
Recovering Temporally Rewiring Networks: A Model-based Approach

Fan Guo
Steve Hanneke
Wenjie Fu
Eric P. Xing

FANGUO@CS.CMU.EDU
SHANNEKE@CS.CMU.EDU
WENJIEF@CS.CMU.EDU
EPXING@CS.CMU.EDU

School of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213 USA

Abstract

A plausible representation of relational information among entities in dynamic systems such as a living cell or a social community is a stochastic network which is topologically rewiring and semantically evolving over time. While there is a rich literature on modeling static or temporally invariant networks, much less has been done toward modeling the dynamic processes underlying rewiring networks, and on recovering such networks when they are not observable. We present a class of *hidden temporal exponential random graph models* (htERGMs) to study the yet unexplored topic of modeling and recovering temporally rewiring networks from time series of node attributes such as activities of social actors or expression levels of genes. We show that one can reliably infer the latent time-specific topologies of the evolving networks from the observation. We report empirical results on both synthetic data and a *Drosophila* lifecycle gene expression data set, in comparison with a static counterpart of htERGM.

1. Introduction

In many problems arising in biology, social sciences and various other fields, it is often necessary to analyze populations of entities such as molecules or individuals that are interconnected by a set of relationships (e.g., physical interactions or functional regulations in biological contexts, friendships or communications in social contexts). Studying networks of these kinds can

reveal a wide range of information, such as how biological entities or individuals organize themselves into groups (Airoldi et al., 2005), which individuals are in positions of power or which molecules or genes are the key orchestrators of cellular functions (Barabási & Albert, 1999), and how patterns of biological regulations or social interactions evolve over time (Sarkar & Moore, 2006). While there is a rich literature on modeling a static network or time-invariant networks, much less has been done towards modeling the dynamic processes underlying networks that are *topologically rewiring* and *semantically evolving* over time, and on developing inference and learning techniques, especially in a dynamic context, for recovering unobserved network topologies from observed attributes of entities constituting the network. In this paper, we propose a new formalism for modeling network evolution over time on a fixed set of nodes, and an algorithm for recovering unobserved temporally rewiring networks from time series of entity attributes.

A classical model of network analysis, which will be generalized in this paper, is the *exponential random graph model* (ERGM) (Frank & Strauss, 1986; Strauss & Ikeda, 1990). According to the Hammersley-Clifford theorem, the distribution of the topology (denoted by A) of a graph can be modeled by an ERGM defined as follows:

$$P(A) = \frac{1}{Z(\boldsymbol{\theta})} \exp \left\{ \sum_c \theta_c \Phi_c(A) \right\}, \quad (1)$$

where c denotes a clique in A , Φ_c is a *potential function* defined on c which corresponds to the *network statistics* of clique c , θ_c is the parameter corresponding to clique c , and $Z(\boldsymbol{\theta})$ is a normalization constant, also known as the *partition function*. The general class of stochastic block models (Fienberg et al., 1985) is an example of dyadic ERGMs, which only take into account singleton and pairwise cliques. More general depen-

Appearing in *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, OR, 2007. Copyright 2007 by the author(s)/owner(s).

dency assumptions also lead to richer network models, such as the Markov random graphs (Wasserman & Pattison, 1996; Robins et al., 2007). These statistical formalisms, along with recent developments in related areas such as latent space models (Hoff et al., 2002), Bayesian networks (Heckerman, 1995) and graph mining techniques (Hu et al., 2005), have provided rich methodological foundation for a wide range of analysis of social, biological, or other kinds of networks.

While these advances have offered important insight and tools for analyzing network data, a limitation of most of current methods is that they usually assume that networks are topologically static and, in many cases, fully observable. In many complex dynamic systems, such as a developing biological organism, the interactions between network entities such as genes and regulatory proteins depend on various external and internal conditions. Therefore the underlying network of these entities may exhibit large topological changes over time (Luscombe et al., 2004). In most of such circumstances, it is technically impossible to experimentally determine time-specific network topologies for a series of time points. Resorting to standard computational inference methods such as structural learning algorithms for Bayesian networks is also difficult because we can only obtain a single snapshot of the node states at each time point. Indeed, if we follow the naïve assumption that each snapshot is independently distributed, this task would be statistically impossible because our estimator (from only one sample) would suffer from extremely high variance. For this reason, to our knowledge, in current network inference literature, samples from all time points are often pooled together to infer a single time-invariant network, e.g. (Friedman et al., 2000), which means they choose to ignore network rewiring and simply assume that the network snapshots are independently and identically distributed¹. There is a recent paper (Lawrence et al., 2007) using Gaussian processes to model the dynamics of a very simple transcriptional regulatory network motif, which is a small biological subnetwork. The approach they took focused on the continuous-time dynamics of node attributes rather than the dynamics of interactions studied in this paper.

