Biostatistics, 2025, **26**(1), kxae018
https://doi.org/10.1093/biostatistics/kxae018
Article

Stochastic EM algorithm for partially observed stochastic epidemics with individual heterogeneity

Fan Bu[1], Allison E. Aiello[2], Alexander Volfovsky [3,*,†], Jason Xu [3,†]

[1]Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109, USA
[2]Mailman School of Public Health, Columbia University, 722 West 168th Street, New York, NY 10032, USA
[3]Department of Statistical Science, Duke University, 214 Old Chemistry, Box 90251, Durham, NC 27708, USA

[*]Corresponding author: Department of Statistical Science, Duke University, 214 Old Chemistry, Box 90251, Durham, NC 27708, USA. Email: alexander.volfovsky@duke.edu
[†]Joint last authors.

SUMMARY

We develop a stochastic epidemic model progressing over dynamic networks, where infection rates are heterogeneous and may vary with individual-level covariates. The joint dynamics are modeled as a continuous-time Markov chain such that disease transmission is constrained by the contact network structure, and network evolution is in turn influenced by individual disease statuses. To accommodate partial epidemic observations commonly seen in real-world data, we propose a stochastic EM algorithm for inference, introducing key innovations that include efficient conditional samplers for imputing missing infection and recovery times which respect the dynamic contact network. Experiments on both synthetic and real datasets demonstrate that our inference method can accurately and efficiently recover model parameters and provide valuable insight at the presence of unobserved disease episodes in epidemic data.

KEYWORDS: contact tracing; data-augmented inference; SEIR models; stochastic EM; stochastic epidemic models.

1. INTRODUCTION

Modern epidemiological studies seek to understand disease dynamics, evaluate intervention strategies and differentiate between population level and individual level effects. A traditional approach to modeling infectious diseases relies on mechanistic compartmental models, where only the summary of disease statuses of individuals in the population plays a role in understanding the disease dynamics. Examples of such mechanistic compartmental models abound in the epidemiology and mathematical biology literature, e.g. the susceptible-infectious-recovered (SIR) model (Kermack and McKendrick 1927). The majority of these summarize disease transmission as a population-level process, and are poised to answer questions related to aggregate dynamics (e.g. "what is the effective reproduction number?," "will the outbreak end?"). However, they cannot address finer grained heterogeneity given individual characteristics and contact behavior (e.g. "what is *my* risk of infection?," "does social distancing help *me*?"). This is exemplified by the well-mixed assumption that underpins many of these models, which posits that any infectious individual can transmit to any susceptible individual, and that all individuals within the same compartment are interchangeable. However, it is clear that the contact network plays an integral role in disease transmission, and that interventions on individual behavior can change the overall dynamics

Received: August 21, 2023. **Revised:** May 22, 2024. **Accepted:** May 24, 2024

of an outbreak (Eames and Keeling 2003; Kiss et al. 2006; Lunz et al. 2021). These social and behavioral aspects are highlighted in recent works that emerged during the SARS-CoV-2 pandemic (e.g. Ferguson et al. 2020; Soriano-Arandes et al. 2021; Ball and Britton 2022; see Supplementary Section 1 for a detailed discussion on related works).

In this paper, we develop a mechanistic stochastic model and a likelihood-based inferential framework that can account for individual heterogeneity in disease transmission. We specifically consider two main aspects of individual heterogeneity: people have differential contact patterns—characterized by a dynamic contact network that can adapt to disease transmission— and baseline characteristics (such as hygiene and immunization) which associate with differential risks of infection. We propose a continuous-time Markov chain (CTMC) model that jointly captures the dynamics of epidemic and contact network processes, building on an individualized stochastic SEIR model, where the Exposed (E) compartment accounts for incubation periods. These contributions are motivated by emerging infectious disease studies that collect high-resolution contact tracing data and surveys on individual health and social behavior (Aiello et al. 2016).

A key challenge arises from partial observations of the process. Even with high resolution contact tracing, individual-level infection times and recovery times are often unavailable, due to incubation periods and lack of follow-up. We address this by developing a data-augmented inference algorithm under the stochastic expectation–maximization (sEM) framework (Nielsen 2000) that accommodates *and* leverages contact network dynamics. The data augmentation steps in our iterative algorithm efficiently sample unobserved infection and recovery times via carefully designed conditional samplers amenable to parallel implementation (Section 3.2). Our approach is much more computationally tractable than exact inference using the marginal likelihood of observed data, which is numerically challenging and delicate even without the complexity of contact networks (Ho et al. 2018*b*,*a*; Ju et al. 2021). Given the high-dimensional latent space containing all individual-level, time-varying disease statuses in our context, an sEM approach is also more efficient and numerically stable than simulation-based methods relying on matching summary statistics with observed trajectories (He et al. 2010; Andrieu et al. 2010; Pooley et al. 2015). Compared to recent data-augmented Markov chain Monte Carlo (MCMC) approaches (Bu et al. 2020; Rose et al. 2020; Fintzi et al. 2022; Morsomme and Xu 2022; Wang and Walker 2023), an sEM framework enables substantially faster convergence and reduced computing time, while established results admit uncertainty quantification through conservative variance estimation (Nielsen 2000).

2. MODEL FRAMEWORK

We adopt a stochastic compartmental model for epidemics, where all members of the target population are divided into non-overlapping subsets related to their disease statuses, and the mechanism of disease spread is described by the transition between disease statuses for each individual. We base our epidemic model on the SEIR model with four disease statuses: S (susceptible), E (exposed), I (infectious), and R (recovered or removed). An S individual may get exposed (and thus become an E person) upon contact with an I individual, and an infectious (I) person will eventually recover and transition to the R status. In this model, the E status resembles an incubation period that does not entail transmissibility, and a recovered person acquires immunity to the disease and therefore no longer contributes to more infections.

