

# Supplementary Material for Generalizing Tree Probability Estimation via Bayesian Networks

## A The SBN Representation of The Rooted Tree from Figure 1

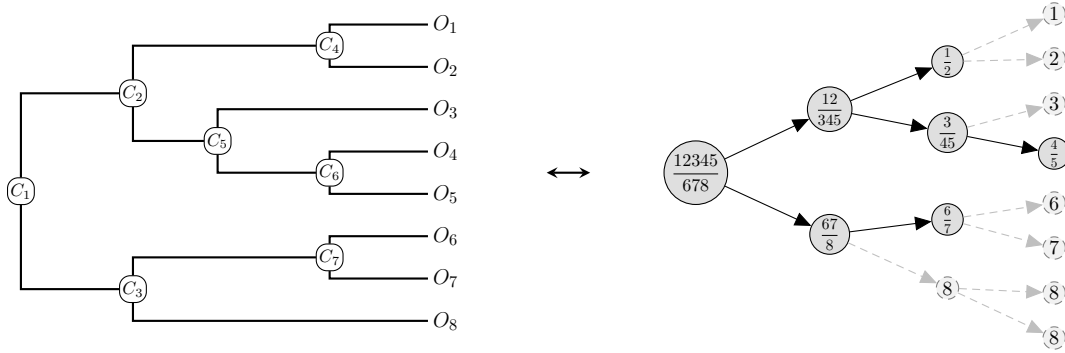


Figure A.1: The SBN representation of the rooted tree from Figure 1. For ease of presentation, we use the indexes of the leaf labels for subsplits. We also omit the fake splits after depth 4.

## B Proof of Lemma 1

*Proof.* It suffices to prove the lemma for the minimum SBN  $\mathcal{B}_{\mathcal{X}}^*$ . For any rooted tree  $T$ , with natural indexing, a compatible assignment is easy to obtain by following the splitting process and satisfying the compatibility requirement defined in Definition 2. Moreover, the depth of the splitting process is at most  $N - 1$ , where  $N$  is the size of  $\mathcal{X}$ . To show this, consider the maximum size of clades,  $d_i$ , in the  $i$ -th level of the process. The first level is for the root split, and  $d_1 \leq N - 1$  since the subclades need to be proper subset of  $\mathcal{X}$ . For each  $i > 1$ , if  $d_i > 1$ , the splitting process continues and the maximum size of clades in the next level is at most  $d_i - 1$ , that is  $d_{i+1} \leq d_i - 1$ . Note that the split process ends when  $d_i = 1$ , its depth therefore is at most  $N - 1$ . The whole process is deterministic, so each rooted tree can be uniquely represented as a compatible assignment of  $\mathcal{B}_{\mathcal{X}}^*$ . On the other hand, given a compatible assignment, we can simply remove all fake splits. The remaining net corresponds to the true splitting process of a rooted tree, which then can be simply restored.  $\square$

## C Proof of Proposition 1

*Proof.* Denote the set of all SBN assignments as  $\mathcal{A}$ . For any assignment  $a = \{S_i = s_i^a\}_{i \geq 1}$ , we have

$$p_{\text{sbn}}(a) = p(S_1 = s_1^a) \prod_{i > 1} p(S_i = s_i^a | S_{\pi_i} = s_{\pi_i}^a) > 0 \Rightarrow a \text{ is compatible.} \quad (1)$$

As a Bayesian network,

$$\sum_{a \in \mathcal{A}} p_{\text{sbn}}(a) = \sum_{s_1} p(S_1 = s_1) \prod_{i > 1} \sum_{s_i} p(S_i = s_i | S_{\pi_i} = s_{\pi_i}) = 1.$$

By Lemma 1,

$$\sum_T p_{\text{sbn}}(T) = \sum_{a \sim \text{compatible}} p_{\text{sbn}}(a) = \sum_{a \in \mathcal{A}} p_{\text{sbn}}(a) = 1.$$

□

## D Proof of Proposition 2

*Proof.*

$$\begin{aligned} \sum_{T^u} p_{\text{sbn}}(T^u) &= \sum_{T^u} \sum_{S_1 \sim T^u} p(S_1) \prod_{i>1} p(S_i | S_{\pi_i}) \\ &= \sum_T p_{\text{sbn}}(T) \\ &= 1. \end{aligned}$$