In this paper, we describe a new framework for statistical modeling of the evolution of networks based on time series of node attributes such as levels of gene expressions, or activities of social actors. We study the

¹It is worth noting that the *dynamic Bayesian network* often used for modeling time series data is also a *topologically static* network model because it employs a time-invariant graph over nodes of every pair of adjacent time-points to define the distribution of node attributes.

yet unexplored topic of recovering unobserved temporally rewiring networks from samples of node states based on this new framework. We concern ourselves with a dynamic network over a fixed set of nodes, and posit that the network topology evolves over time according to a Markov process characterized by global topological features such as density, and local features such as transitivity. Such a process is modeled by a hidden *temporal exponential random graph model* (tERGM). Conditioned on the topology of latent networks at each time point, the observations are generated from an emission model that can be tailored for various data characteristics. For concreteness, the proposed model is illustrated in the context of recovering gene-coexpression networks scenario, but our approach is generalizable to a broad range of contexts.

The rest of the paper is organized as follows. In Sec. 2 we described a temporal exponential random graph model (tERGM) for longitudinal network analysis to lay the ground work of the rest of the paper². In Sec. 3 we extend tEGRMs to hidden tEGRMs for modeling time series of observed node-attributes. We develop a Gibbs sampling algorithm for posterior inference of the latent network topologies in Sec. 4. And we present in Sec. 5 small scale empirical results on both synthetic data and a *Drosophila* lifecycle gene expression data set (Arbeitman et al., 2002), in comparison with methods based on a single time-invariant network assumption. We conclude with a discussion on the difficulty of the network inference problem studied here and the limitations of our inference algorithms.

2. Temporal Exponential Random Graph Model

As mentioned earlier, the exponential random graph models (ERGMs) provide a general framework for descriptively modeling a static network. However, in their present form, ERGMs are unable to describe the behavior of a network evolving over time. Here we extended ERGMs to *temporal* ERGMs for modeling networks evolving over discrete timesteps. We begin with tERGMs over an observed sequence of networks. In the next section we will extend tERGMs to *hidden* tERGMs for modeling a sequence of node attribute observations.

Let \mathbf{A}^t represent the network topology observed at time t , for $t \in \{1, 2, \dots, T\}$ (the upper chain in Fig. 1). We make a Markov assumption over time, and specify a probability distribution over this sequence of T

²This model was initially announced in a paper presented at an ICML'06 workshop (Hanneke & Xing, 2006).

networks as:

$$P(\mathbf{A}^1, \mathbf{A}^2, \dots, \mathbf{A}^T) = P(\mathbf{A}^1) \prod_{t=2}^T P(\mathbf{A}^t | \mathbf{A}^{t-1}), \quad (2)$$

where

$$P(\mathbf{A}^t | \mathbf{A}^{t-1}) = \frac{1}{Z(\boldsymbol{\theta}, \mathbf{A}^{t-1})} \exp \{ \boldsymbol{\theta}' \boldsymbol{\Psi}(\mathbf{A}^t, \mathbf{A}^{t-1}) \}, \quad (3)$$

and $P(\mathbf{A}^1)$ is the prior distribution for the network at the initial time point. Here $\boldsymbol{\theta}$ is a parameter vector and $\boldsymbol{\Psi}$ is a vector-valued function. Note that the temporal evolution model, $P(\mathbf{A}^t | \mathbf{A}^{t-1})$, also takes the form of an ERGM conditioned on \mathbf{A}^{t-1} , and the potential functions or *features*, $\boldsymbol{\Psi}$, for specifying this ERGM can be designed to capture various dynamic properties governing the network rewiring over time. Appropriate choices of these features can provide an expressive framework to model the network rewiring in adjacent timesteps at a modest parameterization cost (Hanneke & Xing, 2006). Unlike Bayesian networks over nodes, which would usually employ numerous local conditional distributions, one over each node and its parents, and hence requires a large number of parameters, in tERGMs, features can typically be defined as functions of a small number of simple summary statistics over pairs of attendant networks. For example, one can explore simple features of tERGMs by characterizing a distribution in terms of “density”, “stability”, and “transitivity” features:

$$\begin{aligned} \Psi_1(\mathbf{A}^t, \mathbf{A}^{t-1}) &= \sum_{ij} A_{ij}^t, \\ \Psi_2(\mathbf{A}^t, \mathbf{A}^{t-1}) &= \sum_{ij} \mathbb{I}(A_{ij}^t = A_{ij}^{t-1}), \\ \Psi_3(\mathbf{A}^t, \mathbf{A}^{t-1}) &= \frac{\sum_{ijk} A_{ij}^t A_{ik}^{t-1} A_{kj}^{t-1}}{\sum_{ijk} A_{ik}^{t-1} A_{kj}^{t-1}}, \end{aligned}$$

where $\mathbb{I}(\cdot)$ is the indicator function. More specifically, Ψ_1 captures the *density* of the network at current timestep, which measures the relative number of interactions in the network; Ψ_2 captures network *stability* over time, which measures the tendency of a network to stay the same from one timestep to the next at the edge level; Ψ_3 represents a temporal notion of *transitivity*, which measures the tendency of node i having an interaction with node k which has an interaction with node j at timestep $t-1$ to result in node i having an interaction with node j at timestep t . The natural parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ can be used to adjust the strength of each of these features and thereby network evolution dynamics. The degree of randomness is implicitly determined by the magnitude of $\boldsymbol{\theta}$.