These disease spread dynamics evolve as a continuous-time Markov chain (CTMC) defined through exponentially distributed waiting times between consecutive events (Guttorp 2018). This implies that the disease process progresses as a series of competing Poisson processes at the individual level. For example, suppose β_{ij} is the rate of exposure between an I person i and an S person j who are in contact at time t. Then the probability of j getting exposed (thus becoming an E person) at time $t+h$ for $h>0$ is

$$Pr(j \text{ gets exposed by } i \text{ by } t+h \mid i, j \text{ in contact at } t) = \beta_{ij}h + o(h). \tag{2.1}$$

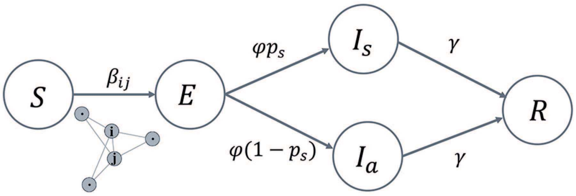

Figure 1. Diagram of the epidemic process: an extension of the stochastic SEIR model, with heterogeneous exposure rates and two sub-types of infectives. Disease transmission (exposure) is conditioned on pairwise contact status in the dynamic contact network.

Since the contact structure of the population also changes in time, we extend the CTMC model to the dynamic contact network setting. For any pair of individuals i and j, at time $t > 0$, they either share an undirected contact link ("connected") or they do not ("disconnected"). The contact network at time point t can thus be represented by a binary symmetric matrix W_t called the adjacency matrix. Its dynamics are described at the pairwise level, where each entry $W_{ij,t}$ evolves as a CTMC that takes values in $\{0, 1\}$. For example, if individuals i and j are disconnected at time t, the probability of them engaging in contact by time $t + h$ $(h > 0)$ is

$$Pr(W_{ij,t+h} = 1 \text{ at time } t + h \mid W_{ij,t} = 0 \text{ at time } t) = \alpha_{ij} h + o(h), \tag{2.2}$$

where the link activation rate α_{ij} can depend on the disease statuses of i and j (details later).

We further accommodate different types of individual heterogeneity in the disease transmission dynamics: (i) people may exhibit different levels of susceptibility that can be explained by individual characteristics such as health conditions, hygiene habits, and behavioral choices; (ii) those who are infectious might not be equally contagious for the susceptible population; (iii) contact rates in the network may vary in time to reflect phases of social intervention and/or behavioral changes as response to an epidemic; (iv) contact rates in the network may vary between pairs of individuals based on their healthy (denoted by H, the collection of S, E, and R statuses) versus infectious (denoted by I) statuses.

By the superposition property of Poisson processes, the population-level process combines all the competing individual-level processes. In particular, combining all pairwise infection processes naturally accounts for competing infectors for each susceptible person (Kenah et al. 2008).

2.1. Model specification

At any time point t in the process, we represent the disease statuses of all individuals by $\mathcal{X}_t$, where each entry $\mathcal{X}_{i,t} \in \{S, E, I, R\}$ denotes the status of person i. The states of the networked epidemic process at time t can be summarized by $\mathcal{Z}_t = \{\mathcal{X}_t, W_t\}$. Conditioned on $\mathcal{Z}_t$, at the individual level, the next possible event after time t falls into five categories: (i) *exposure* – an S person gets exposed and becomes E); (ii) *manifestation* – an E person becomes I; (iii) *recovery* – an I person recovers and becomes an R person; (iv) *link activation* – an unconnected pair get connected in the network; and (v) *link termination* – a connected pair break off their contact. For simple exposition, we assume a closed population with no immigration or emigration, but the framework can be naturally extended to account for external infections (see Section 4). For any members i and j in the population, we specify the instantaneous rates for the five possible events as follows (see Figure 1 for a model diagram):

(i) Exposure ($S \to E$). For i infectious and j susceptible who are in contact at time t, i exposes j with instantaneous rate β_{ijt} that can be decomposed as

$$\log \beta_{ijt} = \log \beta + \eta_i(t) + b_S^T x_j. \tag{2.3}$$

Here β is the baseline exposure rate, $\eta_i(t)$ represents i's infectiousness level at time t (details in the next item), and b_S are coefficients on j's individual characteristics x_j that account for additional heterogeneity in susceptibility. For example, x_j could encode if j is vaccinated which may reduce exposure risk, or could represent unobserved covariates for possible extensions to a mixed effects model.

(ii) Manifestation $(E \to I)$. For an exposed person, they become infectious with rate φ to be either I_s ("symptomatic") or I_a ("asymptomatic") with probability p_s and $(1 - p_s)$, respectively. With the two-type infective setup, the $\eta_i(t)$ term in (2.3) can be written as

$$\eta_i(t) = \eta \mathbb{I}(i \text{ is } I_s \text{ at } t), \tag{2.4}$$

which means an I_s person is on average e^{η} times more infectious than I_a. Given this framework, it is straightfoward to introduce more sub-types of infectives or include continuous and even time-varying explanatory variables for the function $\eta_i(t)$, if more intricate modeling of heterogeneous transmissibility is necessary.

(iii) Recovery $(I \to R)$. An infectious person i recovers with rate γ, where γ could differ across individuals with similar parametrization to that in (2.3); here for model parsimony we assume this rate is shared across the population.

(iv) Link activation. For unconnected pair i and j, they activate their contact link with rate α_{ijt}, where $\alpha_{ijt} = \alpha_{A_{it}A_{jt}0}\mathbb{I}(t \in \mathcal{T}_0) + \alpha_{A_{it}A_{jt}1}\mathbb{I}(t \in \mathcal{T}_1)$. Here A_{it} is the healthy (H, meaning S, I or R) or infectious (I) status of person i at time t and α_{ABk} stands for the activation rate of link type $A \sim B$ in phase $\mathcal{T}_k$ ($A, B \in \{H, I\}$ and $k \in \{0, 1\}$). For brevity we assume there are two social phases $\mathcal{T}_0$ and $\mathcal{T}_1$ in time, which could represent intermittent lockdown and no-lockdown phases with different baseline contact rates.

(v) Link termination. For connected pair i and j, they break off their contact link with rate ω_{ijt}, where $\omega_{ijt} = \omega_{A_{it}A_{jt}0}\mathbb{I}(t \in \mathcal{T}_0) + \omega_{A_{it}A_{jt}1}\mathbb{I}(t \in \mathcal{T}_1)$, and ω_{ABk} stands for the termination rate of link type $A \sim B$ in phase $\mathcal{T}_k$.

