Variational Inference with Agent-Based Models

Wen Dong
Department of Computer Science and Engineering &
Institute of Sustainable Transportation and Logistics
State University of New York at Buffalo
wendong@buffalo.edu

ABSTRACT

In this paper, we develop a variational method to track and make predictions about a real-world system from continuous imperfect observations about this system, using an agent-based model that describes the system dynamics. By combining the power of big data with the power of model-thinking in the stochastic process framework, we can make many valuable predictions. We show how to track the spread of an epidemic at the individual level and how to make short-term predictions about traffic congestion. This method points to a new way to bring together modelers and data miners by turning the real world into a living lab.

Keywords

Social simulation, interactive simulation, novel agent and multiagent applications, epidemic dynamics, short term traffic forecasting, discrete event simulation, stochastic kinetic model, variational methods, expectation propagation, Bethe variational principle, Markov process.

1. INTRODUCTION

Agent-based modeling has been employed by researchers in many disciplines to specify the elements of a complex system and their interactions, to check their understandings of a system, to conduct thought experiments and to inform design and analysis [16, 47, 6, 21]. With the availability of big data in recent years [10, 30, 12], we hope to track and make predictions about a real world system from the data that represent the continuous observations of this system and an agent-based model that specifies how the system evolves, and consequently to turn our world into a living lab. In this paper, we identify the agent-based model as a discrete-event Markov process, and develop a variational inference method that searches the latent state trajectories of the elements of the system in the probability space that are most compatible with the noisy observations by minimizing the Bethe variational principle [4].

Data have traditionally been used by agent-based modelers to calibrate model parameters, drive model execution, and validate the model. The inquiry in this paper is instead about how continuous imperfect observations about a real-

world system can help us make inferences about the system here and now. Instead of simulating traffic jams at rush hours using road network data and agent trips synthesized from census data [32, 42], we are more interested in predicting whether today's traffic jams will be formed earlier or last longer from the trajectories of probe vehicles, and how the knowledge about future traffic states will help drivers use the road network more efficiently. Instead of constructing the S-shaped curve of an infectious population from simulation [8, 31], we are more interested in who got a sniffle from his dynamic social network and how we can prevent epidemics from further spreading [29, 14, 18]. Instead of showing the emergence of cities and roads from how people explore and exploit resources [2, 44, 19], We are more interested in identifying poverty and extracting census information from how people make phone calls [5, 10, 38]. Predictions with an agent-based model about real-world data are interpretable in terms of how agents interact with one another and change states, and are amenable to reason regarding non-recurrent scenarios. This transparency about the predictions is lacking in non-parametric approaches.

Our approach is to identify an agent-based simulator as a Markov process, and to search in the probability space specified by the simulator for agent behaviors and interactions that best match the continuous observations about our real-world system. The key observation behind this approach is that an agent-based simulator generates different sample paths with different probabilities — it therefore defines a stochastic process with a probability measure assigned to the space of the sample paths that describe the interactions among the elements of the system. In this stochastic process, the system state as a function of time is composed of the states of its elements. This stochastic process is driven by a number of events that change the system state and happen with event rates that are functions of the current system state. A sample path of the stochastic process is defined by a sequence of events and the corresponding times when those events happened, from which we can unambiguously recover the system state as a function of time. An agent-based simulator therefore iteratively samples the next event according to event rates then changes the world state according to the sampled state starting from the initial state, until the required amount of simulated time has passed.

To find out the maximum likelihood probability distribution of the system state X_t for $t_1 \leq t \leq t_2$ from observed system state x_{t_1} at time t_1 and x_{t_2} at time t_2, we follow the forward-backward algorithm: we first let the probability mass diffuse from x_{t_1} in the forward step according to how

this system evolves from t_1 to t_2, and then we iteratively trace backwards from t_2 to t_1 in the backward step how the probability mass ended at system state x_{t_2} instead of another state. After the forward step and the backward step, we get the probability distribution of system state X_t for $t_1 \leq t \leq t_2$ conditioned on both of its states x_{t_1} and x_{t_2}.

The challenge in making probabilistic inferences about an agent-based model is that we have to deal with an exploding state space — for just a simple task of tracking the binary states of 50 agents, we must cope with 2^{50} combinatorial states because the agents interact with one another. And, of course, a real-world system is much larger. To cope with this exploding state space, we use mean field approximation: the probabilistic evolution of an agent state is determined by the mean field (average) effect of the states of the other agents. The variational framework for making inferences about stochastic processes was developed in the field of machine learning [45] as minimizing Bethe variational principle [4] with applications to expectation propagation [33, 24] and loopy belief propagation [34].

This paper therefore advocates that we should combine the power of big data and the power of model-thinking in the stochastic process framework. Agent-based modeling is a physicist's approach for modeling human societies when data are unavailable and experiments impossible [16], and we believe that big data will transform agent-based modeling from speculation into a physical science. This paper also offers a solution that fits to big time-series data any agent-based model defined by a production rule system based on mean-field approximation. Hence, this system brings together modelers and data miners. We have benchmarked our solution on systems of hundreds of agents, and our benchmarking gives meaningful results.

The rest of this paper is organized as follows. In Section 2 we introduce a probabilistic production (rule) system to describe the microscopic dynamics of a generative model, and identify the production system as a stochastic process. In Section 3 we derive a mean-field solution to the generative model under the constraint of data. In other words, given a simulator and noisy observational data about a process generated by the simulator logic, our algorithm infers the probabilities on a per-agent basis of all possible outcomes. In Section 4 we give examples and benchmark this algorithm against other algorithms. We summarize what we have accomplished and offer our speculation about big data in Section 5.

2. STOCHASTIC PROCESS INDUCED BY AGENT-BASED MODELS

In this section, we introduce the stochastic kinetic model described by the Gillespie algorithm [22] to make inferences about social dynamics from information about individuals in the social system. A "stochastic kinetic model" is a chemist's way of describing the temporal evolution of a system with M agent species driven by V events (or chemical reactions) parameterized by rate constants $c = (c_1, \ldots, c_V)$. At any specific time t, the populations of the species are $x_t = (x_t^{(1)}, \ldots, x_t^{(M)})$. An event v happens with rate $h_v(x_t, c_v)$, changing the populations by Δ_v. The V events are mutually independent.

Gillespie algorithm

1. Initialize the system at time $t = 0$ with rate constants $c_1, \ldots, c_V$ and initialize the populations of the species as $x^{(1)}, \ldots x^{(M)}$.

2. Simulate the time τ to the next event according to exponential distribution $\tau \sim \text{Exponential}(h_0(x, c) = \sum_{v=1}^{V} h_v(x, c_v))$.

3. Simulate the event v according to categorical distribution $v \sim \text{Categorical}(\frac{h_1}{h_0}, \ldots, \frac{h_V}{h_0})$.

