Supplemental Material

1 Tamura-Nei can be parameterized by an open set

For the Tamura-Nei model, we will show how the model can be parameterized by an open set. The model has rates α_0, α_1 and β and time t. For $i, j \in \{0, 1, 2, 3\}$, the transition probabilities are the following (see (13.11) in Felsenstein, 2004):

$$\mathbf{Pr}(j \mid i, t) = \exp(-(\alpha_{\ell} + \beta)t)\delta(i = j)$$

$$+ \exp(-\beta t)(1 - \exp(-\alpha_{\ell} t))\left(\frac{\pi(j)\epsilon(i, j)}{\sum_{k} \epsilon(j, k)\pi(k)}\right) + (1 - \exp(-\beta t))\pi(j),$$

where $\ell = \lfloor i/2 \rfloor$, δ is the standard Kronecker delta function, and $\epsilon(a,b)$ is an indicator function that is 1 if $a,b \in \{0,1\}$ or $a,b \in \{2,3\}$, and 0 otherwise. Setting

$$x_0 = \exp(-\alpha_0 t), \quad x_1 = \exp(-\alpha_1 t), \quad y = \exp(-\beta t),$$

we have

$$\mathbf{Pr}(j \mid i, t) = x_{\ell} y \delta(i = j) + y(1 - x_{\ell}) \left(\frac{\pi(j) \epsilon(i, j)}{\sum_{k} \epsilon(j, k) \pi(k)} \right) + (1 - y) \pi(j).$$

Hence, the transition probabilities can be expressed as multi-linear polynomials in x_0, x_1 and y.

2 Main theorems about maximum likelihood

Here we prove the following theorem about the Jukes-Cantor model.

Theorem 1. Consider the Jukes-Cantor model. For the two trees in Figure 2, let μ denote the mixture distribution defined by a uniform mixture (i.e., $p_1 = p_2$) of these two trees. Note μ is from the class U^2 (the set of uniform mixtures of size 2). Recall the maximum likelihood $\mathcal{L}_{\mu}(T)$ is over a homogenous model (i.e., the class M^1).

For all $C \in (0, 1/4)$, there exists x_0 , for all $x < x_0$, there exists $\alpha_c = \alpha_c(C, x) \in (0, 1/4)$ such that:

1. For $\alpha = \alpha_c$, the maximum likelihood of T_1 and T_3 are the same, i.e.,

$$\mathcal{L}_{T_1}(\mu) = \mathcal{L}_{T_2}(\mu).$$

2. For all $\alpha < \alpha_c$, T_3 is the maximum-likelihood tree, i.e.,

$$\mathcal{L}_{T_3}(\mu) > \mathcal{L}_{T_1}(\mu) \text{ and } \mathcal{L}_{T_3}(\mu) > \mathcal{L}_{T_2}(\mu).$$

For the CFN model we can prove ambiguity of likelihood at the critical point α_c for all x, and the maximum likelihood is on the "wrong tree" when $\alpha < \alpha_c$ and x is sufficiently small. Here is the formal statement of the theorem.

Theorem 2. Consider the CFN model. For the two trees in Figure 2, let μ denote the mixture distribution defined by a uniform mixture (i.e., $p_1 = p_2$) of these two trees. Note μ is from the class U^2 (the set of uniform mixtures of size 2). Recall the maximum likelihood $\mathcal{L}_{\mu}(T)$ is over a homogenous model (i.e., the class M^1).

For all $C \in (0, 1/2)$, there exists x_0 , for all $x < x_0$, there exists $\alpha_c = \alpha_c(C, x) \in (0, 1/2)$ such that:

1. For $\alpha = \alpha_c$, the maximum likelihood of T_1 and T_3 are the same, i.e.,

$$\mathcal{L}_{T_1}(\mu) = \mathcal{L}_{T_3}(\mu).$$

2. For all $\alpha < \alpha_c$, T_3 is the maximum-likelihood tree, i.e.,

$$\mathcal{L}_{T_3}(\mu) > \mathcal{L}_{T_1}(\mu) \text{ and } \mathcal{L}_{T_3}(\mu) > \mathcal{L}_{T_2}(\mu).$$

Remark 3. For the CFN model, we in fact prove part (1) of Theorem 2 holds for all $x \in (0, \min\{C, 1/2 - C\})$. This follows from Theorem 5.