We assume that link rates α_{ijt} and ω_{ijt} are dependent on individual disease statuses H or I since we wish to characterize the social adaptation behavior in response to epidemics – for instance, a healthy-infectious $(H \sim I)$ pair that are in contact might be more likely to disconnect from each other than a pair of healthy individuals to avoid disease transmission (Shukla et al. 2022; Rieger et al. 2022; Bor et al. 2023). Further, since we assume an undirected contact network, the link rates satisfy $\alpha_{HI_k} = \alpha_{IH_k}$ and $\omega_{HI_k} = \omega_{IH_k}$ for $k = 0, 1$ between time phases $\mathcal{T}_0$ and $\mathcal{T}_1$. One may also choose to include individual-level covariates in the link activation and termination rates (α_{ijt} and ω_{ijt}) to reflect more heterogeneity in the network dynamics (Supplementary Section 2.2); for exposition purposes, here we highlight the regression formulation in (2.3) for the infection process.

3. INFERENCE WITH PARTIAL OBSERVATIONS: THE DANCE FRAMEWORK

We take a data augmentation approach for inference in the partial observations setting. As such, we begin by describing key likelihood-based inferential terms related to the complete data setting, and then develop a sEM algorithm for partial observations, **D**ata-**A**ugmented **N**etwork **C**ontagion **EM** (DANCE).

3.1. Inference with complete data

A complete dataset refers to the fully observed event sequence between time 0 and maximum time T (> 0) of one realization from the generative model. That is, a hypothetical continuous observer of the networked epidemic would have access to (1) exact times of exposure ($t_i^{(E)}$), manifestation ($t_i^{(I)}$), recovery, and link activation and termination; (2) individual identities in each event;

Table 1. Explanation of notation.

Notation	Explanation
N	Total population size (assumed fixed)
$n_{I_s}, n_{I_a}, n_E, n_I, n_R$	Total number of I_s, I_a, exposed (E), infectious (I) and recovered (R) cases
$I_i^a(t), I_i^s(t)$	Total number of I_a and I_s neighbors of i at time t
$I^a(t), I^s(t), E(t)$	Total number of status I_a, I_s, and E individuals in the population at time t
$t_i^{(E)}, t_i^{(I)}$	Exposure time and manifestation time for individual i (set to T if never exposed/manifested)
C_{ABk}, D_{ABk}	Total number of link activation & termination events among type $A \sim B$ pairs in phase $\mathcal{T}_k$
$M_{AB}^c(t), M_{AB}^d(t)$	Number of connected & disconnected type $A - B$ pairs at time t

(3) I_s or I_a subtype for each individual who becomes infectious; and (4) the initial contact network structure and all initial disease statuses at time 0.

Given the complete data (or equivalently, sufficient statistics summarizing the data) and all individual characteristics $\{x_i\}$, we can write down the complete data likelihood with respect to the model parameters $\Theta = \{\beta, \varphi, \gamma, \eta, b_S, \boldsymbol{\alpha}, \boldsymbol{\omega}\}$. Here $\boldsymbol{\alpha} = (\alpha_{ABk})_{k\in\{0,1\},(A,B)\in\mathcal{S}}$, $\boldsymbol{\omega} = (\omega_{ABk})_{k\in\{0,1\},(A,B)\in\mathcal{S}}$ denote the network link rates for different link types between social phases, indexed by $\mathcal{S} = \{(H,H),(H,I),(I,I)\}$ the set of all pair types. The likelihood takes the form (see Table 1 for a summary of all notation):

$$
\begin{aligned}
&L(\Theta; \text{complete data}) \\
&\quad = \beta^{n_E}\gamma^{n_R}\varphi^{n_I}p_s^{n_{I_s}}(1-p_s)^{n_{I_a}} \prod_{i: i \text{ got exposed}} e^{b_S^T x_i}\left[I_i^a(t_i^{(E)}) + I_i^s(t_i^{(E)})e^{\eta}\right] \\
&\qquad \times \prod_{k=0,1}\prod_{(A,B)\in\mathcal{S}} \left[(\alpha_{ABk})^{C_{ABk}}(\omega_{ABk})^{D_{ABk}}\right] \\
&\qquad \times \exp\left(-\int_0^T \left[\beta\sum_{i=1}^N e^{b_S^T x_i}\left[I_i^a(t) + I_i^s(t)e^{\eta}\right]\mathbb{I}(i \text{ is susceptible at } t)\right.\right. \\
&\qquad\qquad\qquad \left.\left. + \gamma(I^a(t) + I^s(t)) + \varphi E(t)\right]dt\right) \\
&\qquad \times \exp\left(-\int_0^T \sum_{k=0,1}\sum_{(A,B)\in\mathcal{S}}\left[\alpha_{ABk}M_{AB}^d(t) + \omega_{ABk}M_{AB}^c(t)\right]\mathbb{I}(t\in\mathcal{T}_k)dt\right).
\end{aligned} \tag{3.5}
$$

Since the generative model is a CTMC comprised of individual-level Poisson processes, the above likelihood can be decomposed into epidemic-related components (1st and 3rd lines above) and network-related components (2nd and 4th lines). Evaluation of this seemingly lengthy likelihood function involves either bookkeeping of population-level quantities (such as $n_E =$ total number of exposed cases), or parallelizable computation of individual-level quantities (such as $I_i^s(t) =$ number of I neighbors for i at time t).

When complete data are available, we can obtain closed-form maximum likelihood estimates (MLEs) for most of the parameters, and find the remaining MLEs for parameters β, η and b_S through simple numerical procedures, which can be implemented by fitting conditional Poisson regression models (See Supplementary Section 2 for full derivations). This suggests that likelihood-based inference given completely observed data is easily implementable and can be modularized toward inference in the missing data setting.

3.2. Inference with partial observations

We now discuss our inferential framework for partial observations, the **D**ata-**A**ugmented **N**etwork **C**ontagion **E**M (DANCE) algorithm. Since real-world epidemic data rarely include measurements of the full event sequence, our goal is to utilize the simplicity of complete data inference (described above) through data augmentation, based on the stochastic EM approach (Celeux 1985).