4. Update and output time $t \leftarrow t + \tau$ and populations $x \leftarrow x + \Delta_v$.

5. Repeat steps 2-5 until the termination condition is satisfied.

The stochastic kinetic model specified by the Gillespie algorithm assigns a probability measure to a sample path of the system induced by a sequence of events $v_1, \ldots, v_n$, happening between time 0 and time T, $0 = t_0 < t_1 < \cdots < t_n < T$, which is

$$P(v_1, \ldots, v_n, t_1, \ldots, t_n, x) \qquad (1)$$
$$= \prod_{i=1}^{n} h_{v_i}(x_{t_{i-1}}, c_{v_i}) \exp\left(-\sum_{i=1}^{n} h_0(x_{t_{i-1}}, c)(t_i - t_{i-1})\right)$$
$$= \prod_{i=1}^{n} h_{v_i}(x_{t_{i-1}}, c_{v_i}) \exp\left(-\int d\,t\, h_0(x_t, c)\right).$$

The stochastic kinetic model was designed to explain the macroscopic properties of a system via the microscopic interactions among particles in the system [47]. If we are able to observe particles, we expect not only to improve our estimation of the system properties but also to make inferences about the particles. When this model is applied to a social system, the particles are the individuals in the system, and the capability to make inferences about these particles becomes even more important. An event in the Gillespie algorithm looks like the following:

$$\alpha_v^{(1)} X^{(1)} + \cdots + \alpha_v^{(M)} X^{(M)} \to \beta_v^{(1)} X^{(1)} + \cdots + \beta_v^{(M)} X^{(M)}, \qquad (2)$$

$$h_v(x, c_k) = c_v g_v(x) = c_v \prod_{m=1}^{M} g_v^{(m)}(x^{(m)}) = c_v \prod_{m=1}^{M} (x^{(m)})^{\alpha_v^{(m)}}, \qquad (3)$$

$$\Delta_v = (\beta_v^{(1)} - \alpha_v^{(1)}, \cdots, \beta_v^{(M)} - \alpha_v^{(M)}).$$

When $\alpha_v^{(1)}$ individuals of species 1, $\alpha_v^{(2)}$ individuals of species 2 ... meet, they trigger event v with rate constant c_v, which result in $\beta_v^{(1)}$ individuals of species 1, $\beta_v^{(2)}$ individuals of species 2, and so on. The rate $h_v(x, c_v)$ for this event to happen is rate constant c_v times a total of $\prod_{m=1}^{M} \left(x^{(m)}\right)^{\alpha_v^{(m)}}$ different ways for the individuals to meet.

When the components of the agent-based model $X^{(1)}, \cdots, X^{(M)}$ lose meaning as the populations of agent species, we can find the probability distribution of a component state among a finite partition of the component's state space. When the event rates cannot be expressed as the multiplications of component contributions in the form of Eq. 3, we

take Taylor expansions of the event rates around the mean value of the system state

$$h_v(x, c_v) \quad = \quad \sum_{|\alpha| \geq 0} \frac{(x - \mathbf{E}x)^\alpha}{\alpha!} \partial^\alpha h_v(\mathbf{E}x, c_v),$$

with each term being in the form of Eq. 3. α is a multi-index and E is the expectation operator.

Although the stochastic kinetic model is a continuous time model, we work with a discrete time stochastic model in the rest of this paper, because our goal is to track stochastic kinetic dynamics from observations of populations or individuals with countably many computational steps. There are two ways to turn a continuous stochastic process into a discrete one, both of which involve Jensen's uniformization/randomization method [23].

The first method for discretizing a continuous time stochastic system is through approximating the continuous time process with a discrete time process on a countable set of equally spaced time points $0, \tau, 2\tau, \ldots$, with a time interval so small that the probability of more than one event happening in the interval τ is negligible. This approximation works because the state transition kernel from time 0 to time τ is $p(x_0 \rightarrow x_\tau) = \sum_{n=0}^{\infty} \left(I + \frac{Q}{\gamma} \right)^n \exp(-\gamma\tau)\frac{(\gamma\tau)^n}{n!}$ according to the uniformization method, where γ is a uniformization rate, I is the identity matrix and Q is the infinitesimal generator defined by $h_k, k = 1, \ldots, V$. With $\gamma \rightarrow \infty$ and $\gamma\tau = 1$, we get a first-order approximation of the state transition kernel $I + Q \cdot \tau$.

Specifically, let $v_1, \ldots, v_T$ be a sequence of events in the discrete time stochastic kinetic system, $x_1, \ldots, x_T$ be a sequence of states (populations of species), and $y_1, \ldots, y_T$ be a set of observations about the populations. Our goal is to make inferences about $\{v_t, x_t : t = 1, \ldots T\}$ from $\{y_t : t = 1, \ldots, T\}$ according to the following probability measure, where indicator function $1(x_t - x_{t-1} = \Delta_{v_t})$ is 1 if the previous state is x_{t-1} and the current state is $x_t = x_{t-1} + \Delta_{v_t}$, and 0 otherwise.

$$P\left(v_{1,\ldots,T}, x_{1,\ldots,T}, y_{1,\ldots,T}\right) = \prod_{t=1}^{T} P(x_t, y_t, v_t | x_{t-1}), \qquad (4)$$

where $P(x_t, y_t, v_t | x_{t-1})$

$$= P(v_t | x_{t-1}) 1(x_t - x_{t-1} = \Delta_{v_t}) P(y_t | x_t), \quad (5)$$

and

$$P(v_t | x_{t-1}) = \begin{cases} c_k \tau \cdot g_k (x_{t-1}) & \text{if } v_t = k \\ 1 - \sum_j c_j \tau g_j (x_{t-1}) & \text{if } v_t = \emptyset \end{cases} . \quad (6)$$

The second way to discretize a continuous time stochastic system is by introducing a uniformization rate γ that is faster than all event rates in Q and inspecting a discrete time Markov chain defined by the state transition matrix $I + \frac{Q}{\lambda}$, with the transitions happening at time $t_1, t_2, \ldots$, sampled according to a uniform Poisson process with rate γ. This works because according to the uniformization method the uniformised continuous time process has the same probability measure as the original process.

We employ a stochastic kinetic model to simplify the state space transition kernel for several reasons. First, the stochastic kinetic model already successfully describes the time evolution of reaction systems in many areas, including chemistry and cell biology [1, 22]. It is therefore a more natural model for describing and tracking the spatio-temporal process driven by events. Second, the event based transition kernel is more general and flexible–we can define the number of events based on the complexity of real transitions.

3. MAKING INFERENCES WITH AN AGENT-BASED MODEL

In this section, we derive a mean-field solution to infer the probabilities on a per-agent basis of all possible paths of system evolution, given a simulator and noisy observational data about this system generated by the simulator logic.

3.1 Variational Inference

Recall the forward-backward algorithm to make inferences with a state space model [39]. Let X_t be the hidden states and y_t be the observations of a discrete-time state-space model (Kalman filter and hidden Markov model) identified by a transition probability $P(X_{t+1} | X_t)$ and an observation model $P(Y_t | X_t)$, where $t = 1, \cdots, T$. The forward-backward algorithm for making inferences about hidden states X_t from observations y_t is comprised of a forward/filtering sweep to compute the forward statistics $\alpha(x_t) = P(x_t | y_1, \cdots, y_T)$ and a backward/smoothing sweep to estimate the one-slice statistics $\gamma(y_t) = P(x_t | y_1, \cdots, y_T)$. From the forward statistics and the one-slice statistics we can extract the backward statistics $\beta(x_t) = \gamma(x_t)/\alpha(x_t)$ and the two-slice statistics $\xi(x_t, x_{t+1}) = \alpha(x_t) P(y_{t+1}, x_{t+1} | x_t) \beta(x_{t+1}) P(y_{t+1} | y_{1,\cdots,t})$. Here we follow the tradition, use upper case letters for random variables and use lower case letters for the values of random variables.