Remark 4. We prove there exists at least one critical point α_c where part 1 holds (i.e., $\mathcal{L}_{T_1}(\mu) = \mathcal{L}_{T_3}(\mu)$), but there may be many such points α_c . In Section 2.1 we prove that

$$\alpha_c = \frac{x^2}{\frac{1}{4} - C - C^2 + x^2} \tag{1}$$

is such a critical point. (In fact, at that particular critical point the distribution is also non-identifiable.) Since there may be multiple critical points, Part 2 holds with respect to the smallest critical α_c , which we can not determine the exact value of.

We expect that the α_c in (1) is the unique critical point. However our proof methodology only uses the highest order terms of the likelihood function. Hence it is not detailed enough to prove the uniqueness of α_c .

Our proofs of the above theorems rely on technical tools developed in Štefankovič and Vigoda (2006). We first describe the relevant technical lemmas in Section 2.2, and then prove Theorems 1 and 2 in Sections 2.4 and 2.3 respectively. Before going into the proofs we formally present the non-identifiability result at the critical point.

2.1 Non-identifiabile mixture for CFN at α_c

We next present the proof of a non-identifiable mixture distribution in the CFN model. The result is a generalization of the example from Section , and also shows there exists non-uniform mixtures which are non-identifiable, as mentioned in Section . The theorem implies Part 1 of Theorem 2 as a special case.

In the following we describe the branch lengths on a 4-taxon tree T as a 5-dimensional vector w. For $1 \le i \le 4$, the i-th coordinate of w is the branch length of the edge incident to the leaf labeled i. The final coordinate of w is

the branch length of the internal edge of T.

In the following theorem, p is the mixing parameter. When $p \neq 1/2$ (i.e., it is not a uniform mixture), then the branch length of the internal edge will differ between the two trees.

Now we can formally state the theorem.

Theorem 5. Consider the CFN model. For any 0 < a, b < 1/2 and 0 , let

$$w = \frac{1}{2} - (a, b, b, a, \gamma)$$
 and $w' = \frac{1}{2} - (b, a, a, b, \delta),$

where

$$\gamma = \eta/p,$$

$$\delta = \eta/(1-p), \text{ and}$$

$$\eta = \frac{ab}{2(a^2+b^2)}.$$

Consider the following mixture distribution which is in the class M^2 :

$$\mu = p\mu(T_1, w) + (1 - p)\mu(T_1, w')$$

The distribution μ is invariant under the swapping of leaves 1 and 3. In

particular, for the mixture distribution (which is also in M^2)

$$\widehat{\mu} = p\mu(T_3, w) + (1 - p)\mu(T_3, w')$$

we have

$$\mu = \widehat{\mu}$$
.

Hence, whenever γ and δ satisfy $0 < \gamma, \delta < 1/2$ then μ and $\widehat{\mu}$ are valid distributions and the topology is non-identifiable (since there is a mixture μ on T_1 which is identical to a distribution $\widehat{\mu}$ on T_3). Note for every $0 , there exists a and b where <math>\gamma$ and δ are valid, and hence the above construction defines a non-identifiable mixture distribution.

Since the distribution μ is invariant under the relabeling of leaves 1 and 3, likelihood maximized over M^1 is the same for topology T_1 and T_3 , i.e.,

$$\mathcal{L}_{T_1}(\mu) = \mathcal{L}_{T_3}(\mu)$$

Part 1 of Theorem 2 is the special case when p=1/2, and a and b are rephrased as a=1/2-(C+x) and b=1/2-(C-x). Note, when p=1/2, $\delta=\gamma$ and thus the internal edge has the same branch length in the two trees

In Stefankovič and Vigoda (2006) (see Proposition 17) we give a relatively simple proof of the above theorem. That proof relies on some symmetry properties of the model which are introduced as a precursor to the proof of the duality theorem. Since we have not introduced these properties here, we

instead present a more "brute-force" style proof.