The EM algorithm offers an approach to efficiently carry out maximum likelihood estimation for continuous-time Markov chain models in missing data settings (Doss et al. 2013; Xu et al. 2015; Guttorp 2018). Imputing the missing data in the E-step requires access to the conditional expectation, and sEM is a variant that approximates the conditional expectation using augmented data obtained via conditional simulation. To be more precise, let X denote the observed data and Z be the missing data; a general outline of sEM for estimating parameter θ is as follows: For $s = 1 :$ maxIter, do

- (E-step) draw one sample of missing data, $Z^{(s)}$ from its conditional distribution $p(Z \mid X, \theta^{(s-1)})$, and then let
$$Q(\theta \mid \theta^{(s-1)}) = \log L(\theta; X, Z^{(s)});$$
- (M-step) maximize with respect to target function $Q(\theta \mid \theta^{(s)})$ to update θ:
$$\theta^{(s)} = \arg\max_{\theta} Q(\theta \mid \theta^{(s-1)}).$$

There are two advantages of this approach. First, in the E-step, integrating to obtain an expected log-likelihood (as in the traditional EM algorithm) is replaced by sampling, which avoids the often intractable marginalization step in the case of complex models (Renshaw 2015; Xu and Minin 2015; Stutz et al. 2022). Second, the M-step simply requires solving for the MLEs given a version of the complete data, which is often straightforward, as discussed previously for the present setting.

These advantages come at the cost of a potential challenge: we have to conditionally sample the missing data given our observed data and current parameter estimates. In our framework, this is equivalent to sampling event times of a continuous-time Markov chain conditioned on end-points, a notably difficult problem (Hobolth and Stone 2009; Rao and Teg 2013).

In the case of our motivating study, eX-FLU, true exposure times are not available even though the data contain daily symptom reports, due to the incubation period. Exact recovery times are not available either, with recoveries discernible only at a weekly resolution from epidemic surveys. Therefore, we need to consider inference with partially observed epidemic data, in particular with exposure times and recovery times unknown. Data augmentation under sEM thus involves conditionally simulating these missing event times, while preserving consistency between epidemic events and the dynamic contact network.

Let $\mathbf{t}^{(\mathbf{E})}$ and $\mathbf{t}^{(\mathbf{R})}$ denote all missing exposure times and recovery times, respectively. We assume that (i) that all manifestation times $\{t_i^{(I)}\}$ are observed, available through daily symptom monitoring or routine testing, and (ii) the contact network events are fully observed with high-resolution contact-tracing. Thus, our DANCE framework for partial observations is outlined as follows. For $s = 1 :$ maxIter, do

1. sample missing exposure times $\mathbf{t}^{(\mathbf{E})^{(s)}}$ from their joint conditional distribution $p(\mathbf{t}^{(\mathbf{E})} \mid$ observed events, $\mathbf{t}^{(\mathbf{R})^{(s-1)}}, \Theta^{(s-1)})$;
2. sample missing recovery times $\mathbf{t}^{(\mathbf{R})^{(s)}}$ from their joint conditional distribution $p(\mathbf{t}^{(\mathbf{R})} \mid$ observed events, $\mathbf{t}^{(\mathbf{E})^{(s)}}, \Theta^{(s-1)})$;
3. form an augmented dataset by combining sampled event times in Steps 1 and 2 with observed data, then solve for the complete data MLEs to obtain updated parameter estimates $\Theta^{(s)}$.

Since Step 3 is already addressed in the previous section, we derive conditional sampling algorithms for Steps 1 and 2. One essential consideration is that the conditional samplers must respect the dynamic contact network constraints while leveraging dynamic contact information.

Step 1: conditional sampling of missing exposure times. We derive a rejection sampler for all individual exposure times, conditional on parameter values and recovery times. In fact, it is sufficient to *separately* sample exposure time $t_i^{(E)}$ for each individual i who has ever become infectious. This is because each person i's exposure time is independent from other individuals' exposure times conditional on all other event times, as implied by the form of the complete data likelihood in (3.5).

We consider sampling the missing exposure time $t_i^{(E)}$ within a plausible interval, $L_i = (t^i_{\min}, t^i_{\max})$, possibly informed by prior knowledge or computational capacity. For example, we may set $t^i_{\min} = \max(0, t_i^{(I)} - 14)$ and $t^i_{\max} = \max(0, t_i^{(I)} - 2)$, if we believe the incubation period should be longer than 2 days but shorter than 2 weeks. Here, for generosity, we consider $L_i = (0, t_i^{(I)})$, meaning exposure could occur any time before the start of infectiousness.

The target we wish to sample from is the conditional density for i's exposure time, which can be written as

$$\begin{aligned} &p_i(t \mid t_i^{(I)}, \beta, \delta_i, \eta, \varphi, \text{network events}) \\ &\quad = \frac{\lambda_i(t)\exp\left(-\int_{t^i_{\min}}^{t} \lambda_i(u)du\right) \times \varphi\exp(-\varphi(t_i^{(I)} - t))\mathbb{I}(t^i_{\min} < t < t^i_{\max})}{C_i(\lambda_i(t), \varphi; t^i_{\min}, t^i_{\max})}. \end{aligned} \tag{3.6}$$

Here $\lambda_i(t)$ is i's time-varying total exposure risk, which is a step-constant function with change points fully determined when all other event times are known (see Supplementary Section 3 for full details). The normalizing constant $C_i(\lambda_i(t), \varphi; t^i_{\min}, t^i_{\max})$ can be explicitly evaluated since $\lambda_i(t)$ is a step function.

Consider the density

$$q_i(t) = \frac{\lambda_i(t)\exp\left(-\int_0^t \lambda_i(u)du\right)\mathbb{I}(0 < t < t_i^{(I)})}{1 - \exp\left(-\int_0^{t_i^{(I)}} \lambda_i(u)du\right)}, \tag{3.7}$$

which is the density function of a truncated inhomogeneous Exponential distribution with rate $\lambda_i(t)$. It is straightforward to show that $p_i(t)/q_i(t) \leq M$ for a constant $M > 1$ (Supplementary Section 3), which suggests we can use $q_i(t)$ as a proposal for a rejection sampling scheme for sampling from the conditional density $p_i(t)$.