The challenge with making inferences about a non-trivial agent-based model is that we have to search in a formidable state space — $X_t = (X_t^{(1)}, X_t^{(2)}, \cdots, X_t^{(M)})$, where the superscripts $1, \cdots, M$ represent the states of the interacting elements of the system. We therefore estimate the state distributions of the hidden states in an amenable state space with mean field approximation $\gamma_t(x_t) = \prod_m \gamma_t^{(m)}(x_t^{(m)})$:

minimize over $\xi_t(x_{t-1,t}, v_t)$:

$$\sum_{t; x_{t-1,t}; v_t} \xi_t(x_{t-1,t}, v_t) \log \frac{\xi_t(x_{t-1,t}, v_t)}{P(x_t, v_{t-1}, y_t | x_{t-1})} \qquad (7)$$
$$- \sum_{t; x_t} \prod_m \gamma_t^{(m)}(x_t^{(m)}) \log \prod_m \gamma_t^{(m)}(x_t^{(m)})$$

subject to:

$$\sum_{v_t; x_{t-1,t}: \text{ fixed } x_t^{(m)}} \xi_t(x_{t-1}, x_t, v_t) = \gamma_t^{(m)}(x_t^{(m)}), \text{ for all } t, m, x_t^{(m)}, \quad (8)$$

$$\sum_{v_t; x_{t-1,t}: \text{ fixed } x_{t-1}^{(m)}} \xi_t(x_{t-1}, x_t, v_t) = \gamma_{t-1}^{(m)}(x_{t-1}^{(m)}), \text{ for all } t, m, x_{t-1}^{(m)}, \quad (9)$$

$$\sum_{x_t: \text{ fixed } x_t^{(m)}} \gamma_t^{(m)}(x_t^{(m)}) = 1, \text{ for all } t, m, x_t^{(m)}. \qquad (10)$$

We apply the method of Lagrange multipliers to solve this optimization problem, which begins with forming the Lagrange function to be optimized:

$$\sum_{t; x_{t-1,t}; v_t} \xi_t(x_{t-1,t}, v_t) \log \frac{\xi_t(x_{t-1,t}, v_t)}{P(x_t, v_{t-1}, y_t | x_{t-1})} \qquad (11)$$
$$- \sum_{t; x_t} \prod_m \gamma_t^{(m)}(x_t^{(m)}) \log \prod_m \gamma_t^{(m)}(x_t^{(m)})$$

$$+ \sum_{t;m;x_t^{(m)}} \beta_{t,x_t^{(m)}}^{(m)} (\sum_{v_t; x_{t-1,t}: \text{ fixed } x_t^{(m)}} \xi_t(x_{t-1}, x_t, v_t) - \gamma_t^{(m)}(x_t^{(m)}))$$

$$+ \sum_{t;m;x_{t-1}^{(m)}} \alpha_{t-1,x_{t-1}^{(m)}}^{(m)} (\sum_{v_t; x_{t-1,t}: \text{ fixed } x_{t-1}^{(m)}} \xi_t(x_{t-1}, x_t, v_t) - \gamma_{t-1}^{(m)}(x_{t-1}^{(m)})).$$

Taking the derivative of the expression involving Lagrange multipliers over $\xi_t(x_{t-1}, x_t, v_t)$ and $\gamma_t^{(m)}(x_t^{(m)})$, we see that $\alpha_t^{(m)}(x_t^{(m)}) = \exp(\sum_i \alpha_{t,i}^{(m)} \cdot 1(x_t^{(m)}=i))$ is associated with the marginalized forward probabilities, $\beta_t^{(m)}(x_t^{(m)}) = \exp(\sum_i \beta_{t,i}^{(m)} \cdot 1(x_t^{(m)}=i))$ is associated with the marginalized backward probabilities, with $\gamma_t^{(m)}(x_t^{(m)}) = \alpha_t^{(m)}(x_t^{(m)})\beta_t^{(m)}(x_t^{(m)})$. The dual optimization problem is to find the marginal forward statistics $\alpha_t^{(m)}(x_t^{(m)})$ and the marginal backward statistics $\beta_t^{(m)}(x_t^{(m)})$ to maximize the approximate partition function given by Eq. 12, and the solution is the fixed point of Eq. 13, where normalization constant $Z_t = P(y_t | y_{1,\cdots,t-1})$:

$$\log P(y_1,\cdots,T) = \sum_t \log \sum_{x_{t-1,t}} \prod_m \alpha_{t-1}^{(m)}(x_{t-1}^{(m)}) P(x_t, v_t | x_{t-1}) \prod_m \beta_t^{(m)}(x_t^{(m)}) \tag{12}$$

$$\xi_t(x_{t-1}, x_t, v_t) = \frac{1}{Z_t} P(x_t, v_t | x_{t-1}) \tag{13}$$

$$\cdot \prod_m \alpha_{t-1}^{(m)}\left(x_{t-1}^{(m)}\right) \cdot \prod_m P\left(y_t^{(m)} | x_t^{(m)}\right) \prod_m \beta_t^{(m)}\left(x_t^{(m)}\right),$$

where $P(x_t, v_t | x_{t-1}) = \tag{14}$

$$\begin{cases} c_k \tau \prod_m g_k^{(m)}\left(x_{t-1}^{(m)}\right) \cdot \prod_m 1(x_t^{(m)} - x_{t-1}^{(m)} = \Delta_k^{(m)}) & v_t = k \neq \emptyset \\ (1 - \sum_k c_k \tau \prod_m g_k^{(m)}(x_{t-1}^{(m)})) \cdot \prod_m 1(x_t^{(m)} - x_{t-1}^{(m)} = 0) & v_t = \emptyset. \end{cases}$$

Marginalizing Eq. 13 over all chains $X_t^{(m')}$ for $m' \neq m$ and t, we find that the solution to the Bethe variational principle is the mean field approximation of the original dynamics with marginal two-slice statistic given by Eq. 15. From the mean field approximation, $\alpha_{t,i}^{(m)}$ and $\beta_{t,i}^{(m)}$ can be solved by forward-backward algorithm and fixed point iteration.

$$\xi_t(x_{t-1}^{(m)}, x_t^{(m)}, v_t) = \frac{1}{Z_t} P(x_t^{(m)}, v_t | x_{t-1}^{(m)}) \tag{15}$$

$$\cdot \alpha_{t-1}^{(m)}(x_{t-1}^{(m)}) \cdot P(y_t^{(m)}|x_t^{(m)})\beta_t^{(m)}(x_t^{(m)}),$$

where $P(x_t^{(m)}, v_t | x_{t-1}^{(m)}) \propto \tag{16}$

$$\begin{cases} c_k \tau g_k^{(m)}(x_{t-1}^{(m)}) \prod_{m' \neq m} \tilde{g}_{k,t-1}^{(m')} \cdot 1(x_t^{(m)} - x_{t-1}^{(m)} = \Delta_k^{(m)}) & v_t^{(m)} = k \neq \emptyset \\ \left(1 - \sum_k c_k \tau g_k^{(m)}(x_{t-1}^{(m)}) \prod_{m' \neq m} \hat{g}_{k,t-1}^{(m')}\right) 1(x_t^{(m)} - x_{t-1}^{(m)} = 0) & v_t^{(m)} = \emptyset, \end{cases}$$

$$\tilde{g}_{k,t-1}^{(m')} = \frac{\sum\limits_{\substack{x_t^{(m')} - x_{t-1}^{(m')} \equiv \Delta_k^{(m')}}} \alpha_{t-1}^{(m')}(x_{t-1}^{(m')}) P(y_t^{(m')}|x_t^{(m')})\beta_t^{(m')}(x_t^{(m')}) g_k^{(m')}(x_{t-1}^{(m')})}{\sum\limits_{\substack{x_t^{(m')} - x_{t-1}^{(m')} \equiv 0}} \alpha_{t-1}^{(m')}(x_{t-1}^{(m')}) P(y_t^{(m')}|x_t^{(m')})\beta_t^{(m')}(x_t^{(m')})},$$

$$\hat{g}_{k,t-1}^{(m')} = \frac{\sum\limits_{\substack{x_t^{(m')} - x_{t-1}^{(m')} \equiv 0}} \alpha_{t-1}^{(m')}(x_{t-1}^{(m')}) P(y_t^{(m')}|x_t^{(m')})\beta_t^{(m')}(x_t^{(m')}) g_k^{(m')}(x_{t-1}^{(m')})}{\sum\limits_{\substack{x_t^{(m')} - x_{t-1}^{(m')} \equiv 0}} \alpha_{t-1}^{(m')}(x_{t-1}^{(m')}) P(y_t^{(m')}|x_t^{(m')})\beta_t^{(m')}(x_t^{(m')})},$$

$$Z_t = \sum_j c_j \tau \prod_m \tilde{g}_{j,t-1}^{(m)} + 1 - \sum_j c_j \tau \prod_m \hat{g}_{j,t-1}^{(m)}.$$

The above is a factorized stochastic kinetic model. The marginal two-slice probability $\xi_t(x_{t-1}^{(m)}, x_t^{(m)}, v_t)$ in Eq. 15

takes the same form as the coupled two-slice probability $\xi_t(x_{t-1}, x_t, v_t)$ in Eq. 13. The marginal state transition kernel $P(x_t^{(m)}, v_t | x_{t-1}^{(m)})$ in Eq. 16 consists of choosing an event (or no event) v_t according to event probability $P(v_t | x_{t-1}^{(m)})$ and changing the state $x_t^{(m)}$ in a deterministic way, similar to the joint state transition kernel $P(x_t, v_t | x_{t-1})$ in Eq. 14, except that we marginalize over all $x^{(m')}$ for $m' \neq m$.