The second equality is due to the one-to-one correspondence between rooted trees and unrooted trees with specific rooting edges (compatible splits):

$$T \xLeftrightarrow{1\text{-to-1}} (T^u, S_1), S_1 \sim T^u.$$

□

## E Figure for Shared Parent-child Subsplit Pairs in SBNs

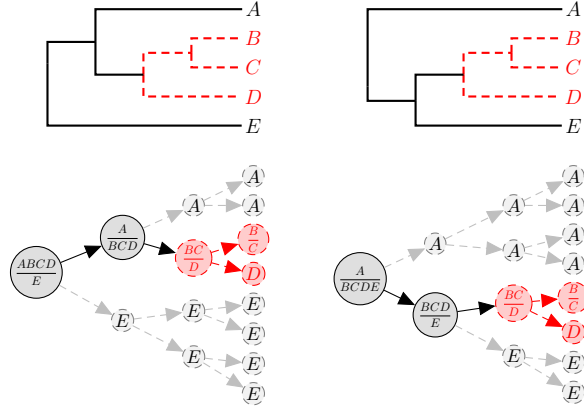


Figure E.1: Subsplit pairs shared among different nodes in SBNs. Top panels show the rooted trees that share the same local parent-child subsplit pair  $((B, C), D)$ . Bottom panels are their corresponding representations in SBNs. Just like in the trees, the same pair could appear at multiple locations in SBNs.

## F Related Work

The idea of using conditional probabilities associated with local tree structures (e.g., clades) to approximate the whole posterior distribution on tree topologies started with Höhna and Drummond [2012], where the authors extended the additive binary (AB) coding scheme [Farris et al., 1970, Brooks, 1981] to a new algorithm that approximates the posterior probabilities of trees by a product of conditional clade probabilities (CCP). However, the CCP method does not provide a true probability distribution on trees and requires renormalization to obtain the estimated probabilities. Therefore, CCP is not tractable for trees with many taxa except for restricting to a small subset of high probability trees. Larget [2013], further extended the CCP method by allowing joint modeling of all children clades of the parent clade in the conditional clade probabilities. The resulting conditional clade distribution (CCD) approach not only improves the approximation accuracy of CCP but also leads to

a valid probability distribution on the tree space. However, CCD is still not flexible enough to capture the complexity of inferred posterior distributions on real data [Whidden and Matsen, 2015].

SBNs also leverage parameter sharing to reduce the number of free parameters in the model and promote the generalization performance. Similar to weight sharing used in convolutional networks for detecting *translationally-invariant structure* of images (e.g., edges, corners), our heuristic parameter sharing in SBNs is for identifying *conditional splitting patterns* of leaf-labeled trees. Parameter sharing has been widely employed in probability graphical models such as hidden Markov models [Baum et al., 1970], dynamic Bayesian networks [Murphy, 2002], conditional random fields [Lafferty et al., 2001] and statistical relational models [Getoor and Taskar, 2007]. Other than specifying the tied parameters a priori, recent work [Chou et al., 2016, 2018] also utilizes quantization for automatic parameter tying.

## G On Expectation Maximization for SBNs

Using the conditional probabilities of the missing root node, we can construct the following adaptive lower bound

$$LB^{(n)}(T^u; p) = \sum_{S_1 \sim T^u} p(S_1 | T^u, \hat{p}^{EM, (n)}) \log \left( \frac{p(S_1) \prod_{i>1} p(S_i | S_{\pi_i})}{p(S_1 | T^u, \hat{p}^{EM, (n)})} \right) \leq \log L(T^u; p). \quad (2)$$

The above adaptive lower bound  $LB^{(n)}(T^u; p)$  contains a constant term that only depends on the current estimates, and another term that is the expected complete log-likelihood with respect to the conditional distribution of  $S_1$  given  $T^u$  and the current estimates:

$$Q^{(n)}(T^u; p) = \sum_{S_1 \sim T^u} p(S_1 | T^u, \hat{p}^{EM, (n)}) \left( \log p(S_1) + \sum_{i>1} \log p(S_i | S_{\pi_i}) \right). \quad (3)$$

