Continuous-Time Belief Propagation

Tal El-Hay[1] TALE@CS.HUJI.AC.IL
Ido Cohn[1] IDO_COHN@CS.HUJI.AC.IL
Nir Friedman[1] NIR@CS.HUJI.AC.IL
Raz Kupferman[2] RAZ@MATH.HUJI.AC.IL
[1]School of Computer Science and Engineering, [2]Institute of Mathematics, Hebrew University, Jerusalem 91904, Israel

Abstract

Many temporal processes can be naturally modeled as a stochastic system that evolves continuously over time. The representation language of *continuous-time Bayesian networks* allows to succinctly describe multi-component continuous-time stochastic processes. A crucial element in applications of such models is (approximate) inference. Here we introduce a variational approximation scheme, which is a natural extension of Belief Propagation for continuous-time processes. In this scheme, we view messages as inhomogeneous Markov processes over individual components. This leads to a relatively simple procedure that allows to easily incorporate adaptive ordinary differential equation (ODE) solvers to perform individual steps. We provide the theoretical foundations for the approximation, and show how it performs on a range of networks. Our results demonstrate that our method is quite accurate on singly connected networks, and provides close approximations in more complex ones.

1. Introduction

The dynamics of many real-life processes are naturally modeled in terms of continuous-time stochastic processes, allowing for a wide range of time scales within the same process. Examples include biological sequence evolution (Felsenstein, 2004), computer systems (Xu & Shelton, 2008; Simma et al., 2008), and social networks (Fan & Shelton, 2009).

While the mathematical foundations of continuous-time stochastic processes are well understood (Chung, 1960), the study of efficient computer representations, inference, and learning of complex continuous-time processes is still in early stages. *Continuous-time Bayesian networks* (CTBNs) (Nodelman et al., 2002) provide a sparse representation of complex multi-component processes by describing how the dynamics of an individual component depends on the state of its neighbors. A major challenge is translating the structure of a CTBN to computational gains in inference problems—answering queries about the process from partial observations.

As exact inference in a CTBN is exponential in the number of components, we have to resort to approximations. Broadly speaking, these fall into two main categories. The first category is stochastic approximations (Fan & Shelton, 2008; El-Hay et al., 2008), which sample trajectories of the process. While these can be asymptotically exact, they can be computationally expensive and incur computational penalties when sampling rapidly evolving processes. The second category of approximations is variational methods. Nodelman et al. (2005) and Saria et al. (2007) developed an approach based on *expectation propagation* (Minka, 2001; Heskes & Zoeter, 2002), where the posterior distribution over a process is approximated by *piecewise homogeneous* factored processes. This involves an elaborate message passing scheme between the approximations for different components, and an adaptive procedure for determining how to segment each time interval. More recently, Cohn et al. (2009) introduced a mean-field approximation (Jordan et al., 1998), which uses factored *inhomogeneous* processes (Opper & Sanguinetti, 2007). This allowed them to build on the rich literature of adaptive ODE solvers. While the mean-field approximation provides a lower-bound on the likelihood, it suffers from the expected drawbacks when approximating highly coupled processes.

Here we introduce a variational approximation that combines insights from both previous approaches for variational inference in CTBNs. Our approximation is a natural extension of the successful Bethe approximation (Yedidia et al., 2005) to CTBNs. Alternatively, it can be viewed

Appearing in *Proceedings of the 27^{th} International Conference on Machine Learning*, Haifa, Israel, 2010.

applying the approach of Nodelman et al where the segment length diminishes to zero. Our approximation finds a collection of inhomogeneous processes over subsets of components, which are constrained to be locally consistent over single components. We show that this approximation is often accurate on tree-networks, and provides good approximations for more complex networks. Importantly, the approximation scheme is simple and allows to easily exploit the large suites of computational tools offered in the field of ODEs.

2. Continuous-Time Bayesian Networks

Consider a d-component Markov process $\boldsymbol{X}^{(t)} = (X_1^{(t)}, X_2^{(t)}, \dots X_d^{(t)})$ with state space $S = S_1 \times S_2 \times \cdots \times S_d$. A notational convention: vectors are denoted by boldface symbols, e.g., $\boldsymbol{X}$, and matrices are denoted by blackboard style characters, e.g., $\mathbb{Q}$. The states in S are denoted by vectors of indexes, $\boldsymbol{x} = (x_1, \dots, x_d)$. We use indexes $1 \le i, j \le d$ for enumerating components and $\boldsymbol{X}^{(t)}$ and $X_i^{(t)}$ to denote the random variable describing the state of the process and its i'th component at time t.

The dynamics of a *time-homogeneous continuous-time Markov process* are fully determined by the *Markov transition function*,

$$p_{\boldsymbol{x},\boldsymbol{y}}(t) = \Pr(\boldsymbol{X}^{(t+s)} = \boldsymbol{y} | \boldsymbol{X}^{(s)} = \boldsymbol{x}),$$

where time-homogeneity implies that the right-hand side does not depend on s. These dynamics are captured by a matrix $\mathbb{Q}$—the *rate matrix*, with non-negative off-diagonal entries $q_{\boldsymbol{x},\boldsymbol{y}}$ and diagonal entries $q_{\boldsymbol{x},\boldsymbol{x}} = -\sum_{\boldsymbol{y} \neq \boldsymbol{x}} q_{\boldsymbol{x},\boldsymbol{y}}$. The rate matrix is related to the transition function by

$$\frac{d}{dt} p_{\boldsymbol{x},\boldsymbol{y}}(t)\Big|_{t=0} = q_{\boldsymbol{x},\boldsymbol{y}}.$$

The probability of being in state $\boldsymbol{x}$ at time t satisfies the *master equation* (Chung, 1960)

$$\frac{d}{dt} Pr(\boldsymbol{X}^{(t)} = \boldsymbol{x}) = \sum_{\boldsymbol{y}} q_{\boldsymbol{y},\boldsymbol{x}} Pr(\boldsymbol{X}^{(t)} = \boldsymbol{y}).$$