Hence the solution to the above Bethe variational principle through Legendre-Fenchel transform [40] is one in which the interacting elements of the system evolve their states marginally according to the average effects of the other elements. As such, instead of searching the joint probability space of of $(X_1, \cdots, X_T)$, we search the marginal probability spaces of $(X_1^{(m)}, \cdots, X_T^{(m)})$.

3.2 Graphical Model Representation

The stochastic kinetic model with its distinct graphical model structure is more suitable than traditional models for modeling complex interactions in social dynamics. To illustrate this point, we compare its graphical model with the coupled hidden Markov model.

A coupled hidden Markov model (CHMM, Figure 1(a)) combines a number of conventional hidden Markov models (HMMs) to model the dynamics of interacting processes [7, 35]. In CHMM the latent state of HMM at time t depends on the latent states of *all* HMMs at time $t - 1$. Traditional ways to simplify the state transition kernel include the factorial hidden Markov model which decouples the inter-chain probability dependence [28] and the hidden Markov decision tree which assumes fixed and sparse inter-chain probability dependence [20].

A stochastic kinetic model has a graphical model representation different from CHMM, as shown in Figure 1(b). First, we define a set of stochastic events to summarize the complex interactions and decouple direct dependencies between nodes. Second, while the system can move from any state to any other state in a CHMM, in any infinitesimal time interval no more than one out of V possible events is happening in a stochastic kinetic model. Third, conditioned on the system state $x_{t-1}^{(1)}, \ldots, x_{t-1}^{(M)}$ describing the populations of species $1, \ldots, M$, the latent state at the next time step $x_t^{(1)}, \ldots, x_t^{(M)}$ could be dependent. (Consider an event that changes population m_1 and population m_2 simultaneously.) In contrast, in a CHMM the states $x_t^{(1)}, \ldots, x_t^{(M)}$ at time t are conditionally independent given the states $x_{t-1}^{(1)}, \ldots, x_{t-1}^{(M)}$ at time $t - 1$. Thus the inference algorithms of CHMM are not applicable for modeling the complex interactions in social dynamics driven by events.

The factorial stochastic kinetic model has a graphical model similar to the stochastic kinetic model except that it factorizes stochastic events to individual species (Figure 1(c)). The new graphical model further simplifies the inference algorithm.

3.3 Parameter Learning

In order to find the rate constants c_v in a stochastic kinetic model and in a factorial stochastic kinetic model, we maximize the expected log likelihood and the Bethe entropy approximation respectively over these rate constants.

The likelihood of rate constants c_v in a continuous time stochastic kinetic model with respect to a sample path iden-
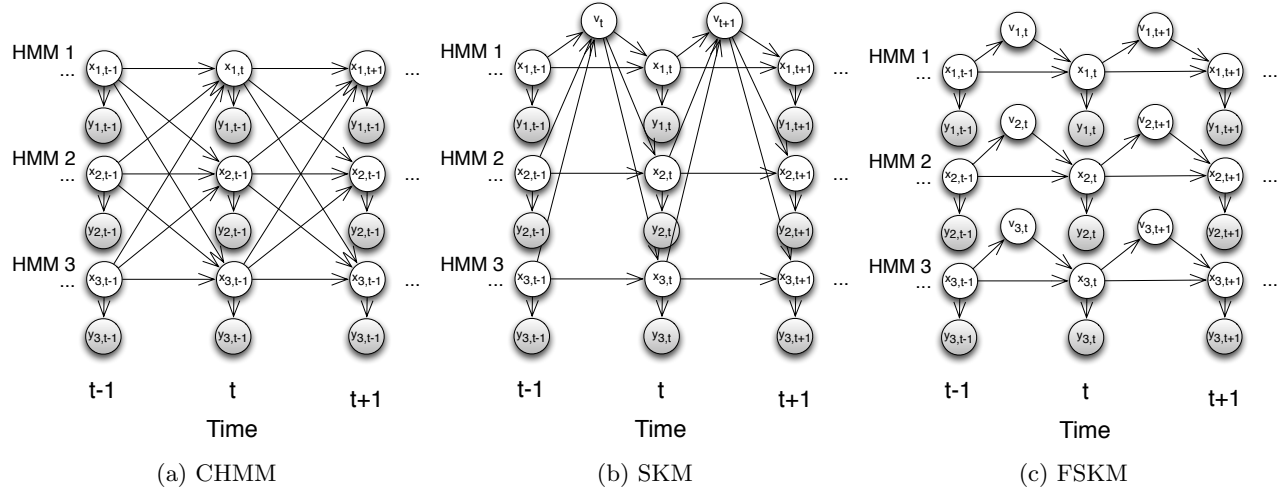
(a) CHMM (b) SKM (c) FSKM

Figure 1: In comparison of coupled hidden markov model (a), a stochastic kinetic model (b) decouples inter-chain interactions with events v and allows factorization (c).

tified by events $v_1, \cdots, v_n$, system states to $x_{t_1}, \cdots, x_{t_n}$ and times $t_1, \cdots, t_n$ is given in Eq. 1, and the event rates are given in Eq. 3. To find the maximum likelihood estimate of the rate constants, we set the partial derivatives of the log-likelihood over the rate constants to 0:

$$\log P = \sum_i \log c_{v_i} + \log g_{v_i}(x_{t_{i-1}}) - \sum_i \sum_j c_j g_j(x_{t_{i-1}}) \cdot (t_i - t_{i-1}),$$

$$\frac{\partial}{\partial c_v} \log P = \frac{\sum_i 1(v_i = v)}{c_v} - \sum_i g_v(x_{t_{i-1}}) \cdot (t_i - t_{i-1}),$$

$$c_v = \frac{\sum_i 1(v_i = v)}{\sum_i g_v(x_{t_{i-1}})(t_i - t_{i-1})} = \frac{\sum_i 1(v_i = v)}{\int_0^T d\,t\,g_v(x(t))}.$$

Hence, the maximum likelihood estimate of the rate constants c_v is such that the numbers of events that are expected to happen according to the event rates along the sample path $\int_0^T dt\, c_v g_v(x(t))$ match the numbers of events $\sum_i 1(v_i = v)$ that happened in the sample path. Indicator function $1(v_i = v)$ takes value 1 if $v_i = v$ and 0 if $v_i \neq v$.

