VARIATIONAL INFERENCE FOR LATENT SPACE MODELS FOR DYNAMIC NETWORKS

Yan Liu and Yuguo Chen

University of Illinois at Urbana-Champaign

Supplementary Material

S1 Proof of Theorem 1

First, by Theorem 3.1 of Yang et al. (2020), we have the following variational risk bound for the $\alpha < 1$ case:

$$\begin{aligned} & \int \frac{1}{n(n-1)T} D_\alpha^{(n)}(\theta, \theta^*) \hat{q}_{\theta,\alpha}(d\theta) \\ \leq \; & \frac{\alpha}{n(n-1)T(1-\alpha)} \Psi_{n,\alpha}(q_\theta, q_{\boldsymbol{\chi}}) + \frac{1}{n(n-1)T(1-\alpha)} \log\left(\frac{1}{\zeta}\right), \end{aligned}$$

with probability at least $(1-\zeta)$ for any $\zeta \in (0,1)$. This result establishes a connection between the variational Bayes risk and the α-VB objective function, which implies that minimizing the α-VB objective function $\Psi_{n,\alpha}$ will also minimize the variational Bayes risk.

The next step of the proof is to further simplify the above upper bound

based on a certain choice of the variational family of the latent variable $\boldsymbol{\mathcal{X}}$ and the model parameter θ. Recall that $\theta = (\beta, \sigma^2, \tau^2)$ denotes all the model parameters, and $\theta^* = (\beta^*, \sigma^{*2}, \tau^{*2})$ is their true values. From now on, we use $\pi = (\sigma^2, \tau^2)$ to denote the parameters that characterize the distribution of latent variables, and use $\pi^* = (\sigma^{*2}, \tau^{*2})$ to denote their true values.

We still use the mean-field decomposition $q(\boldsymbol{\mathcal{X}}, \theta) = q(\boldsymbol{\mathcal{X}})q(\theta)$. However, the dependence structure between the observations and the latent variables in our model is different from the simplifying assumptions in Yang et al. (2020), since we no longer have i.i.d. observations or observation-specific latent variables. In our case, for any fixed q_θ, we choose the variational distribution $q(\boldsymbol{\mathcal{X}})$ in the following way:

$$q(\boldsymbol{\mathcal{X}}) \propto \left(\prod_{t=1}^{T} \prod_{i \neq j} p(Y_{ijt} | \beta^*, \boldsymbol{X}_{it}, \boldsymbol{X}_{jt}) \right) \times \left(\prod_{i=1}^{n} \left(p(\boldsymbol{X}_{i1} | \pi^*) \prod_{t=1}^{T} p(\boldsymbol{X}_{it} | \pi^*, \boldsymbol{X}_{i(t-1)}) \right) \right),$$

where the normalizing constant is $p(\boldsymbol{\mathcal{Y}} | \theta^*)$.

With this choice of variational family, the α-VB objective function becomes

$$\begin{aligned}
\Psi_{n,\alpha}(q_\theta, q_{\boldsymbol{\mathcal{X}}}) &= -\int_\Theta \left(l_n(\theta) - l_n(\theta^*) \right) q(d\theta) + \Delta_J(q_\theta, q_{\boldsymbol{\mathcal{X}}}) + \frac{1}{\alpha} D(q_\theta || p_\theta) \\
&= -\int_\Theta \left(l_n(\theta) - l_n(\theta^*) + \hat{l}_n(\theta) - l_n(\theta) \right) q(d\theta) + \frac{1}{\alpha} D(q_\theta || p_\theta) \\
&= -\int_\Theta \left(\hat{l}_n(\theta) - l_n(\theta^*) \right) q(d\theta) + \frac{1}{\alpha} D(q_\theta || p_\theta),
\end{aligned}$$

where the first term on the right hand side is

$$
\begin{aligned}
&-\int_{\Theta}\left(\hat{l}_n(\theta)-l_n(\theta^*)\right)q(d\theta)\\
=&-\int_{\Theta}\left(\int_{\boldsymbol{\mathcal{X}}}\log\frac{p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{X}},\beta)p(\boldsymbol{\mathcal{X}}|\pi)}{q(\boldsymbol{\mathcal{X}})}q(d\boldsymbol{\mathcal{X}})-\log\left(\int_{\boldsymbol{\mathcal{X}}}\frac{p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{X}},\beta^*)p(\boldsymbol{\mathcal{X}}|\pi^*)}{q(\boldsymbol{\mathcal{X}})}q(d\boldsymbol{\mathcal{X}})\right)\right)q(d\theta)\\
=&-\int_{\Theta}\left(\int_{\boldsymbol{\mathcal{X}}}\sum_{t=1}^{T}\sum_{i\neq j}\log\frac{p(Y_{ijt}|\beta,\boldsymbol{X}_{it},\boldsymbol{X}_{jt})}{p(Y_{ijt}|\beta^*,\boldsymbol{X}_{it},\boldsymbol{X}_{jt})}q(d\boldsymbol{\mathcal{X}})-D(p(\boldsymbol{\mathcal{X}}|\pi^*)||p(\boldsymbol{\mathcal{X}}|\pi))\right)q(d\theta)\\
&+\left(\int_{\Theta}\int_{\boldsymbol{\mathcal{X}}}p(\boldsymbol{\mathcal{Y}}|\theta^*)q(d\boldsymbol{\mathcal{X}})q(d\theta)-\int_{\Theta}\log\left(\int_{\boldsymbol{\mathcal{X}}}\frac{p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{X}},\beta^*)p(\boldsymbol{\mathcal{X}}|\pi^*)}{q(\boldsymbol{\mathcal{X}})}q(d\boldsymbol{\mathcal{X}})\right)q(d\theta)\right).
\end{aligned}
$$

Note that the last expression in the equation above contains two terms and the second term is 0. Therefore, the variational risk bound becomes

$$
\begin{aligned}
&\int\frac{1}{n(n-1)T}D_{\alpha}^{(n)}(\theta,\theta^*)\hat{q}_{\theta,\alpha}(d\theta)\\
\leq&\frac{\alpha}{n(n-1)T(1-\alpha)}\Psi_{n,\alpha}(q_{\theta},q_{\boldsymbol{\mathcal{X}}})+\frac{1}{n(n-1)T(1-\alpha)}\log\left(\frac{1}{\zeta}\right)\\
=&-\frac{\alpha}{n(n-1)T(1-\alpha)}\int_{\Theta}\int_{\boldsymbol{\mathcal{X}}}\Bigg(\sum_{t=1}^{T}\sum_{i\neq j}\log\frac{p(Y_{ijt}|\boldsymbol{X}_{it},\boldsymbol{X}_{jt},\beta)}{p(Y_{ijt}|\boldsymbol{X}_{it},\boldsymbol{X}_{jt},\beta^*)}\\
&\qquad\qquad-\sum_{i=1}^{n}D(p(\boldsymbol{X}_{it}|\pi^*)||p(\boldsymbol{X}_{it}|\pi))\Bigg)q(d\boldsymbol{\mathcal{X}})q(d\theta)\\
&+\frac{1}{n(n-1)T(1-\alpha)}\left(D(q(\theta)||p(\theta))+\log(1/\zeta)\right)\\
:=&-\frac{\alpha}{n(n-1)T(1-\alpha)}W_1+\frac{1}{n(n-1)T(1-\alpha)}\left(D(q(\theta)||p(\theta))+\log(1/\zeta)\right).
\end{aligned}
\tag{S1.1}
$$

The variational risk bound (S1.1) can be viewed as an analogue of Corollary 3.2 in Yang et al. (2020). Now in order to apply Chebyshev's inequality to obtain the desired result, we need to bound the first and

second moments of the first term of the right-hand side of (S1.1).