Finally, to complete any tERGM, we need to specify a prior distribution for the initial network \mathbf{A}^1 . An informed prior distribution could be more helpful than a uniform one and speed up the convergence of the

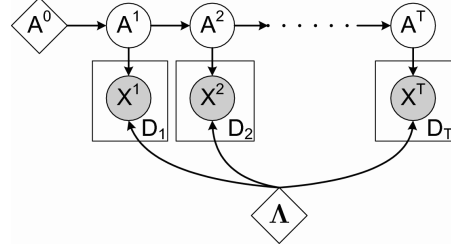


Figure 1. The graphical structure of a htERGM. $\mathbf{A}^{1:T}$ represent latent network topologies over T timesteps. Shaded nodes $\mathbf{x}^{1:T}$ represent observed node attributes. At timestep t there are D_t iid observations. \mathbf{A}^0 and Λ are fixed and estimated *a priori*. They are related to the prior on the initial network topology and global activation function respectively.

inference algorithm. We let the prior distribution have the same ERGM representation as Eq. 3 conditioned on some fixed network \mathbf{A}^0 to obtain $P(\mathbf{A}^1 | \mathbf{A}^0)$. \mathbf{A}^0 can be estimated from the static counterpart of the proposed model, abbreviated as sERGM.

3. Hidden Temporal Exponential Random Graph Model

In the previous section, we assume that the evolving sequence of networks is available to fit a tERGM. But where do we obtain such a sequence during a dynamic biological/socialological process? Now we describe an approach that systematically explores the possible dependencies of an unobserved rewiring network underlying biological/socialological temporal processes, and leads to the inference algorithm that can reconstruct temporally rewiring networks from a sequence of node-attribute snapshots.

Our proposed statistical model would allow a time-specific network topology of each timestep to be inferred based on node-attribute samples measured over the entire time series. The intuition underlying this scheme is as follows. Although the whole network is not likely to be repeated over time, its building blocks, such as motifs and subgraphs do recur over time, as suggested by a number of empirical studies (Hu et al., 2005). This suggests that a network can be assembled from small subgraphs, and each subgraph can be estimated from relevant node attributes observed at those time points *at which the subgraph is present*. In the simplest scenario, we can tie the sequence of unobserved rewiring networks, from which node-attribute sequences are sampled, by a latent Markov chain, to facilitate the aforementioned information sharing across time. The tERGM described in Sec. 2 provides a useful building block to formulate this model. Also, as a starting point, we only consider recurring subgraphs at

the *dyadic level* – recurring edges, and design pairwise features in the emission model detailed in Sec. 3.1.

Let $\mathbf{x}^t = \{x_{1:N}^t\}$ denote observed attributes over all N nodes of a network \mathbf{A}^t at timestep t . And let \mathbf{A} denote a time-invariant global “activation function” that specifies distributions of node states (e.g., discrete gene expression levels) under specific pairwise node interactions (e.g., gene a activates or suppresses gene b and vice versa). Our model takes the following form, which we would refer to as a hidden temporal ERGM (htERGM):

$$P(\mathbf{A}^{1:T}, \mathbf{x}^{1:T} | \mathbf{A}^0, \mathbf{A}) = \prod_{t=1}^T P(\mathbf{A}^t | \mathbf{A}^{t-1}) P(\mathbf{x}^t | \mathbf{A}^t, \mathbf{A}). \quad (4)$$

Figure 1 gives a graphical illustration of a htERGM. Note that this graphical model is conceptually similar to a hidden Markov model, however, in our model each node represents a *mega* variable of a set of all possible pairwise interactions or a full node-attribute profile, rather than a simple random vector or variable. The activation function parameterized by \mathbf{A} represents information beyond the topology of each timestep-specific network, which is assumed to be invariant over all timesteps. We made this set of time-independent parameters explicit in Fig. 1 to provide a fuller picture of the emission model. Depending on the specific modeling assumptions and data characteristics, one can explore various instantiations of htERGMs, resulting from different choices of the transition model $P(\mathbf{A}^t | \mathbf{A}^{t-1})$ and the emission model $P(\mathbf{x}^t | \mathbf{A}^t, \mathbf{A})$.