The likelihood of rate constants c_v in a discrete time stochastic kinetic model with respect to a sample path identified by events $v_1, \cdots, v_T$, system states $x_1, \cdots, x_T$ and observations $y_1, \cdots, y_T$ is given in Eq. 4. To find the maximum likelihood estimate of the rate constants, we similarly set the partial derivatives of the log-likelihood over the rate constants to 0:

$$\log P = \sum_t \log p(y_t|x_t) + \sum_t \log p(v_t|x_{t-1}) 1_{(x_t - x_{t-1} = \Delta_{v_t})}$$

$$\frac{\partial}{\partial c_v} \log P = \frac{\sum_t 1(v_t = v)}{c_v} - \sum_t \frac{\tau g_v(x_{t-1}) 1(v_t = \emptyset)}{1 - \sum_j c_j \tau g_j(x_{t-1})}$$

$$c_v = \frac{\sum_t 1(v_t = v)}{\sum_t \frac{\tau g_v(x_{t-1}) \cdot 1(v_t = \emptyset)}{1 - \sum_j c_j \tau g_j(x_{t-1})}}$$

$$\xrightarrow{\tau \to 0} \frac{\sum_t 1(v_t = v)}{\sum_t \tau g_v(x_{t-1})}.$$

Hence, the maximum likelihood estimate of the rate constants c_v in a discrete-time stochastic kinetic model similarly matches the numbers of events that are expected to happen according to the event rates ($\sum_t \frac{c_v \tau g_v(x_{t-1}) 1(v_t = \emptyset)}{1 - \sum_j c_j \tau g_j(x_{t-1})}$,

extrapolated from the times of null events) with the number of events that happened ($\sum_t 1(v_t = v)$). As the interval τ approaches 0, the probability of a null event ($1 - \sum_j c_j \tau g_j(x_{t-1})$) and the fraction of null events ($\sum_{t=1}^T 1(v_t = \emptyset)/T$) both approach 1, and the maximum likelihood estimate of the rate constants in the discrete time stochastic kinetic model approaches the maximum likelihood estimate in the continuous time stochastic kinetic model.

When the events $v_1, \cdots, v_T$ and the system states $x_1, \cdots, x_T$ are unobserved latent variables, we use the expectation maximization (EM) algorithm to iteratively search for the rate constants that maximize the expected log likelihood over the probability distribution of the latent variables. EM is an iterative method for finding the maximum likelihood estimate of the parameters in statistical models involving unobserved latent variables [11]. It alternates performing the expectation (E) step, which constructs the expected log likelihood as a function of the parameters over the probability distribution of the latent variables using the current estimate for the parameters, with the maximization (M) step, which computes the parameters to maximize the expected log likelihood function constructed in the E step. The estimated parameters are used to determine the probability distribution of the latent variables in the next E step.

The expected log likelihood over the posterior probability of events $v_1, \cdots, v_T$ and system states $x_1, \cdots, x_T$ conditioned on the observations $y_1, \ldots, y_T$ takes the form in Eq. 17. Maximizing this expected log likelihood by setting its partial derivatives over the rate constants gives the updated estimate of rate constants in Eq. 18.

$$E(\log P) = \sum_{t; x_{t-1,t}; v_t} \xi_t(x_{t-1,t}, v_t; c^{\text{old}}) \cdot \log P(x_t, y_t, v_t | x_{t-1}; c) \qquad (17)$$

$$\frac{\partial E(\log P)}{\partial c_v} = \sum_t \frac{\xi(v_t = v)}{c_v} - \sum_{t; x_{t-1}} \frac{\tau g_v(x_{t-1}) \xi(x_{t-1}, v_t = \emptyset)}{1 - \sum_j c_j \tau g_j(x_{t-1})} \stackrel{\text{set}}{=} 0,$$

$$c_v = \frac{\sum_t \xi(v_t = v)}{\sum_{t; x_{t-1}} \frac{\tau g_v(x_{t-1}) \cdot \xi(x_{t-1}, v_t = \emptyset)}{1 - \sum_j c_j \tau g_j(x_{t-1})}} \qquad (18)$$

858

$$\xrightarrow{\tau \to 0} \frac{\sum\limits_{t} \xi(v_t = v)}{\sum\limits_{t;x_{t-1}} \tau \gamma(x_{t-1}) g_v(x_{t-1})}.$$

As such, the rate constant c_v for event v matches the expected number of times this event could have happened ($\sum_t \frac{c_v \tau g_v(x_{t-1}) \xi(x_{t-1}, v_t = \emptyset)}{1 - \sum_j c_j \tau g_j(x_{t-1})}$) according to the event rates ($c_j \tau g_j(x_{t-1})$) along the sample path with the expected number of times the events happened ($\sum_i \xi(v_i = v)$).

Using Bethe entropy approximation, $\alpha_{t-1}(x_{t-1}) = \prod_m \alpha_{t-1}^{(m)}(x_{t-1}^{(m)})$ and $\beta_t(x_t) = \prod_m \beta_t^{(m)}(x_t^{(m)})$, and setting the discretization time interval τ to be small enough, the rate constants c_v can be updated according to Eq. 19.

$$c_v = \frac{\sum\limits_{t} \xi(v_t = v)}{\sum\limits_{t} \tau \prod_m \sum\limits_{x_{t-1}^{(m)}} \gamma^{(m)}(x_{t-1}^{(m)}) g_v^{(m)}(x_{t-1}^{(m)})}, \qquad (19)$$

where $\xi(v_t = v) = \dfrac{c_v^{\text{old}} \prod_m \tilde{g}_{v,t-1}^{(m)}}{\sum_j c_j^{\text{old}} \tau \prod_m \tilde{g}_{j,t-1}^{(m)} + 1 - \sum_j c_j^{\text{old}} \tau \prod_m \hat{g}_{j,t-1}^{(m)}}.$

Therefore, the rate constant for event v is the expected number of occurrences of this event summed over all times, divided by the total cross-section of this event also summed over all times.

To summarize, we provide the variational agent-based inference algorithm below.

Variational Inference with Gillespie Algorithm

Given observations $y_t^{(m)}$ for $t = 1, \ldots, T$ and $m = 1, \ldots, M$, and the stochastic kinetic model of a complex system defined by Eq. 4, find $x_t^{(m)}$, $v_t^{(m)}$ and rate constants c_k for $k = 1, \ldots, V$.

- Latent state inference. Iterate through the following forward pass and backward pass until convergence, where $P(x_t^{(m)}, v_t | x_{t-1}^{(m)})$ is given by Eq. 16.

    - Forward pass. For $t = 1, \ldots, T$ and $m = 1, \ldots, M$, update $\alpha_t^{(m)}(x_t^{(m)})$ according to

$$\alpha_t^{(m)}(x_t^{(m)}) \leftarrow \frac{1}{Z_t} \sum_{x_{t-1}^{(m)}, v_t} \alpha_{t-1}^{(m)}(x_{t-1}^{(m)}) P(x_t^{(m)}, v_t | x_{t-1}^{(m)}) P(y_t^{(m)} | x_t^{(m)}).$$

    - Backward pass. For $t = T, \ldots, 1$ and $m = 1, \ldots, M$, update $\beta_{t-1}^{(m)}(x_{t-1}^{(m)})$ according to

$$\beta_{t-1}^{(m)}(x_{t-1}^{(m)}) \leftarrow \frac{1}{Z_t} \sum_{v_t, x_t^{(m)}} P(x_t^{(m)}, v_t | x_{t-1}^{(m)}) P(y_t^{(m)} | x_t^{(m)}) \beta_t^{(m)}(x_t^{(m)}).$$

- Parameter estimation. Iterate through latent state inference (above) and rate constants estimate of c_k according to Eq. 19, until convergence.

---

4. EXPERIMENTAL RESULTS

In this section, we evaluate the performance of variational agent-based inference for two applications: epidemic dynamics and traffic dynamics. We selected these because they are important applications with significant practical value.

4.1 Epidemic Dynamics

In this section, we infer the progression of common cold in a dynamic social network using an agent-based susceptible-infectious-susceptible (SIS) model at the individual level through a small numbr of volunteers who report their symptoms, and estimate the total number of infectious individuals. Being able to estimate the outbreak of an epidemic in advance and determine who has the highest probability of infection is important for health-care providers and health policy researchers who must optimize limited medical resources. Conventional agent-based epidemic simulators [17, 26, 41, 27, 15] lack the capability to infer epidemic spreading with symptoms observations in the social network, and thus the sample paths given by these simulators can differ significantly from the truth.