Recall the following definition of the KL-neighborhoods of the true model parameters given in the statement of Theorem 1:

$$\mathcal{B}_n(\pi^*, \epsilon_\pi) := \left\{\pi : D\left(p(\boldsymbol{\mathcal{X}}_1|\pi^*)||p(\boldsymbol{\mathcal{X}}_1|\pi)\right) \leq \epsilon_\pi^2, \quad V\left(p(\boldsymbol{\mathcal{X}}_1|\pi^*)||p(\boldsymbol{\mathcal{X}}_1|\pi)\right) \leq \epsilon_\pi^2\right\},$$

$$\begin{aligned}\mathcal{B}_n(\beta^*, \epsilon_\beta) :=& \Bigg\{\beta : \sup_{\boldsymbol{X}_{11}, \boldsymbol{X}_{21}} D(p(Y_{121}|\beta^*, \boldsymbol{X}_{11}, \boldsymbol{X}_{21})||p(Y_{121}|\beta, \boldsymbol{X}_{11}, \boldsymbol{X}_{21})) \leq \epsilon_\beta^2, \\ & \sup_{\boldsymbol{X}_{11}, \boldsymbol{X}_{21}} V(p(Y_{121}|\beta^*, \boldsymbol{X}_{11}, \boldsymbol{X}_{21})||p(Y_{121}|\beta, \boldsymbol{X}_{11}, \boldsymbol{X}_{21})) \leq \epsilon_\beta^2\Bigg\},\end{aligned}$$

where $V(p||q) := \int p \log^2(\frac{p}{q}) d\mu$. Then we choose $q_\theta(\theta)$ as the probability density function (pdf) q_θ^*, which is the pdf of the product measure of restrictions of the priors of the model parameters to KL-neighborhoods $\mathcal{B}_n(\pi^*, \epsilon_\pi)$ and $\mathcal{B}_n(\beta^*, \epsilon_\beta)$.

By Fubini's Theorem,

$$\begin{aligned}&\mathbb{E}_{\theta^*}[W_1] \\ =&\mathbb{E}_{\theta^*}\left[\int_\Theta q_\theta^*(\theta) \int_{\boldsymbol{\mathcal{X}}} \left(\sum_{t=1}^T \sum_{i \neq j} \log \frac{P(Y_{ijt}|\boldsymbol{X}_{it}, \boldsymbol{X}_{jt}, \beta)}{P(Y_{ijt}|\boldsymbol{X}_{it}, \boldsymbol{X}_{jt}, \beta^*)} - \sum_{i=1}^n D(p(\boldsymbol{\mathcal{X}}_i|\pi^*)||p(\boldsymbol{\mathcal{X}}_i|\pi))\right) q(d\boldsymbol{\mathcal{X}}) d\theta\right] \\ =&\int_\Theta \left\{\mathbb{E}_{\theta^*}\left[\int_{\boldsymbol{\mathcal{X}}} \left(\sum_{t=1}^T \sum_{i \neq j} \log \frac{P(Y_{ijt}|\boldsymbol{X}_{it}, \boldsymbol{X}_{jt}, \beta)}{P(Y_{ijt}|\boldsymbol{X}_{it}, \boldsymbol{X}_{jt}, \beta^*)} - \sum_{i=1}^n D(p(\boldsymbol{\mathcal{X}}_i|\pi^*)||p(\boldsymbol{\mathcal{X}}_i|\pi))\right) q(d\boldsymbol{\mathcal{X}})\right]\right\} q_\theta^*(\theta) d\theta \\ =&\int_\Theta \Bigg\{ - \sum_{i=1}^n D(p(\boldsymbol{\mathcal{X}}_i|\pi^*)||p(\boldsymbol{\mathcal{X}}_i|\pi)) \\ & \quad - n(n-1)T \int_{\boldsymbol{\mathcal{X}}} D(p(\cdot|\boldsymbol{X}_{11}, \boldsymbol{X}_{21}, \beta^*)||p(\cdot|\boldsymbol{X}_{11}, \boldsymbol{X}_{21}, \beta)) q(d\boldsymbol{\mathcal{X}})\Bigg\} q_\theta^*(\theta) d\theta.\end{aligned}$$

Since q_θ^* is the restriction of $p(\theta)$ into the KL-neighborhoods defined above,

we have

$$-\frac{\alpha}{n(n-1)T(1-\alpha)}\mathbb{E}_{\theta^*}[W_1] \leq \frac{\alpha\left(n(n-1)T\epsilon_\beta^2 + n\epsilon_\pi^2\right)}{n(n-1)T(1-\alpha)}.$$

Furthermore, the variance

$$\begin{aligned}
&\mathrm{Var}_{\theta^*}[W_1] \\
=&\mathrm{Var}_{\theta^*}\left[\int_\Theta q_\theta^*(\theta)\int_{\boldsymbol{\mathcal{X}}}\left(\sum_{t=1}^T\sum_{i\neq j}\log\frac{P(Y_{ijt}|\boldsymbol{X}_{it},\boldsymbol{X}_{jt},\beta)}{P(Y_{ijt}|\boldsymbol{X}_{it},\boldsymbol{X}_{jt},\beta^*)} - \sum_{i=1}^n D(p(\boldsymbol{\mathcal{X}}_i|\pi^*)||p(\boldsymbol{\mathcal{X}}_i|\pi))\right)q(d\boldsymbol{\mathcal{X}})d\theta\right] \\
=&\mathrm{Var}_{\theta^*}\left[\mathbb{E}_{q(\boldsymbol{\mathcal{X}}),q_\theta^*(\theta)}\left[\sum_{t=1}^T\sum_{i\neq j}\log\frac{P(Y_{ijt}|\boldsymbol{X}_{it},\boldsymbol{X}_{jt},\beta)}{P(Y_{ijt}|\boldsymbol{X}_{it},\boldsymbol{X}_{jt},\beta^*)} - \sum_{i=1}^n D(p(\boldsymbol{\mathcal{X}}_i|\pi^*)||p(\boldsymbol{\mathcal{X}}_i|\pi))\right]\right] \\
=&\mathrm{Var}_{\theta^*}\left[\sum_{t=1}^T\sum_{i\neq j}\mathbb{E}_{q(\boldsymbol{\mathcal{X}}),q_\theta^*(\theta)}\left[\log\frac{P(Y_{ijt}|\boldsymbol{X}_{it},\boldsymbol{X}_{jt},\beta)}{P(Y_{ijt}|\boldsymbol{X}_{it},\boldsymbol{X}_{jt},\beta^*)}\right] - \sum_{i=1}^n \mathbb{E}_{q_\theta^*(\theta)}\left[D(p(\boldsymbol{\mathcal{X}}_i|\pi^*)||p(\boldsymbol{\mathcal{X}}_i|\pi))\right]\right] \\
=&\sum_{t,s=1}^T\sum_{i\neq j}\sum_{k\neq l}\mathrm{Cov}\left(\mathbb{E}_{q(\boldsymbol{\mathcal{X}}),q_\theta^*(\theta)}\left[\log\frac{P(Y_{ijt}|\boldsymbol{X}_{it},\boldsymbol{X}_{jt},\beta)}{P(Y_{ijt}|\boldsymbol{X}_{it},\boldsymbol{X}_{jt},\beta^*)}\right],\ \mathbb{E}_{q(\boldsymbol{\mathcal{X}}),q_\theta^*(\theta)}\left[\log\frac{P(Y_{kls}|\boldsymbol{X}_{ks},\boldsymbol{X}_{ls},\beta)}{P(Y_{kls}|\boldsymbol{X}_{ks},\boldsymbol{X}_{ls},\beta^*)}\right]\right) \qquad \text{(S1.2)} \\
&+\sum_{i,j=1}^n\mathrm{Cov}\left(\mathbb{E}_{q_\theta^*(\theta)}\left[D(p(\boldsymbol{\mathcal{X}}_i|\pi^*)||p(\boldsymbol{\mathcal{X}}_i|\pi))\right],\ \mathbb{E}_{q_\theta^*(\theta)}\left[D(p(\boldsymbol{\mathcal{X}}_j|\pi^*)||p(\boldsymbol{\mathcal{X}}_j|\pi))\right]\right) \qquad \text{(S1.3)} \\
&-\sum_{t=1}^T\sum_{i\neq j}\sum_{k=1}^n\mathrm{Cov}\left(\mathbb{E}_{q(\boldsymbol{\mathcal{X}}),q_\theta^*(\theta)}\left[\log\frac{P(Y_{ijt}|\boldsymbol{X}_{it},\boldsymbol{X}_{jt},\beta)}{P(Y_{ijt}|\boldsymbol{X}_{it},\boldsymbol{X}_{jt},\beta^*)}\right],\ \mathbb{E}_{q_\theta^*(\theta)}\left[D(p(\boldsymbol{\mathcal{X}}_k|\pi^*)||p(\boldsymbol{\mathcal{X}}_k|\pi))\right]\right). \qquad \text{(S1.4)}
\end{aligned}$$

