01:XXX:XXX - Homework n

Pranav Tikkawar

June 29, 2025

1 Model Formulation

Consider a directed graph $G(V, E)$ with components $\{C_i\}_{i=1}^n$ over discrete time $t \in \{1, \ldots, T\}$. Each component C_i has:

- State space $S^{(i)} = \{s_1^{(i)}, \ldots, s_{m_i}^{(i)}\}$
- Time-homogeneous transition matrix $P^{(i)}$ where $P_{jk}^{(i)} = \mathbb{P}(C_i(t) = k \mid C_i(t-1) = j)$

Edge weights $w_{ij} \in \mathbb{R}$ quantify dependency strength. The **aggregated influence** on C_j is:

$$W_j = \sum_{i \in \mathcal{I}_j} w_{ij}, \quad \mathcal{I}_j = \{i : (C_i, C_j) \in E\} \tag{1}$$

The adjusted transition matrix $\widetilde{P}^{(j)}$ is defined per row k:

$$\widetilde{P}_{kk}^{(j)} = \sigma\left(\text{logit}\left(P_{kk}^{(j)}\right) + W_j\right) \tag{2}$$

$$\widetilde{P}_{kl}^{(j)} = P_{kl}^{(j)} \cdot \frac{1 - \widetilde{P}_{kk}^{(j)}}{1 - P_{kk}^{(j)}}, \quad l \neq k \tag{3}$$

where $\sigma(z) = (1 + e^{-z})^{-1}$ and $\text{logit}(x) = \log(x/(1-x))$.

2 Parameter Estimation Framework

2.1 Data Requirements

The estimation requires:

- Component state trajectories $\mathcal{D} = \{\mathbf{s}(t)\}_{t=0}^T$
- Graph topology $G(V, E)$
- Transition records with $T \gg \max_j |S^{(j)}|^2$ for reliability

2.2 Maximum Likelihood Estimation

2.2.1 Base Transition Probabilities ($P^{(i)}$)

The MLE for $P_{jk}^{(i)}$ is derived from the categorical likelihood:

$$\mathcal{L}(P^{(i)}) = \prod_{j=1}^{m_i} \prod_{k=1}^{m_i} (P_{jk}^{(i)})^{N_{jk}^{(i)}} \tag{4}$$

where $N_{jk}^{(i)}$ counts $j \to k$ transitions. Maximizing under $\sum_k P_{jk}^{(i)} = 1$ yields:

$$\hat{P}_{jk}^{(i)} = \frac{N_{jk}^{(i)}}{\sum_{l=1}^{m_i} N_{jl}^{(i)}} \tag{5}$$

Proof: The Lagrangian is:

$$\mathcal{L} = \sum_{j,k} N_{jk}^{(i)} \ln P_{jk}^{(i)} + \sum_j \lambda_j \left(1 - \sum_k P_{jk}^{(i)}\right) \tag{6}$$

Setting $\partial\mathcal{L}/\partial P_{jk}^{(i)} = 0$ gives $P_{jk}^{(i)} = N_{jk}^{(i)}/\lambda_j$. The constraint implies $\lambda_j = \sum_l N_{jl}^{(i)}$.

2.2.2 Edge Weights (w_{ij})

The complete-data log-likelihood is:

$$\mathcal{L}(\mathbf{w}) = \sum_{t=1}^{T-1} \sum_{j=1}^{|V|} \ln \widetilde{P}_{s_j(t),s_j(t+1)}^{(j)}(\mathbf{w}) \tag{7}$$

The gradient components are:

$$\frac{\partial \mathcal{L}}{\partial w_{ij}} = \sum_{t=1}^{T-1} \begin{cases} 1 - \widetilde{P}_{k_t k_t}^{(j)}(\mathbf{w}) & \text{if } s_j(t+1) = k_t \\ -\dfrac{\widetilde{P}_{k_t k_t}^{(j)}(\mathbf{w}) \widetilde{P}_{k_t l_t}^{(j)}(\mathbf{w})}{P_{k_t l_t}^{(j)}} & \text{otherwise} \end{cases} \tag{8}$$

where $k_t = s_j(t)$. Optimization uses gradient ascent:

$$\mathbf{w}^{(n+1)} = \mathbf{w}^{(n)} + \gamma_n \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}^{(n)}) \tag{9}$$

2.3 Expectation-Maximization Algorithm for Incomplete Data

2.3.1 Setup

When states are partially observed:

- Observed data: $\mathbf{Y} = \{\mathbf{s}(t)\}_{t \in \mathcal{O}}$
- Latent variables: $\mathbf{Z} = \{\mathbf{s}(t)\}_{t \notin \mathcal{O}}$
- Complete-data likelihood: $\mathcal{L}_c(\theta) = \mathbb{P}(\mathbf{Y}, \mathbf{Z} \mid \theta)$

2.3.2 Algorithm

1. **Initialize**: $\theta^{(0)} = (\hat{P}^{(i)}, \mathbf{w}^{(0)})$
2. **E-step**: Compute
$$Q(\theta \mid \theta^{(n)}) = \mathbb{E}_{\mathbf{Z}|\mathbf{Y},\theta^{(n)}}[\ln \mathcal{L}_c(\theta)] \tag{10}$$
using forward-backward algorithm for expected transition counts:
$$\bar{N}_{kl}^{(j)} = \sum_{t=1}^{T-1} \mathbb{P}(s_j(t) = k, s_j(t+1) = l \mid \mathbf{Y}, \theta^{(n)}) \tag{11}$$

3. **M-step**: Update parameters

$$\hat{P}_{jk}^{(i)} = \frac{\bar{N}_{jk}^{(i)}}{\sum_l \bar{N}_{jl}^{(i)}} \tag{12}$$

$$\mathbf{w}^{(n+1)} = \underset{\mathbf{w}}{\operatorname{argmax}}\ Q(\theta \mid \theta^{(n)}) \tag{13}$$

4. **Iterate** until $|\mathcal{L}(\theta^{(n+1)}) - \mathcal{L}(\theta^{(n)})| < \epsilon$

2.3.3 Theoretical Guarantees

Theorem 1 (Monotonicity)**.** *$\mathcal{L}(\theta^{(n+1)}) \geq \mathcal{L}(\theta^{(n)})$ with equality iff $\theta^{(n)}$ is stationary point.*

Proof: By Jensen's inequality and properties of conditional expectation.

3 Theoretical Properties

3.1 Consistency

Under regularity conditions:

$$\hat{\theta}_T \xrightarrow{p} \theta_0 \quad \text{as} \quad T \to \infty \tag{14}$$

where θ_0 is the true parameter vector.

3.2 Asymptotic Normality

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}^{-1}(\theta_0)) \tag{15}$$

where $\mathcal{I}(\theta)$ is the Fisher information matrix.

4 Implementation Considerations

4.1 Regularization

- **Base probabilities**: Dirichlet prior with parameter α

$$\tilde{P}_{jk}^{(i)} = \frac{N_{jk}^{(i)} + \alpha}{\sum_l N_{jl}^{(i)} + m_i \alpha} \tag{16}$$

- **Edge weights**: L_2 regularization

$$\mathcal{L}_{\text{reg}}(\mathbf{w}) = \mathcal{L}(\mathbf{w}) - \lambda \|\mathbf{w}\|_2^2 \tag{17}$$

4.2 Identifiability Conditions

For unique parameter identification:

1. $\widetilde{P}^{(j)}(\mathbf{w}_1) = \widetilde{P}^{(j)}(\mathbf{w}_2)\, \forall j \implies \mathbf{w}_1 = \mathbf{w}_2$
2. Each component has $\deg_{\text{in}}(C_j) \geq 1$
3. $G(V, E)$ is directed acyclic

5 Conclusion

This framework provides statistically rigorous estimation for stochastic graph models. The MLE and EM algorithms offer efficient parameter recovery with theoretical guarantees, enabling application to reliability analysis and networked systems.

6 Model Formulation

Consider a directed graph $G(V, E)$ representing a system of n components over discrete time $t = 1, 2, \ldots, T$. Each component $C_i \in V$ has:

- State space $S^{(i)} = \{s_1^{(i)}, s_2^{(i)}, \ldots, s_{m_i}^{(i)}\}$
- Time-homogeneous transition matrix $P^{(i)} = [P_{jk}^{(i)}]$ where:

$$P_{jk}^{(i)} = \mathbb{P}(C_i(t) = s_k^{(i)} \mid C_i(t-1) = s_j^{(i)}) \tag{18}$$

Edges $(C_i, C_j) \in E$ carry weights $w_{ij} \in \mathbb{R}$ quantifying dependency strength. The **aggregated influence** on component C_j is:

$$W_j = \sum_{i \in \mathscr{I}_j} w_{ij}, \quad \mathscr{I}_j = \{i : (C_i, C_j) \in E\} \tag{19}$$

The adjusted transition matrix $\widetilde{P}^{(j)}$ is defined per row k as:

$$\widetilde{P}_{kk}^{(j)} = \sigma\left(\text{logit}\left(P_{kk}^{(j)}\right) + W_j\right) \tag{20}$$

$$\widetilde{P}_{kl}^{(j)} = P_{kl}^{(j)} \cdot \frac{1 - \widetilde{P}_{kk}^{(j)}}{1 - P_{kk}^{(j)}}, \quad l \neq k \tag{21}$$

where $\sigma(z) = (1 + e^{-z})^{-1}$ is the sigmoid function and $\text{logit}(x) = \log(x/(1-x))$.