In the SIS dynamics, each individual is either infectious (I) or susceptible (S), and the system has three events: a) an in infectious individual in the network infects a susceptible individual and turns that person infectious in the network with rate constant c_1 (probability per unit time), b) an infectious individual recovers and becomes susceptible again with rate constant c_2, and 3) a susceptible individual becomes infectious by contacting an infectious individual from outside the system with rate constant c_3.

$$I + S \to 2 \times I, \text{ infection, rate constant} = c_1,$$
$$I \to S, \text{ recover, rate constant} = c_2,$$
$$S \to I, \text{ infection from outside, rate constant} = c_3.$$

To model the SIS dynamics at the individual level with Gillespie algorithm, we assign two "species" to each person p: $I^{(p)} \in \{0, 1\}$, $S^{(p)} \in \{0, 1\}$ and $I^{(p)} + S^{(p)} = 1$. The probability for a susceptible person p to become infectious through one unit time of contact with an infectious person q is thus $h(x, c_1) = c_1 \cdot s^{(p)} \cdot i^{(q)} = c_1$. The mean field probability for the susceptible person p to become infectious is thus

$$\sum_{q \in \text{neighbor of } p} c_1 \cdot s^{(p)} \cdot \mathbf{E} I^{(q)} = c_1 \cdot s^{(p)} \sum_{q \in \text{neighbor of } p} \mathbf{E} I^{(q)},$$

i.e., the average total number of infectious neighbors in the individual's social network times the probability of infection per infectious neighbor. If infection happens, we change $S^{(p)}$ from 1 to 0 and change $I^{(p)}$ from 0 to 1.

We benchmark the performance of the variational agent-based inference algorithm using the Dartmouth College campus data set [29]. This data set contains the locations of 13,888 on-campus WiFi users from April 2001 to June 2004. On top of this dynamic social network we synthesized epidemic progression using the SIS model and set parameters such that an individual is on average infected twice per year and takes one week on average to recover. We randomly select 10% of individuals as volunteers who report their daily symptoms and from these we infer the daily infectious/susceptible state of the other 90% individuals. As far as we know, there is no real data set with both a large amount of sensor data and symptom reports; we hope our research encourages data collection and analysis in this direction.

Figure 2(a) compares the receiver operating characteristic (ROC) curve of predicting whether an individual is infectious using either a variational agent-based inference algorithm or a support vector classifier [37]. The support vec-
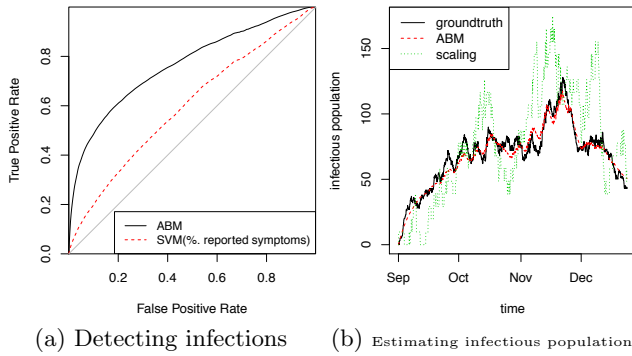
(a) Detecting infections   (b) Estimating infectious population

Figure 2: Statistical inference in Dartmouth data

tor classifier (SVC) estimates the probability that an individual is infectious based on the percentage of his contacts reporting symptoms. Since only 10% of individuals are reporting symptoms in the experiment setup, the likelihood that an individual is infected can be only roughly estimated by SVC. In addition, infectious individuals make different contributions to epidemic progression because infectious diseases from randomly infected individuals go first to the hubs of a social network then spread to the other nodes from those hubs [9], and SVC has difficulty in capturing such network-related features. Variational agent-based inference, on the other hand, can correctly predict 60% of infections with only a 20% false positive rate.

Figure 2(b) compares the performance in estimating the number of infectious persons in the 90% of individuals who do not report their daily symptoms from daily symptoms reported by the 10%. A scaling-based method missed the rapid increase of infectious individuals in early September and overestimated the number of infectious individuals in mid-October and November. This occurred because not all infectious individuals contribute to epidemic progression the same way.

4.2 Traffic Dynamics

In this experiment, we predict road traffic up to one hour ahead of time from a large set of tracked vehicle locations in conjunction with an agent-based transportation simulator called MATSIM [32]. While tracked vehicle locations from car telematics systems are already being exploited to provide drivers with real-time traffic information, the chaotic nature of transportation networks means that an incident at one location might affect the traffic condition of another location up to a hundred miles away. A decision made according to current travel times might therefore be suboptimal, and can in certain cases even lead to global system breakdown.

Researchers employ a transportation simulator to explain the macroscopic phenomena of transportation dynamics by simulating how individuals travel in a real-world transportation network. Such a simulator takes three primary components as its input: a road network like the one used in a GPS navigator, a population specification that lists the location and travel of individuals on a typical day synthesized from census data and trip surveys with the number of simulated vehicles matching the number of vehicles in the real world, and a control file specifying how daily itineraries are scored and how individuals improve their daily itineraries (and the parameters specifying the modeling details). From

these data, the simulator will proceed to execute travel, to score daily itineraries, and to perturb daily itineraries in an attempt to improve them, and then repeat the three steps until equilibrium is reached. Simulation without continuous observtations about a real-world transportation network however doesn't tell us whether today's traffic jams will be formed earlier or last longer.

Many existing algorithms to track and predict real-time traffic dynamics on the other hand — vector ARIMA [3], state space models [46], neural networks [48], and Bayesian network models [25, 13] — have difficulty in coping with noisy and missing data, with making predictions in non-recurrent scenarios, and with explaining predictions in terms of agent trips. According to Vlahogianni [43], the challenge in short-term traffic forecasting is not only to predict but also to explain phenomena at the city network level — to fuse new data sources such as those from telematics units and to easily incorporate the effects of non-recurrent conditions.

To join the event model of a discrete-event simulator with continued observations about real-world social systems, we make use of the fact that all discrete event simulators (at least, to the best of our knowledge) have a way to dump the events happening in a simulation run. As such, we can reconstruct simulation runs according to the event sequences and so reconstruct the stochastic discrete-event model from simulation runs outside the simulator, instead of hacking the source code of a specific simulator over many man-months. MATSIM, for example, has about 140 thousand lines of code, and hacking its source code to make real-time inferences with real-world data wouldn't be easy.

In this way, we dump four events: vehicle leaving a building, vehicle entering a link, vehicle leaving a link and vehicle entering a building. From these four events, we have constructed a data frame representing continued observations of the locations (the link or building) of all vehicles at equally spaced time steps. From the data frame we constructed a state transition matrix to represent vehicle dynamics (with each row and column representing a link/building), along with entries giving the state transition probabilities according to how long a vehicle stays at a link/building and how frequently a vehicle chooses the next link/building. By uniformly sampling a given fraction of tracked vehicles, we have constructed an observation model that provides the probability distribution of observed vehicles at a location given the total number of vehicles there. This system has only one event $p_i \circ l_j \rightarrow p_i \circ l_k$, a vehicle i moving from link/building j to link/building k with rate constant p_{l_j,l_k}, changing the location of vehicle from $X_t^{(p_i)} = l_j$ to $X_{t+1}^{(p_i)} = l_k$, changing the number of vehicles on link l_i from $X_t^{(l_i)} = x_t^{(l_i)}$ to $X_{t+1}^{(l_i)} = x_t^{(l_i)} - 1$, and changing the number of vehicles on link l_k from $X_t^{(l_k)} = x_t^{(l_k)}$ to $X_{t+1}^{(l_k)} = x_t^{(l_k)} + 1$. But the inference is much more complex due to the interactions between vehicles and links.

