Probabilistic Program Inference in Network-based Epidemiological Simulations

Niklas Smedemark-Margulies*[1], Robin Walters*[1], Heiko Zimmermann*[1], Lucas Laird[2], Christian van der Loo[2], Neela Kaushik[2], Rajmonda Caceres[2], Jan-Willem van de Meent[1]

[1] Northeastern University, Boston, MA [2] MIT Lincoln Laboratory, Lexington, MA *Equal Contribution

{smedemark-margulie.n,r.walters,zimmermann.h,j.vandemeent}@northeastern.edu, {christian.vanderloo, lucas.laird,Neela.Kaushik,Rajmonda.Caceres}@ll.mit.edu, j.vandemeent@northeastern.edu

Goal

We seek to produce accurate simulations of disease transmission dynamics using regional mobility data. We use probabilistic programming to fit the parameters of these simulations to real world epidemiological data. These simulations can help policy makers and epidemiologists design intervention strategies for pandemics.

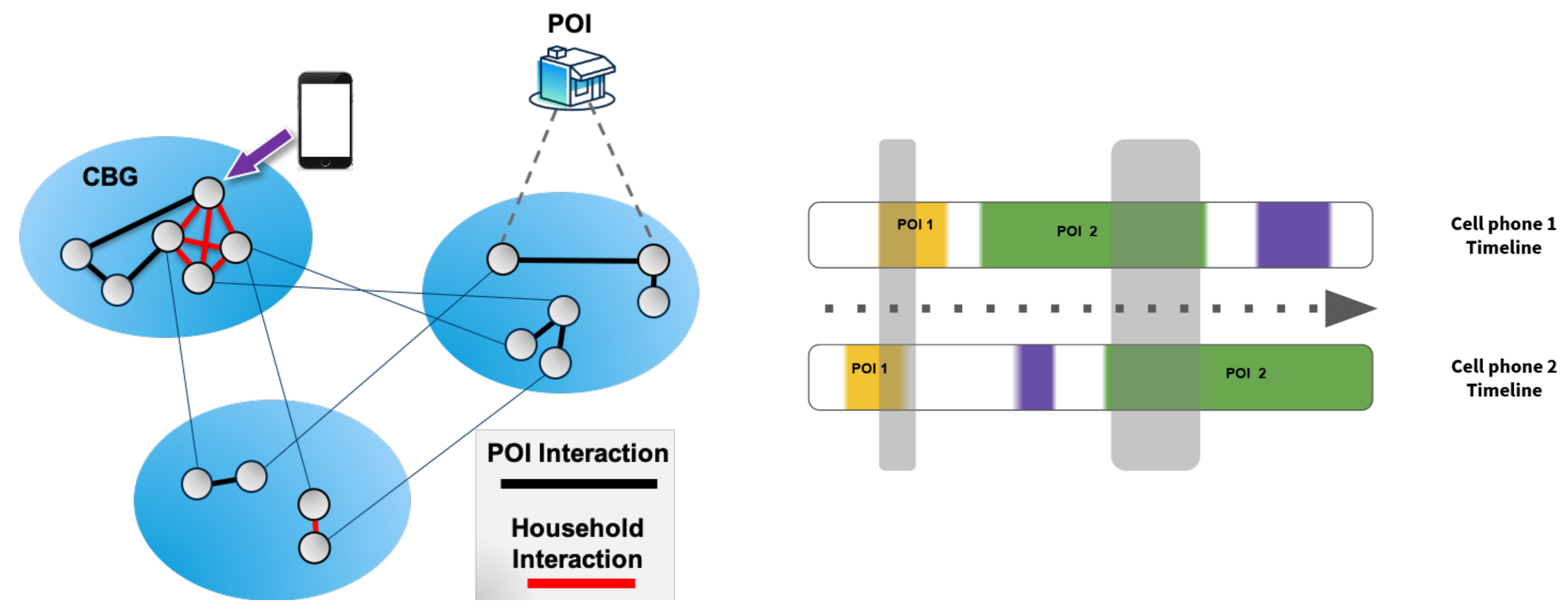
Our Results

We construct a region-specific disease simulation model that is capable of reproducing local transmission dynamics more accurately than baseline models and closely matches historical infection data from that region. To achieve this, we:

- Build mobility networks to represent individual counties by fitting a degree-corrected stochastic block model to cellphone GPS data from that region.
- Implement an agent-based model, Network-SEIR, with SEIR-like disease state dynamics to simulate the spread of disease in these mobility networks.
- Use probabilistic programming and condition our model on regional infection counts in order to calibrate the parameters of our disease model for a given county.

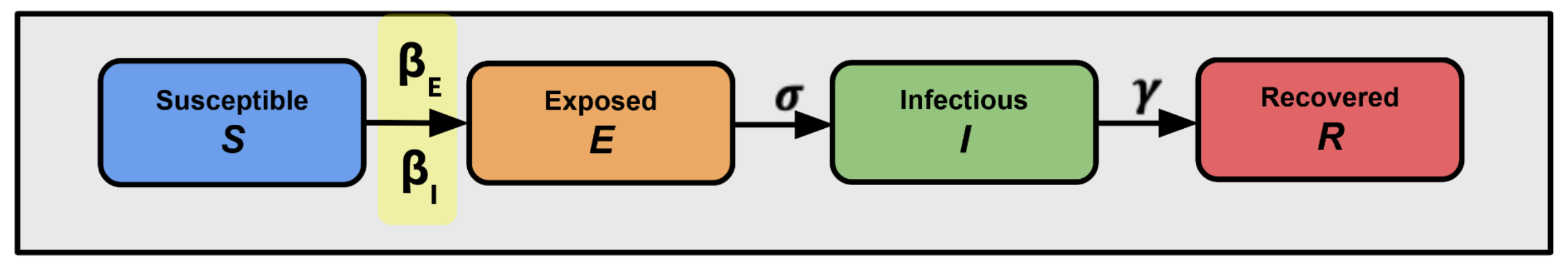
Constructing Mobility Networks

We construct regional mobility graphs using anonymized cell-phone GPS data obtained from SafeGraph. This data assigns phones to a home census block group (CBG), and tracks their visits to points of interest (POIs). We fit a degree-corrected stochastic block model (DCSBM) to the original county data. The probability of forming an edge between two CBGs is proportional to the number of shared visits to POIs. We set the edge weight between two CBGs based on the length of overlap during shared visits to POIs.



Compartmental and Network Disease Simulation

Traditional SEIR models use aggregated disease compartments and simple ordinary differential equations describing the dynamics between compartments.



We design a stochastic network-SEIR model to capture the effect of social and mobility networks on transmission rates, as well as variation due to regional policies, public health differences, and changes in disease properties.

Our model is parametrized by: (1) a regional graph $\mathcal{G}$, (2) initial exposure rate α for each community $1 \ldots C$, (3) K control points for time-varying disease parameters (β_E controls how *exposed* individuals transmit; β_I controls how *infected* individuals transmit), (4) dwell probabilities γ for the exposed state and λ for the infectious state, and (5) simulation duration T.

Each simulation produces an estimate of cumulative counts of infected individuals at each day.

This material is also supported by the United States Air Force under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force.

