account for network topology. There has also been work that applies planning-as-inference methods for probabilistic programs (van de Meent et al., 2016) to disease models (Wood et al., 2020; de Witt et al., 2020). This work also applies to simulation-based models, but focuses on planning rather than parameter estimation.

B Background

B.1 Compartmental SEIR model

In a traditional global compartmental SEIR model, the population is separated into 4 compartments representing Susceptible (S), Exposed (E), Infected (I), and Removed (R) individual, where the total population is $N = S + E + I + R$. Dynamics are modeled by the differential equations,

$$\frac{dS}{dt} = \quad \frac{IS}{N}; \qquad \frac{dE}{dt} = \frac{IS}{N} \quad E; \qquad \frac{dI}{dt} = \quad E \quad I; \qquad \frac{dR}{dt} = \quad I: \tag{11}$$

In some cases, this model includes a natural birth rate and death rate for the population. We do not account for these effects in our network model for two reasons. Newborn individuals will not contribute meaningfully to the spread of infection; since they cannot interact with other individuals alone. Individuals who die of natural causes make the overall mobility network slightly more sparse, and omitting this effect causes only a slight, conservative error in our predictions.

B.2 Stochastic Variational Inference

ELBO Objective. We use Stochastic Variational Inference (SVI) to approximate the intractable posterior $p(z|x)$ by optimizing the parameters of a variational distribution $q(z; \ )$ (usually denoted $q \ $), which we are able to directly draw samples from. A common approach is to find the parameters that minimize the exclusive KL divergence between the variational approximation and the posterior: $= \arg\min \ \mathrm{KL}(q \ (z) \| p(z \ j \ x))$,

$$\mathrm{KL}(q \ \| p) = \mathrm{E}_{z \ q} \quad \log \frac{q \ (z)}{p(z|x)} \tag{12}$$

$$= \mathrm{E}_{z \ q} \quad \log \frac{q \ (z)}{p(x; z)} + \log p(x) \tag{13}$$

$$= \mathrm{E}_{z \ q} \quad \log \frac{q \ (z)}{p(x; z)} \quad + \log p(x): \tag{14}$$

Minimizing this exclusive KL divergence is equivalent to maximizing a lower bound on the log marginal likelihood of the data. Since the marginal likelihood $p(x)$ is also referred to as the "evidence", this bound is called the Evidence Lower BOund (ELBO)

$$L = \underbrace{\mathbb{E}_{z \sim q}\left[\log \frac{p(x,z)}{q(z)}\right]}_{\text{ELBO}} = \log p(x) - \underbrace{KL(q \,||\, p)}_{0} \quad \underbrace{\log p(x)}_{\text{log evidence}}. \tag{15}$$

Score Function ELBO Gradient. Variational Autoencoders (Kingma & Welling, 2013; Rezende et al., 2014) and related methods optimize the ELBO by computing reparameterized gradient estimates, which require the generative model $p(x,z)$ to be differentiable with respect to the latent variables z. However, generative models that incorporate discrete variables and control flow will not be differentiable, and it may be infeasible or undesirable to find a continuous approximation to make such a model differentiable. We instead maximize the ELBO using a so-called "score-function" gradient estimator that does not require reparameterization,

$$\begin{aligned}
\nabla L &= \nabla \mathbb{E}_{z \sim q}\left[\log \frac{p(x,z)}{q(z)}\right] && (16)\\
&= \int dz\, \nabla q(z) \log \frac{p(x,z)}{q(z)} && (17)\\
&= \int dz\, \nabla q(z) \log \frac{p(x,z)}{q(z)} + q(z) \nabla \log \frac{p(x,z)}{q(z)} && (18)\\
&= \int dz\, q(z) \log \frac{p(x,z)}{q(z)} \nabla \log q(z) + q(z) \nabla \log \frac{p(x,z)}{q(z)} && (19)\\
&= \int dz\, q(z) \log \frac{p(x,z)}{q(z)} \nabla \log q(z) - q(z) \nabla \log q(z) && (20)\\
&= \mathbb{E}_{z \sim q}\left[\log \frac{p(x,z)}{q(z)} - 1 \nabla \log q(z)\right] && (21)\\
&= \mathbb{E}_{z \sim q}\left[\log \frac{p(x,z)}{q(z)} \nabla \log q(z)\right] - \underbrace{\mathbb{E}_{z \sim q}[\nabla \log q(z)]}_{=0} && (22)\\
&\approx \frac{1}{S}\sum_{s=1}^{S} \log \frac{p(x,z_s)}{q(z_s)} \nabla \log q(z_s); \quad z_s \sim q(z). && (23)
\end{aligned}$$

Here $\hat{b}$ is known as a baseline estimator, which in the Gen implementation is simply $\hat{b} = 0$.