Summing (3) over  $T^u \in \mathcal{D}^u$  together with conditional probability sharing, we get the complete data score function

$$\begin{aligned} Q^{(n)}(\mathcal{D}^u; p) &= \sum_{k=1}^K Q^{(n)}(T_k^u; p) \\ &= \sum_{k=1}^K \sum_{e \in E(T_k^u)} q_k^{(n)}(S_1 = s_{1,k}^e) \left( \log p(S_1 = s_{1,k}^e) + \sum_{i>1} \log p(S_i = s_{i,k}^e | S_{\pi_i} = s_{\pi_i,k}^e) \right) \\ &= \sum_{s_1 \in \mathcal{C}_r} \bar{m}_{s_1}^{u, (n)} \log p(S_1 = s_1) + \sum_{s|t \in \mathcal{C}_{ch|pa}} \bar{m}_{s,t}^{u, (n)} \log p(s|t) \end{aligned} \quad (4)$$

where

$$\begin{aligned} \bar{m}_{s_1}^{u, (n)} &= \sum_{k=1}^K \sum_{e \in E(T_k^u)} q_k^{(n)}(S_1 = s_1) \mathbb{I}(s_{i,k}^e = s_1) \\ \bar{m}_{s,t}^{u, (n)} &= \sum_{k=1}^K \sum_{e \in E(T_k^u)} q_k^{(n)}(S_1 = s_{1,k}^e) \sum_{i>1} \mathbb{I}(s_{i,k}^e = s, s_{\pi_i,k}^e = t) \end{aligned}$$

are the expected frequency counts.

### G.1 Proof of Theorem 1

*Proof.*

$$\begin{aligned} Q^{(n)}(T^u; p) - Q^{(n)}(T^u; \hat{p}^{EM, (n)}) &= LB^{(n)}(T^u; p) - LB^{(n)}(T^u; \hat{p}^{EM, (n)}) \\ &\leq \log L(T^u; p) - \log L(T^u; \hat{p}^{EM, (n)}) \end{aligned}$$

since  $LB^{(n)}(T^u; p) \leq \log L(T^u; p)$  and

$$\begin{aligned}
LB^{(n)}(T^u; \hat{p}^{\text{EM},(n)}) &= \sum_{S_1 \sim T^u} p(S_1|T^u, \hat{p}^{\text{EM},(n)}) \log \left( \frac{p(T^u, S_1|\hat{p}^{\text{EM},(n)})}{p(S_1|T^u, \hat{p}^{\text{EM},(n)})} \right) \\
&= \sum_{S_1 \sim T^u} p(S_1|T^u, \hat{p}^{\text{EM},(n)}) \log p(T^u|\hat{p}^{\text{EM},(n)}) \\
&= \log p(T^u|\hat{p}^{\text{EM},(n)}) \\
&= \log \sum_{S_1 \sim T^u} p(T^u, S_1|\hat{p}^{\text{EM},(n)}) \\
&= \log L(T^u; \hat{p}^{\text{EM},(n)}).
\end{aligned}$$

□

With Theorem 1, it is clear that

$$Q^{(n)}(\mathcal{D}^u; p) - Q^{(n)}(\mathcal{D}^u; \hat{p}^{\text{EM},(n)}) \leq \log L(\mathcal{D}^u; p) - \log L(\mathcal{D}^u; \hat{p}^{\text{EM},(n)}).$$

As before, the maximum estimates of  $Q^{(n)}$  have closed form expressions which lead to the following updating formula for the CPDs at the M-step

$$\hat{p}^{\text{EM},(n+1)}(S_1 = s_1) = \frac{\bar{m}_{s_1}^{u,(n)}}{\sum_{s \in \mathbb{C}_r} \bar{m}_s^{u,(n)}} = \frac{\bar{m}_{s_1}^{u,(n)}}{K}, \quad s_1 \in \mathbb{C}_r \quad (5)$$

$$\hat{p}^{\text{EM},(n+1)}(s|t) = \frac{\bar{m}_{s,t}^{u,(n)}}{\sum_s \bar{m}_{s,t}^{u,(n)}}, \quad s|t \in \mathbb{C}_{\text{ch}|pa}. \quad (6)$$

