MT-BN: Multi-Scale Topological Bayesian Networks

Hierarchical Structure Learning with Topological Priors

Aaryan Arora

MT-BN Research Project

October 19, 2025

Talk Roadmap

Preliminaries: Bayesian networks, scores, and why flat learning fails at scale

Problem Introduction: $p \gg n$, interpretability gaps, and missed topology

MT-BN Pipeline (Novelty): multi-layer partitions $\Rightarrow$ inter/intra DAGs $\Rightarrow$ VB CPDs & posteriors

Visuals throughout: hierarchy diagrams, priors, and an end-to-end inputs$\rightarrow$outputs schematic

Bayesian Networks: Definition

Let $X \in \mathbb{R}^{n \times p}$ be the data (rows=samples, cols=variables). A BN defines

$$p(X_1, \ldots, X_p) \; = \prod_{i=1}^{p} p\big(X_i \mid X_{\mathrm{Pa}(i)}\big)\,, \quad X_i = \sum_{j \in \mathrm{Pa}(i)} \beta_{ji} X_j + \varepsilon_i,\; \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2).$$

Scores (continuous data):

- BIC: local MLE + penalty; add/remove/reverse edges to increase total score.
- BGe: Bayesian linear-Gaussian marginal likelihood integrating out (β, σ^2).

Why Flat BN Learning is Intractable

- Number of possible directed graphs on p nodes is $2^{\binom{p}{2}}$ (acyclic subset still super-exponential).
- Heuristics (HC, GES, MCMC) struggle when $p \gtrsim 500$; mixing degrades rapidly.
- In the common regime $p \gg n$, local tests and model selection are noisy $\Rightarrow$ spurious edges.

Flat Search over p nodes
$2^{\binom{p}{2}}$ candidates

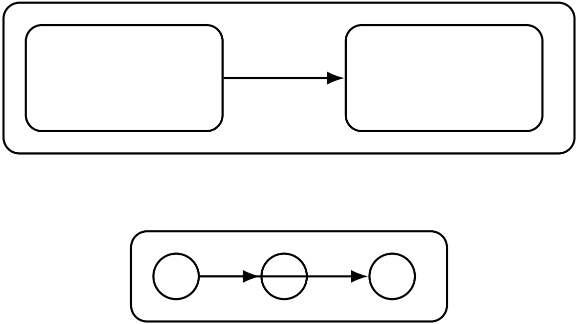
Natural Topology in Real Systems

Modularity: variables cluster into functional groups (pathways, sectors, communities).

Hubs: degree is heterogeneous (few high-degree regulators, many low-degree nodes).

Multi-scale: modules of modules $\Rightarrow$ layered structure.

Layer 1 (modules)



Layer 2 (inside a module)

Topological Priors We Will Use

nCRP for hierarchical partitions: data-driven layers & module membership.

SBM prior for edges: $\theta_{\text{in}} \gg \theta_{\text{out}}$ favors dense intra-/sparse inter-module wiring.

Log-normal degree prior: encourages hub heterogeneity.

Depth prior on L: geometric/Poisson to avoid over-fragmentation.

Notation (aligned to this project)

$X \in \mathbb{R}^{n \times p}$, p variables, n samples.

Layers $\ell = 1, \ldots, L$; partition at layer ℓ: $P^{(\ell)} = \{C_1^{(\ell)}, \ldots, C_{M_\ell}^{(\ell)}\}$.

Graphs: $G^{(\ell)}$ over modules at layer ℓ; within-module $G_m^{(\ell+1)}$ over $C_m^{(\ell)}$.

Module signals: $S^{(\ell)} \in \mathbb{R}^{n \times M_\ell}$ via PC1 (or mean) per module.

Scoring: BGe/BIC; Indegree cap d_{max}.

Priors: $\text{SBM}(\theta_{\text{in}}, \theta_{\text{out}}) + \text{LogNormal}(\mu, \sigma) +$ depth prior $p(L)$.

Limits of Traditional BN Learning

Computational: Super-exponential state space; infeasible beyond modest p.