A *continuous-time Bayesian network* is a structured multi-component continuous-time Markov process. It is defined by assigning each component i a set of components $\mathbf{Pa}_i \subseteq \{1, \dots, d\} \setminus \{i\}$, which are its parents in the network (Nodelman et al., 2002). With each component i we then associate a family of rate matrices $\mathbb{Q}^{i|\mathbf{Pa}_i}_{\cdot|\boldsymbol{u}_i}$, with entries $q^{i|\mathbf{Pa}_i}_{x_i,y_i|\boldsymbol{u}_i}$, that describe the rates of change of the i'th component given the state $\boldsymbol{u}_i$ of the parents $\mathbf{Pa}_i$. The dynamics of $\boldsymbol{X}^{(t)}$ are defined by a rate matrix $\mathbb{Q}$ with entries $q_{\boldsymbol{x},\boldsymbol{y}}$ that combines the conditional rate matrices as follows:

$$q_{\boldsymbol{x},\boldsymbol{y}} = \begin{cases} q^{i|\mathbf{Pa}_i}_{x_i,y_i|\boldsymbol{u}_i} & \delta_{\boldsymbol{x},\boldsymbol{y}} = \{i\} \\ \sum_i q^{i|\mathbf{Pa}_i}_{x_i,x_i|\boldsymbol{u}_i} & \boldsymbol{x} = \boldsymbol{y} \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where $\delta_{\boldsymbol{x},\boldsymbol{y}} = \{i | x_i \neq y_i\}$. This definition implies that changes occur one component at a time.

Given a continuous-time Bayesian network, we would like to evaluate the likelihood of evidence, to compute the probability of various events given the evidence (e.g., that the state of the system at time t is $\boldsymbol{x}$), and to compute conditional expectations (e.g., the expected amount of time X_i was in state x_i). Direct computations of these quantities involve matrix exponentials of the rate matrix $\mathbb{Q}$, whose size is exponential in the number of components, making this approach infeasible beyond a modest number of components. We therefore have to resort to approximations.

3. A Variational Principle

Variational inference methods pose the inference task in terms of an optimization problem. The objective is to maximize a functional which lower-bounds the log probability of the evidence by introducing an auxiliary set of *variational parameters* (Wainwright & Jordan, 2008). Recently, Cohn et al. (2009) introduced a variational formulation of inference in continuous-time Markov processes. We start by reviewing the relevant results of Cohn et al.

For convenience we restrict our treatment to an interval $[0, T]$ with boundary evidence $\boldsymbol{X}^{(0)} = \boldsymbol{e}_0$ and $\boldsymbol{X}^{(T)} = \boldsymbol{e}_T$. The posterior distribution of a homogeneous Markov process given evidence $\boldsymbol{e} = \{\boldsymbol{e}_0, \boldsymbol{e}_T\}$ on the two boundaries is a *non-homogeneous Markov process*. Such a process can be represented using a *time varying rate matrix* $\mathbb{Q}(t)$ that describe the instantaneous transition rates. However, such a representation is unwieldy, since as t approaches T the transition rates from $\boldsymbol{x} \neq \boldsymbol{e}_T$ to $\boldsymbol{e}_T$ approach infinity.

To circumvent the problem of unbounded values near the boundaries, Cohn et al introduce *marginal density sets* which represent the posterior process in terms of univariate and joint pairwise distributions. More formally, if Pr denotes the posterior distribution, its *marginal density set* is the following family of continuous functions:

$$\begin{aligned} \mu_{\boldsymbol{x}}(t) &= \Pr(\boldsymbol{X}^{(t)} = \boldsymbol{x}) \\ \gamma_{\boldsymbol{x},\boldsymbol{y}}(t) &= \lim_{h \downarrow 0} \frac{\Pr(\boldsymbol{X}^{(t)} = \boldsymbol{x}, \boldsymbol{X}^{(t+h)} = \boldsymbol{y})}{h}, \quad \boldsymbol{x} \neq \boldsymbol{y}. \end{aligned} \tag{2}$$

In addition to providing a bounded representation to the posterior, this representation allows to easily compute ex-

pected sufficient statistics using numerical integration:

$$\mathbf{E}\left[T_{\boldsymbol{x}}(t)\right] = \int_0^t \mu_{\boldsymbol{x}}(s)ds, \quad \mathbf{E}\left[M_{\boldsymbol{x},\boldsymbol{y}}(t)\right] = \int_0^t \gamma_{\boldsymbol{x},\boldsymbol{y}}(s)ds,$$

where $T_{\boldsymbol{x}}(t)$ is the residence time in state $\boldsymbol{x}$ in the interval $[0,t]$, and $M_{\boldsymbol{x},\boldsymbol{y}}(t)$ is the number of transitions from $\boldsymbol{x}$ to $\boldsymbol{y}$ in the same interval. Thus, this representation is analogous to sets of *mean parameters* that are employed in variational approximations over exponential families with a finite dimensional parametrization (Wainwright & Jordan, 2008; Koller & Friedman, 2009).

Families of functions μ, γ that satisfy (2) for some Pr, must satisfy self-consistent relations imposed by the master equation.

Definition 3.1 : (Cohn et al., 2009) A family $\eta = \{\mu_{\boldsymbol{x}}(t), \gamma_{\boldsymbol{x},\boldsymbol{y}}(t) : 0 \le t \le T\}$ of continuous functions is a *Markov-consistent density set* if the following constraints are fulfilled:

$$\begin{aligned} \mu_{\boldsymbol{x}}(t) &\geq 0, \quad \sum_{\boldsymbol{x}} \mu_{\boldsymbol{x}}(0) = 1, \\ \gamma_{\boldsymbol{x},\boldsymbol{y}}(t) &\geq 0 \qquad \forall \boldsymbol{y} \neq \boldsymbol{x}, \\ \frac{d}{dt}\mu_{\boldsymbol{x}}(t) &= \sum_{\boldsymbol{y}\neq\boldsymbol{x}} \left(\gamma_{\boldsymbol{y},\boldsymbol{x}}(t) - \gamma_{\boldsymbol{x},\boldsymbol{y}}(t)\right). \end{aligned}$$

and $\gamma_{\boldsymbol{x},\boldsymbol{y}}(t) = 0$ whenever $\mu_{\boldsymbol{x}}(t) = 0$. For convenience, we define $\gamma_{\boldsymbol{x},\boldsymbol{x}} = -\sum_{\boldsymbol{y}\neq\boldsymbol{x}} \gamma_{\boldsymbol{x},\boldsymbol{y}}$. ∎

The evidence at the boundaries impose additional constraints on potential posterior processes. Specifically, the representation η corresponding to the posterior distribution $P_{\mathbb{Q}}(\cdot|\boldsymbol{e}_0, \boldsymbol{e}_T)$ is in the set $\mathcal{M}_{\boldsymbol{e}}$ that contains Markov-consistent density sets $\{\mu_{\boldsymbol{x}}(t), \gamma_{\boldsymbol{x},\boldsymbol{y}}(t)\}$, that satisfy $\mu_{\boldsymbol{x}}(0) = \mathbf{1}_{\boldsymbol{x}=\boldsymbol{e}_0}$, $\mu_{\boldsymbol{x}}(T) = \mathbf{1}_{\boldsymbol{x}=\boldsymbol{e}_T}$ and $\gamma_{\boldsymbol{x}\boldsymbol{y}}(T) = 0$ for all $\boldsymbol{y} \neq \boldsymbol{e}_T$. In addition, since these sets are posteriors of a CTBN, they also change one component at a time, which implies that $\gamma_{\boldsymbol{x},\boldsymbol{y}}(t) = 0$ if $|\delta_{\boldsymbol{x},\boldsymbol{y}}| > 1$.