We employ the mobility traces of more than 500 taxi cabs collected over 30 days in the San Francisco metropolitan area [36] to benchmark the advantages of considering possible future traffic conditions for individual transportation planning. We extract the road network from OpenStreetMap, a collaborative project to create a free editable map of the world. We obtain population distribution and daily trip statistics from the U.S. Census, and obtain the state transi-
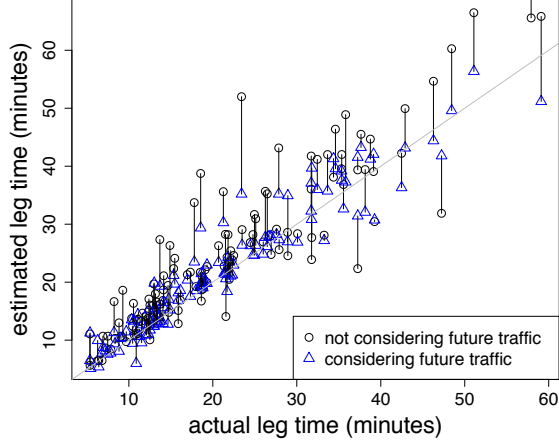
Figure 3: Considering future traffic reduces the relative error of travel time versus actual travel time from 35% when future traffic is not considered to 29% when it is.

tion matrices from one link/building to another at equilibrium from simulations. We map the latitudes and longitudes of tracked vehicle locations to links and buildings because our stochastic inference is at the link/building level.

To benchmark how the estimation of travel time can be improved by considering future traffic conditions even though traffic conditions vary significantly, we extract the 20% of trips with the highest difference between actual travel time and estimated travel time according to traffic at the time of departure, then estimate the average travel time according to our probabilistic model. Such travel often occurs at the rising edge of overall traffic volume, during bad weather, and on less developed roads.

Figure 3 compares these estimated travel times according to only traffic at the time of departure with the same data considering possible future traffic estimations through a random sample of 128 trips. The travel estimations that do *not* project future traffic exhibit on average a 35% relative error (in comparison with a 29% when projecting future traffic), and therefore occasionally differ significantly from the actual travel time.

As such, we can combine the sporadically observed vehicle locations with the large compilation of typical trip plans to continuously estimate current and future traffic conditions. Starting from the number and behavior of tracked vehicles in a road link, we can determine the total number of vehicles in the link by scaling and estimating traffic conditions. If we trace the origins and destinations of the estimated number of vehicles through the factorial stochastic process model (filling any gaps with prior individual travel behaviors), we can extract information about the traffic at other road links. If we then iterate estimations between the traffic at links and the trip choices of simulated vehicles, we improve our estimation of both.

5. CONCLUSIONS AND DISCUSSIONS

In this paper, we have developed a variational method to make inferences about a real-world system from continuous

imperfect observations about the system, using an agent-based model that describes the dynamics of this system. To demonstrate the value of combining the power of big data and the power of model-thinking in the stochastic process framework, we show how we can track epidemics at the individual level from only a small number of volunteers who report their symptoms, and we make short-term predictions about road traffic from sporadically observed probe vehicles. This is just a taste of what this powerful combination of approaches can do, and we expect to see further applications and theoretical development to test the bounds of this methodology.

APPENDIX

The duality between Eq. 7 and Eq. 12 is a duality between maximum relative entropy and maximum pseudo-likelihood. To get the dual form of the Bethe variational problem in Eq. 12, we set the derivatives of Eq. 11 over $\xi_t(x_{t-1,t}v_t)$ and $\gamma_t^{(m)}(x_t^{(m)})$ to 0.

$$\frac{\partial L}{\partial \xi_t(x_{t-1,t},v_t)} = \log \frac{\xi_t(x_{t-1,t},v_t)}{p(x_t,v_t,y_t|x_{t-1})} + 1 - \sum_m \beta_{t,x_t^{(m)}}^{(m)} - \sum_m \alpha_{t-1,x_{t-1}^{(m)}}^{(m)} \overset{\text{set}}{=} 0$$

$$\Rightarrow \xi_t(x_{t-1,t},v_t) \propto \exp\left(\sum_m \alpha_{t-1,x_{t-1}^{(m)}}^{(m)}\right) p(x_t,v_t,y_t|x_{t-1}) \exp\left(\sum_m \beta_{t,x_t^{(m)}}^{(m)}\right),$$

$$\frac{\partial L}{\partial \gamma_t^{(m)}(x_t^{(m)})} = \log \gamma_t^{(m)}(x_t^{(m)}) + 1 - \beta_{t,x_t^{(m)}}^{(m)} - \alpha_{t,x_t^{(m)}}^{(m)} \overset{\text{set}}{=} 0$$

$$\Rightarrow \gamma_t^{(m)}(x_t^{(m)}) \propto \exp\left(\alpha_{t,x_t^{(m)}}^{(m)} + \beta_{t,x_t^{(m)}}^{(m)}\right).$$

After taking $\xi_t(x_{t-1,t},v_t)$ and $\gamma_t^{(m)}(x_t^{(m)})$ into Eq 7, we get Eq. 12.

To derive Eq. 15, we marginalize Eq. 13 over all $x_{t-1,t}^{(m')}$ for $m' \neq m$.

$$\xi(x_{t-1}^{(m)}, v_t, x_t^{(m)})$$
$$= \frac{1}{Z_t} \sum_{x_{t-1,t},v_t:\text{fix } x_{t-1,t}^{(m)},v_t} p(x_t,v_t|x_{t-1}) \prod_m \alpha^{(m)}(x_{t-1}^{(m)})\beta^{(m)}(x_t^{(m)})p^{(m)}(y_t^{(m)}|x_t^{(m)})$$
$$\propto c_k\tau \prod_{m'\neq m} \sum_{x_{t-1,t}^{(m')}} \alpha_{t-1}^{(m')}g^{(m')}(x_{t-1}^{(m')})\beta_t^{(m')}p(y_t^{(m')}|x_t^{(m')})1(\Delta x_t^{(m')}=\Delta_k^{(m')})$$
$$\cdot\alpha_{t-1}^{(m)}(x_{t-1}^{(m)})g^{(m)}(x_{t-1}^{(m)})\beta_t^{(m)}(x_t^{(m)})p(y_t^{(m)}|x_t^{(m)})1(\Delta x_t^{(m)}=\Delta_k^{(m)})$$
$$\text{if } v_t = k \neq \emptyset, \qquad (20)$$
$$\propto \prod_{m'\neq m} \sum_{x_{t-1,t}^{(m')}} \alpha_{t-1}^{(m')}(x_{t-1}^{(m')})\beta_t^{(m')}(x_t^{(m')})p(y_t^{(m')}|x_t^{(m')})1(\Delta x_t^{(m')}=0)$$
$$\cdot\alpha_t^{(m)}(x_t^{(m)})\beta_t^{(m)}(x_t^{(m)})p(y_t^{(m)}|x_t^{(m)})1(\Delta x_t^{(m)}=0)$$
$$- \sum_k c_k\tau \prod_{m'\neq m} \sum_{x_{t-1,t}^{(m')}} \alpha_{t-1}^{(m')}g^{(m')}(x_{t-1}^{(m')})\beta_t^{(m')}p(y_t^{(m')}|x_t^{(m')})1(\Delta x_t^{(m')}=0)$$
$$\cdot\alpha_{t-1}^{(m)}(x_{t-1}^{(m)})g^{(m)}(x_{t-1}^{(m)})\beta_t^{(m)}(x_t^{(m)})p(y_t^{(m)}|x_t^{(m)})1(\Delta x_t^{(m)}=0)$$
$$\text{if } v_t = \emptyset. \qquad (21)$$

After deviding Eq. 20 and Eq. 21 by

$$\prod_{m'\neq m} \sum_{x_{t-1,t}^{(m')}} \alpha_{t-1}^{(m')}(x_{t-1}^{(m')})\beta_t^{(m')}(x_t^{(m')})p(y_t^{(m')}|x_t^{(m')})1(\Delta x_t^{(m')}=0)$$

we get Eq. 15, $\tilde{g}_{k,t-1}^{(m)}$, $\hat{g}_{k,t-1}^{(m)}$ and Z_t.

REFERENCES

[1] Adam Arkin, John Ross, and Harley H McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected escherichia coli cells. *Genetics*, 149(4):1633–1648, 1998.