7 Parameter Estimation Framework

7.1 Data Requirements

The estimation requires:

- **State trajectories**: $\mathscr{D} = \{\mathbf{s}(t)\}_{t=0}^{T}$ where $\mathbf{s}(t) = (s_1(t), \ldots, s_n(t))$ and $s_j(t) \in S^{(j)}$
- **Graph topology**: Directed graph $G(V, E)$
- **Transition records**: Documented state transitions with $T \gg \max_j |S^{(j)}|^2$ for reliability

7.2 Maximum Likelihood Estimation

7.2.1 Base Transition Probabilities ($P^{(i)}$)

The likelihood for component C_i's transitions follows a categorical distribution:

$$\mathscr{L}(P^{(i)}) = \prod_{j=1}^{m_i} \prod_{k=1}^{m_i} (P_{jk}^{(i)})^{N_{jk}^{(i)}} \tag{22}$$

where $N_{jk}^{(i)}$ counts observed transitions from state j to k.

Theorem 2 (MLE for $P^{(i)}$)**.** *The maximum likelihood estimator is:*

$$\hat{P}^{(i)}_{jk} = \frac{N^{(i)}_{jk}}{\sum_{l=1}^{m_i} N^{(i)}_{jl}} \tag{23}$$

Proof. Maximize the log-likelihood $\ell(P^{(i)}) = \sum_{j,k} N^{(i)}_{jk} \ln P^{(i)}_{jk}$ subject to $\sum_k P^{(i)}_{jk} = 1$. Introduce Lagrange multipliers λ_j:

$$\mathscr{L} = \sum_{j,k} N^{(i)}_{jk} \ln P^{(i)}_{jk} + \sum_j \lambda_j \left(1 - \sum_k P^{(i)}_{jk}\right) \tag{24}$$

Taking derivatives:

$$\frac{\partial \mathscr{L}}{\partial P^{(i)}_{jk}} = \frac{N^{(i)}_{jk}}{P^{(i)}_{jk}} - \lambda_j = 0 \implies P^{(i)}_{jk} = \frac{N^{(i)}_{jk}}{\lambda_j} \tag{25}$$

$$\sum_k P^{(i)}_{jk} = \frac{1}{\lambda_j} \sum_k N^{(i)}_{jk} = 1 \implies \lambda_j = \sum_k N^{(i)}_{jk} \tag{26}$$

Substituting yields (23). The Hessian matrix is negative semi-definite, confirming a maximum. □

7.2.2 Edge Weights (w_{ij})

The complete-data log-likelihood is:

$$\ell(\mathbf{w}) = \sum_{t=1}^{T-1} \sum_{j=1}^{n} \ln \widetilde{P}^{(j)}_{s_j(t), s_j(t+1)}(\mathbf{w}) \tag{27}$$

where $\widetilde{P}^{(j)}$ depends on $\mathbf{w} = \{w_{ij}\}_{(i,j)\in E}$ through (20) and (21).

Theorem 3 (Gradient of Log-Likelihood)**.** *The gradient component for edge (i, j) is:*

$$\frac{\partial \ell}{\partial w_{ij}} = \sum_{t=1}^{T-1} \begin{cases} 1 - \widetilde{P}^{(j)}_{k_t k_t}(\mathbf{w}) & \text{if } s_j(t+1) = k_t \\ -\dfrac{\widetilde{P}^{(j)}_{k_t k_t}(\mathbf{w}) \widetilde{P}^{(j)}_{k_t l_t}(\mathbf{w})}{P^{(j)}_{k_t l_t}} & \text{otherwise} \end{cases} \tag{28}$$

where $k_t = s_j(t)$, $l_t = s_j(t+1)$.