Using this representation, the variational formulation of Cohn et al is reminiscent of similar formulations for discrete probabilistic models (Jordan et al., 1998).

Theorem 3.2: (Cohn et al., 2009) Let $\mathbb{Q}$ be a rate matrix and $\boldsymbol{e} = (\boldsymbol{e}_0, \boldsymbol{e}_T)$ be states of $\boldsymbol{X}$. Then

$$\ln P_{\mathbb{Q}}(\boldsymbol{e}_T|\boldsymbol{e}_0) = \max_{\eta\in\mathcal{M}_{\boldsymbol{e}}} \mathcal{F}(\eta;\mathbb{Q}),$$

where

$$\mathcal{F}(\eta;\mathbb{Q}) = \mathcal{E}(\eta;\mathbb{Q}) + \mathcal{H}(\eta),$$

is the *free energy functional* which is a sum of an *average energy functional*

$$\mathcal{E}(\eta;\mathbb{Q}) = \int_0^T \sum_{\boldsymbol{x}} \left[\mu_{\boldsymbol{x}}(t) q_{\boldsymbol{x},\boldsymbol{x}} + \sum_{\boldsymbol{y}\neq\boldsymbol{x}} \gamma_{\boldsymbol{x},\boldsymbol{y}}(t) \ln q_{\boldsymbol{x},\boldsymbol{y}}\right] dt,$$

and an *entropy functional*

$$\mathcal{H}(\eta) = \int_0^T \sum_{\boldsymbol{x}} \sum_{\boldsymbol{y}\neq\boldsymbol{x}} \gamma_{\boldsymbol{x},\boldsymbol{y}}(t)[1 + \ln\mu_{\boldsymbol{x}}(t) - \ln\gamma_{\boldsymbol{x},\boldsymbol{y}}(t)]dt \ .$$

To illustrate this principle, we can examine how to derive an exact inference procedure. We can find the optimum of $\mathcal{F}(\eta;\mathbb{Q})$ by introducing Lagrange multipliers that enforce the consistency constraint, and then find the stationary point of the corresponding Lagrangian. Since we are dealing with a continuous-time formula, we need to use the Euler-Lagrange method (Gelfand & Fomin, 1963). As Cohn et al. (2009) show, the maximum satisfies a system of differential equations:

$$\begin{aligned} \frac{d}{dt}\rho_{\boldsymbol{x}} &= -\sum_{\boldsymbol{y}} q_{\boldsymbol{x},\boldsymbol{y}}\rho_{\boldsymbol{y}} & \rho_{\boldsymbol{x}}(T) &= \mathbf{1}_{\boldsymbol{x}=\boldsymbol{e}_T} \\ \frac{d}{dt}\mu_{\boldsymbol{x}} &= \sum_{\boldsymbol{y}\neq\boldsymbol{x}} (\gamma_{\boldsymbol{y},\boldsymbol{x}} - \gamma_{\boldsymbol{x},\boldsymbol{y}}), & \mu_{\boldsymbol{x}}(0) &= \mathbf{1}_{\boldsymbol{x}=\boldsymbol{e}_0} \\ \gamma_{\boldsymbol{x},\boldsymbol{y}} &= \mu_{\boldsymbol{x}} q_{\boldsymbol{x},\boldsymbol{y}} \frac{\rho_{\boldsymbol{y}}}{\rho_{\boldsymbol{x}}}, & \boldsymbol{y} &\neq \boldsymbol{x}, \rho_{\boldsymbol{x}} \neq 0, \end{aligned} \tag{3}$$

where we omit the (t) argument for clarity. The auxiliary functions $\rho_{\boldsymbol{x}}(t)$ are Lagrange multipliers.

These equations have a simple intuitive solution that involves backward integration of $\rho_{\boldsymbol{x}}(t)$, as we have a boundary condition at time T and $\rho_{\boldsymbol{x}}(t)$ does not depend on $\mu_{\boldsymbol{x}}(t)$. This integration results in

$$\rho_{\boldsymbol{x}}(t) = \Pr(\boldsymbol{e}_T|\boldsymbol{X}^{(t)} = \boldsymbol{x})$$

Once we solve for $\rho_{\boldsymbol{x}}(t)$, we can forward integrate $\mu_{\boldsymbol{x}}(t)$ from the boundary conditions at 0 to get the solution for $\mu_{\boldsymbol{x}}$ and $\gamma_{\boldsymbol{x},\boldsymbol{y}}$. This analysis suggests that this system of ODEs is similar to forward-backward propagation, except that unlike classical forward propagation, here the forward propagation takes into account the backward messages to directly compute the posterior. Note that applying this exact solution to a multi-component process results in an exponential (in d) number of coupled differential equations.

4. Continuous-Time Expectation Propagation

Approximate variational inference procedures are derived by posing an optimization problem that is an approximate version of the original one. Different approximations differ in terms of whether they approximate the objectives, limit or relax the allowed set of solutions, or combine several such approaches. Here, we follow a strategy which is based on the approach of *expectation propagation*, in which the set of admissible solutions is extended to ones that are consistent only on the expectations of statistics of interest, and in addition, use an approximate functional.

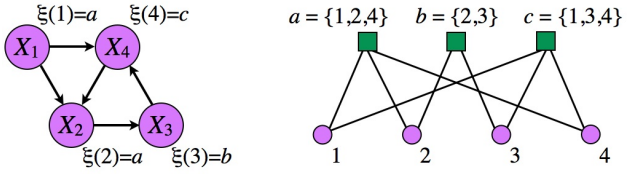

Figure 1. A CTBN and a corresponding factor graph.