This approach is generally referred to as Blackbox Variational Inference (BBVI) (Ranganath et al., 2014) and was originally proposed specifically (in the context of probabilistic programming systems) by Wingate & Weber (2013). Note that this gradient estimate does not require the generative model to be differentiable; we only need to sample from the variational distribution and evaluate both the generative model and variational models pointwise. While score function gradient estimators are known to have higher variance and require smaller step sizes than other approaches, they are unbiased estimates of the gradient, and hence our inference will converge assuming Robbins and Monro conditions for stochastic approximation algorithms (Robbins & Monro, 1951).

C Network Modeling

We use mobility data from SafeGraph (SafeGraph, 2020), a data company that provides aggregated information on foot-traffic to points-of-interest (POIs) across the United States. POIs in the data include establishments such as grocery stores, restaurants, and schools. The foot-traffic data provided can be considered a sample of the population, and is based on location information from a panel of anonymous, opt-in mobile devices. Devices in this data are assigned to a home Census Block Group (CBG), the smallest geographical unit for which population data is reported in the U.S. Census. Given a pair of devices in a CBG, we use visit data to POIs as well information about time duration at that POI to model person-to-person contact patterns.

We use a Degree-Corrected Stochastic Block model (DCSBM) to generate a synthetic contact network, considering that such a model allows us to represent community structure as well as heterogeneous degree distribution, both important characteristics of social interaction and mobility patterns (Karrer & Newman, 2011). The DCSBM has two parameters: 1) a partition of vertices $\{V^1, \ldots, V^C\}$ into C communities, 2) a symmetric matrix $P \in \mathbb{R}^{C \times C}$ of edge probabilities, where element p_{rs} gives the probability of an edge existing between any two vertices $u \in V^r$ and $v \in V^s$.

Let $G = (V, E)$ represent the network instance generated using DCSBM, where V is the vertex set and $E \subseteq V \times V$ is the edge set. Each node v in the topology represents an individual that resides in a community, i.e. a census block group $\mathrm{CBG}_i \in \{1, \ldots, C\}$. Each edge $e \in E$ is associated with a type in $\{h, c\}$, representing household and community edges respectively. POI co-location data are used to generate the edge probability matrix P that models community edge types. The existence of an edge between $u \in V^r$ and $v \in V^s$ is given by the cross-correlation score:

$$p_{rs} = \frac{\sum_i (l_{ri} - \bar{l}_r)(l_{si} - \bar{l}_s)}{\sqrt{\sum_i (l_{ri} - \bar{l}_r)^2 \sum_i (l_{si} - \bar{l}_s)^2}}, \tag{24}$$

where l_r is a vector of length n_{POI} of weekly visit count estimates from devices in CBG_r to each POI. We perform a degree-correction procedure on this network to yield a heterogeneous degree distribution between nodes in different blocks. The parameter for degree-correction is selected such that the node with the largest degree in each block has a degree in the range of 50-100, which is consistent with other realistic social contact networks of similar sizes (Eagle & Pentland, 2005; Génois & Barrat, 2017; Salathé et al., 2010).

Household edges are added to the resultant DCSBM based on census survey data on the number of single-person up to seven-person households. Each household is represented as a subset of fully connected nodes. Nodes in the same CBG can be connected with both household and community edges, while nodes in different blocks can only be connected with community edges.