Proof. Consider two cases for each transition $s_j(t) \to s_j(t+1)$:

Case 1: $s_j(t+1) = s_j(t) = k_t$

$$\frac{\partial}{\partial w_{ij}} \ln \widetilde{P}^{(j)}_{k_t k_t} = \frac{1}{\widetilde{P}^{(j)}_{k_t k_t}} \cdot \frac{\partial}{\partial w_{ij}} \sigma(\text{logit}(P^{(j)}_{k_t k_t}) + W_j) \tag{29}$$

$$= \frac{1}{\widetilde{P}^{(j)}_{k_t k_t}} \cdot \sigma'(\cdot) \cdot 1 \tag{30}$$

$$= 1 - \widetilde{P}^{(j)}_{k_t k_t} \tag{31}$$

since $\sigma'(z) = \sigma(z)(1 - \sigma(z))$.

Case 2: $s_j(t+1) = l_t \neq k_t$

$$\frac{\partial}{\partial w_{ij}} \ln \widetilde{P}^{(j)}_{k_t l_t} = \frac{1}{\widetilde{P}^{(j)}_{k_t l_t}} \cdot P^{(j)}_{k_t l_t} \frac{\partial}{\partial w_{ij}} \left(\frac{1 - \widetilde{P}^{(j)}_{k_t k_t}}{1 - P^{(j)}_{k_t k_t}} \right) \tag{32}$$

$$= \frac{1}{\widetilde{P}^{(j)}_{k_t l_t}} \cdot P^{(j)}_{k_t l_t} \cdot \frac{-\frac{\partial \widetilde{P}^{(j)}_{k_t k_t}}{\partial w_{ij}}}{1 - P^{(j)}_{k_t k_t}} \tag{33}$$

$$= -\frac{\widetilde{P}^{(j)}_{k_t k_t} \widetilde{P}^{(j)}_{k_t l_t}}{P^{(j)}_{k_t l_t}} \tag{34}$$

Summing over time points gives (28). □

Optimization is performed via gradient ascent:

$$\mathbf{w}^{(n+1)} = \mathbf{w}^{(n)} + \gamma_n \nabla_{\mathbf{w}} \ell(\mathbf{w}^{(n)}) \tag{35}$$

with adaptive step size γ_n.

7.3 Expectation-Maximization Algorithm for Incomplete Data

7.3.1 Problem Setup

When state observations are incomplete:

- Observed data: $\mathscr{Y} = \{\mathbf{s}(t)\}_{t \in \mathscr{O}}$
- Latent variables: $\mathscr{Z} = \{\mathbf{s}(t)\}_{t \notin \mathscr{O}}$
- Complete data: $(\mathscr{Y}, \mathscr{Z})$

7.3.2 Algorithm Derivation

The complete-data log-likelihood is:

$$\ell_c(\theta) = \sum_{t=1}^{T-1} \sum_{j=1}^{n} \ln \widetilde{P}^{(j)}_{s_j(t), s_j(t+1)}(\mathbf{w}) \tag{36}$$

1. **Initialization**: Set $\theta^{(0)} = (\hat{P}^{(i)}, \mathbf{w}^{(0)})$
2. **E-step**: Compute expected log-likelihood

$$Q(\theta \mid \theta^{(n)}) = \mathbb{E}_{\mathscr{Z} \mid \mathscr{Y}, \theta^{(n)}}[\ell_c(\theta)] \tag{37}$$

Key quantities are expected transition counts:

$$\bar{N}_{kl}^{(j)} = \sum_{t=1}^{T-1} \mathbb{P}(s_j(t) = k, s_j(t+1) = l \mid \mathscr{Y}, \theta^{(n)}) \tag{38}$$

Computed via the forward-backward algorithm:

$$\alpha_t(j,k) = \mathbb{P}(\mathscr{Y}_{1:t}, s_j(t) = k) \tag{39}$$

$$\beta_t(j,k) = \mathbb{P}(\mathscr{Y}_{t+1:T} \mid s_j(t) = k) \tag{40}$$

$$\mathbb{P}(s_j(t) = k, s_j(t+1) = l \mid \mathscr{Y}) \propto \alpha_t(j,k) \widetilde{P}_{kl}^{(j)}(\mathbf{w}^{(n)}) \beta_{t+1}(j,l) \tag{41}$$

3. **M-step**: Update parameters

$$\hat{P}_{jk}^{(i)} = \frac{\bar{N}_{jk}^{(i)}}{\sum_l \bar{N}_{jl}^{(i)}} \tag{42}$$

$$\mathbf{w}^{(n+1)} = \underset{\mathbf{w}}{\operatorname{argmax}} Q(\theta \mid \theta^{(n)}) \tag{43}$$

solved via gradient ascent on Q

4. **Convergence**: Stop when $|\ell(\theta^{(n+1)}) - \ell(\theta^{(n)})| < \epsilon$