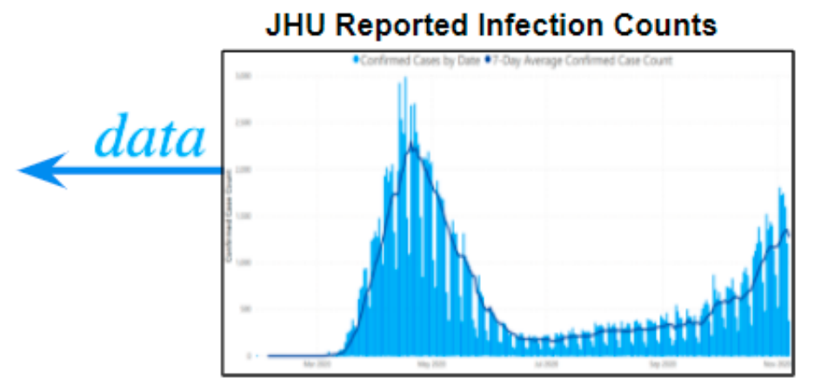
Stochastic Disease Simulator

```
Function $f_{\text{SEIR}}\,(\mathcal{G} = (\mathcal{V}, W),\ \rho,\ \beta^E,\ \beta^I,\ \gamma,\ \lambda,\ \tau,\ T)$:
  for $c \leftarrow 1$ to $C$ do                                                     // Initial exposure
    for $v \in \mathcal{V}_c$ do  if $\textsc{Uniform}(0,1) < \rho_c$ then $v \rightarrow E_1$ else $v \rightarrow S_1$
  for $t \leftarrow 1, \ldots, T-1$ do                                                // Simulate T days
    $\beta^E_t \leftarrow \text{INTERPOLATE}\big((\tau_1, \beta^E_1), \ldots, (\tau_N, \beta^E_N)\big)$ ; $\beta^I_t \leftarrow \text{INTERPOLATE}\big((\tau_1, \beta^I_1), \ldots, (\tau_N, \beta^I_N)\big)$
    for $v \in S_t$ do                                                                // New exposures
      $E^{\text{pressure}} \leftarrow \sum_{u \in N^E_t(v)} W_{uv}\beta^E_t$ ; $I^{\text{pressure}} \leftarrow \sum_{u \in N^I_t(v)} W_{uv}\beta^I_t$
      if $\textsc{Uniform}(0,1) < 1 - \exp(-E^{\text{pressure}} - I^{\text{pressure}})$ then $v \rightarrow E_{t+1}$
    for $v \in E_t$ do if $\textsc{Uniform}(0,1) < \gamma$ then $v \rightarrow I_{t+1}$              // Symptoms begin
    for $v \in I_t$ do if $\textsc{Uniform}(0,1) < \lambda$ then $v \rightarrow R_{t+1}$              // Infection ends
  return $\left\{ \sum_{t=1}^{j} |I_t| \right\}_{j=1}^{T}$                                  // List of Cumulative Infections
```

Stochastic Variational Inference

We seek to learn a posterior distribution over disease transmission parameters for our disease simulator given observed infection data $p(\beta|\text{data})$. Computing this posterior directly is intractable because it would require a high-dimensional integral with respect to all latent variables of the model.

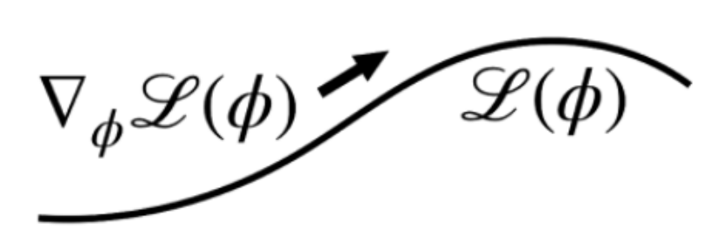
$$p(\beta \mid \mathit{data}) = \frac{p(\mathit{data} \mid \beta)\, p(\beta)}{\underbrace{\int p(\mathit{data} \mid \beta)\, p(\beta)\, d\beta}_{\textbf{intractable}}} \xleftarrow{\mathit{data}}$$



JHU Reported Infection Counts

Instead, we define a variational distribution $q_\phi(\alpha, \beta_E, \beta_I, \gamma, \lambda)$ which approximates this posterior, and try to minimize the KL divergence between q and p. We achieve this by optimizing a surrogate ELBO objective as a function of the parameters of q.