The weight of edge uv represents a modeled estimate of the total time (in minutes) that individuals u and v might spend in proximity to one another – that is, at the same POI at the same time. Weights are aggregated over 1 week duration. SafeGraph provides the median dwell time d_p of every visit to POI p. This is used to construct a dwell time matrix D of size $C \times n_{\mathrm{POI}} \times M$ where M is the total number of minutes POIs are open (assumed to be 10 hours per day, 7 days per week = 4,200 minutes). For each POI p we estimate l_{rp} total visits from CBG_r, where each visit starts at a randomized minute in the interval $(0, M)$ and lasts for d_p minutes. An entry D_{rpt} indicates the number of visitors from CBG_r present at POI p at minute t. The weight on a community edge between nodes $u \in V^r$ and $v \in V^s$ is given as

$$W^c_{uv} = \int_0^M dt D_r(t) D_s(t). \tag{25}$$

The weight on household edges are set to the maximum weekly interaction time such that $W^h = 10,080$. We model networks from three diverse geographical regions in the US: Middlesex County (MA), Los Angeles County (CA) and Miami-Dade County (FL) and select CBGs that contribute to the top 10% of mobility data available for each county during the month of February 2020.

D Analytic R_t-matched Parameters.

We derive constant SEIR parameters β_E, β_I, so that each lead to COVID-19 R_t values reported in literature. We compute an effective reproductive number R_t as

$$R_t(\beta_E, \beta_I) = \mathbb{E}_{u \in V}\left[\sum_{v \in N(v)} \frac{W_{uv}\beta_E}{1 - (1 - W_{uv}\beta_E)(1 - \beta_I)}\right]. \tag{26}$$

COVID-19 R_t values across the world have been reported to be roughly between 1.4 and 5 (Adam, 2020; Li et al., 2020). These values reflect primarily asymptomatic transmission, therefore a value

of R_t in this range is suitable for the estimation of $_E$. For our calculations we set $R_t = 3$ and $= 0.048$. We approximate the expectation over neighbors by considering the top 25% highest degree nodes. Note that this solution implicitly assumes uniform edge probabilities and constant average degree, a deviation from our network generation modeling choices, where we explicitly model community structured interactions and heterogeneous degree distribution.

Note that $_E$ and $_I$ cannot be simultaneously identified from R_t alone. To estimate $_I$, we again apply (26) by solving $R_t(\ _I;\ ) = 0.25$ for $_I$. For $_I$, we use an R_t of 1=12 the size, since this ratio holds for the R_t values computed from the compartmental model fit to data using CE-EM. Specifically, we found the expected number of exposures caused by each infected individual is 1/12 the number caused by each exposed individual.

Since the R_t-Analytic method provides only a point estimate of parameters, to fairly compare this method to our method as in Table 2, we set the variance of our variational posterior q to 0.

E Additional Experimental Details

E.1 Hyperparameter Tuning

We list the hyperparameter values we tried in Table 3. We selected the bold parameters based on training stability and goodness of fit.

Hyperparameter	Range
Observational Noise Scaling ( )	0.00025, 0.00035, **0.0005**
Learning Rate	10^{-6},10^{-5},5 10^{-4}, 10^{-4}
Samples Per Iteration	20,50,75,**120**, 200
Prior Initial Exposed Logit Mean ( )	7, 6.5, 6
Prior Initial Exposed Logit StdDev ( )	0.0, 1.0, 1.6
Prior $_E$ Logit Mean ($_E$)	2.2, 1.39, 0.85
Prior shared disease parameter StdDev ($_E$; $_I$; ;)	0.0, 0.47, 1.16

Table 3: Hyperparameter tuning ranges. Bold values correspond to best performance.

E.2 Setting Initial Exposed for Baselines

Both the compartmental SEIR model and topological-SEIR model require an initial exposure percentage. However, neither the CE-EM nor the R_t-analytic method fits this value. Thus to compare our models fairly, we have used the best MDAE over six different initial exposures: 1) uniform exposure of 0.5%, 2) uniform exposure of 1%, 3) uniform exposure of 2%, 4) uniform exposure of %5, 5) inferred initial exposure using BBVI with high observational noise $= 0.0005$, 6) inferred initial exposure using BBVI with low observational noise $= 0.00025$.

E.3 Data Processing

For infected individual counts we use data from Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) dashboard (Dong et al., 2020). We use data February 29, 2020 until August 9, 2020. We process data using a rolling median to remove spikes from late reporting and negative counts inserted to correct totals. The cumulative infection count was within 0.6% of the original total.

Our network topologies contain between 2000 and 15000 nodes to allow for reasonable computation times. The true total infection counts are thus too low to resolve the infection dynamics on the resolution our of graph. We thus multiply the infection counts by a factor C to improve resolution. For Los Angeles and Miami-Dade counties we use $C = 10$ and for Middlesex county, Massachusetts $C = 20$.