First, note that for any $1 \leq i \neq j \leq n$, $t = 1, \ldots, T$,

$$\begin{aligned}\log \frac{P(Y_{ijt}|\boldsymbol{X}_{it}, \boldsymbol{X}_{jt}, \beta)}{P(Y_{ijt}|\boldsymbol{X}_{it}, \boldsymbol{X}_{jt}, \beta^*)} =& Y_{ijt}(\beta - ||\boldsymbol{X}_{it} - \boldsymbol{X}_{jt}||^2) - Y_{ijt}(\beta^* - ||\boldsymbol{X}_{it} - \boldsymbol{X}_{jt}||^2) \\ & - \log\left(1 + e^{\beta - ||\boldsymbol{X}_{it} - \boldsymbol{X}_{jt}||^2}\right) + \log\left(1 + e^{\beta^* - ||\boldsymbol{X}_{it} - \boldsymbol{X}_{jt}||^2}\right) \\ =& Y_{ijt}(\beta - \beta^*) - \log\left(\frac{1 + e^{\beta - ||\boldsymbol{X}_{it} - \boldsymbol{X}_{jt}||^2}}{1 + e^{\beta^* - ||\boldsymbol{X}_{it} - \boldsymbol{X}_{jt}||^2}}\right).\end{aligned}$$

Thus, after taking expectation with respect to $\boldsymbol{\mathcal{X}}$ and β, only the first term is random. Let

$$\mathbb{E}_{q(\boldsymbol{\mathcal{X}}), q_\theta^*(\theta)}\left[\log \frac{P(Y_{ijt}|\boldsymbol{X}_{it}, \boldsymbol{X}_{jt}, \beta)}{P(Y_{ijt}|\boldsymbol{X}_{it}, \boldsymbol{X}_{jt}, \beta^*)}\right] := c_1 Y_{ijt} + c_2(i, j, t),$$

where c_1 and c_2 are constants that depend on the variational distribution q.

Then the term (S1.2) becomes

$$\sum_{t,s=1}^{T} \sum_{i \neq j} \sum_{k \neq l} \mathrm{Cov}\left[c_1 Y_{ijt} + c_2(i, j, t),\ c_1 Y_{kls} + c_2(k, l, s)\right] = c_1^2 \sum_{t,s=1}^{T} \sum_{i \neq j} \sum_{k \neq l} \mathrm{Cov}\left[Y_{ijt}, Y_{kls}\right].$$

The number of terms in the summation is $n^2(n-1)^2T^2$, but for any t and s, $\mathrm{Cov}\left[Y_{ijt}, Y_{kls}\right] = 0$ when $i \neq k$ and $j \neq l$. Also, for any $i, j = 1, \ldots, n$

and $t = 1, \ldots, T$, the variance term can be bounded in the following way:

$$
\begin{aligned}
&\mathrm{Var}_{\theta^*}\left[\mathbb{E}_{q(\boldsymbol{\mathcal{X}}),q^*_\theta(\theta)}\left[\log\frac{P(Y_{ijt}|\boldsymbol{X}_{it},\boldsymbol{X}_{jt},\beta)}{P(Y_{ijt}|\boldsymbol{X}_{it},\boldsymbol{X}_{jt},\beta^*)}\right]\right]\\
=&\mathrm{Var}_{\theta^*}\left[\int_\Theta q^*_\theta(\theta)\int_{\boldsymbol{\mathcal{X}}} q_{\boldsymbol{\mathcal{X}}}(\boldsymbol{\mathcal{X}})\log\frac{P(Y_{ijt}|\boldsymbol{X}_{it},\boldsymbol{X}_{jt},\beta)}{P(Y_{ijt}|\boldsymbol{X}_{it},\boldsymbol{X}_{jt},\beta^*)}d\boldsymbol{\mathcal{X}}d\theta\right]\\
\leq&\mathbb{E}_{\theta^*}\left[\int_\Theta q^*_\theta(\theta)\int_{\boldsymbol{\mathcal{X}}} q_{\boldsymbol{\mathcal{X}}}(\boldsymbol{\mathcal{X}})\log\frac{P(Y_{ijt}|\boldsymbol{X}_{it},\boldsymbol{X}_{jt},\beta)}{P(Y_{ijt}|\boldsymbol{X}_{it},\boldsymbol{X}_{jt},\beta^*)}d\boldsymbol{\mathcal{X}}d\theta\right]^2\\
\leq&\mathbb{E}_{\theta^*}\left[\int_\Theta q^*_\theta(\theta)\left[\int_{\boldsymbol{\mathcal{X}}} q_{\boldsymbol{\mathcal{X}}}(\boldsymbol{\mathcal{X}})\log\frac{P(Y_{ijt}|\boldsymbol{X}_{it},\boldsymbol{X}_{jt},\beta)}{P(Y_{ijt}|\boldsymbol{X}_{it},\boldsymbol{X}_{jt},\beta^*)}d\boldsymbol{\mathcal{X}}\right]^2 d\theta\right]\\
=&\int_\Theta \mathbb{E}_{\theta^*}\left[\int_{\boldsymbol{\mathcal{X}}} q_{\boldsymbol{\mathcal{X}}}(\boldsymbol{\mathcal{X}})\log\frac{P(Y_{ijt}|\boldsymbol{X}_{it},\boldsymbol{X}_{jt},\beta)}{P(Y_{ijt}|\boldsymbol{X}_{it},\boldsymbol{X}_{jt},\beta^*)}d\boldsymbol{\mathcal{X}}\right]^2 q^*_\theta(\theta)d\theta\\
\leq&\int_\Theta \mathbb{E}_{\theta^*}\left[\int_{\boldsymbol{\mathcal{X}}} q_{\boldsymbol{\mathcal{X}}}(\boldsymbol{\mathcal{X}})\log^2\frac{P(Y_{ijt}|\boldsymbol{X}_{it},\boldsymbol{X}_{jt},\beta)}{P(Y_{ijt}|\boldsymbol{X}_{it},\boldsymbol{X}_{jt},\beta^*)}d\boldsymbol{\mathcal{X}}\right] q^*_\theta(\theta)d\theta\\
=&\int_\Theta\left[\int_{\boldsymbol{\mathcal{X}}} V\left[p(\cdot|\beta^*,\boldsymbol{X}_{11},\boldsymbol{X}_{21})||p(\cdot|\beta,\boldsymbol{X}_{11},\boldsymbol{X}_{21})\right]q_{\boldsymbol{\mathcal{X}}}(\boldsymbol{\mathcal{X}})d\boldsymbol{\mathcal{X}}\right] q^*_\theta(\theta)d\theta\\
\leq&\epsilon^2_\beta,
\end{aligned}
$$

where the second and third inequalities are due to Jensen's inequality, and the second equality is due to Fubini's theorem. Thus, by Cauchy-Schwarz inequality,

$$
\begin{aligned}
&\text{term } (S1.2)\\
\leq&\sum_{t,s=1}^{T}\sum_{i\neq j}\sum_{k\neq l}\sqrt{\mathrm{Var}_{\theta^*}\left[\mathbb{E}\left[\log\frac{P(Y_{ijt}|\boldsymbol{X}_{it},\boldsymbol{X}_{jt},\beta)}{P(Y_{ijt}|\boldsymbol{X}_{it},\boldsymbol{X}_{jt},\beta^*)}\right]\right]\cdot\mathrm{Var}_{\theta^*}\left[\mathbb{E}\left[\log\frac{P(Y_{kls}|\boldsymbol{X}_{ks},\boldsymbol{X}_{ls},\beta)}{P(Y_{kls}|\boldsymbol{X}_{ks},\boldsymbol{X}_{ls},\beta^*)}\right]\right]}\\
\leq&(n(n-1)T^2+n(n-1)(n-2)T^2)\epsilon^2_\beta\\
\leq&n(n-1)^2T^2(\epsilon^2_\beta+\epsilon^2_\pi).
\end{aligned}
$$

The other two terms: term $(S1.3)$ = term $(S1.4) = 0$. By Chebyshev's

inequality, for any fixed $(\epsilon_\beta, \epsilon_\pi) \in (0,1)$ and any $D > 1$,

$$\begin{aligned}\mathbb{P}_{\theta^*}\left(W_1 \le -Dn(n-1)T(\epsilon_\beta^2+\epsilon_\pi^2)\right) &\le \mathbb{P}_{\theta^*}\left(W_1 - \mathbb{E}[W_1] \le -(D-1)n(n-1)T(\epsilon_\beta^2+\epsilon_\pi^2)\right)\\ &\le \frac{\mathrm{Var}\left[W_1\right]}{(D-1)^2n^2(n-1)^2T^2(\epsilon_\beta^2+\epsilon_\pi^2)^2}\\ &\le \frac{1}{(D-1)^2n(\epsilon_\beta^2+\epsilon_\pi^2)}. \qquad \text{(S1.5)}\end{aligned}$$