Therefore, we have the following rejection sampler for $t_i^{(E)}$ that runs in two steps:

1. Sample t from $q_i(t)$, an inhomogeneous Exponential with step-constant rate $\lambda_i(t)$ truncated on L_i (sampling details included in Supplementary Section 3).
2. Compute the acceptance probability for t by (here $M > 1$ is a constant, see Supplementary Section 3)

$$\frac{p_i(t)}{Mq_i(t)} = \exp(-\varphi(t_i^{(I)} - t)), \tag{3.8}$$

   and draw $U \sim Unif(0, 1)$; accept t as a sample of $t_i^{(E)}$ if $U < \exp(-\varphi(t_i^{(I)} - t))$, and otherwise go back to Step 1 and repeat.

A full derivation of the above (importantly showing that $M > 1$) and other technical details are provided in Section 3 of the Supplementary Material. This step is also fully parallelizable across individuals, as the conditional sampling is performed separately for each person *i*. Through simulation experiments (see Supplementary Section 6), we see that the rejection sampler is very efficient, with an average acceptance rate of approximately 45%.

Step 2: conditional sampling of missing recovery times The conditional samples of missing recovery times should satisfy two conditions: first, an individual *i* cannot recover when they are still known to be infectious; second, *i* cannot recover either if they should serve as the infector of another exposure case. This amounts to conditionally sampling event times with endpoints restricted by low-resolution epidemic data and high-resolution contact data.

This challenge was previously addressed by the DARCI algorithm developed in Bu et al. (2020) (Proposition 4.2) for a simpler epidemic model with only one type of infectives. Here, we can adapt and modify DARCI for our two-type infective setting, conditional on the value of p_s (the proportions of I_s among all I individuals) and the sampled exposure times in Step 1. For brevity, we leave technical details to the Supplementary Material (Section 4).

Uncertainty quantification. For estimates produced by DANCE, we can quantify uncertainty by leveraging expressions for their asymptotic variances, using results established in Nielsen (2000). We further implement a multiple-chain strategy to reduce variance, by (i) averaging the last m iterations in one chain, or (ii) averaging m independent chains. As derived in Nielsen (2000), for example, averaging $m = 10$ independent chains of DANCE would provide a conservative variance estimate of $1.05(I(\hat{\Theta}))^{-1}$ where $\hat{\Theta}$ are the parameter estimates and $I(\cdot)$ denotes the Fisher information matrix. This allows us to produce conservative Wald-type confidence intervals. See full details in Supplementary Section 5.

Validation via simulations. We perform comprehensive simulation studies to validate the DANCE inference framework, by first validating the complete data inference procedure (Supplementary Section 6.1), and then testing the data augmentation component of DANCE (Step 1 and Step 2) by first taking out all simulated exposure times followed by removing all exposure and recovery times from simulated datasets (Supplementary Section 6.2). Across 40 independent simulations for each scenario, our inference algorithm is able to accurately recover the parameter values and produce confidence intervals with good coverage rates. We present a detailed description and all results of the simulation studies in Supplementary Section 6.

Remarks on computing time. As a stochastic EM algorithm, DANCE enjoys fast computation. With moderate efforts of parallelized implementation, on a regular 4-core laptop, each iteration typically takes a few seconds and the algorithm usually converges in about 100 iterations for a 100-200 person population (similar in size to our motivating dataset), amounting to total computing time only in the order of minutes.

4. CASE STUDY: FLU SEASON ON A UNIVERSITY CAMPUS

To illustrate our model and inference framework, we present a case study on transmissions of influenza-like illnesses among students on a university campus, where high-resolution contact tracing was performed to track physical proximity between study subjects and individual-level baseline characteristics were collected.

This dataset was collected over a 10-week epidemiological study, eX-FLU (Aiello et al. 2016), where inter-personal physical contacts of study participants were surveyed to investigate the effect of social intervention on respiratory infection transmissions. 590 university students enrolled in the study and were asked to respond to weekly surveys on influenza-like illness symptoms and social interactions; they also completed a comprehensive entry survey about demographic information, lifestyles, immunization history, health-related habits, and tendencies of behavioral changes during a flu season or a hypothetical pandemic. 103 individuals among the study population were further recruited to participate in a sub-study in which each study subject was provided a smartphone equipped with an application, iEpi. This application pairs smartphones with other nearby study devices via Bluetooth and thus can record individual-level contacts (i.e. physical proximity) at five-minute intervals. Bluetooth signals are pre-processed based on signal strengths to identify sufficiently intimate pairwise physical proximity which is treated as a contact link (Supplementary Section 7.2).

The iEpi sub-study took place from January 28, 2013 to April 15, 2013 (that is, from week 2 until after week 10 in the main study). Between weeks 6 and 7, there was a one-week spring break (March 1 to March 7), during which epidemic data collection was paused and volume of recorded contacts also dropped considerably. In our application case study, we use data obtained on the $N = 103$ sub-study population from January 28 to April 4 (week 2 to week 10), and treat the two periods before and after the spring break as two different social behavior phases. That is, we regard weeks 2-6 as $\mathcal{T}_0$ and weeks 7-10 as $\mathcal{T}_1$ in our analysis.

We consider two types of "infectious" (status I) members within the study population: (1) **multi-symptomatic (I_s)** – a case with a cough AND one of these three symptoms: fever or feverishness, chills, or body aches (definition of "influenza-like-illnesses"); (2) **uni-symptomatic (I_a)** – a case with a cough, a non-specific but important symptom for influenza.

For each infection case, we set the reported symptom onset time as the manifestation time (denoted by $t_i^{(I)}$ in previous sections), and treat the exposure time ($t_i^{(E)}$) and recovery time ($t_i^{(R)}$) as unobserved. Since $t_i^{(E)} < t_i^{(I)}$ (implied by the assumed SEIR mechanism), we set the plausible incubation interval as $L_i = (0, t_i^{(I)})$. Using weekly surveys (which asked each participant if they felt sick in the past week), we know that the missing recovery times must lie within a 7-day interval for each individual, where the lower and upper bounds are the start and end of a week. Moreover, we assume that all the contact network events are fully observed, as the high-resolution contact tracing can provide timepoints of activation and termination of all individual-level contacts. This suggests that our proposed DANCE algorithm is applicable to this dataset.