Statistical: $p \gg n$ yields unstable tests and weakly identified structures.

Interpretability: Flat graphs ignore hierarchy $\Rightarrow$ hard to read, low modular insight.

Missed topology: No native encoding of modularity or hub structure.

Design Goals for a Better Learner

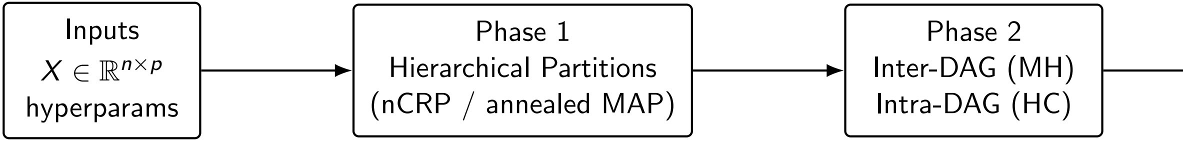
Exploit hierarchy: learn layers & modules from data (not pre-fixed).

Shrink search: operate on small graphs (modules; small within-module DAGs).

Encode topology: priors for blocks and degrees prune implausible edges.

Quantify confidence: posterior edge probabilities & parameter uncertainty.

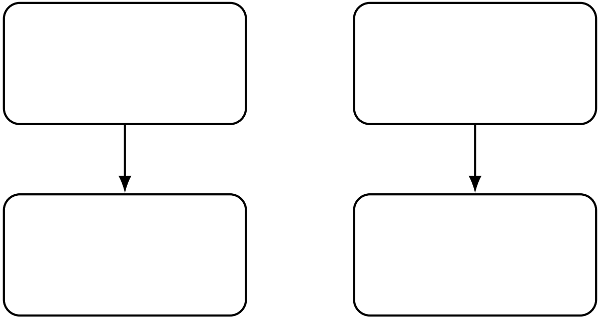
Big Picture: From Data to Multi-Scale DAGs



Phase 1: Hierarchical Partitions (nCRP)

Goal: infer L and $P^{(\ell)}$ jointly from data.

- Variables choose paths down a tree: "rich-get-richer" with concentration α_ℓ.
- Practical inference: annealed MAP or collapsed Gibbs; enforce min module size.
- Stop when $K_\ell = M_\ell \leq 1$ or modules too small / marginal gain $< \varepsilon$.



Phase 2 (a): Inter-Module Graph by MH

Data: module signals $S^{(\ell)}$ (PC1 or mean).
Posterior for a candidate inter-graph A:

$$\log p(A \mid S) \;=\; \underbrace{\log p(S \mid A)}_{\text{BGe/BIC}} + \underbrace{\log p_{\mathsf{SBM}}(A)}_{\theta_{\mathsf{in}},\theta_{\mathsf{out}}} + \underbrace{\sum_i \log p_{\mathsf{LN}}(d_i)}_{\mu,\sigma}.$$

MH update: toggle $u \to v$ if acyclic and indegree $\leq d_{\max}$; accept by $\min\{1, \exp(\Delta \log p)\}$.

Phase 2 (b): Within-Module Graphs by HC

For each module $C_m^{(\ell)}$ (small $k = |C_m^{(\ell)}|$):

- Enumerate legal single-edge add/remove/reverse moves (acyclic, indegree cap).
- Score each by $\Delta S = \Delta(\mathrm{BGe}/\mathrm{BIC}) + \Delta \log p_{\mathsf{SBM}} + \Delta \log p_{\mathsf{LN}}$.
- Take best positive move; stop after *patience* iterations without improvement.

Phase 3: Variational Bayes for CPDs

With graphs fixed, fit linear-Gaussian CPDs via mean-field VB.

$$q(\beta, \sigma^2) = \prod_i \mathcal{N}(\beta_i \mid m_i, \Sigma_i)\, \mathrm{InvGamma}(\sigma_i^2 \mid a_i, b_i),$$

$$\mathrm{ELBO} = \mathbb{E}_q[\log p(X, \beta, \sigma^2 \mid G)] - \mathbb{E}_q[\log q(\beta, \sigma^2)].$$

Outputs:

Posterior means/covariances for coefficients; $\mathbb{E}[\sigma_i^2]$ per node.