4.1. Approximate Optimization Problem

To represent potential solutions, we follow methods used in recent approximate inference procedures that use factor graph representations (Yedidia et al., 2005; Koller & Friedman, 2009). Specifically, we keep only marginal density sets over smaller clusters of components.

We start with definitions and notations. A *factor graph* is an undirected bipartite graph. One layer in the graph consists of *component nodes* that are labeled by component indexes. The second layer consists of *clusters nodes* $\mathcal{A}$, where each cluster $\alpha \in \mathcal{A}$, is a subset of $\{1, \ldots, d\}$. The edges in the graph are between a component node i to a cluster node α if and only if $i \in \alpha$. Thus, the neighbors of α are $N(\alpha) = \{i : i \in \alpha\}$ and the neighbors of i are $N(i) = \{\alpha : i \in \alpha\}$.

A factor graph is *family preserving*, with respect to a given CTBN, if there exists an assignment function $\xi(i)$ that maps components to clusters, such that for every i, we have that $\{i\} \cup \mathbf{Pa}_i \subseteq \xi(i)$. We denote by $A(\alpha)$ the set of components i for which $\xi(i) = \alpha$. From now on, we assume that we deal only with family preserving factor graphs.

Example 4.1: Figure 1 shows a simple CTBN and a corresponding factor graph. In this specific factor graph, $\mathcal{A}(a) = \{1, 2\}$, $\mathcal{A}(b) = \{3\}$ and $\mathcal{A}(c) = \{4\}$. ∎

Given a factor graph, we use its structure to define an approximation for a distribution. Instead of describing the distribution over all the components, we use a family of density sets $\tilde{\boldsymbol{\eta}} = \{\eta^i : i = 1, \ldots, d\} \cup \{\eta^\alpha : \alpha \in \mathcal{A}\}$. A family of marginal density sets can be inconsistent. We do not require full consistency, but only consistency between neighboring nodes in the following sense.

Definition 4.2: A family of density sets $\tilde{\boldsymbol{\eta}}$ is said to be *locally consistent* if for all $\alpha \in \mathcal{A}$ and all $i \in N(\alpha)$ we have $\mu^i = \mu^\alpha|_i$ where

$$\left(\mu^\alpha|_i\right)_{x_i} = \sum_{\boldsymbol{x}_{\alpha\setminus i}} \mu^\alpha_{[\boldsymbol{x}_{\alpha\setminus i}, x_i]} \tag{4}$$

and $[\boldsymbol{x}_{\alpha\setminus i}, x_i]$ is the assignment to $\boldsymbol{x}_\alpha$ composed from $\boldsymbol{x}_{\alpha\setminus i}$ and x_i. Likewise, $\gamma^i = \gamma^\alpha|_i$ where

$$\left(\gamma^\alpha|_i\right)_{x_i, y_i} = \sum_{\boldsymbol{x}_{\alpha\setminus i}} \gamma^\alpha_{[\boldsymbol{x}_{\alpha\setminus i}, x_i], [\boldsymbol{x}_{\alpha\setminus i} y_i]}. \tag{5}$$

Let $\tilde{\mathcal{M}}_{\boldsymbol{e}}$ be the set of locally consistent densities that correspond to evidence $\boldsymbol{e}$ ∎

The local consistency of η^α and η^i does not imply that the distribution $\Pr_{\eta^i}(X_i)$ is equal to the marginal distribution $\Pr_{\eta^\alpha}(X_i)$, as marginalization of a Markov process is not necessarily a Markov process. Rather, $\Pr^i_\eta$ is the projection of $\Pr_{\eta^\alpha}(X_i)$ to a Markov process with the matching expectations of $\mathbf{E}\left[T_{x_i}(t)\right]$ and $\mathbf{E}\left[M_{x_i,y_i}(t)\right]$ (Koller & Friedman, 2009).

Such locally consistent sets allow us to construct a tractable approximation to the variational optimization problem by introducing the *continuous-time Bethe functional*

$$\tilde{\mathcal{F}}(\tilde{\boldsymbol{\eta}}; \mathbb{Q}) = \sum_i \mathcal{E}_i(\eta^{\alpha(i)}; \mathbb{Q}^{i|\mathbf{Pa}_i}) + \sum_\alpha \mathcal{H}(\eta^\alpha) - \sum_i c_i \mathcal{H}(\eta^i)$$

where

$$\mathcal{E}_i(\eta^\alpha; \mathbb{Q}^{i|\mathbf{Pa}_i}) = \int_0^T \sum_{\boldsymbol{x}_\alpha} \left[\mu^\alpha_{\boldsymbol{x}_\alpha}(t) q^{i|\mathbf{Pa}_i}_{x_i, x_i|\boldsymbol{u}_i} + \sum_{\boldsymbol{y} \neq \boldsymbol{x}} \gamma^\alpha_{\boldsymbol{x},\boldsymbol{y}}(t) \ln q^{i|\mathbf{Pa}_i}_{x_i, y_i|\boldsymbol{u}_i} \right] dt,$$

and $c_i = N(i) - 1$ ensure that the total weight of sets containing component i sums up to 1. This functional is analogous to the well-known Bethe approximation for discrete models (Yedidia et al., 2005).

Combining the two approximations the approximate optimization problem becomes:

$$\max_{\tilde{\boldsymbol{\eta}} \in \tilde{\mathcal{M}}_{\boldsymbol{e}}} \tilde{\mathcal{F}}(\tilde{\boldsymbol{\eta}}; \mathbb{Q}) \tag{6}$$

Once the optimal parameters are found, we can use the relevant marginal density set to answer queries.

4.2. Stationary Point Characterization

To characterize the stationary points of the approximate optimization problem (6) we use again the Euler-Lagrange method, where we introduce Lagrange multiplier functions to enforce the cluster-wise constraints, $\frac{d}{dt}\mu^\alpha_{\boldsymbol{x}_\alpha} = \sum_{\boldsymbol{y} \neq \boldsymbol{x}} (\gamma^\alpha_{\boldsymbol{y}_\alpha, \boldsymbol{x}_\alpha} - \gamma^\alpha_{\boldsymbol{x}_\alpha, \boldsymbol{y}_\alpha})$ as well as the local consistency constraints defined in equations (4) and (5). Differentiating the Lagrangian, equating the derivatives to zero, and performing some algebra, which we omit for the lack of space, we obtain fixed-point equations that consist of the initial constraints and two classes of coupled equations.

The first class consists of equations similar to (3), which refer to the dynamics within each cluster. To simplify the presentation, we introduce some definitions.