4.1. Inference with external infection sources

Since the 103 individuals in the dataset are sub-sampled from the 590 study participants, which are also sub-sampled from the entire university campus population, we have to treat the data as observed from an open population instead of a closed one. Therefore, some slight modifications should be made to the model. Specifically, individuals in our target population may get infected from outside infection sources, whom we refer to as "external infectors."

For simplicity, we represent the joint forces of all external infectors by a single infector that exists outside of the population and exhibits a constant level of transmissibility over time, and this external force of infection is exerted uniformly on all members of the target population.

For each susceptible individual j, let the rate of disease onset (i.e. manifestation) due to external infectors be ξ_j, and let this onset rate depend on individual characteristics x_j, similar to our treatment of the internal exposure rate β_{ij}: $\log \xi_j = \log \xi + x_j^T b_E$, where ξ denotes the population average external onset rate, and coefficients b_E represent coefficients to explain associations between individual characteristics x_j and subject j's deviations of susceptibility from the average level.

Here ξ_j is the rate of moving from status S directly to either I_a or I_s, rather than from S to E, and that is why we are naming it the "external onset rate" instead of "external exposure/infection rate." We are *not* introducing both an exposure rate (like β_{ij}) and a manifestation rate (like φ) for external infection cases because of identifiability concerns: since all susceptible people are exposed to the same external infector with time-invariant transmissibility, the exposure rate and manifestation rate would not be identifiable at the same time when the exposure times are not observed. Thus, to ensure identification, we choose to include only one rate instead of two, and the "onset rate" can be thought of as the rate of any susceptible individual developing contagiousness due to external infection forces.

Now the set of parameters is extended to $\tilde{\Theta} = \{\beta, \varphi, \gamma, \eta, b_S, \xi, b_E, \boldsymbol{\alpha}, \boldsymbol{\omega}\}$, and we can write down a complete data likelihood by slightly modifying Eq. (3.5), where the term related to the new parameters ξ and b_E are separate from the other terms (Supplementary Section 7.4). This means

Table 2. Estimates of key epidemic parameters, with conservative estimates of asymptotic standard errors.

Parameter	Estimate	Standard error
β (internal exposure)	4.497	2.005
ξ (external onset)	0.00445	0.00114
φ (latency)	0.221	0.0591
γ (recovery)	0.161	0.0279
e^{η} (I_s v.s. I_a infectiousness)	0.0622	0.0526
p_s (proportion of I_s)	0.382	0.0854

Table 3. Estimates of epidemic coefficients on individual characteristics, with conservative asymptotic standard deviations in the parentheses.

	(flushot)	(wash_opt)	(change_behavior)	(prevention)
b_S (internal exposure)	−0.105 (0.671)	−2.42 (0.817)	−0.201 (0.326)	−0.0541 (0.273)
b_E (external onset)	−0.805 (0.597)	−0.139 (0.471)	0.257 (0.263)	−0.0362 (0.273)

that introducing external cases does not affect estimation of the other parameters at all, and that we can still use the DANCE algorithm detailed in Section 3.2.

4.2. Data analysis

We first discuss how we identify internal and external infection cases and describe the individual characteristics used in the analysis. If an infected person had any infectious contact (within the 103-person population) up to 2 weeks prior to symptom onset, then we label this case as "internal," and otherwise this case is labeled as "external." This procedure gives us 18 internal cases and 16 external cases in total. Moreover, among all 34 cases, 13 are multi-symptomatic (I_s) and 21 are uni-symptomatic (I_a). We provide a summary of the breakdown of all infection cases in Supplementary Section 7.

We consider the following four individual-level characteristics collected from the entry survey that have previously been linked to disease transmission risk (the original survey questions used to calculate the derived covariates "change_behavior" and "prevention" are provided in Supplementary Section 7): (i) **flushot** – whether or not the study subject has taken a flu shot for this year; (ii) **wash_opt** – whether or not the study subject's hand-washing habit is considered "optimal," derived from survey questions about how long and how frequently one usually washes their hands; (iii) **change_behavior** – a derived numeric score measuring how willingly the study subject would change their lifestyle during a hypothetical pandemic, where a higher score represents more willingness in changing one's lifestyle in response to a pandemic; (iv) **prevention** – a derived score measuring one's belief in the effectiveness of different preventative practices in reducing the risk of catching the flu; a higher score represents stronger belief in the effectiveness of preventative practices.

We perform 20 independent runs of the stochastic-EM inference procedure on the dataset, each time with a different random initialization and 60 burn-in steps. For each run, we take the average of the last 20 iterations (after burn-in) and then average over the 20 averages (across runs) to produce estimates of the parameters. Convergence is assessed by examining traceplots and Geweke diagnostics and model fit is validated by simulation-based predictive checks (Supplementary Section 7.8). Conservative asymptotic standard errors are obtained using the method described in Section 3.2, setting $m = 20$ and upper-bounding the asymptotic variance matrix by $1.025 I(\hat{\hat{\Theta}})^{-1}$, where $\hat{\hat{\Theta}}$ are the final parameter estimates produced by averaging.

Tables 2 and 3 present estimates of key epidemic parameters. Here we take one day as 1 unit of time. For this population, the baseline exposure rate is quite high, indicating fast disease exposure

Table 4. Estimates of link activation and termination rates for different link types in the two phases ($\mathcal{T}_0$ spans from week 2 to week 6, and $\mathcal{T}_1$ from week 7 to week 10), with estimates of standard deviations in the parentheses.