F ADDITIONAL RESULTS

F.1 DAILY COUNT SEIR CURVE

We characterize the dynamics of our learned models by tracking the daily size of each disease compartment ($jS^tj; jE^tj; jI^tj; jR^tj$) over time. This representation allows us to more easily look for phenomena such as multiple peaks.

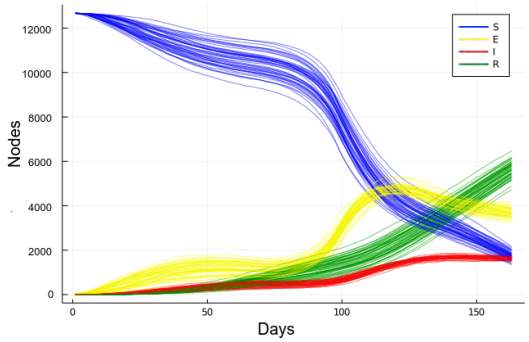

Figure 4: Los Angeles daily infection states (12,703 nodes) from 50 simulations using the mean inferred disease parameters. Here we learn a single initial exposure fraction for the whole network. Randomness occurs due to our stochastic blackbox simulator model. Note that multiple peaks are clearly visible.

F.2 VALIDATION ON SIMULATED DATA

We present additional results from our validation experiments, where we generate data using fixed parameters, and then condition our inference model on these results. We observe that our inference procedure converges accurately for a variety of geographic regions.

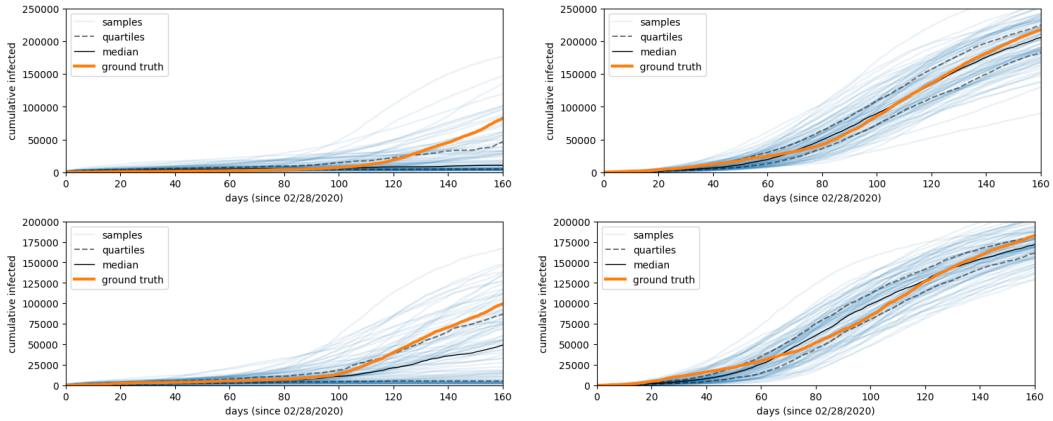

Figure 5: Validation on a simulated model over Los Angeles (top) and Miami-Dade (bottom). Left: Generated disease trajectory using "low" $_E = 0.1$. Right: Generated disease trajectory using "high" $_E = 0.45$.

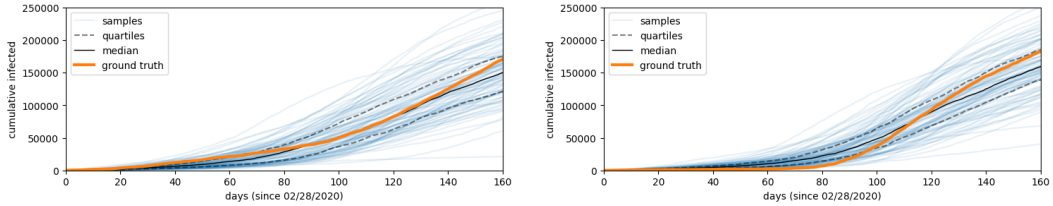

Figure 6: Validation on a simulated model over Los Angeles (top) and Miami-dade (bottom) topology. Left: Generated disease trajectory using "high-low-high" $_E = 0.45; 0.1; 0.45$. Right: "low-high-low" $_E = 0.1; 0.45; 0.1$