Definition 4.3: Assume we are given a time-varying matrix function $\mathbb{G}(t)$, and boundary conditions $\boldsymbol{x}_0$ and $\boldsymbol{x}_T$. Define the operator $\eta = \mathcal{R}(\mathbb{G}, \boldsymbol{x}_0, \boldsymbol{x}_T)$ to return $\eta = (\mu, \gamma)$, the unique solution of the following ODEs

$$\begin{aligned}
\frac{d}{dt}\rho_{\boldsymbol{x}} &= -\sum_{\boldsymbol{y}} g_{\boldsymbol{x},\boldsymbol{y}}\rho_{\boldsymbol{y}}, & \rho_{\boldsymbol{x}}(T) &= \mathbf{1}_{\boldsymbol{x}=\boldsymbol{x}_T} \\
\frac{d}{dt}\mu_{\boldsymbol{x}} &= \sum_{\boldsymbol{y}\neq\boldsymbol{x}} (\gamma_{\boldsymbol{y},\boldsymbol{x}} - \gamma_{\boldsymbol{x},\boldsymbol{y}}), & \mu_{\boldsymbol{x}}(0) &= \mathbf{1}_{\boldsymbol{x}=\boldsymbol{x}_0} \\
\gamma_{\boldsymbol{x},\boldsymbol{y}} &= \mu_{\boldsymbol{x}} g_{\boldsymbol{x},\boldsymbol{y}} \frac{\rho_{\boldsymbol{y}}}{\rho_{\boldsymbol{x}}}, & \rho_{\boldsymbol{x}} &\neq 0, \boldsymbol{y} \neq \boldsymbol{x}.
\end{aligned}$$

∎

Note that this set of equations is identical to (3), but replaces the constant rate matrix $\mathbb{Q}$ by a time varying matrix function $\mathbb{G}(t)$. Using this terminology, the first part of the fixed-point equations is

$$\eta^\alpha = \mathcal{R}(\mathbb{G}^\alpha, \boldsymbol{e}_0|_\alpha, \boldsymbol{e}_T|_\alpha), \tag{7}$$

where $\mathbb{G}^\alpha(t)$ is the time-dependent matrix with entries

$$g^\alpha_{\boldsymbol{x}_\alpha,\boldsymbol{y}_\alpha} = \tag{8}$$
$$\begin{cases}
(q^{i|\mathbf{Pa}_i}_{x_i y_i|\boldsymbol{u}_i})^{\mathbf{1}_{i\in A(\alpha)}} \cdot n^{i\to\alpha}_{x_i,y_i} & \delta_{\boldsymbol{x}_\alpha,\boldsymbol{y}_\alpha} = \{i\} \\
\sum_{i\in N(\alpha)} \left(\mathbf{1}_{i\in A(\alpha)} q^{i|\mathbf{Pa}_i}_{x_i x_i|\boldsymbol{u}_i} + n^{i\to\alpha}_{x_i,x_i}\right) & \boldsymbol{x}_\alpha = \boldsymbol{y}_\alpha \\
0 & \text{otherwise,}
\end{cases}$$

and $n^{i,\alpha}$ are time-dependent functions that originate from the Lagrange multipliers that enforce local consistency constraints,

$$\begin{aligned}
\prod_{\alpha\in N(i)} n^{i\to\alpha}_{x_i,y_i} &= \left(\frac{\gamma^i_{x_i y_i}}{\mu^i_{x_i}}\right)^{c_i}, \quad x_i \neq y_i \\
\sum_{\alpha\in N(i)} n^{i\to\alpha}_{x_i,x_i} &= c_i \frac{\gamma^i_{x_i x_i}}{\mu^i_{x_i}}.
\end{aligned} \tag{9}$$

These equations together with, (4) and (5) form the second set of equations that couple different clusters.

Equation (7) suggests that the matrix $\mathbb{G}^\alpha$ plays the role of a rate matrix. Unlike $\mathbb{Q}$, $\mathbb{G}^\alpha$ is not guaranteed to be a rate matrix as its rows do not necessarily sum up to zero. Nonetheless, even though it is not a rate matrix, this system of equations has a unique solution that can be found using a backward-forward integration. Thus, since the matrix function $\mathbb{G}^\alpha$ corresponds to a unique density set, we say that $\mathbb{G}^\alpha$ is an *unnormalized parametrization* of the process P_{η^α}.

At this point, it is tempting to proceed to construct a message passing algorithm based on this fixed point characterization. However, we are faced with a problem. Note that $\lim_{t\to T} \frac{\gamma_{x_i e_i}}{\mu_{x_i}} = \infty$. Therefore, according to Equation (9), when t approaches T, there exists some $\alpha \in N(i)$ for which $n^{i,\alpha}_{x_i e_{T,i}}(t)$ approaches ∞ as $t \to T$. This implies that a simple-minded message passing procedure is susceptible to unbounded values and numerical difficulties.

4.3. Gauge Transformation

To overcome these numerical difficulties, we now derive an alternative characterization, which does not suffer from unbounded values. We start with a basic result.

Proposition 4.4: *Let $\mathbb{G}$ be a unnormalized rate matrix function, and let $\omega_{\boldsymbol{x}}(t)$ be a smooth positive vector-valued function, where $\omega_{\boldsymbol{x}}(t) > 0$ in $[0, T)$. Let $\mathbb{G}^\omega$ to be the matrix function with*

$$g^\omega_{\boldsymbol{x}\boldsymbol{y}} = \begin{cases} g_{\boldsymbol{x}\boldsymbol{y}} \cdot \frac{\omega_{\boldsymbol{x}}}{\omega_{\boldsymbol{y}}} & \boldsymbol{y} \neq \boldsymbol{x} \\ g_{\boldsymbol{x}\boldsymbol{x}} - \frac{d}{dt}\log\omega_{\boldsymbol{x}} & \boldsymbol{y} = \boldsymbol{x}. \end{cases} \tag{10}$$

Then, $\mathcal{R}(\mathbb{G}, \boldsymbol{x}_0, \boldsymbol{x}_T) = \mathcal{R}(\mathbb{G}^\omega, \boldsymbol{x}_0, \boldsymbol{x}_T)$.

Proof sketch: Let ρ, η satisfy the system of equations of Def. 4.3 with $\mathbb{G}$. Define $\rho^\omega = \rho \cdot \omega$, and show that ρ^ω, η satisfy the same system of equations with $\mathbb{G}^\omega$. ∎

This result characterizes transformations of (8–9) that do not change the fixed point solutions for cluster density sets. We seek transformations that reweigh the functions $n^{i,\alpha}$ so that they remain bounded using the following result.