Since the variational family of θ is the restriction of the prior on the KL-neighborhoods $\mathcal{B}_n(\pi^*, \epsilon_\pi)$ and $\mathcal{B}_n(\beta^*, \epsilon_\beta)$, we have

$$D(q_\theta^*(\theta)||p_\theta(\theta)) = -\log P_\pi\left[\mathcal{B}_n(\pi^*, \epsilon_\pi)\right] - \log P_\beta\left[\mathcal{B}_n(\beta^*, \epsilon_\beta)\right],$$

where P_π and P_β denote the probability measures corresponding to the priors of π and β, respectively. This together with inequality (S1.5) implies that for any fixed $(\epsilon_\pi^2, \epsilon_\beta^2) \in (0,1)^2$ and $D > 1$, it holds with probability at least $1 - \frac{2}{(D-1)^2n(\epsilon_\beta^2+\epsilon_\pi^2)}$ that

$$\begin{aligned}\int \frac{1}{n(n-1)T} D_\alpha^{(n)}(\theta, \theta^*)\hat{q}_{\theta,\alpha}(d\theta) \le& \frac{D\alpha}{1-\alpha}(\epsilon_\pi^2+\epsilon_\beta^2) - \frac{1}{n(n-1)T(1-\alpha)}\log P_\pi(\mathcal{B}_n(\pi^*, \epsilon_\pi))\\ &- \frac{1}{n(n-1)T(1-\alpha)}\log P_\beta(\mathcal{B}_n(\beta^*, \epsilon_\beta)).\end{aligned}$$

S2 Proof of Theorem 2

For any $i, j = 1, \ldots, n$ and $t = 1, \ldots, T$, the KL divergence

$$\begin{aligned}D\left(p(\cdot|\beta^*, \boldsymbol{X}_{it}, \boldsymbol{X}_{jt})||p(\cdot|\beta, \boldsymbol{X}_{it}, \boldsymbol{X}_{jt})\right) &= \mathbb{E}\left[Y_{ijt}\right](\beta^* - \beta) - \log\left(\frac{1+e^{\beta^* - ||\boldsymbol{X}_{it}-\boldsymbol{X}_{jt}||^2}}{1+e^{\beta-||\boldsymbol{X}_{it}-\boldsymbol{X}_{jt}||^2}}\right)\\ &\le |\beta-\beta^*| - \log e^{-|\beta-\beta^*|} \le 2|\beta-\beta^*|,\end{aligned}$$

and the V-divergence

$$
\begin{aligned}
&V\left(p(\cdot|\beta^*, \boldsymbol{X}_{it}, \boldsymbol{X}_{jt})||p(\cdot|\beta, \boldsymbol{X}_{it}, \boldsymbol{X}_{jt})\right) \\
=&\mathbb{E}_{\beta^*}\left[\left(Y_{ijt}(\beta^*-\beta) - \log\left(\frac{1+e^{\beta^*-||\boldsymbol{X}_{it}-\boldsymbol{X}_{jt}||^2}}{1+e^{\beta-||\boldsymbol{X}_{it}-\boldsymbol{X}_{jt}||^2}}\right)\right)^2\right] \\
=&\mathbb{E}\left[Y_{ijt}^2\right]|\beta^*-\beta|^2 + \log^2\left(\frac{1+e^{\beta^*-||\boldsymbol{X}_{it}-\boldsymbol{X}_{jt}||^2}}{1+e^{\beta-||\boldsymbol{X}_{it}-\boldsymbol{X}_{jt}||^2}}\right) - (\beta^*-\beta)\log\left(\frac{1+e^{\beta^*-||\boldsymbol{X}_{it}-\boldsymbol{X}_{jt}||^2}}{1+e^{\beta-||\boldsymbol{X}_{it}-\boldsymbol{X}_{jt}||^2}}\right)\mathbb{E}[Y_{ijt}] \\
\leq&|\beta^*-\beta|^2 + |\beta^*-\beta|^2 = 2|\beta^*-\beta|^2.
\end{aligned}
$$

Note that $(\beta^*-\beta)\log\left(\frac{1+e^{\beta^*-||\boldsymbol{X}_{it}-\boldsymbol{X}_{jt}||^2}}{1+e^{\beta-||\boldsymbol{X}_{it}-\boldsymbol{X}_{jt}||^2}}\right)\mathbb{E}[Y_{ijt}] \geq 0$, and the last inequality is based on the following fact

$$
e^{-|x-y|} \leq \frac{1+e^x}{1+e^y} \leq e^{|x-y|}.
$$

This implies that the KL neighborhood $\mathcal{B}_n(\beta^*, \epsilon_\beta)$ contains the set $\left\{\beta : |\beta-\beta^*| \leq \frac{c}{2}\epsilon_\beta^2\right\}$ for some constant c, and the volume of this set is at most of the order $\mathcal{O}(\epsilon_\beta^2)$. Consequently, by the thick prior assumption, the prior mass of this set $P_\beta(\left\{\beta : |\beta-\beta^*| \leq \frac{c}{2}\epsilon_\beta^2\right\})$ is at least of the order $\mathcal{O}(1/\epsilon_\beta^2)$.

Similarly, for any $i = 1, \ldots, n$,

$$
\begin{aligned}
D(p(\boldsymbol{\mathcal{X}}_i|\sigma^{*2}, \tau^{*2})||p(\boldsymbol{\mathcal{X}}_i|\sigma^2, \tau^2)) = &\left[-\frac{d}{2}\log\left(\frac{\sigma^{*2}}{\sigma^2}\right) - d\left(\frac{\sigma^{*2}}{2\sigma^{*2}} - \frac{\sigma^{*2}}{2\sigma^2}\right)\right] \\
&+ \left[-\frac{d}{2}\log\left(\frac{\tau^{*2}}{\tau^2}\right) - d\left(\frac{\tau^{*2}}{2\tau^{*2}} - \frac{\tau^{*2}}{2\tau^2}\right)\right] \cdot (T-1).
\end{aligned}
$$

Since we restrict model parameters in a compact set, by Lipschitz con-

tinuity, there exists some constant C, such that

$$\mathcal{B}_n(\pi^*, \epsilon_\pi) \supset \left\{(\sigma^2, \tau^2) : |\sigma^2 - \sigma^{*2}| \leq C\epsilon_\pi^2, \quad |\tau^2 - \tau^{*2}| \leq C\epsilon_\pi^2\right\}.$$

The volume of the set on the right-hand side is $(2C\epsilon_\pi^2)^2$. Thus, for any $D > 1$, we have the following risk bound with probability tending to 1 as $n \to \infty$,

$$\begin{aligned} & \int \frac{1}{n(n-1)T} D_\alpha^{(n)}(\theta, \theta^*) \hat{q}_{\theta,\alpha}(d\theta) \\ \leq \; & \frac{D\alpha}{1-\alpha}(\epsilon_\pi^2 + \epsilon_\beta^2) - \frac{2\log\epsilon_\beta}{n(n-1)T(1-\alpha)} - \frac{2\log(2C\epsilon_\pi^2)}{n(n-1)T(1-\alpha)}. \end{aligned}$$

Choosing $\epsilon_\beta = \epsilon_\pi = \frac{1}{\sqrt{n}}$, then the risk bound is of the order $\mathcal{O}(\frac{1}{n})$.

S3 Proof of Theorem 3

As stated in the main text, showing theoretical properties of the regular VB requires extra conditions. We first restate the two assumptions proposed by Yang et al. (2020).

Assumption 1. For some $\epsilon_n > 0$ and any $\epsilon > \epsilon_n$, there exist a subset of the parameter space $\mathcal{F}_{n,\epsilon} \subset \Theta$ and a test function $\phi_{n,\epsilon}$ such that

$$P_\theta(\mathcal{F}_{n,\epsilon}^c) \leq e^{-cn(n-1)\epsilon^2},$$

$$\mathbb{E}_{\theta^*}\left[\phi_{n,\epsilon}\right] \leq e^{-cn(n-1)\epsilon_n^2},$$

$$\mathbb{E}_\theta\left[1 - \phi_{n,\epsilon}\right] \leq e^{-cn(n-1)h^2(\theta||\theta^*)}, \quad \forall\ \theta \in \mathcal{F}_{n,\epsilon} \quad \text{such that} \quad h^2(\theta||\theta^*) \geq \epsilon^2.$$

Assumption 2. There exists a constant $C > 0$ such that

$$P_\theta \left[\mathcal{B}_n(\pi^*, \epsilon_n)\right] \geq e^{-Cn\epsilon_n^2},$$

$$P_\theta \left[\mathcal{B}_n(\beta^*, \epsilon_n)\right] \geq e^{-Cn\epsilon_n^2}.$$

Assumption 1 is the statistical identifiability condition characterized by the test function condition (see Ghosal and Van Der Vaart (2007)). Since we restrict the parameter space to a compact set, such a test function exists and Assumption 1 is automatically satisfied.