Also, there is a more general case of temporal network learning than we discussed above, where we concern ourselves with “epoch-specific” networks. Each epoch t may consist of one or multiple timesteps. The model could accommodate non-uniform observation intervals and network evolution rate by changing the granularity of an epoch and allowing epoches to be of different lengths. To simplify the discussion in the sequel, we do not explicitly differentiate between an epoch and a timestep, but allow multiple observations to be generated *iid* from the same network topology as illustrated in Fig. 1. Also, the static version of the model could be interpreted as a special case with $T = 1$.

3.1. Energy-based Emission Model for Gene Expression Data

For gene network modeling, a popular approach of modeling discrete gene expression levels given the network topology is to adapt a Bayesian network (BN) formalism. In this BN framework, the local conditional probability tables (CPTs) need to be specified besides the network topology to generate expression data. The

BN approach could model directional regulation interactions and computation is usually tractable as a result of factorization of the joint probability distribution. However, it cannot be simply integrated into our model because the configurations of local CPTs depend on the network topology. When the network topology is evolving, both the size of and the variables involved in the CPTs are different from different timesteps. So it is difficult to allow the timestep-invariant information to be incorporated into a BN-based emission model.

An energy-based model, on the other hand, places soft constraints on the variables usually simplifies the parameterization via summary statistics. It is difficult to find an intuitive interpretation on how to draw samples from an energy-based model because the probability distribution is generally not factorized. Our design follows an exponential family distribution as in the transition model and pairwise features can be defined to reflect our assumption of recurring edges:

$$P(\mathbf{x}^t | \mathbf{A}^t) = \frac{1}{Z(\boldsymbol{\eta}, \mathbf{A}^t)} \exp \left\{ \boldsymbol{\eta}' \sum_{ij} \Phi(x_i^t, x_j^t, A_{ij}^t, \Lambda_{ij}) \right\}, \quad (5)$$

where Λ_{ij} is the global activation for expression levels of gene i and gene j , which ranges in $[-1, 1]$. $\Lambda_{ij} = 1, -1$ indicate a perfect mutual activation (positive correlation) or suppression (negative correlation).

We propose only one feature for the emission model:

$$\Phi(x_i^t, x_j^t, A_{ij}, \Lambda_{ij}) = A_{ij}^t \Lambda_{ij} (2\mathbb{I}(x_i^t = x_j^t) - 1). \quad (6)$$

When $A_{ij}^t = 0$, i.e. at timestep t there is no interaction between gene i and gene j , then whatever their expression levels are, the feature $\Phi = 0$. When $A_{ij}^t = 1$, if the values x_i^t and x_j^t agree with the direction of the correlation specified by Λ_{ij} , $\Phi > 0$; otherwise $\Phi < 0$. The absolute value of Φ depends the absolute value of Λ_{ij} which indicates the strength of the activation/depression. The parameter $\eta > 0$ reflects the level of randomness in the emission model.

4. Inference and Learning

4.1. The Inference Algorithm

The posterior distribution $P(\mathbf{A}^{1:T} | \mathbf{x}^{1:T})$ in a htERGM is intractable because there is no conjugate relationship between the local probabilistic distributions. Moreover, the derivation of an exact MCMC sampling scheme is non-trivial and involves evaluation of the ratios of two partition functions. This is a doubly-intractable problem. Similar problems also arise in Bayesian learning of undirected graphical models, and approximate MCMC algorithms have been presented in (Murray & Ghahramani, 2004) and (Murray et al.,

2006). However, neither of them can be simply applied to our setting because in practice we can only make at most a few observations per timestep (so the approximation in (Murray & Ghahramani, 2004) is subject to high variance), and we also have a high-dimensional space of latent variables (so the MCMC sampler in (Murray et al., 2006) would have unacceptably low acceptance rates). In this section, we apply the Gibbs sampling algorithm and derive the sampling formulae for hidden variables of binary interactions plausible for relatively small networks.

In order to obtain the sampling formula for updating a binary hidden variable A_{ij}^t ,

$$A_{ij}^t \sim \text{Bernoulli}(1/(1 + \exp(-\mu_{ij}^t))), \quad (7)$$

we compute the log-odds:

$$\mu_{ij}^t = \log \frac{P(A_{ij}^t = 1 | \mathbf{A}^{t-1}, \mathbf{A}_{-ij}^t, \mathbf{A}^{t+1}, \mathbf{x}^t)}{P(A_{ij}^t = 0 | \mathbf{A}^{t-1}, \mathbf{A}_{-ij}^t, \mathbf{A}^{t+1}, \mathbf{x}^t)} \quad (8)$$

where we use \mathbf{A}_{-ij}^t to denote the set of variables $\{A_{kl}^t | kl \neq ij\}$. To simplify the notation, we define $\mathbf{A}_1^t[ij] \equiv (A_{ij}^t = 1, \mathbf{A}_{-ij}^t)$ and $\mathbf{A}_0^t[ij] \equiv (A_{ij}^t = 0, \mathbf{A}_{-ij}^t)$, then