Edge posteriors: thin snapshots over outer iterations to estimate inclusion probs.

Acyclicity & Constraints

No self-loops; reject proposals that create cycles (reachability check).

Global indegree cap $d_{\max}$ for identifiability and speed.

Cross-layer consistency: within-module edges only among members; inter-edges only among modules.

Search-Space Reduction (Key Theorem)

Let $n = \sum_{k=1}^{K} |C_k|$, $m = \max_k |C_k|$. Then the MT-BN structure space factorizes:

$$|\mathcal{G}_{\mathrm{MT}}(K, m)| \;\le\; |\mathcal{G}_K| \cdot |\mathcal{G}_m|^K,$$

and

$$\frac{|\mathcal{G}_{\mathrm{MT}}(K, m)|}{|\mathcal{G}_n|} \;\le\; \exp\Big\{ -\tfrac{1}{2}\big[n(n-1) - K(K-1) - Km(m-1)\big] \log 2 \Big\},$$

showing an exponential shrinkage vs. flat BN when $m \ll n$ and $K \ll n$.

Putting It Together: The MT-BN Orchestrator

Outer iterations $t = 1, \dots, T$:

- Infer/update hierarchy up to $L_{\max}$ (stop if $K_\ell \leq 1$ or modules too small).
- For each layer: MH on inter-graph; HC within modules.
- Score log-joint; save thinned snapshots post burn-in.

End: VB parameter refinement + edge posterior aggregation.

Configuration Knobs (from our implementation)

Depth prior: geometric/Poisson; $\alpha_{\text{per-layer}}$; `L_max`; `min_module_size`.

Likelihood: BGe (ridge hyperparams) or BIC; `max_indegree`.

Priors: $(\theta_{\text{in}}, \theta_{\text{out}})$, (μ, σ).

Signals: `pc1` or `mean`; schedules: MH steps, HC patience.

Complexity & Parallelism (Intuition)

Never search over all p at once: sum of small subproblems across layers/modules.

Inter-graphs: $O(M_\ell)$ proposals per sweep; within-modules: roughly $O(k^2)$ per HC iter.

Module-wise independence $\Rightarrow$ parallel within layers; caching local scores further speeds up.

Visual: Inputs $\rightarrow$ MT-BN $\rightarrow$ Outputs

Inputs
$X \in \mathbb{R}^{n \times p}$
Hyperparams

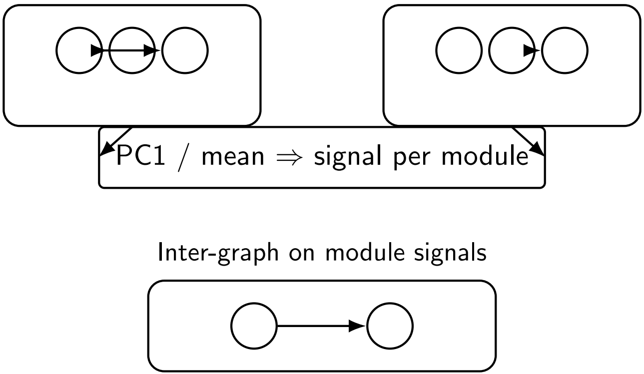
$\longrightarrow$

MT-BN Model
nCRP + SBM + LogNormal
BGe/BIC+ VB

$\longrightarrow$

Output
$L,\ P^{(\ell)},\ G^{(\ell)}$
Edge posteriors, CP

Visual: Module Signals and Inter-Graph



What You Take Away

MT-BN **discovers** hierarchy and **learns** structure at every layer.

Topological priors (SBM, log-normal degree) **prune** the search space.

Hybrid inference (**nCRP** $\rightarrow$ **MH** $\rightarrow$ **HC** $\rightarrow$ **VB**) **scales** and yields **edge posteriors**.

Outputs are **interpretable**: modules, cross-module flow, local hubs, and quantified confidence.

Questions?