Proposition 4.5: *Assume $\mathbb{G}$ is a unnormalized rate matrix function such that $g_{\boldsymbol{x},\boldsymbol{y}}(t) \neq 0$ for all $\boldsymbol{x}, \boldsymbol{y}$, $g_{\boldsymbol{x},\boldsymbol{y}}(t)$ is continuously differentiable in $[0, T]$, and $\eta = \mathcal{R}(\mathbb{G}, \boldsymbol{x}_0, \boldsymbol{x}_T)$. If $\omega(t)$ is a family of smooth functions satisfying $\omega_{\boldsymbol{x}}(T) = \mathbf{1}_{\boldsymbol{x}=\boldsymbol{x}_T}$ and $\frac{d}{dt}\omega_{\boldsymbol{x}}(T) < 0$ for $\boldsymbol{x} \neq \boldsymbol{x}_T$, then*

$$\lim_{t\to T} \frac{\gamma_{\boldsymbol{x},\boldsymbol{y}}(t)}{\mu_{\boldsymbol{x}}(t)} \frac{\omega_{\boldsymbol{x}}(t)}{\omega_{\boldsymbol{y}}(t)} < \infty, \qquad \forall \boldsymbol{x} \neq \boldsymbol{y}$$

and

$$\lim_{t\to T} \left(\frac{\gamma_{\boldsymbol{x},\boldsymbol{x}}(t)}{\mu_{\boldsymbol{x}}(t)} - \frac{d}{dt}\log\omega_{\boldsymbol{x}}(t)\right) < \infty, \qquad \forall \boldsymbol{x}.$$

Example 4.6: One function that satisfies the conditions of Proposition 4.5 is $\omega_{\boldsymbol{x}}(t) = 1 - t/T, \forall \boldsymbol{x} \neq \boldsymbol{e}_T$ and $\omega_{\boldsymbol{e}_T}(t) = 1$. ∎

Using this result, we introduce weight functions $\omega^i_{x_i}$ (as above) and define $\omega^\alpha_{\boldsymbol{x}_\alpha} = \prod_{i\in N(\alpha)} (\omega^i_{x_i})^{c_i/(c_i+1)}$. Using these weight functions, we define $m^{i\to\alpha}_{x_i,y_i} = n^{i\to\alpha}_{x_i,y_i} (\frac{\omega^i_{x_i}}{\omega^i_{y_i}})^{c_i/(c_i+1)}$ and $m^{i\to\alpha}_{x_i,x_i} = n^{i\to\alpha}_{x_i,x_i} - \frac{c_i}{c_i+1}\frac{d}{dt}\log\omega^i_{x_i}$.

Algorithm 1 Continuous-Time Belief Propagation

Initialize messages: for all α and all $i \in N(\alpha)$
  Choose $\eta^\alpha \in \mathcal{M}_{\boldsymbol{e}}^\alpha$
  Compute $\eta^{\alpha\to i}$ using (14)
  Set $m_{x_i,y_i}^{i\to\alpha} = 1\ \forall x_i \neq y_i$, $m_{x_i,x_i}^{i\to\alpha} = 0$
repeat
  Choose a cluster α:
  1. $\forall i \in N(\alpha)$, set $m_{x_i,y_i}^{i\to\alpha}$ using (15)
  2. Update $\tilde{\mathbb{G}}^\alpha$ using (11)
  3. Compute η^α from $\tilde{\mathbb{G}}^\alpha$ using (12)
  4. $\forall i \in N(\alpha)$ compute $\eta^{\alpha\to i}$ using (14)
until convergence

Now if we define the time-dependent matrix $\tilde{\mathbb{G}}^\alpha$ with entries

$$\tilde{g}^\alpha_{\boldsymbol{x}_\alpha,\boldsymbol{y}_\alpha} = \begin{cases} (q^{i|\mathbf{Pa}_i}_{x_i y_i|\boldsymbol{u}_i})^{\mathbf{1}_{i\in A(\alpha)}} \cdot m^{i\to\alpha}_{x_i,y_i} & \delta_{\boldsymbol{x}_\alpha,\boldsymbol{y}_\alpha} = \{i\} \\ \sum_{i\in N(\alpha)} \left(\mathbf{1}_{i\in A(\alpha)} q^{i|\mathbf{Pa}_i}_{x_i x_i|\boldsymbol{u}_i} + m^{i\to\alpha}_{x_i,x_i} \right) & \boldsymbol{x}_\alpha = \boldsymbol{y}_\alpha \\ 0 & \text{otherwise,} \end{cases} \tag{11}$$

then $\tilde{\mathbb{G}}^\alpha = (\mathbb{G}^\alpha)^{\omega^\alpha}$. By Proposition 4.4,

$$\eta^\alpha = \mathcal{R}(\tilde{\mathbb{G}}^\alpha, \boldsymbol{e}_0|_\alpha, \boldsymbol{e}_T|_\alpha). \tag{12}$$

Plugging the definition of $m^{i\to\alpha}_{x_i,y_i}$ and $m^{i\to\alpha}_{x_i,x_i}$ into (9) we get

$$\begin{aligned} \prod_{\alpha\in N(i)} m^{i\to\alpha}_{x_i,y_i} &= \left(\frac{\gamma^i_{x_i y_i}}{\mu^i_{x_i}} \frac{\omega^i_{x_i}}{\omega^i_{y_i}} \right)^{c_i}, \quad x_i \neq y_i \\ \sum_{\alpha\in N(i)} m^{i\to\alpha}_{x_i,x_i} &= c_i \left(\frac{\gamma^i_{x_i x_i}}{\mu^i_{x_i}} - \frac{d}{dt} \log \omega^i_{x_i} \right). \end{aligned} \tag{13}$$

If the preconditions of Proposition 4.5 are satisfied, the terms in (13) are bounded. Together (11)–(13) provide an alternative characterization of the fixed point(s) of the optimization problem.

4.4. Message Passing Scheme

We now use the above characterization as justification for a message passing scheme, that if converged, will satisfy the fixed point equations. While (11) and (12) are readily transformed into assignments, (13) poses a challenge.

We start by noting that (13) contains the terms $\mu^i_{x_i}$ and $\gamma^i_{x_i,y_i}$. We can get these terms from η^α for *any* $\alpha \in N(i)$. Thus, for $\alpha \in N(i)$, we define

$$\mu^{\alpha\to i} = \mu^\alpha|_i \qquad \gamma^{\alpha\to i} = \gamma^\alpha|_i \tag{14}$$

We view these as the messages from cluster α to the component i. At convergence, $\mu^{\alpha\to i} = \mu^{\beta\to i}$ for $\alpha, \beta \in N(i)$, but this is not true before convergence.