$$\begin{aligned} \mu_{ij}^t = \log \frac{P(\mathbf{A}_1^t[ij] | \mathbf{A}^{t-1})}{P(\mathbf{A}_0^t[ij] | \mathbf{A}^{t-1})} + \log \frac{P(\mathbf{A}^{t+1} | \mathbf{A}_1^t[ij])}{P(\mathbf{A}^{t+1} | \mathbf{A}_0^t[ij])} \\ + \log \frac{P(\mathbf{x}^t | \mathbf{A}_1^t[ij])}{P(\mathbf{x}^t | \mathbf{A}_0^t[ij])} \end{aligned} \quad (9)$$

The first two terms on the right side are related to the local probability distributions in the transition model. Evaluation of the first term is straightforward:

$$\begin{aligned} \log \frac{P(\mathbf{A}_1^t[ij] | \mathbf{A}^{t-1})}{P(\mathbf{A}_0^t[ij] | \mathbf{A}^{t-1})} = \\ \theta_1 + \theta_2(2A_{ij}^{t-1} - 1) + \theta_3 \frac{\sum_k A_{ik}^{t-1} A_{jk}^{t-1}}{\sum_{klm} A_{lk}^{t-1} A_{mk}^{t-1}}. \end{aligned}$$

For the second term,

$$\begin{aligned} \log \frac{P(\mathbf{A}^{t+1} | \mathbf{A}_1^t[ij])}{P(\mathbf{A}^{t+1} | \mathbf{A}_0^t[ij])} = \\ \theta' (\Psi(\mathbf{A}^{t+1}, \mathbf{A}_1^t[ij]) - \Psi(\mathbf{A}^{t+1}, \mathbf{A}_0^t[ij])) - \log \frac{Z(\theta, \mathbf{A}_1^t[ij])}{Z(\theta, \mathbf{A}_0^t[ij])}, \end{aligned}$$

All the features $\Psi(\mathbf{A}^{t+1}, \mathbf{A}^t)$ factorize over each binary variable A_{ij}^{t+1} , therefore the binary interactions of a timestep are conditionally independent of each other given all the interactions of the previous timestep. The log partition function becomes tractable:

$$\begin{aligned} \log Z(\theta, \mathbf{A}^t) = \\ \sum_{ij} \log \left(1 + \exp \left(\theta_1 + \theta_2(2A_{ij} - 1) + \theta_3 \frac{\sum_k A_{ik} A_{kj}}{\sum_{klm} A_{lk} A_{km}} \right) \right). \end{aligned}$$

so we know that this second term can be computed efficiently. Its further derivation is omitted here.

The third term on the right side of Eq. 9 is related to the local probability distributions in the emission model,

$$\log \frac{P(\mathbf{x}^t | \mathbf{A}_1^t[ij])}{P(\mathbf{x}^t | \mathbf{A}_0^t[ij])} = \eta \Lambda_{ij} (2\mathbb{I}(x_i^t = x_j^t) - 1) - \frac{Z(\eta, \mathbf{A}_1^t[ij])}{Z(\eta, \mathbf{A}_0^t[ij])}.$$

Now this ratio of two partition functions is intractable because components of \mathbf{x}^t are generally not conditionally independent of each other given \mathbf{A}^t . However, \mathbf{A}^t does determine the conditional independence relationship within \mathbf{x}^t . To make it clearer, we rewrite the emission model as follows:

$$P(\mathbf{x}^t | \mathbf{A}^t) = \frac{1}{Z(\eta, \mathbf{A}^t)} \exp \left\{ \sum_{(i,j) \in E^t} \eta \Lambda_{ij} (2\mathbb{I}(x_i^t = x_j^t) - 1) \right\},$$

where $E^t = \{(i,j) | A_{ij}^t = 1\}$ is the set of undirected edges in the network at timestep t . The conditional distribution in the equation above induces a (conditional) random field whose graphical structure \mathcal{G}^t consists of \mathbf{x}^t as the node and the edge set E^t . Conditioned on \mathbf{A}^t , the Markov property holds for the undirected graph \mathcal{G}^t . Therefore, we can perform the standard variable elimination procedure on \mathcal{G}^t to eliminate all the variables \mathbf{x}^t in order to exactly compute the partition function $Z(\eta, \mathbf{A}^t)$. The worst case complexity of this computation is exponential. However, since the \mathbf{A}^t is usually sparse, in practice we can find pretty efficient elimination orderings by applying some heuristics such as “min-fill” method which finds the vertex due to whose elimination a minimum number of edges need to be added to the induced graph for variable elimination.

4.2. The Learning Algorithm

The unknown parameters in a htERG model are θ and η , which govern the transition and emission model respectively. Usually they are of very low dimension, e.g. in the example of the model, the number of unknown parameters is only 4. Therefore, it is possible to use grid search to find a good starting point for any iterated parameter estimation algorithm.