## G.2 Regularization

For the score function in (4), the conjugate prior distributions have to be Dirichlet distributions

$$S_1 \sim \text{Dir}(\alpha_{s_1} + 1, s_1 \in \mathbb{C}_r), \quad \cdot|t \sim \text{Dir}(\alpha_{s,t} + 1, s|t \in \mathbb{C}_{\text{ch}|pa})$$

where  $\alpha_{s_1} > 0$ ,  $\alpha_{s,t} > 0$  are some hyper-parameters. The regularized score function, therefore, takes the following form

$$\begin{aligned}
Q^{\text{R},(n)}(\mathcal{D}^u; p) &= Q^{(n)}(\mathcal{D}^u; p) + \sum_{s_1 \in \mathbb{C}_r} \alpha_{s_1} \log p(S_1 = s_1) + \sum_{s|t \in \mathbb{C}_{\text{ch}|pa}} \alpha_{s,t} \log p(s|t) \\
&= \sum_{s_1 \in \mathbb{C}_r} (\bar{m}_{s_1}^{u,(n)} + \alpha_{s_1}) \log p(S_1 = s_1) + \sum_{s|t \in \mathbb{C}_{\text{ch}|pa}} (\bar{m}_{s,t}^{u,(n)} + \alpha_{s,t}) \log p(s|t).
\end{aligned} \quad (7)$$

One can adapt the hyper-parameters according to this decomposition as follows

$$\alpha_{s_1} = \alpha \cdot \tilde{m}_{s_1}^u, \quad \alpha_{s,t} = \alpha \cdot \tilde{m}_{s,t}^u$$

where  $\alpha$  is the global regularization coefficient that balances the effect of the data on the estimates in comparison to the prior, and  $\tilde{m}_{s_1}^u, \tilde{m}_{s,t}^u$  are the equivalent sample counts for the root splits and parent-child subsplit pairs that distribute the regularization across the CPDs. In practice, one can simply use the sample frequency counts as equivalent sample counts as we did in our experiments.

**Remark:** The choice of hyper-parameter setting is not thoroughly studied and what we do here is some modification from the common practice in Bayesian estimation for learning Bayesian networks; see Buntine [1991] for more details.

Theorem 1 also implies that maximizing (or improving) the regularized  $Q$  score is sufficient to improve the regularized log-likelihood. Denote the regularization term as

$$R(p) = \sum_{s_1 \in \mathbb{C}_r} \alpha \tilde{m}_{s_1}^u \log p(S_1 = s_1) + \sum_{s|t \in \mathbb{C}_{\text{ch}|pa}} \alpha \tilde{m}_{s,t}^u \log p(s|t)$$

we have

$$\begin{aligned} Q^{R,(n)}(\mathcal{D}^u; p) - Q^{R,(n)}(\mathcal{D}^u; \hat{p}^{\text{EM},(n)}) &= Q^{(n)}(\mathcal{D}^u; p) - Q^{(n)}(\mathcal{D}^u; \hat{p}^{\text{EM},(n)}) + R(p) - R(\hat{p}^{\text{EM},(n)}) \\ &\leq \log L(\mathcal{D}^u; p) + R(p) - (\log L(\mathcal{D}^u; \hat{p}^{\text{EM},(n)}) + R(\hat{p}^{\text{EM},(n)})). \end{aligned}$$

Similarly, the regularized score  $Q^{R,(n)}$  can be maximized in the same manner as  $Q^{(n)}$

$$\hat{p}^{\text{EM},(n+1)}(S_1 = s_1) = \frac{\bar{m}_{s_1}^{u,(n)} + \alpha \tilde{m}_{s_1}^u}{\sum_{s_1 \in \mathbb{C}_r} (\bar{m}_{s_1}^{u,(n)} + \alpha \tilde{m}_{s_1}^u)} = \frac{\bar{m}_{s_1}^{u,(n)} + \alpha \tilde{m}_{s_1}^u}{K + \alpha \sum_{s_1 \in \mathbb{C}_r} \tilde{m}_{s_1}^u}, \quad s_1 \in \mathbb{C}_r \quad (8)$$

$$\hat{p}^{\text{EM},(n+1)}(s|t) = \frac{\bar{m}_{s,t}^{u,(n)} + \alpha \tilde{m}_{s,t}^u}{\sum_s (\bar{m}_{s,t}^{u,(n)} + \alpha \tilde{m}_{s,t}^u)}, \quad s|t \in \mathbb{C}_{\text{ch|pa}}. \quad (9)$$