Assumption 2 is the prior concentration condition. With Assumptions 1 and 2 in the main text, it can be shown that Assumption 2 here is satisfied.

Under Assumption 2, with a similar proof as the proof of (S1.5) in Theorem 1, we can show that for any $D > 1$, there exists an event $\mathcal{A}_n$ such that

$$P_{\theta^*}(\mathcal{A}_n) \geq 1 - \frac{1}{2(D-1)^2 n\epsilon_n^2},$$

and there exist variational distributions $(q_\theta^*, q_{\boldsymbol{\mathcal{X}}}^*)$, such that under event $\mathcal{A}_n$,

$$\begin{aligned}\Psi_n(q_\theta^*, q_{\boldsymbol{\mathcal{X}}}^*) &\leq 2Dn(n-1)T\epsilon_n^2 - \log P_\pi\left(\mathcal{B}_n(\pi^*, \epsilon_\pi)\right) - \log P_\beta\left(\mathcal{B}_n(\beta^*, \epsilon_\beta)\right) \\ &\leq 2Dn(n-1)T\epsilon_n^2 + 2C\epsilon_n^2,\end{aligned}$$

where Ψ_n denotes the regular VB objective function.

Under Assumption 1, by Theorem 3.5 of Yang et al. (2020), for any

$\epsilon \geq \epsilon_n$, there exists an event $\mathcal{B}_\epsilon$, such that

$$P_{\theta^*}(\mathcal{B}_\epsilon) \geq 1 - 2e^{-cn(n-1)T\epsilon_n^2},$$

and under event $\mathcal{B}_\epsilon$, we have the following upper bound to the variational Bayes risk for any $(q_\theta, q(\boldsymbol{\mathcal{X}}))$ in the variational family

$$\begin{aligned}
&\hat{Q}_\theta(\mathcal{F}_{n,\epsilon}^c) \log \frac{\hat{Q}_\theta(\mathcal{F}_{n,\epsilon}^c)}{P_\theta(\mathcal{F}_{n,\epsilon}^c)} + (1 - \hat{Q}_\theta(\mathcal{F}_{n,\epsilon}^c)) \log \frac{1 - \hat{Q}_\theta(\mathcal{F}_{n,\epsilon}^c)}{1 - P_\theta(\mathcal{F}_{n,\epsilon}^c)} \\
&+ cn(n-1)T \int_{\theta \in \mathcal{F}_{n,\epsilon}, h^2(\theta||\theta^*) \geq \epsilon^2} h^2(\theta||\theta^*) \hat{Q}_\theta(d\theta) \\
\leq &\Psi_n(q_\theta, q_{\boldsymbol{\mathcal{X}}}) + \frac{cn(n-1)T\epsilon_n^2}{2} + \log 2, \qquad \text{(S3.6)}
\end{aligned}$$

where $\hat{Q}_\theta(\cdot)$ is the probability measure corresponding to the VB solution $\hat{q}_\theta$.

Thus, under the event $\mathcal{A}_n \cap \mathcal{B}_\epsilon$, we have

$$\begin{aligned}
&\hat{Q}_\theta(\mathcal{F}_{n,\epsilon}^c) \log \frac{\hat{Q}_\theta(\mathcal{F}_{n,\epsilon}^c)}{P_\theta(\mathcal{F}_{n,\epsilon}^c)} + (1 - \hat{Q}_\theta(\mathcal{F}_{n,\epsilon}^c)) \log \frac{1 - \hat{Q}_\theta(\mathcal{F}_{n,\epsilon}^c)}{1 - P_\theta(\mathcal{F}_{n,\epsilon}^c)} \\
&+ cn(n-1)T \int_{\theta \in \mathcal{F}_{n,\epsilon}, h^2(\theta||\theta^*) \geq \epsilon^2} h^2(\theta||\theta^*) \hat{Q}_\theta(d\theta) \\
\leq &Cn(n-1)T\epsilon_n^2, \qquad \text{(S3.7)}
\end{aligned}$$

where $C > 0$ is a constant.

Note that both the sum of the first terms and the third term in the left hand side of (S3.7) are nonnegative, so there exist constants C', C'', such

that

$$\hat{Q}_\theta(\theta \in \mathcal{F}_{n,\epsilon}, h^2(\theta||\theta^*) \geq \epsilon^2) \leq \frac{1}{\epsilon^2} \int_{\theta \in \mathcal{F}_{n,\epsilon}, h^2(\theta||\theta^*) \geq \epsilon^2} h^2(\theta||\theta^*) \hat{Q}_\theta(d\theta) \leq C' \frac{\epsilon_n^2}{\epsilon^2},$$
$$\hat{Q}_\theta(\mathcal{F}_{n,\epsilon}^c) \leq C'' \frac{\epsilon_n^2}{\epsilon^2}.$$

The inequality in the second expression above is due to the following facts:

$$\hat{Q}_\theta(\mathcal{F}_{n,\epsilon}^c) \log \hat{Q}_\theta(\mathcal{F}_{n,\epsilon}^c) + (1 - \hat{Q}_\theta(\mathcal{F}_{n,\epsilon}^c)) \log(1 - \hat{Q}_\theta(\mathcal{F}_{n,\epsilon}^c)) \geq -\log 2,$$
$$-\hat{Q}_\theta(\mathcal{F}_{n,\epsilon}^c) \log \hat{P}_\theta(\mathcal{F}_{n,\epsilon}^c) - (1 - \hat{Q}_\theta(\mathcal{F}_{n,\epsilon}^c)) \log(1 - \hat{P}_\theta(\mathcal{F}_{n,\epsilon}^c)) \geq -\hat{Q}_\theta(\mathcal{F}_{n,\epsilon}^c) \log \hat{P}_\theta(\mathcal{F}_{n,\epsilon}^c)$$
$$\geq cn(n-1)T\epsilon^2.$$

Let $\epsilon = k\epsilon_n$ for $k = 1, 2, \ldots, \left\lfloor e^{cn(n-1)T\epsilon_n^2/4} \right\rfloor$, we can show that the following inequality holds with probability at least $1 - \frac{1}{2(D-1)^2 n\epsilon_n^2} - 2e^{-cn(n-1)T\epsilon_n^2/4} \geq 1 - \frac{1}{(D-1)^2 n\epsilon_n^2}$,

$$\hat{Q}_\theta(h^2(\theta||\theta^*) \geq \epsilon^2) \leq \hat{Q}_\theta(\theta \in \mathcal{F}_{n,\epsilon}, h^2(\theta||\theta^*) \geq \epsilon^2) + \hat{Q}_\theta(\mathcal{F}_{n,\epsilon}^c) \leq (C' + C'') \frac{\epsilon_n^2}{\epsilon^2}.$$

Let $\epsilon = n^{1/4}\epsilon_n$, then

$$\hat{Q}_\theta(h^2(\theta||\theta^*) \geq \epsilon^2) \leq (C' + C'') \frac{\epsilon_n^2}{\epsilon^2} \to 0.$$

Therefore, for any $R < e^{2cn(n-1)T\epsilon_n^2}$, the variational Bayes risk

$$\begin{aligned}
\int_{\{h^2(\theta||\theta^*)\le R^2\}} h^2(\theta||\theta^*)\hat{q}_\theta(d\theta) &= \int_0^{R^2} \hat{Q}_\theta(h^2(\theta||\theta^*) \ge t)dt \\
&= \int_0^{\epsilon_n^2} \hat{Q}_\theta(h^2(\theta||\theta^*) \ge t)dt + \int_{\epsilon_n^2}^{R^2} \hat{Q}_\theta(h^2(\theta||\theta^*) \ge t)dt \\
&\le \epsilon_n^2 + \int_{\epsilon_n^2}^{R^2} \hat{Q}_\theta(h^2(\theta||\theta^*) \ge t)dt \\
&= \epsilon_n^2 + 2\int_{\epsilon_n}^{R} s\hat{Q}_\theta(h^2(\theta||\theta^*) \ge s^2)ds \\
&\le \epsilon_n^2 + 2\int_{\epsilon_n}^{R} s \cdot (C' + C'')\frac{\epsilon_n^2}{s^2}ds \\
&\le C\epsilon_n^2(1 + \log\frac{R}{\epsilon_n}).
\end{aligned}$$

S4 Additional Simulation: VB vs MCMC

In this simulation we compare the performance of the proposed VB algorithm with a Metropolis within Gibbs MCMC sampling algorithm. We used parallel tempering to facilitate the mixing of MCMC. Details of this MCMC algorithm are given in the next section. To make the computation feasible for MCMC, we simulated 20 dynamic networks, each with only $n = 50$ nodes and $T = 10$ time points. We considered two cases for the variance of the transition distribution: $\tau^2 = 0.0004$ for the small transition case and $\tau^2 = 0.01$ for the large transition case. We considered networks with different edge density by setting $\beta = 0.5, -0.5, -1.5$ for the dense,

moderate and sparse cases, respectively.