A natural algorithm for parameter estimation follows from the Monte Carlo expectation-maximization framework. The algorithm iteratively draws samples from the posterior distribution of the network topologies using the Gibbs sampling algorithm in Sec. 4.1 (E step) followed by one step of gradient update of the parameter values using the posterior samples (M step). The detailed update schemes are akin to those presented in (Hanneke & Xing, 2006) and are omitted here due to space limitations. Convergence is usually fast by virtue of a good choice of starting point.

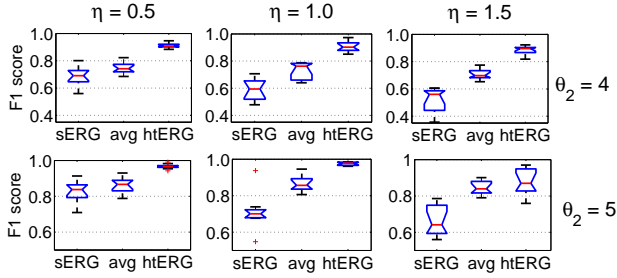


Figure 2. Empirical results for simulated data with varying values of θ_2 and η . θ_1 (density parameter) = -0.5 , θ_3 (transitivity parameter) = 4 . Transitivity features tend to increase the number of interactions in the network, so $\theta_1 < 0$ to balance this effect. Five iid samples are generated from the emission model for each timestep.

5. Empirical Analysis

5.1. Simulation-based Evaluation

We simulated network evolving processes from tERGM and generated *in silico* time series data conditioned on the network topology of each timestep and global activation functions in the emission model. Exact samples can be obtained because both partition functions encountered can be computed exactly as discussed in Sec. 4.1. The simulation was repeated under different combinations of parameters to mimic different scenarios of network evolution and noise in observation. To evaluate network reconstruction performance, we compared htERGM against its static version, sERG, as well as an “average of the truth” or “avg” method, which uses the information from the ground truth and outputs the average of underlying networks over all timesteps. This is approximately the optimal static network under most performance measures. Therefore, it provides an estimate of the modeling limitation of all static network reconstruction methods.

Now we are ready to describe our experimental settings and evaluation scheme. \mathbf{A}^0 was initialized as an $N = 10$ node network with 14 edges. The total number of timesteps T was fixed to be 50, but we may vary the number of observations per timestep. The parameters θ and η may also be different in different experimental settings. For each setting, we did 10 repetitions of simulation and reconstruction. The inference algorithm initialized all the latent network structures to be equal to \mathbf{A}^0 and performed 100k iterations of Gibbs sampling to ensure convergence. The harmonic mean of sensitivity and specificity of the reconstructed network known as F1 score was used as the measure of performance. The score presented is the mean of F1 scores obtained by comparing the inferred network and the ground truth for every timestep.

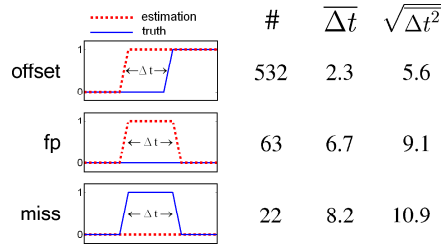


Figure 3. Empirical results on edge-switching events in simulated and reconstructed networks. Listed are the three cases: *offset*, *fp*, *miss*. For each case, the number of occurrences, mean absolute value $\overline{\Delta t}$ and root mean square value $\sqrt{\overline{\Delta t^2}}$ were shown. The numbers listed were averaged over 10 repetitions. Parameter settings are similar to those in Fig. 2 (therefore $T = 50$), with $\theta_2 = 4, \eta = 1$.

We first compared performances of three approaches under varying stability parameter θ_2 in the transition model and varying η which governs the noise level of the emission. Fig. 2 shows F1 scores in box plots. Plots within each row share the same stability parameter, and the bottom row has the larger θ_2 value, which suggests a slower evolution pace of the network topology. Empirically, the averaged proportion of edge switching between adjacent timesteps in the simulated networks are around 5 percent when $\theta_2 = 4$ and around 2 percent when $\theta_2 = 5$ out of all possible 45 pairwise interactions. Downwards each column in Fig. 2 as θ_2 increases, all the three algorithms get higher F1 scores. For sERG, it is because networks in different timesteps are more similar, therefore the modeling assumption becomes a better approximate. The effect of larger θ_2 on htERGM is that the network topology becomes more dependent on the observed node attributes of adjacent timesteps. This additional information increases the quality of the network reconstruction. htERGM consistently outperforms “avg” as well as sERG which suggests that it could get better performance than any static network reconstruction methods which have limited modeling power. And the margin is larger when the problem is harder (smaller θ_2). On the other hand, plots within the same column share the same emission parameter η , and the right-most row corresponds to the largest η value. Larger η suggests smaller noises in the observation and peaks posterior probability distributions, but this does not necessarily boost the performance of network recovery because those posterior distributions are generally multi-modal, and large η may cause overfitting.