Event type	Activation ($\boldsymbol{\alpha}$, $\times 10^{-4}$)		Deletion ($\boldsymbol{\omega}$, $\times 10^{0}$)	
Phase	$\mathcal{T}_0$	$\mathcal{T}_1$	$\mathcal{T}_0$	$\mathcal{T}_1$
$H \sim H$	181 (1.77)	8.68 (1.29)	11.6 (0.132)	5.27 (0.0783)
$H \sim I$	153 (6.67)	0.653 (0.0420)	16.6 (0.725)	8.71 (0.589)

upon contact – it takes approximately 0.22 days on average for an $H - I$ contact to lead to infection if the susceptible individual is not vaccinated, does not wash hands properly and has neutral attitudes about disease prevention. On average, the incubation period lasts slightly less than 5 days, while recovery takes about 6 days. The total external infection force experienced by the entire $N = 103$-person population is on the scale of $0.00445 \times 103 \approx 0.458$, indicating on average there would be a disease onset due to external sources every other day if nobody in the study population had a flu shot or washed their hands optimally. In terms of the coefficients for individual-level covariates, we note that the estimates are associated with relatively large standard errors (indicated in the parentheses), potentially due to the small sample size reflected by the moderate number of infection cases. Nevertheless, hand-washing ("wash_opt") seems to be a considerably influential mitigation measure, given that there is a 11-fold reduction ($1/e^{-2.42} \approx 11.2$) in the exposure risk if one washes their hands optimally compared to suboptimal hand-washing; such a statistical association appears significant, with a 95% Wald confidence interval of $(-4.054, -0.786)$ that does not contain zero.

In Table 4 we include estimates of key parameters related to the contact network process. Here we emphasize the difference between the change rates of $H \sim H$ (healthy-healthy) links and $H \sim I$ (healthy-ill) links, as well as the difference between the two social phases ($\mathcal{T}_0$ before spring break and $\mathcal{T}_1$ after). The link termination rates for $H \sim I$ links are higher than those of $H \sim H$ links in both phases, suggesting that the duration of contact between a healthy-infectious pair is on average shorter than the contact between two healthy people; this might be because infected students avoided social activities as they felt unwell, or susceptible individuals interacted less frequently with peers who seemed sick in order to avoid infection. Moreover, the level of network activity seems much higher (both in terms of establishing and breaking contact) in $\mathcal{T}_0$ (weeks 2 to 6, before spring break) compared to $\mathcal{T}_1$ (weeks 7 to 10, after spring break) when we compare the rates for phase $\mathcal{T}_0$ and phase $\mathcal{T}_1$, possibly due to increased outdoor activities (thus less contacts via close physical proximity) after the spring break. Such findings are enabled by our model design which allows for different levels of network activities by introducing different time phases.

Through our data analysis, we have found quantitative evidence that proper hand-washing is significantly associated with reduced risks of flu infection, and that there is a considerable external force of infection for the study population. Moreover, study participants exhibit adaptive contact behavior to flu transmission with less frequent and shorter-lasting contacts between healthy and infectious individuals. These findings are consistent with intuition and are also reflected in the dataset where optimal hand-washers seem less prone to infections and infectious individuals tend to lose contact links (Supplementary Section 7.9).

5. DISCUSSION

In this paper, we present a data-augmented stochastic EM inference algorithm for partially observed epidemics on a dynamic contact network while accounting for heterogeneous infection risks associated with individual characteristics. The design of a likelihood-based inferential framework is challenged by *and* benefits from the availability of high-resolution contact tracing data – the state space of latent variables is expanded to all unobserved individual epidemic event times, but at the same time largely reduced thanks to the knowledge of dynamic contact links.

It is important to note that the modeling framework we propose is flexible beyond our choice of underlying compartments. That is, our approach can be easily adapted to incorporate notions of reinfection (by allowing some individuals to reenter the susceptible population) or to distinguish between more than two types of infections. In pursuing generalizations of the methodology, introducing additional parameters requires careful consideration of the uncertainty quantification from the stochastic EM algorithm. Although in our setting, the estimated confidence intervals perform well empirically compared to their nominal coverage, they rely on variance approximation formulas, and it is crucial to conduct similar validations in more complex models.

There are, however, several limitations in our model assumptions that can motivate future research. First, for mathematical convenience we assume a Markovian model, but extensions could be made toward non-Markovian infection or recovery processes, using Gamma or Weibull distributions for inter-event wait times. Second, we assume all dynamic contact links are observed, but for a larger population where high resolution contact tracing is less feasible, one could use a social network model (stochastic block models or latent factor models) to account for unobserved contacts; our assumed binary contact links could be extended to categorical or continuous weighted links using the signal strengths of mobile device contact tracing. Lastly, we made a couple of pragmatic compromises in the real data case study: we identified external infections based on lack of internal contacts, but if more population-level data were to become available (e.g. contact surveys or viral sequencing data) we could estimate external infections in a joint statistical model; similarly, we used the asymptomatic or less symptomatic compartment (I_a) to identify non-specific symptomatic cases rather than truly asymptomatic cases, but one could introduce additional latent compartments to infer asymptomatic infections based on differential contact patterns or other data sources such as surveillance testing.

Our analysis of the iEpi data provides further evidence of the importance of personal hygiene and health habits on the reduction of the spread of influenza-like-illness. Through a careful analysis of real observational epidemiological data, we found a considerable association between hand-washing and the transmission rate of a disease in an active population with dynamically changing contact patterns. We hope that this development encourages greater data collection of high-frequency individual-level data in this area to gain better understanding of other pharmaceutical and non-pharmaceutical interventions. For example, future studies will be able to estimate the effectiveness of vaccination in preventing transmission under different social interaction rates and population densities, and assess claims about the efficacy of mask-wearing and active social distancing. Importantly, such data can be collected discretely in closed populations and provide invaluable insight into the deployment of public health interventions (Motta et al. 2021).

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Biostatistics Journal* online.

FUNDING

A.V. and J.X. were partially supported by DMS-2230074. A.V. was partially supported by DMS-2046880.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

Software in the form of `R` and `Python` code, together with a sample input data set and complete documentation is available on Github at https://github.com/fanbu1995/EpiNetHetero

REFERENCES

Aiello AE, Simanek AM, Eisenberg MC, Walsh AR, Davis B, Volz E, Cheng C, Rainey JJ, Uzicanin A, Gao H, et al. Design and methods of a social network isolation study for reducing respiratory infection transmission: the eX-FLU cluster randomized trial. Epidemics. 2016:15:38–55.

Andrieu C, Doucet A, Holenstein R. Particle Markov chain Monte Carlo methods. J R Stat Soc Ser B (Stat Methodol). 2010:72(3):269–342.