The parallel tempering algorithm took about half an hour on average to obtain 100,000 samples (10,000 for burn-in and 90,000 for inference), while the proposed VB algorithm only took several seconds. The performance was evaluated by the AUC values of in-sample predictions. The results of the two algorithms are summarized in Figure 1. Their performances in terms of AUC values are close. Compared with MCMC, the variational algorithm achieves similar performance with much less computation time.

S5 Details of the Parallel Tempering Algorithm

We used the following parallel tempering algorithm in the comparison between VB and MCMC in Section S4.

- For $k = 1, \ldots, K$, construct target distributions $\pi_k \propto \exp\left\{\frac{\log(p(\boldsymbol{\mathcal{Y}}|\beta,\boldsymbol{\mathcal{X}})p(\boldsymbol{\mathcal{X}}|\sigma^2,\tau^2)p(\beta))}{T_k}\right\}$, where $T_K > \cdots > T_1$ are the temperatures. The lowest temperature $T_1 = 1$ corresponds to the target distribution we are interested in.
- Initialize $\boldsymbol{\mathcal{X}}_k^{(0)}$ and $\beta_k^{(0)}$ randomly for $k = 1, \ldots, K$. Here the subscript k indicates the sample in the k-th temperature level.
- Suppose the sample at step t is $(\boldsymbol{\mathcal{X}}_k^{(t)}, \beta_k^{(t)})$, $k = 1, \ldots, K$.
  - Draw $u \sim \text{Uniform}(0, 1)$.

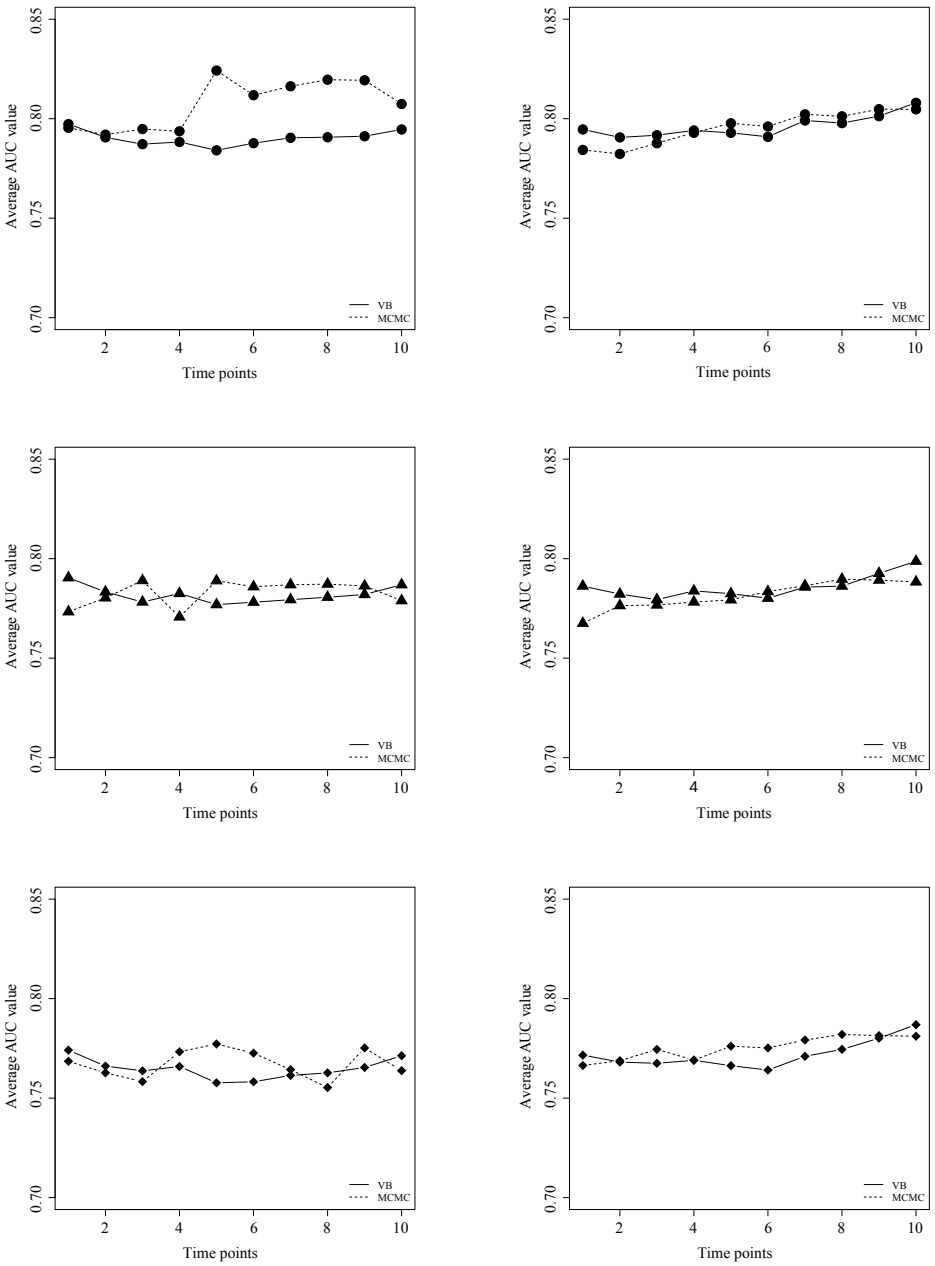

Figure 1: The average in-sample AUC values for VB and MCMC on simulated networks. (Left column: small transition; right column: large transition. First row: dense networks; second row: moderate networks; third row: sparse networks.)

– If $u \leq a_0$, for $k = 1, \ldots, K$, draw $(\boldsymbol{\mathcal{X}}_k^{(t+1)}, \beta_k^{(t+1)})$ using the Metropolis-Hastings within Gibbs algorithm (given below) with target distribution π_k.

– Otherwise, randomly choose a neighboring pair of temperatures T_i and T_{i+1}, and swap $(\boldsymbol{\mathcal{X}}_i^{(t)}, \beta_i^{(t)})$ and $(\boldsymbol{\mathcal{X}}_{i+1}^{(t)}, \beta_{i+1}^{(t)})$ with probabilty $\min\left\{1, \frac{\pi_i\left(\boldsymbol{\mathcal{X}}_{i+1}^{(t)}, \beta_{i+1}^{(t)}\right)\pi_{i+1}\left(\boldsymbol{\mathcal{X}}_i^{(t)}, \beta_i^{(t)}\right)}{\pi_i\left(\boldsymbol{\mathcal{X}}_i^{(t)}, \beta_i^{(t)}\right)\pi_{i+1}\left(\boldsymbol{\mathcal{X}}_{i+1}^{(t)}, \beta_{i+1}^{(t)}\right)}\right\}$.

In our simulations, we set the number of temperatures $K = 3$, and the three temperatures are $T_1 = 1$, $T_2 = 10$ and $T_3 = 20$. The tuning parameter $a_0 = 0.9$.