Besides statistics on network reconstruction averaged over all timesteps, we also investigated the edge-switching events in simulated truth and networks reconstructed by htERGM. An edge-switching happens when $A_{ij}^{t+1} \neq A_{ij}^t$ for some i, j and t . We found the

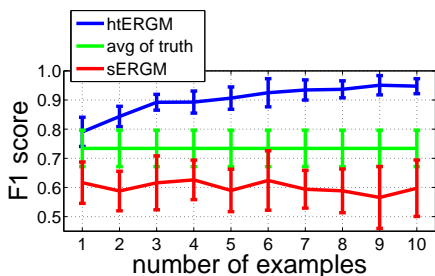


Figure 4. Comparison of F1 scores with different amount of data available. The number of *iid* samples generated from the emission model for each timestep ranges from 1 to 10. Parameter settings are similar to those in Fig. 3. Ten repetitions performed to obtain the error bar.

following three cases of disagreement between the estimation and underlying truth. An *offset* happens when the edge-switching in the estimation occurs a few timesteps before or after that in the truth and their directions agree. An *fp* (false positive) takes place when there are two consecutive edge-switchings in the estimation but none in the truth. A *miss* occurs when there are two consecutive edge-switchings in the truth but none in the estimation. Disagreements of a single edge-switching at the beginning or end of the time series and perfect agreements are treated as occurrences of offset. No other cases of disagreements were discovered in the experimental results. Figure 3 summarizes our findings in one simulation setting over 10 repetitions. htERGM was able to capture most of the edge-switching events with good sensitivity and specificity: the mean missing rate was $63/617 = 11.6\%$ and the mean false positive rate was only $22/617 = 3.6\%$. Most of the offsets are within 4 timesteps, whereas the length of the time series $T = 50$.

Finally, we compared performances varying the number of available observations and summarized the result in Fig. 4. When there were few observations per timestep, htERGM would benefit from the increasing amount of data. However, the performance of its static counterpart did not change significantly because it already pooled the data from T timesteps together, which appeared to be sufficient for inferring a single network. Also it is evident in the plot that no matter what amount of data is available, the modeling assumption limits the performance of sERGM to be no better than the “average of the truth” approach (the middle straight line of error bars). So htERGM has a greater potential to benefit from the increasing microarray experiment data publicly available.

5.2. Results on a Small Biological Network

We applied a htERGM to reconstruct epoch-specific subnetworks for the *Drosophila* lifecycle gene expres-

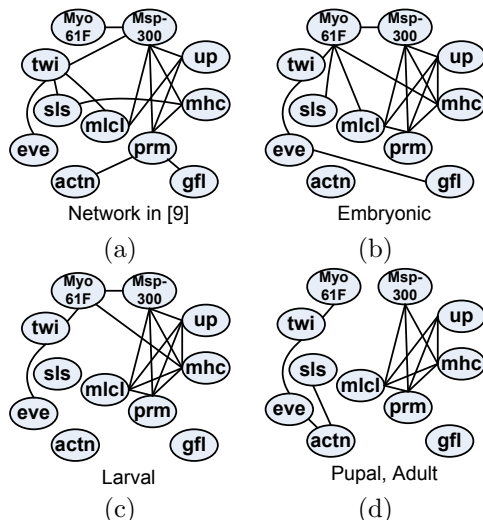


Figure 5. (a) The regulatory network of *Drosophila* muscle development genes reported in (Zhao et al., 2006); (b)-(d) reconstructed muscle development regulatory networks for (b) embryonic stage; (c) larval stage; (d) pupal stage and adulthood stage.

sion data (Arbeitman et al., 2002). This data set measures 66 timesteps for 4028 *Drosophila* genes through four stages of the life cycle, including embryonic, larval, pupal and the first 30 days of adulthood. We focused on the muscle development subnetwork which consists of 11 genes that were reported to be related to muscle development. Most recently, (Zhao et al., 2006) employed information theoretic algorithms to infer the same subnetwork from the same data set and reported their system level regulatory diagram which was assumed to be static over different timesteps. Although they studied 19 genes, 8 of them were separated from the main network. We only compared our results with their network built upon the 11 overlapping genes.