### G.3 Computational Complexity

Compared to those for rooted trees, the expected frequency counts  $m_{s_1}^u, m_{s,t}^u, \bar{m}_{s_1}^{u,(n)}, \bar{m}_{s,t}^{u,(n)}$  for unrooted trees involves additional summation over all the edges. However, we can precompute all the accumulated sums of the root probabilities on the edges in a postorder tree traversal for each tree and the computation cost remains  $\mathcal{O}(KN)$ . Each E-step in EM would in general cost  $\mathcal{O}(KN^2)$  time since for each tree,  $\mathcal{O}(N)$  probability estimations are required and each takes  $\mathcal{O}(N)$  time. Notice that most of the intermediate partial products involved in those SBN probability estimates are shared due to the structure of trees, we therefore use a two-pass algorithm, similar to the one used in [Schadt et al., 1998], that computes all SBN probability estimates for each tree within two loops over its edges. This reduces the computational complexity of each E-step to  $\mathcal{O}(KN)$ . Overall, the computational complexities of maximum lower bound estimate and each EM iteration are both  $\mathcal{O}(KN)$ , the same as CCD and SBNs for rooted trees.

## H Run Time Comparison

In this section, we present a runtime comparison of different algorithms on simulated data with varying  $K$  and  $\beta$ . We see that SBN based methods significantly improve the approximation accuracy of CCD. With additional time budget, sbn-em and sbn-em- $\alpha$  further improve the performance of sbn-sa, and regularization is useful especially for diffuse distributions with limited sample size (Figure H.1). All experiments were done on a 2016 MacBook Pro (2.9 GHz Intel Core i5).

## References

- L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41(1): 164–171, 1970.
- D. R. Brooks. Hennig’s parasitological method: a proposed solution. *Syst. Zool.*, 30:733–767, 1981.
- W. Buntine. Theory refinement on Bayesian networks. In B. D’Ambrosio, P. Smets, and P. Bonissone, editors, *Proceedings of the 7th Conference on Uncertainty in Artificial Intelligence*, pages 52–60. Morgan Kaufmann, 1991.
- L. Chou, S. Sarkhel, N. Ruozi, and V. Gogate. On parameter tying by quantization. In *Proceedings of the Thirteenth AAAI Conference on Artificial Intelligence*, pages 3241–3247, 2016.
- L. Chou, P. Sahoo, S. Sarkhel, N. Ruozi, and V. Gogate. Automatic parameter sharing: A new approach for regularized parameter learning in Markov networks. In *Proceedings of the Fifteenth AAAI Conference on Artificial Intelligence*, 2018.
- J. S. Farris, A. G. Kluge, and M. J. Eckardt. A numerical approach to phylogenetic systematics. *Syst. Zool.*, 19:172–189, 1970.
- L. Getoor and B. Taskar. *Introduction to Statistical Relational Learning*. MIT Press, 2007.