$$\begin{aligned}\phi^* &= \arg\min_{\phi}\ \mathrm{KL}(q_\phi(\,\cdot\,)\,||\,\underbrace{p(\,\cdot \mid \mathit{data})}_{\text{intractable}}) \\ &= \arg\max_{\phi}\ \mathscr{L}(\phi) \leftarrow \text{tractable surrogate objective (ELBO)}\end{aligned}$$



$$\phi_{t+1} \leftarrow \phi_t + \alpha_t \nabla_\phi \mathscr{L}(\phi_t)$$

By performing stochastic gradient ascent on the ELBO objective, we obtain locally optimal parameters ϕ for our variational distribution. Specifically we use Black Box Variational Inference as implemented in the Gen probabilistic programming package [4, 3, 1].

Validation on Simulated Data

Mean Daily Absolute Error (MDAE). We compare model outputs to true data by summing the area between curves, normalizing for time range and population size.

$$\text{MDAE} \equiv \mathbb{E}_{q_\phi(z)}\left[\frac{\| f_{\text{SEIR}}(z) - x\|_1}{TN}\right] \approx \frac{1}{N}\frac{1}{ST}\sum_{s}\sum_{t}\left|f_{\text{SEIR}}(z_s)^t - x^t\right| \tag{1}$$

Our Method Recovers Parameters for Simulated Data. We perform inference using simulated infection counts with 6 different time-varying patterns for β^t_E such as low-high-low. Synthetic data comes from our disease simulator using fixed disease parameters. Then, we run our inference procedure and compare disease trajectories from our learned variational distribution to the trajectories using the ground truth disease hyperparameters.

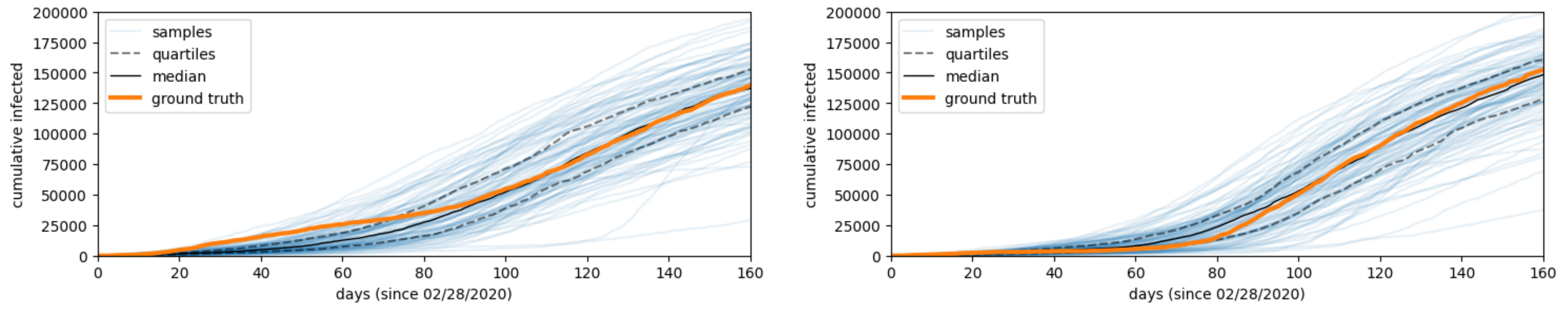


Figure: Validation on a simulated model on Miami-Dade topology. Generated disease trajectory using "high-low-high" $\beta_E = 0.45, 0.1, 0.45$ (left) and "low-high-low" $\beta_E = 0.1, 0.45, 0.1$ (right).

Table: 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

Fitting Parameters in Different Regions

We apply our method to Los Angeles, CA and Miami-Dade, FL. We fit parameters for Network-SEIR by conditioning on the reported cumulative infection counts. Our method fits observed data better (in MDAE) than several alternative baselines: (1) Compartmental CE-EM: a compartmental SEIR model with parameters fit using CE-EM [2], and (2) Network R_t-Analytic: a simplified analytic solution for f_{SEIR} parameters, (3) Metropolis-Hastings: a standard inference strategy based on MCMC, and (4) Likelihood Weighting in which samples the prior are re-weighted according to their likelihood.