Next, we rewrite (13) as an assignment

$$m^{i\to\alpha}_{x_i,y_i} = \begin{cases} \prod\limits_{\substack{\beta\in N(i)\\ \beta\neq\alpha}} \frac{1}{m^{i\to\beta}_{x_i,y_i}} \frac{\gamma^{\beta\to i}_{x_i y_i}}{\mu^{\beta\to i}_{x_i}} \frac{\omega^i_{x_i}}{\omega^i_{y_i}} & x_i \neq y_i \\ \sum\limits_{\substack{\beta\in N(i)\\ \beta\neq\alpha}} \left(\frac{\gamma^{\beta\to i}_{x_i x_i}}{\mu^{\beta\to i}_{x_i}} - \frac{d}{dt} \log \omega^i_{x_i} - m^{i\to\beta}_{x_i,x_i} \right) & x_i = y_i \end{cases} \tag{15}$$

where we write $\frac{\gamma^i_{x_i y_i}}{\mu^i_{x_i}} = \frac{\gamma^{\beta\to i}_{x_i y_i}}{\mu^{\beta\to i}_{x_i}}$ once for each β.

The algorithm is summarized in Algorithm 1. The implementation of these steps involve a few details. We start with the initialization of messages. The only free parameter is the initial values of η^α. To ensure that these initial choices are in $\mathcal{M}_{\boldsymbol{e}}^\alpha$, we choose initial rates, and perform computations to get a valid posterior for the clusters. Another degree of freedom is the order of cluster updates. We use a randomized strategy, choosing a cluster at random, and if one of its neighbors was updated since it was last chosen, we update it.

The computation in Step 3, involves reverse integration followed by forward integration (as explained in Section 3). We gain efficiency by using adaptive numerical integration procedures. Specifically, we use the Runge-Kutta-Fehlberg (4,5) method (Press et al., 1992). This method chooses temporal evaluation points on the fly, and returns values at these points. The computations of Step 2 is done on demand only at the evaluation points. To allow efficient interpolation, we use a piecewise linear approximation of η whose boundary points are determined by the evaluation points that are chosen by the adaptive integrator. Finally, as might be expected, we do not have convergence guarantees. However, if the algorithm converges, the fixed point equations are satisfied, hence giving a stationary point (hopefully a local maximum) of problem (6).

5. Experiments

We tested our method on three representative network structures: a directed tree, a directed toroid, and a bidirectional ring (Fig. 2). The tree network does not have any cycles. The toroid network has cycles, but these are fairly long, whereas the bidirectional ring has multiple short cycles. All networks are parametrized as *dynamic Ising models*, in which neighboring components prefer to be in the same state. Specifically, we use the conditional rates

$$q^{i|\mathbf{Pa}_i}_{x_i,y_i|\boldsymbol{u}_i} = \tau \left(1 + \exp\left(-2 y_i \beta \textstyle\sum_{j\in\mathbf{Pa}_i} x_j \right) \right)^{-1}$$

where $x_j \in \{-1, 1\}$, β is a *coupling parameter* and τ is a *rate parameter*. Small values of β correspond to weak coupling, whereas τ determines how fast components tend to switch states. For each experiment we set evidence at times 0 and 1 (Fig. 2, left panel).

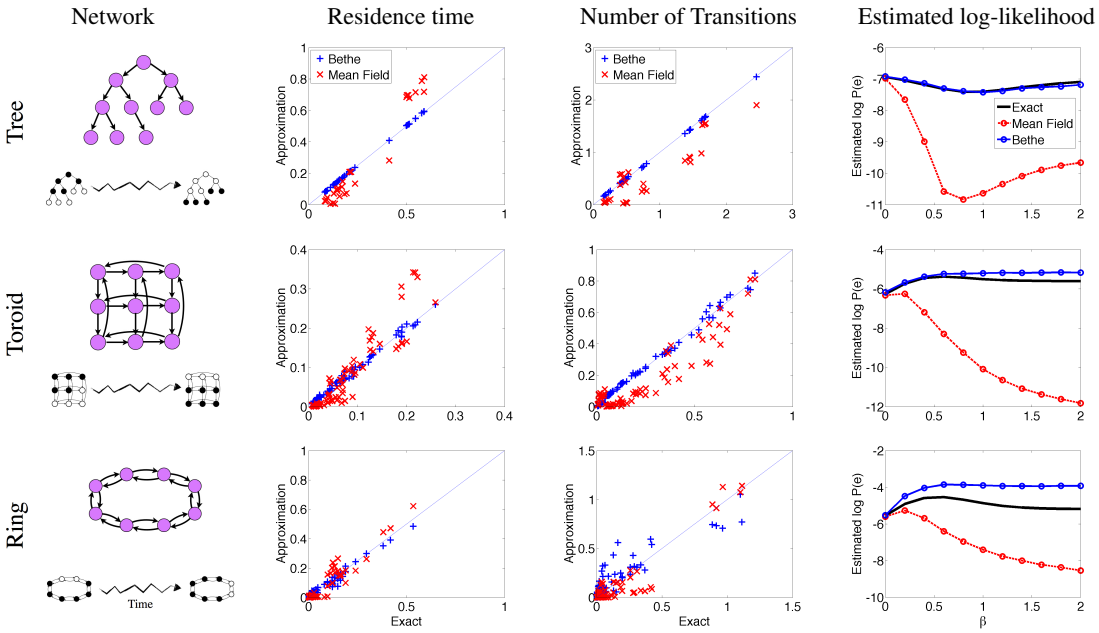

Figure 2. Simulation results for a tree network (top row), a toroid network (middle), and a bidirectional chain (bottom). **Left** network structure and the evidence at start and end points; black is $+1$ and white is -1. **Middle-left**: scatter plot of expected conditional residence times for networks with $\beta = 1$, $\tau = 8$. Each point corresponds to a single statistic, the x-axis is the exact value and the y-axis is the approximate value. **Middle-right**: same for expected conditional transition times. **Right**: exact and approximations of log-likelihood as function of β, the strength of coupling between components ($\tau = 8$).