[2] M. Batty. *Cities and Complexity: Understanding Cities With Cellular Automata, Agent-Based Models, and Fractals*. Mit Press, 2007.

[3] Moshe Ben-Akiva, Michel Bierlaire, Haris Koutsopoulos, and Rabi Mishalani. Dynamit: a simulation-based system for traffic prediction. In *DACCORS short term forecasting workshop, The Netherlands*. Citeseer, 1998.

[4] HA Bethe. Statistical theory of superlattices. In *Proc. Roy. Soc. London A*, volume 150, pages 552–575, 1935.

[5] Vincent D Blondel, Adeline Decuyper, and Gautier Krings. A survey of results on mobile phone datasets analysis. *EPJ Data Science*, 4(1):1–55, 2015.

[6] Andrei Borshchev. *The Big Book of Simulation Modeling: Multimethod Modeling with AnyLogic 6*. AnyLogic North America, 2013.

[7] Matthew Brand, Nuria Oliver, and Alex Pentland. Coupled hidden markov models for complex action recognition. In *Proc. of CVPR*, pages 994–999, 1997.

[8] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Reviews of Modern Physics*, 81(2):591–646, 2009.

[9] Nicholas A. Christakis and James H. Fowler. The spread of obesity in a large social network over 32 years. *The New England Journal of Medicine*, 357:370–379, 2007.

[10] Yves-Alexandre de Montjoye, Zbigniew Smoreda, Romain Trinquart, Cezary Ziemlicki, and Vincent D Blondel. D4d-senegal: the second mobile phone data for development challenge. *arXiv preprint arXiv:1407.4885*, 2014.

[11] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[12] Wen Dong, Bruno Lepri, and Alex Sandy Pentland. Modeling the co-evolution of behaviors and social relationships using mobile phone data. In *Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia*, pages 134–143. ACM, 2011.

[13] Wen Dong and Alex Pentland. A network analysis of road traffic with vehicle tracking data. In *AAAI Spring Symposium: Human Behavior Modeling*, pages 7–12, 2009.

[14] Wen Dong, Alex Pentland, and Katherine A. Heller. Graph-coupled hmms for modeling the spread of infection. In Nando de Freitas and Kevin P. Murphy, editors, *UAI*, pages 227–236. AUAI Press, 2012.

[15] Stefan B Edlund, Matthew A Davis, and James H Kaufman. The spatiotemporal epidemiological modeler. In *Proceedings of the 1st ACM International Health Informatics Symposium*, pages 817–820. ACM, 2010.

[16] Joshua M. M. Epstein. *Generative Social Science: Studies in Agent-Based Computational Modeling (Princeton Studies in Complexity)*. Princeton University Press, 2007.

[17] Stephen Eubank, Hasan Guclu, VS Anil Kumar, Madhav V Marathe, Aravind Srinivasan, Zoltan Toroczkai, and Nan Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180–184, 2004.

[18] Kai Fan, Marisa Eisenberg, Alison Walsh, Allison Aiello, and Katherine Heller. Hierarchical graph-coupled hmms for heterogeneous personalized health data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 239–248. ACM, 2015.

[19] Jay Wright Forrester and Jay W Forrester. *Urban dynamics*, volume 114. mIt press Cambridge, 1969.

[20] Zoubin Ghahramani and Michael I Jordan. Factorial hidden markov models. *Machine learning*, 29(2-3):245–273, 1997.

[21] Nigel Gilbert. *Agent-Based Models (Quantitative Applications in Social Sciences)*. Sage Publications, Inc., 2007.

[22] Daniel T Gillespie. Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.*, 58:35–55, 2007.

[23] Winfried Karl Grassmann. Transient solutions in markovian queueing systems. *Computers & Operations Research*, 4:47–53, 1977.

[24] Tom Heskes and Onno Zoeter. Expectation propogation for approximate inference in dynamic bayesian networks. In Adnan Darwiche and Nir Friedman, editors, *UAI*, pages 216–223. Morgan Kaufmann, 2002.

[25] Eric J Horvitz, Johnson Apacible, Raman Sarin, and Lin Liao. Prediction, expectation, and surprise: Methods, designs, and study of a deployed traffic forecasting service. *arXiv preprint arXiv:1207.1352*, 2012.

[26] Lars Hufnagel, Dirk Brockmann, and Theo Geisel. Forecast and control of epidemics in a globalized world. *Proceedings of the National Academy of Sciences of the United States of America*, 101(42):15124–15129, 2004.

[27] Lorenzo Isella, Juliette Stehlé, Alain Barrat, Ciro Cattuto, Jean-François Pinton, and Wouter Van den Broeck. What's in a crowd? analysis of face-to-face behavioral networks. *Journal of theoretical biology*, 271(1):166–180, 2011.

[28] Michael I Jordan, Zoubin Ghahramani, and Lawrence K Saul. Hidden markov decision trees. In *Proc. of NIPS*, pages 501–507, 1996.

[29] David Kotz, Tristan Henderson, Ilya Abyzov, and Jihwang Yeo. CRAWDAD data set dartmouth/campus (v. 2007-02-08). Downloaded from http://crawdad.org/dartmouth/campus/, 2007.

[30] Kalev Leetaru and Philip A Schrodt. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA Annual Convention*, volume 2, page 4, 2013.

[31] Thomas Liggett. *Interacting particle systems*, volume 276. Springer Science & Business Media, 2012.

[32] MATSim development team (ed.). MATSIM-T: Aims, approach and implementation. Technical report, IVT, ETH Zürich, Zürich, 2007.

[33] Thomas P. Minka. Expectation propagation for approximate bayesian inference. In Jack S. Breese and

Daphne Koller, editors, *UAI*, pages 362–369. Morgan Kaufmann, 2001.

[34] Kevin P Murphy, Yair Weiss, and Michael I Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 467–475. Morgan Kaufmann Publishers Inc., 1999.

[35] Wei Pan, Wen Dong, Manuel Cebrian, Taemie Kim, James H Fowler, and Alex Sandy Pentland. Modeling dynamical influence in human interaction: Using data to make better inferences about influence within social systems. *Signal Processing Magazine, IEEE*, 29(2):77–86, 2012.

[36] Michal Piorkowski, Natasa Sarafijanovic-Djukic, and Matthias Grossglauser. CRAWDAD data set epfl/mobility (v. 2009-02-24). Downloaded from http://crawdad.org/epfl/mobility/, February 2009.

[37] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[38] Neeti Pokhriyal, Wen Dong, and Venu Govindaraju. Virtual networks and poverty analysis in senegal. *arXiv preprint arXiv:1506.03401*, 2015.

[39] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[40] Ralph Tyrell Rockafellar. *Convex analysis*. Princeton university press, 2015.

[41] 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.

[42] Laron Smith, Richard Beckman, and Keith Baggerly. Transims: Transportation analysis and simulation system. Technical report, Los Alamos National Lab., NM (United States), 1995.

[43] Eleni I Vlahogianni, Matthew G Karlaftis, and John C Golias. Short-term traffic forecasting: Where we are and where we're going. *Transportation Research Part C: Emerging Technologies*, 43:3–19, 2014.

[44] Paul Waddell. Urbansim: Modeling urban development for land use, transportation, and environmental planning. *Journal of the American Planning Association*, 68(3):297–314, 2002.

[45] Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

[46] Yibing Wang, Markos Papageorgiou, and Albert Messmer. Real-time freeway traffic state estimation based on extended kalman filter: Adaptive capabilities and real data testing. *Transportation Research Part A: Policy and Practice*, 42(10):1340–1358, 2008.

[47] Darren James. Wilkinson. *Stochastic modelling for systems biology*. Taylor & Francis, Boca Raton, FL, 2006.

[48] Hongbin Yin, S_C Wong, Jianmin Xu, and CK Wong. Urban traffic flow prediction using a fuzzy-neural approach. *Transportation Research Part C: Emerging Technologies*, 10(2):85–98, 2002.