Theorem 4 (Monotonicity of EM)**.** *$\ell(\theta^{(n+1)}) \geq \ell(\theta^{(n)})$ with equality iff $\theta^{(n)}$ is a stationary point.*

Proof. By Jensen's inequality:

$$\ell(\theta) - \ell(\theta^{(n)}) \geq Q(\theta \mid \theta^{(n)}) - Q(\theta^{(n)} \mid \theta^{(n)}) \tag{44}$$

$$\ell(\theta^{(n+1)}) - \ell(\theta^{(n)}) \geq Q(\theta^{(n+1)} \mid \theta^{(n)}) - Q(\theta^{(n)} \mid \theta^{(n)}) \geq 0 \tag{45}$$

since $\theta^{(n+1)}$ maximizes $Q(\cdot \mid \theta^{(n)})$. □

8 Theoretical Properties

8.1 Consistency

Assumption (Regularity Conditions)**.** A1: Parameter space Θ is compact
A2: True parameter $\theta_0 \in \text{int}(\Theta)$
A3: Model is identifiable
A4: $\ell(\theta)$ is continuous in θ and differentiable in $\text{int}(\Theta)$
A5: Dominated convergence applies to derivatives

Theorem 5 (Consistency)**.** *Under A1-A5, the MLE satisfies:*

$$\hat{\theta}_T \xrightarrow{p} \theta_0 \quad as \quad T \to \infty \tag{46}$$

8.2 Asymptotic Normality

Theorem 6 (Asymptotic Normality)**.** *Under regularity conditions:*

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathscr{I}^{-1}(\theta_0)) \tag{47}$$

where $\mathscr{I}(\theta)$ is the Fisher information matrix:

$$\mathscr{I}(\theta)_{ij} = -\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j}\right] \tag{48}$$

8.3 Identifiability Conditions

For unique parameter identification:

1. **Observational equivalence**: $\widetilde{P}^{(j)}(\mathbf{w}_1) = \widetilde{P}^{(j)}(\mathbf{w}_2)\, \forall j \implies \mathbf{w}_1 = \mathbf{w}_2$
2. **Connectivity**: $\deg_{\text{in}}(C_j) \geq 1$ for all j
3. **Acyclicity**: $G(V, E)$ is directed acyclic

9 Implementation Considerations

9.1 Regularization

- **Base probabilities**: Dirichlet prior

$$\tilde{P}^{(i)}_{jk} = \frac{N^{(i)}_{jk} + \alpha}{\sum_l N^{(i)}_{jl} + m_i \alpha}, \quad \alpha > 0 \tag{49}$$

- **Edge weights**: L_2 regularization

$$\ell_{\text{reg}}(\mathbf{w}) = \ell(\mathbf{w}) - \lambda \|\mathbf{w}\|_2^2, \quad \lambda > 0 \tag{50}$$

9.2 Computational Complexity

- **Gradient computation**: $\mathscr{O}(|E| \cdot T \cdot \max_j m_j^2)$ per iteration
- **Forward-backward**: $\mathscr{O}(T \cdot \max_j m_j^2)$ per component per E-step
- **Parallelization**: Component-wise independence allows distributed computation

9.3 Validation Metrics

- **Brier score**:

$$\text{BS} = \frac{1}{T-1} \sum_{t=1}^{T-1} \sum_{j=1}^{n} \|\mathbf{e}_{s_j(t+1)} - \widetilde{\mathbf{p}}^{(j)}(t)\|^2 \tag{51}$$

  where $\widetilde{\mathbf{p}}^{(j)}(t)$ is predicted state distribution

- **Transition KL-divergence**:

$$D_{KL} = \sum_j \sum_k \hat{\pi}_k^{(j)} \sum_l \hat{P}_{kl}^{(j)} \ln \frac{\hat{P}_{kl}^{(j)}}{\widetilde{P}_{kl}^{(j)}} \tag{52}$$

  with $\hat{\pi}^{(j)}$ empirical state occupancy

10 Conclusion

This framework provides a mathematically rigorous foundation for parameter estimation in stochastic graph transition models. The MLE derivation offers statistically efficient estimators, while the EM algorithm extends applicability to partially observed systems. Theoretical guarantees ensure reliability, and implementation strategies address practical challenges in complex systems.