Now we give the detail of the Metropolis-Hastings within Gibbs algorithm for temperature $T_1 = 1$. The algorithms for other temperatures are similar and will be omitted. Recall the likelihood function of the model and the priors are given by

$$p(\boldsymbol{\mathcal{Y}}|\beta, \boldsymbol{\mathcal{X}}) = \prod_{t=1}^{T}\prod_{i\neq j}\frac{e^{Y_{ijt}(\beta - ||\boldsymbol{X}_{it} - \boldsymbol{X}_{jt}||^2)}}{1 + e^{\beta - ||\boldsymbol{X}_{it} - \boldsymbol{X}_{jt}||^2}},$$

$$p(\boldsymbol{\mathcal{X}}|\sigma^2, \tau^2) \propto \prod_{i=1}^{n}\left(e^{-\frac{||X_{i1}||^2}{2\sigma^2}} \cdot \prod_{t=2}^{T} e^{-\frac{||X_{it} - X_{i(t-1)}||^2}{2\tau^2}}\right), \quad p(\beta) \propto e^{-\frac{(\beta-\xi)^2}{2\psi^2}}.$$

The full conditional distributions of the latent variables and the intercept are given by

- For $t = 1$, $i = 1, \ldots, n$:

$$p(\boldsymbol{X}_{i1}|\cdot) \propto \left(\prod_{i=1}^{n} e^{-\frac{||\boldsymbol{X}_{i1}||^2}{2\sigma^2} - \frac{||\boldsymbol{X}_{i2}-\boldsymbol{X}_{i1}||^2}{2\tau^2}} \right) \cdot \left(\prod_{i \neq j} \frac{e^{Y_{ij1}(\beta - ||\boldsymbol{X}_{i1}-\boldsymbol{X}_{j1}||^2)}}{1 + e^{\beta - ||\boldsymbol{X}_{i1}-\boldsymbol{X}_{j1}||^2}} \right)$$

- For $t = 2, \ldots, T$:

$$p(\boldsymbol{X}_{it}) \propto \left(\prod_{i=1}^{n} e^{-\frac{||\boldsymbol{X}_{it}-\boldsymbol{X}_{i(t-1)}||^2 + ||\boldsymbol{X}_{i(t+1)}-\boldsymbol{X}_{it}||^2}{2\tau^2}} \right) \cdot \left(\prod_{i \neq j} \frac{e^{Y_{ijt}(\beta - ||\boldsymbol{X}_{it}-\boldsymbol{X}_{jt}||^2)}}{1 + e^{\beta - ||\boldsymbol{X}_{it}-\boldsymbol{X}_{jt}||^2}} \right)$$

- $$p(\beta|\cdot) \propto e^{-\frac{(\beta-\xi)^2}{2\psi^2}} \cdot \left(\prod_{t=1}^{T} \prod_{i \neq j} \frac{e^{Y_{ijt}(\beta - ||\boldsymbol{X}_{it}-\boldsymbol{X}_{jt}||^2)}}{1 + e^{\beta - ||\boldsymbol{X}_{it}-\boldsymbol{X}_{jt}||^2}} \right)$$

All of these full conditional distributions are drawn via Metropolis-Hastings with normal random walk proposals. In order to resolve the non-identifiability issue associated with latent space models, we perform a Procrustes transformation after we draw a new set of $\{\boldsymbol{X}_{it}\}$'s.

S6 Details on the Implementation of the Algorithm in Simulation Studies

In the simulation of Section S4, the initial latent positions were drawn from a mixture Gaussian distribution with equal probability on two components centered at $(-0.5, 0)$ and $(0.5, 0)$, respectively, and the variance of both components was set to be $\sigma^2 = 0.5$. In the simulation of Section 5, the initial latent positions were drawn from a mixture Gaussian distribution

with equal probability on two components centered at $(-1.5, 0)$ and $(1.5, 0)$, respectively, and the variance of both components was set to be $\sigma^2 = 0.5$. The variance of the transition distribution was set to be $\tau^2 = 0.01$ for the small transition case and $\tau^2 = 0.16$ for the large transition case.

The edge density of the network can be controlled by the intercept β. In the simulation of Section S4, we set $\beta = 0.5$, -0.5, -1.5 for the dense, moderate and sparse cases, respectively. The corresponding edge density is around 0.24, 0.10 and 0.06, respectively. The prior distribution for β was set to be $\mathcal{N}(0, 2)$. In the simulation of Section 5, the average degree of dense, moderate and sparse networks are approximately 7.5, 4, and 1.8, respectively. The prior for β was set to be $\mathcal{N}(0, 2)$.

The VB algorithm requires initial values for the variational parameters. In the simulation of Section S4, the initial values of $\{\tilde{\boldsymbol{\mu}}_{it}\}$ ($i = 1, \ldots, 50$, $t = 1, \ldots, 10$) were obtained through multi-dimensional scaling (MDS). The initial value for the covariance matrix $\tilde{\Sigma}$ was set to be the identity matrix $\mathbb{I}_2$. The initial values for $\tilde{\xi}$ and $\tilde{\psi}$ were 0 and 2, respectively. In the simulation of Section 5, the variational parameters $\{\tilde{\boldsymbol{\mu}}_{it}\}$ ($i = 1, \ldots, n$, $t = 1, \ldots, T$) were randomly initialized. The initial values of other variational parameters were set to be the same as the simulation of Section S4.

S7 Simulation for Networks with 5000 Nodes

We carried out simulation studies for networks with $n = 5000$ nodes under two settings. To control the density of the networks, we set the intercept $\beta = -2.5$ for the dense case, and $\beta = -4.5$ for the sparse case. All other settings were the same as the simulation in Section 5. The prior for β was set to be $\mathcal{N}(0, 0.01)$. The variational parameters $\{\tilde{\boldsymbol{\mu}}_{it}\}$ $(i = 1, \ldots, n,$ $t = 1, \ldots, T)$ were still randomly initialized. The average AUC values and their standard errors are given in Table 1. We can see that the variational method still performed well with these large networks, and the performance on dense networks is better than sparse ones.

Time	1	2	3	4	5	6	7	8	9	10
Dense,	0.8496	0.8488	0.8488	0.8495	0.8497	0.8489	0.8507	0.8536	0.8545	0.8582
small τ^2	(0.0021)	(0.0034)	(0.0028)	(0.0025)	(0.0031)	(0.0020)	(0.0023)	(0.0026)	(0.0014)	(0.0015)
Sparse,	0.7561	0.7564	0.7580	0.7608	0.7635	0.7663	0.7676	0.7699	0.7725	0.7762
large τ^2	(0.0015)	(0.0025)	(0.0024)	(0.0020)	(0.0018)	(0.0014)	(0.0013)	(0.0011)	(0.0024)	(0.0014)

Table 1: The average AUC values and standard errors (in parentheses) for VB on simulated networks with $n = 5000$ nodes.

S8 Additional Simulation: the Effect of α in α-VB

In this simulation, we studied the effect of α in the α-VB algorithm. While the authors provided some simulation results for the α-VB algorithm in

Yang et al. (2020), we focus on the effect of the choice of α in the dynamic latent space model.

In the α-VB framework, the upper bound of the KL-divergence is given by

$$\begin{aligned} D_\alpha \leq & -\frac{nT}{2}\log(\det(\tilde{\Sigma})) + \left(\frac{n}{2\sigma^2} + \frac{n(T-1)}{\tau^2}\right)\mathrm{tr}(\tilde{\Sigma}) \\ & + \frac{1}{2\sigma^2}\sum_{i=1}^{n}\tilde{\boldsymbol{\mu}}_{i1}^T\tilde{\boldsymbol{\mu}}_{i1} + \frac{1}{2\tau^2}\sum_{t=2}^{T}\sum_{i=1}^{n}(\tilde{\boldsymbol{\mu}}_{it}-\tilde{\boldsymbol{\mu}}_{i(t-1)})^T(\tilde{\boldsymbol{\mu}}_{it}-\tilde{\boldsymbol{\mu}}_{i(t-1)}) \\ & + \frac{1}{2\alpha}\left(\frac{\tilde{\psi}^2}{\psi^2} - \log\frac{\tilde{\psi}^2}{\psi^2} + \frac{(\tilde{\xi}-\xi)^2}{\psi^2}\right) - \sum_{t=1}^{T}\sum_{i\neq j}\Bigg\{Y_{ijt}\left(\tilde{\xi} - 2\,\mathrm{tr}(\tilde{\Sigma}) - ||\tilde{\boldsymbol{\mu}}_{it}-\tilde{\boldsymbol{\mu}}_{jt}||^2\right) \\ & - \log\left(1 + \frac{\exp\{\tilde{\xi}+\frac{1}{2}\tilde{\psi}^2\}}{\det(\mathbb{I}+4\tilde{\Sigma})^{1/2}}\cdot\exp\{-(\tilde{\boldsymbol{\mu}}_{it}-\tilde{\boldsymbol{\mu}}_{jt})^T(\mathbb{I}+4\tilde{\Sigma})^{-1}(\tilde{\boldsymbol{\mu}}_{it}-\tilde{\boldsymbol{\mu}}_{jt})\}\right)\Bigg\} + \textit{constant}. \end{aligned}$$

Note that the update equations for variational parameter $\tilde{\boldsymbol{\mu}}$ and $\tilde{\Sigma}$ stay the same. The update equations for $\tilde{\xi}$ and $\tilde{\psi}^2$ are listed as follows.

- Update of $\tilde{\xi}$:

$$\tilde{\xi}^{(s+1)} \leftarrow \left(1+\alpha\psi^2 f''(\tilde{\xi}^{(s)})\right)^{-1}\left[\xi + \alpha\psi^2\left(\sum_{t=1}^{T}\sum_{i\neq j}Y_{ijt} + f''(\tilde{\xi}^{(s)})\tilde{\xi}^{(s)} - f'(\tilde{\xi}^{(s)})\right)\right].$$

- Update of $\tilde{\psi}^2$: $\tilde{\psi}^{2\ (s+1)} \leftarrow \left(\frac{1}{\psi^2} + 2\alpha f'(\tilde{\psi}^{2\ (s)})\right)^{-1}$.