We compare the Bethe approximation to exact inference and mean-field (Cohn et al., 2009). We start by comparing the value of sufficient statistics (residence time and number of transitions of each component for each state of its parents) computed by each method. For example, for a particular choice of β and τ, (Fig. 2 middle columns) we can see that the Bethe approximation is virtually exact on the tree model and the toroid, but has some bias on the bidirectional ring model. These scatter plots also shed light on the nature of the difference between the two methods. Specifically, in the most likely scenario, two components switch from -1 to 1 near the beginning and the other two switch from 1 to -1 near the end, and so through most of the interval all the components are in the same state. The mean-field algorithm gives a uni-modal solution, focusing on the most likely scenario, resulting in zero residence time for the less likely states. These states are represented by the points close on the x-axis. The Bethe representation on the other hand can capture multiple scenarios.

Another aspect of the approximation is the estimation of the likelihood. In Fig. 2 (right column) we compare these estimations as function of β, the problem hardness. Again, we see that the Bethe approximation is essentially exact on the tree network, and provides close approximations in the two other networks. When we push β and τ to extreme values we do see inaccuracies even in the tree network, showing that the algorithm is an approximation.

While the ODE solvers used here allow adaptive integration error control, we do not have an a-priory control on the propagation of this error. To test this effect on overall accuracy, we repeated these experiments using standard grid refinement. Specifically, we computed integrals using uniformly spaced evaluation points and systematically halving integration intervals until no changes in the output were observed. Final results of these tests were practically the same as those obtained using adaptive integration.

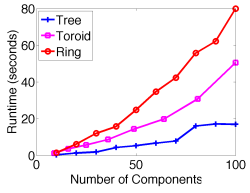

Figure 3. Run time vs. the number components in the three networks types ($\beta = 1$, $\tau = 8$).

Next, we examine how the algorithm scales with the number of components in the networks. In all three networks we see that the magnitude of relative error is essentially independent of the number of components (not shown). Fig. 3 shows that the run time scales linearly with the number of components. In harder networks the algorithm re-

quires more iterations leading to slower convergence.

6. Discussion

Here, we introduce a message passing scheme that provides a variational approximation for CTBNs. This scheme is derived from an approximate free energy functional, based on the principles of expectation-propagation, where we require the posteriors on clusters to be locally consistent in terms of the Markovian projections on individual components. We show that stationary points of this optimization problem are the fixed points of a message passing algorithm, whose structure closely mirrors Belief Propagation.

In contrast to Belief Propagation on discrete (atemporal) networks, our algorithm is *not* guaranteed to be exact on tree CTBNs. The source of inaccuracy is the projection of the marginal distributions over components onto Markov processes. While this projection looses information, our empirical results suggest that this approximation is relatively accurate. In contrast to methods in Dynamic Bayesian Networks (DBNs) (Koller & Friedman, 2009) that approximate the distribution over all components in each time slice, our approach approximates the temporal behavior of a single component over the whole time interval.

The works that are closest to ours are those of Nodelman et al. (2005) and Saria et al. (2007) which are also derived from an expectation-propagation energy functional. The main difference between the two methods is the structure of the approximation. Nodelman et al use a piecewise homogeneous representation, allowing them to represent the rate in each homogeneous segment by a constant (conditional) rate matrix. This, however, requires introducing machinery for deciding how to segment each component. As Saria et al show, this choice can have dramatic impact on the quality of the approximation and the running time. In contrast, our approach uses a (continuously) inhomogeneous representation, which is essentially the limit when segment sizes tend to zero. Surprisingly, rather than making the problem more complex, this choice simplifies the mathematics and also the implementation. In particular, our solution decouples the probabilistic issues (dependencies between components) and numerical issues (adaptive integration) and allows us to build on well-understood methods from numerical integration for efficient and adaptive selection of the number and placement of discretization points.

Our results show how a careful choice of representations and operations over them can narrow the gap between inference methods in discrete and continuous-time graphical models. Our constructions can be naturally generalized to capture more complex dependencies using methods based on Generalized Belief Propagation (Yedidia et al., 2005).

Acknowledgments

We thank the anonymous reviewers for helpful remarks. This research was supported in part by a grant from the Israel Science Foundation. Tal El-Hay is supported by the Eshkol fellowship from the Israeli Ministry of Science.

References

Chung, K.L. *Markov chains with stationary transition probabilities*. 1960.

Cohn, I., El-Hay, T., Friedman, N., and Kupferman, R. Mean field variational approximation for continuous-time Bayesian networks. UAI, 2009.

El-Hay, T., Friedman, N., and Kupferman, R. Gibbs sampling in factorized continuous-time markov processes. UAI, 2008.

Fan, Y. and Shelton, C. R. Learning continuous-time social network dynamics. UAI, 2009.

Fan, Y. and Shelton, C.R. Sampling for approximate inference in continuous time Bayesian networks. In *AI & Math*, 2008.

Felsenstein, J. *Inferring Phylogenies*. Sinauer, 2004.

Gelfand, I. M. and Fomin, S. V. *Calculus of variations*. 1963.

Heskes, T. and Zoeter, O. Expectation propagation for approximate inference in dynamic Bayesian networks. UAI, 2002.

Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., and Saul, L. K. An introduction to variational approximations methods for graphical models. In *Learning in Graphics Models*. 1998.

Koller, D. and Friedman, N. *Probabilistic Graphical Models: Principles and Techniques*. 2009.

Minka, T. P. Expectation propagation for approximate Bayesian inference. UAI, 2001.

Nodelman, U., Shelton, C. R., and Koller, D. Continuous time Bayesian networks. UAI, 2002.

Nodelman, U., Shelton, C.R., and Koller, D. Expectation propagation for continuous time Bayesian networks. UAI, 2005.

Opper, M. and Sanguinetti, G. Variational inference for Markov jump processes. NIPS, 2007.

Press, William H., Teukolsky, Saul A., Vetterling, William T., and Flannery, Brian P. *Numerical Recipes in C*. 1992.

Saria, S., Nodelman, U., and Koller, D. Reasoning at the right time granularity. UAI, 2007.

Simma, A., Goldszmidt, M., MacCormick, M., Barham, M., Black, M., Isaacs, M., and Mortier, R. CT-NOR: Representing and reasoning about events in continuous time. UAI, 2008.

Wainwright, M. J. and Jordan, M. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1:1–305, 2008.

Xu, J. and Shelton, C. R. Continuous time Bayesian networks for host level network intrusion detection. ECML/PKDD, 2008.

Yedidia, J.S., Freeman, W.T., and Weiss, Y. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Trans. Info. Theory*, 51:2282–2312, 2005.