We preprocessed continuous expression data using the same methods as in (Zhao et al., 2006), which included missing data imputation, outlier identification and quantization into binary values. For this instance, a slightly different set of features was employed in the transition model (by replacing the transitivity feature with a density stability feature) for a better fit of the data. We performed a grid search to find a good initialization of parameters, and chose the final result according to the log-likelihood in the single timestep model. Figure 5 (a) and (b-d) show the network reported in (Zhao et al., 2006) and networks inferred by htERGM for each development stage, respectively. As can be seen in Fig. 5, the regulatory network evolves from embryonic stage to larval stage and finally to pupal stage. The network for adulthood stage is the same as that for pupal stage, suggesting that the regulatory network in *Drosophila* muscle development can finish

rewiring before adulthood. This result provides partial evidence that our algorithm is able to capture the evolution of regulatory networks for muscle development pathways. Further biological experiments are necessary for a complete verification of inferred networks.

6. Conclusion and Discussion

We proposed hidden temporal exponential random graph models for recovering temporally rewiring networks from a sequence of node attributes, and presented a sampling-based algorithm for inference and learning on htERGM. Empirical analysis on simulated data from evolving networks demonstrated the superior performance of our approach over methods that recover a single time-invariant network. We also applied our model to a portion of the *Drosophila* lifecycle data and reconstructed a small-scale network of muscle development related genes for each development stage.

One of the appealing properties of a htERGM is that it provides a general framework in which to formalize modeling assumptions in a standard and entirely readable way. Furthermore, the parameters often have useful interpretations corresponding to properties of the problem under investigation (e.g., stability of the network). The specific model we employ in the empirical study has particularly nice properties, in that for relatively small networks we can quickly compute the partition functions for the transition and emission models. However, in general, htERGMs do not have such nice properties. In particular, because the transition and emission models' partition functions depend directly on unobserved network variables, we cannot tractably compute likelihood ratios, so that even such staples as Gibbs sampling can be intractable. Since our current toolbox of approximate inference algorithms cannot be applied here, we are left with the prospect of designing novel inference algorithms for each new problem we wish to model with a htERGM. This problem seems to partially undermine the point of using a unified approach to modeling with htERGMs. We therefore pose as an open problem the task of designing approximate inference algorithms that can be tractably executed on this type of model (or some significant subfamily). A starting point is to apply loopy Metropolis algorithm (Murray & Ghahramani, 2004) to approximate the ratio of partition functions of the emission model while keeping the transition model tractable.

Acknowledgments

We thank Yanxin Shi, Wentao Zhao and Hetunandan Kamisetty for their contribution or help on this

work. We also thank the anonymous ICML reviewers for their insightful comments. This work is partially supported by the NSF under Grant No. DBI-0546594.

References

- Airoldi, E., Blei, D., Xing, E. P., & Fienberg, S. (2005). A latent mixed membership model for relational data. *Proceedings of the 3rd international workshop on Link discovery*, 82–89.
- Arbeitman, et al. (2002). Gene expression during the life cycle of *Drosophila melanogaster*. *Science*, 297, 2270–2275.
- Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286, 509–512.
- Fienberg, S. E., Meyer, M. M., & Wasserman, S. (1985). Statistical analysis of multiple sociometric relations. *Journal of the American Statistical Association*, 80, 51–67.
- Frank, O., & Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, 81, 832–842.
- Friedman, N., Linial, M., Nachman, I., & Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7, 601–620.
- Hanneke, S., & Xing, E. P. (2006). Discrete temporal models of social networks. *Proceedings of the ICML'06 Workshop on Statistical Network Analysis*.
- Heckerman, D. (1995). A tutorial on learning with Bayesian networks. Technical Report MSR-TR-95-06.
- Hoff, P. D., Raftery, A. E., & Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97, 1090–1098.
- Hu, H., et al. (2005). Mining coherent dense subgraphs across massive biological networks for functional discovery. *Bioinformatics*, 21, i213–i221.
- Lawrence, N. D., Sanguinetti, G., & Rattray, M. (2007). Modelling transcriptional regulation using Gaussian processes. *NIPS 19*.
- Luscombe, N. M., et al. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431, 308–312.
- Murray, I., & Ghahramani, Z. (2004). Bayesian learning in undirected graphical models: approximate MCMC scheme. *UAI 2004*, 392–399.
- Murray, I., Ghahramani, Z., & MacKay, D. J. C. (2006). MCMC for doubly-intractable distributions. *UAI 2006*.
- Robins, G., Pattison, P., Kalish, Y., & Lusher, D. (2007). An introduction to exponential random graph (p^*) models for social networks. *Social Networks*, 29, 173–191.
- Sarkar, P., & Moore, A. W. (2006). Dynamic social network analysis using latent space models. *NIPS 18*, 1145–1152.
- Strauss, D., & Ikeda, M. (1990). Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, 85, 204–212.
- Wasserman, S., & Pattison, P. (1996). Logit models and logistic regression for social networks: I. an introduction to markov graphs and p^* . *Psychometrika*, 61, 401–425.
- Zhao, W., Serpedin, E. & Dougherty, E. R. (2006). Inferring gene regulatory networks from time series data using the minimum description length principle. *Bioinformatics*, 22, 2129–2135.