In this simulation study, we tried four different α values: 0.2, 0.5, 0.9, 1.0. The proposed VB algorithm corresponds to the $\alpha = 1.0$ case. For each case, twenty datasets were simulated, each with $n = 100$ and $T = 10$. The data generating process was the same as the one for dense networks with

small transition in the simultion in Section 5. The average AUC values for each choice of α are given in Table 2.

T	1	2	3	4	5	6	7	8	9	10
$\alpha = 0.2$	0.8961	0.8952	0.8960	0.8968	0.8971	0.8965	0.8970	0.8982	0.8984	0.9016
$\alpha = 0.5$	0.8961	0.8952	0.8960	0.8968	0.8971	0.8965	0.8970	0.8982	0.8984	0.9016
$\alpha = 0.9$	0.8961	0.8952	0.8960	0.8968	0.8971	0.8965	0.8970	0.8982	0.8984	0.9017
$\alpha = 1.0$	0.8961	0.8952	0.8960	0.8968	0.8971	0.8965	0.8970	0.8982	0.8984	0.9016

Table 2: The average AUC values given by the α-VB algorithm with four different choices of α.

As shown in Table 2, the performance of the α-VB algorithm for the dynamic latent space model is not very sensitive to the choice of α. This is due to the fact that the α-VB penalization is only used on the intercept β here, and the majority of variational parameters related to the latent positions are not affected.

As mentioned in Yang et al. (2020), in general one may want to choose an α value that is close to 1 in practice (e.g., $\alpha = 0.9$). In this case, the algorithm will enjoy theoretical guarantees without requiring extra assumptions, and at the same time the parameter estimation will be close to the $\alpha = 1$ case.

S9 Addition Simulation: Asymptotic Behaviors

In this simulation study, we first verified the consistency of parameter estimation of the proposed VB algorithm as the number of nodes in the network goes to infinity. Note that both VB and MCMC are schemes for approximating the posterior distribution, but the true posterior distribution is unknown. Our theoretical results indicate that we can compare the VB estimate with the true parameter value.

In all simulations, the true value of the intercept $\beta = -2$. Other than this, the data generating process was the same as the one for dense networks with small transitions in Section 5. The average AUC values and mean squared errors (MSEs) ($\mathbb{E}[(\hat{\beta} - \beta)^2]$) are given in Tables 3 and 4. As the sample size increases, the estimation accuracy of the model parameter β improves.

T	1	2	3	4	5	6	7	8	9	10
$n = 100$	0.6657	0.6297	0.6623	0.6407	0.6282	0.6297	0.6364	0.6724	0.6614	0.6632
$n = 200$	0.7343	0.7299	0.7389	0.7341	0.7346	0.7381	0.7322	0.7478	0.7482	0.7560
$n = 400$	0.7612	0.7572	0.7582	0.7553	0.7576	0.7679	0.7651	0.7663	0.7683	0.7743
$n = 800$	0.7562	0.7583	0.7586	0.7625	0.7648	0.7635	0.7667	0.7682	0.7704	0.7727

Table 3: The average AUC values given by the VB algorithm with increasing number of nodes.

n	100	200	400	800
MSE	0.4567	0.0876	0.0228	0.0052

Table 4: The MSEs of the intercept β with increasing number of nodes.

We also ran another simulation study to explore the behavior of the proposed algorithm when the number of time steps $T \to \infty$. In this simulation, we fixed the number of nodes $n = 50$, and tried several different T values. For each case, we calculated the average AUC value for all snapshots. Results in Table 5 are based on 20 simulations. The performance of the proposed algorithm does not change much as the number of time steps increases.

T	10	20	40	80	160	320
Average AUC	0.8987	0.8979	0.9024	0.9021	0.9078	0.9136

Table 5: The average AUC values given by the VB algorithm with increasing number of time steps.

S10 Teenage Friendship Network Data

Figure 2 shows the networks formed by the 129 pupils who were present at all three measurement time points from the "Teenage Friends and Lifestyle Study" dataset. This network data is analyzed in Section 6.1.

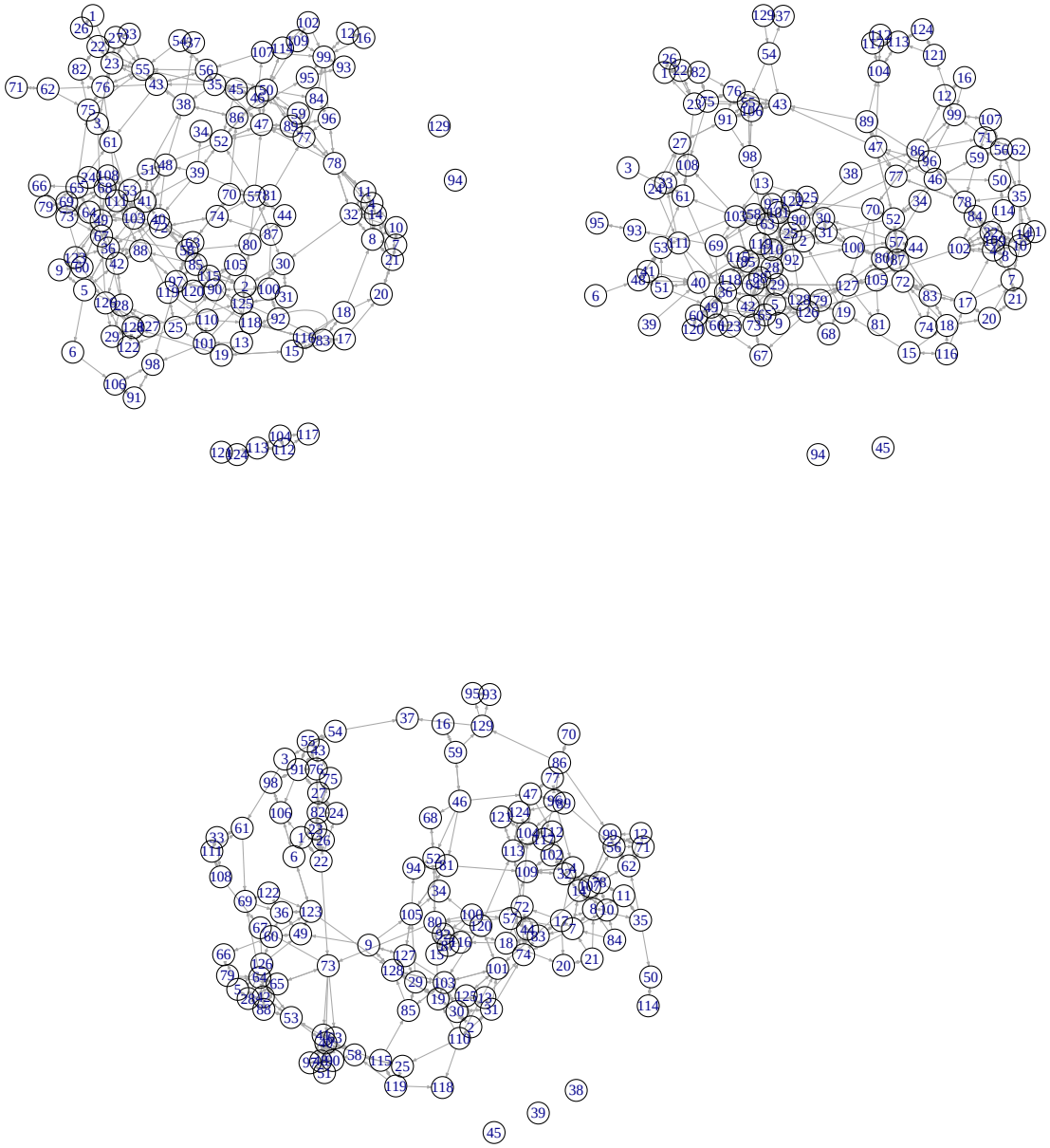

Figure 2: Friendship networks from "Teenage Friends and Lifestyle Study" at times 1, 2 (top) and 3 (bottom).

Bibliography

Ghosal, S. and Van Der Vaart, A. (2007), "Convergence rates of posterior distributions for noniid observations," *The Annals of Statistics*, 35, 192–223.

Yang, Y., Pati, D., and Bhattacharya, A. (2020), "α-Variational Inference with Statistical Guarantees," *The Annals of Statistics*, 48, 886–905.