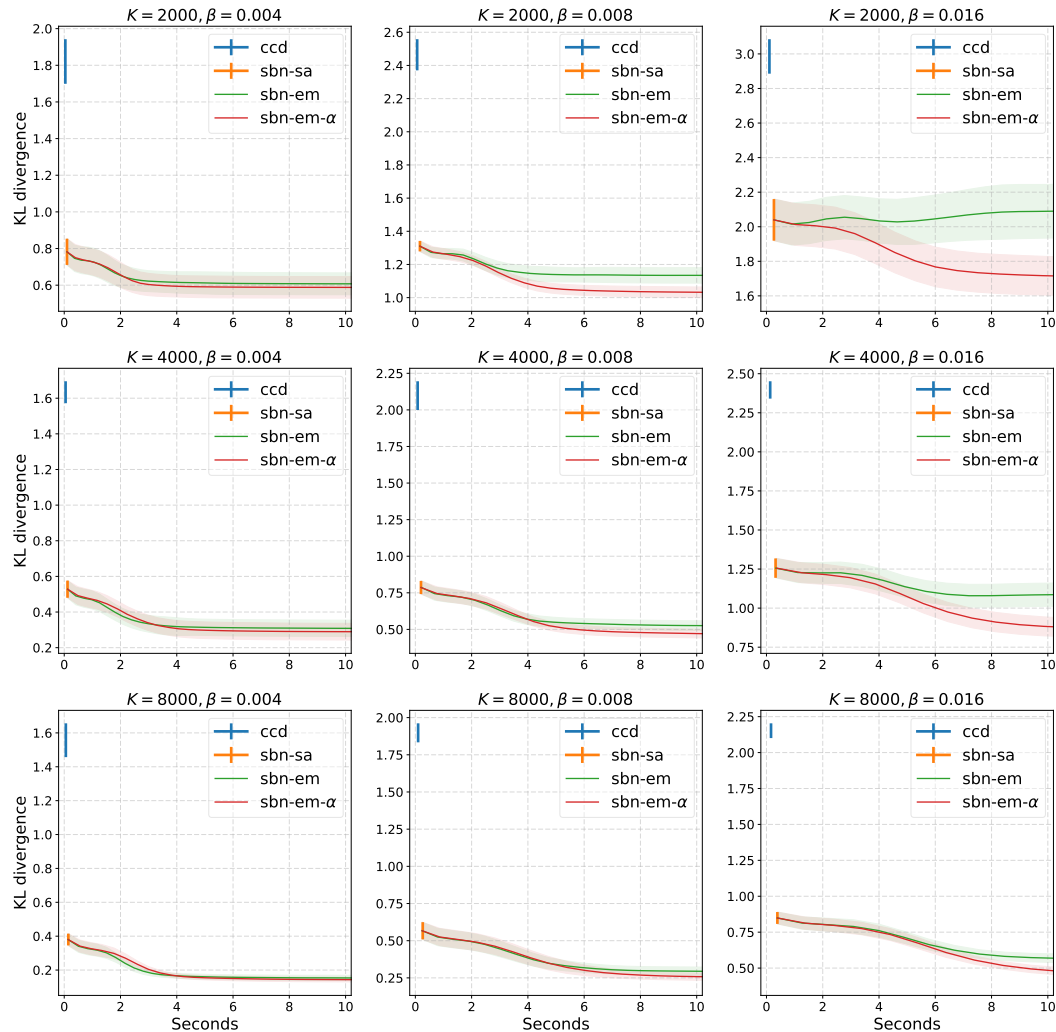


Figure H.1: A runtime comparison of different methods with varying  $K$  and  $\beta$  on simulated data. Error bar shows one standard deviation over 10 runs.

Sebastian Höhna and Alexei J. Drummond. Guided tree topology proposals for Bayesian phylogenetic inference. *Syst. Biol.*, 61(1):1–11, January 2012. ISSN 1063-5157. doi: 10.1093/sysbio/syr074. URL <http://dx.doi.org/10.1093/sysbio/syr074>.

J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, 2001.

Bret Larget. The estimation of tree posterior probabilities using conditional clade probability distributions. *Syst. Biol.*, 62(4):501–511, July 2013. ISSN 1063-5157. doi: 10.1093/sysbio/syt014. URL <http://dx.doi.org/10.1093/sysbio/syt014>.

K. Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, Computer Science Division, UC Berkeley, 2002.

Eric E. Schadt, Janet S. Sinsheimer, and Kenneth Lange. Computational advances in maximum likelihood methods for molecular phylogeny. *Genome Res.*, 8:222–233, 1998. doi: 10.1101/gr.8.3.222.

Chris Whidden and Frederick A Matsen, IV. Quantifying MCMC exploration of phylogenetic tree space. *Syst. Biol.*, 64(3):472–491, May 2015. ISSN 1063-5157, 1076-836X. doi: 10.1093/sysbio/syv006. URL <http://dx.doi.org/10.1093/sysbio/syv006>.