Disease Model	Fitting Method	Los Angeles	Miami-Dade	Middlesex
Compartmental SEIR	CE-EM	0.0127	0.0217	0.0080
Network SEIR	R_t-analytic	0.0103	0.0367	0.0021
Network SEIR	Metropolis Hastings	0.0124	0.0134	0.0076
Network SEIR	Likelihood Weighting	0.0066	0.0090	0.0056
Network SEIR	BBVI	**0.0011**	**0.0036**	**0.0012**

Table: Comparison of the MDAE of different disease models and fitting methods.

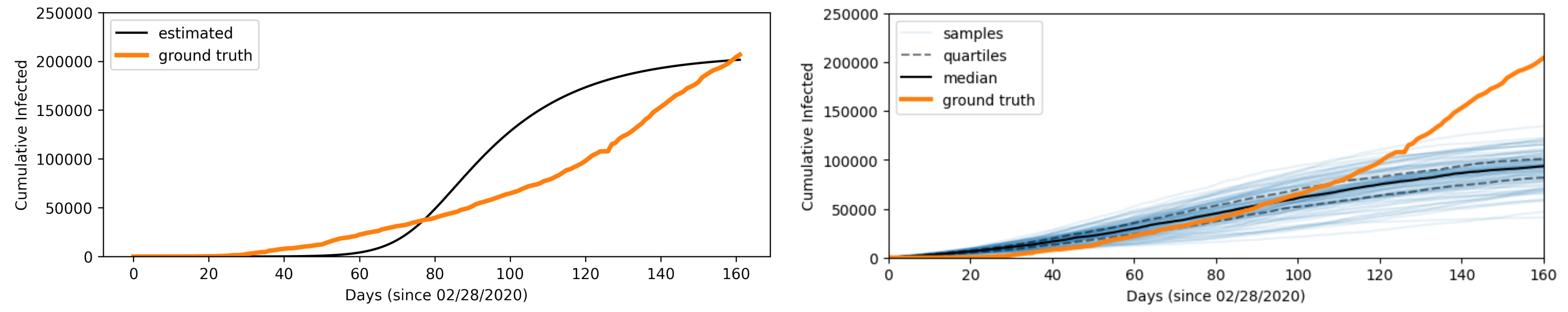


Figure: Inference for Los Angeles using baseline methods. Model vs. true cumulative infections shown for (left) Compartmental CE-EM (right) Network R_t-Analytic.

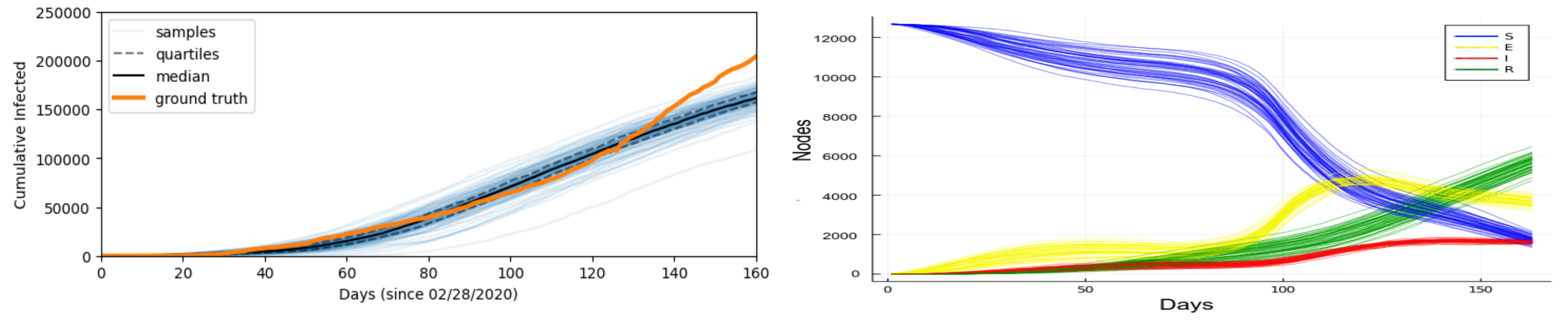


Figure: Inference for Los Angeles: (left) model vs. true cumulative infections; (right) daily SEIR counts output from the fit model. Note that multiple peaks are clearly visible.