Ball F, Britton T. Epidemics on networks with preventive rewiring. Random Struct Algorithms. 2022:61(2):250–297.

Bor A, Jørgensen F, Petersen MB. Discriminatory attitudes against unvaccinated people during the pandemic. Nature. 2023:613(7945):704–711.

Bu F, Aiello AE, Xu J, Volfovsky A. Likelihood-based inference for partially observed epidemics on dynamic networks. J Am Stat Assoc 2022:117(537):510–526.

Celeux G. The sem algorithm: a probabilistic teacher algorithm derived from the em algorithm for the mixture problem. Comput Stat Q. 1985:2:73–82.

Doss CR, Suchard MA, Holmes I, Kato-Maeda M, Minin VN. Fitting birth-death processes to panel data with applications to bacterial dna fingerprinting. Ann Appl Stat. 2013:7(4):2315.

Eames KTD, Keeling MJ. Contact tracing and disease control. Proc R Soc Lond Ser B Biol Sci. 2003:270(1533):2565–2571.

Ferguson NM, Laydon D, Nedjati-Gilani G, Imai N, Ainslie K, Baguelin M, Bhatia S, Boonyasiri A, Cucunubá Z, Cuomo-Dannenburg G et al. Report 9: impact of non-pharmaceutical interventions (npis) to reduce COVID19 mortality and healthcare demand (Vol. 16). London: Imperial College London.

Fintzi J, Wakefield J, Minin VN. A linear noise approximation for stochastic epidemic models fit to partially observed incidence counts. Biometrics. 2022:78(4):1530–1541.

Guttorp P. Stochastic Modeling of Scientific Data (1st ed.). Chapman and Hall/CRC; 1995.

He D, Ionides EL, King AA. Plug-and-play inference for disease dynamics: measles in large and small populations as a case study. J R Soc Interface. 2010:7(43): 271–283.

Ho LST, Crawford FW, Suchard MA. Direct likelihood-based inference for discretely observed stochastic compartmental models of infectious disease. Ann Appl Stat. 2018a:12(3):1993–2021.

Ho LST, Xu J, Crawford FW, Minin VN, Suchard MA. Birth/birth-death processes and their computable transition probabilities with biological applications. J Math Biol. 2018b:76(4):911–944.

Hobolth A, Stone EA. Simulation from endpoint-conditioned, continuous-time Markov chains on a finite state space, with applications to molecular evolution. Ann Appl Stat. 2009:3(3):1204.

Ju N, Heng J, Jacob PE. 2021. Sequential monte carlo algorithms for agent-based models of disease transmission, arXiv, arXiv:arXiv:2101.12156, preprint: not peer reviewed.

Kenah E, Lipsitch M, Robins JM. Generation interval contraction and epidemic data analysis. Math Biosci. 2008:213(1):71–79.

Kermack WO, McKendrick AG. A contribution to the mathematical theory of epidemics. Proc R Soc Lond Ser A. 1927:115(772):700–721.

Kiss IZ, Green DM, Kao RR. Infectious disease control using contact tracing in random and scale-free networks. J R Soc Interface. 2006:3(6):55–62.

Lunz D, Batt G, Ruess J. To quarantine, or not to quarantine: A theoretical framework for disease control via contact tracing. Epidemics. 2021:34:100428.

Morsomme R, Xu J. 2022. Exact inference for stochastic epidemic models via uniformly ergodic block sampling, arXiv, arXiv:2201.09722, preprint: not peer reviewed.

Motta FC, McGoff KA, Deckard A, Wolfe CR, Bonsignori M, Moody MA, Cavanaugh K, Denny TN, Harer J, Haase SB. Assessment of simulated surveillance testing and quarantine in a sars-cov-2–vaccinated population of students on a university campus. JAMA Health Forum. 2021:2(10):e213035.

Nielsen SF. The stochastic em algorithm: estimation and asymptotic results. Bernoulli. 2000:6(3):457–489.

Pooley CM, Bishop SC, Marion G. Using model-based proposals for fast parameter inference on discrete state space, continuous-time Markov processes. J R Soc Interface. 2015:12(107):20150225.

Rao V, Teg YW. Fast mcmc sampling for markov jump processes and extensions. J Mach Learn Res. 2013:14(11):3295–3320

Renshaw E. *Stochastic population processes: analysis, approximations, simulations*. Oxford, New York: Oxford University Press; 2015.

Rieger NS, Worley NB, Ng AJ, Christianson JP. Insular cortex modulates social avoidance of sick rats. Behav Brain Res. 2022:416:113541.

Rose EB, Roy JA, Castillo-Neyra R, Ross ME, Condori-Pino C, Peterson JK, Naquira-Velarde C, Levy MZ. A real-time search strategy for finding urban disease vector infestations. Epidemiol Methods. 2020:9(1):20200001.

Shukla P, Lee M, Whitman SA, Pine, KH. Delay of routine health care during the covid-19 pandemic: a theoretical model of individuals' risk assessment and decision making. Soc Sci Med. 2022:307:115164.

Soriano-Arandes A, Gatell A, Serrano P, Biosca M, Campillo F, Capdevila R, Fàbrega A, Lobato Z, López N, Moreno AM et al. Household sars-cov-2 transmission and children: a network prospective study. Clin Infect Dis Off Public Infect Dis Soc Am. 2021:73(6):e1261–e1269.

Stutz TC, Sinsheimer JS, Sehl M, Xu J. 2022. Computational tools for assessing gene therapy under branching process models of mutation. Bulletin of Mathematical Biology. 84:1–17.

Wang S, Walker SG. Bayesian data augmentation for partially observed stochastic compartmental models. Bayesian Anal. 2023:1(1):1–24.

Xu J, Guttorp P, Kato-Maeda M, Minin VN. Likelihood-based inference for discretely observed birth–death-shift processes, with applications to evolution of mobile genetic elements. Biometrics. 2015:71(4):1009–1021.

Xu J, Minin VN. Efficient transition probability computation for continuous-time branching processes via compressed sensing. In Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence (UAI'15). Arlington, Virginia, USA: AUAI Press. p. 952–961.