Inferring Starting Communities

Our variational distribution includes a mean proportion of initial exposure in each community c. Learned values for these means μ^c_α indicate which communities were likely to have higher initial exposure given the observed disease data. For higher observational noise ν, the inferred parameters are closer to the uniform prior $\mu^c_\alpha = .05$, 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.

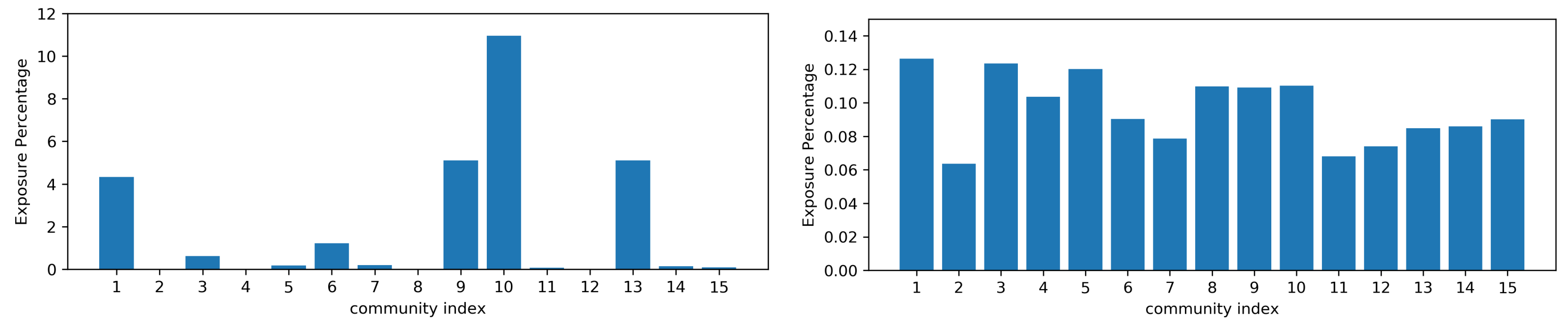


Figure: The network topology of Miami-Dade county is modeled using 15 communities which correspond to actual geographic areas. We plot μ^c_α for $1 \le c \le 15$. We use observational noise $\nu = 2.5 \cdot 10^{-4}$ (left) and $\nu = 5 \cdot 10^{-4}$ (right).

References

[1] M. F. Cusumano-Towner, F. A. Saad, A. K. Lew, and V. 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, pages 221–236, New York, NY, USA, 2019. ACM.

[2] K. R. Menda, L. Laird, M. J. Kochenderfer, and R. S. Caceres. Explaining covid-19 outbreaks with reactive seird models. *medRxiv*, 2021.

[3] R. Ranganath, S. Gerrish, and D. Blei. Black box variational inference. In *Artificial intelligence and statistics*, pages 814–822. PMLR, 2014.

[4] D. Wingate and T. Weber. Automated variational inference in probabilistic programming. *arXiv preprint arXiv:1301.1299*, 2013.

Acknowledgments Walters is supported by a Postdoctoral Fellowship from the Roux Institute and NSF grants 2107256 and 2134178. This work was also supported by the Intel Corporation, the 3M Corporation, NSF award 1835309, startup funds from Northeastern University, the Air Force Research Laboratory (AFRL), and DARPA. Computation was performed using the Northeastern Discovery and MIT Supercloud clusters.