Proof. Swapping the leaves 1 and 3 changes T_1 into T_3 . Let σ denote an assignment of labels from $\{0,1\}$ to the leaves, i.e., $\sigma:\{1,2,3,4\}\to\{0,1\}$. Let $\widehat{\sigma}$ denote the assignment obtained from σ with the assignment for leaves 1 and 3 swapped (i.e., $\widehat{\sigma}(1) = \sigma(3)$, $\widehat{\sigma}(3) = \sigma(1)$ and $\widehat{\sigma}(i) = \sigma(i)$ for i = 2, 4).

An assignment σ has the same probability in (T_1, w) as the assignment $\widehat{\sigma}$ in (T_3, w) . Hence,

$$\mu(\sigma) = \widehat{\mu}(\widehat{\sigma}). \tag{2}$$

Note, many assignments are fixed under swapping the labels for leaves 1 and 3. In particular, for σ as any of the following values:

$$0000,0001,0100,0101,1010,1011,1110,1111$$

we have $\sigma = \hat{\sigma}$ and hence $\mu(\sigma) = \hat{\mu}(\sigma)$. Thus we will can ignore these assignments and prove that the probabilities of the other assignments remain the same when the labels of leaves 1 and 3 are swapped.

We will show that in the mixture distribution μ , the probabilities of assignments 0010 and 1000 are the same:

$$\mu(0010) = \mu(1000) \tag{3}$$

Since for $\sigma = 0010$, we have $\hat{\sigma} = 1000$, then from (3) and (2) it follows that

$$\mu(0010) = \widehat{\mu}(0010)$$
 and $\mu(1000) = \widehat{\mu}(1000)$.

Moreover, since the CFN model is symmetric, we also have

$$\mu(1101) = \widehat{\mu}(1101)$$
 and $\mu(0111) = \widehat{\mu}(0111)$.

Finally, we also show that

$$\mu(0110) = \mu(1100),\tag{4}$$

which implies

$$\mu(0110) = \widehat{\mu}(0110), \mu(1100) = \widehat{\mu}(1100), \mu(1001) = \widehat{\mu}(1001), \mu(0011) = \widehat{\mu}(0011).$$

Hence, (3) and (4) imply that those assignments which are not fixed by swapping leaf labels 1 and 3, have the same probability in μ and $\hat{\mu}$. Thus to complete the proof, we need to show that (3) and (4) hold. These are straightforward to check in any symbolic algebra system, such as Maple. \square

2.2 Proof Tools for Maximum Likelihood Results

In this section we present the technical tools needed to prove Theorems 1 and 2. Before stating precise lemmas, we begin by explaining some of the

intuition of the proofs and how the notation applies to the Theorems.

Let μ_x denote the mixture distribution defined by the example in Figure 2 (in the CFN or Jukes-Cantor model). Our result is for x > 0, but the proof will use properties of the case x = 0. Let $\mu = \mu_0$. Note in the case x = 0, the two trees in the mixture are identical. Hence, $\mu = \mu_0$ is a pure distribution (i.e., generated by a single tree). Moreover, for x = 0, since the internal edge weight is 0 this distribution is achievable on every topology; simply set the terminal edges to taxa 1, 2, 3 and 4 to have weight C, and the other internal edge to have weight 0. Let v denote this assignment of edge weights. For every tree S, this assignment generates distribution μ (i.e., $\mu(S, v) = \mu$). This is also the unique such weight (see Štefankovič and Vigoda (2006) for a proof). Thus this is the assignment of edge weights for S that maximizes the expected log-likelihood under μ .

For x small the following lemma bounds the expected maximum expected log-likelihood in terms of the log-likelihood at x = 0 and the Hessian and Jacobian of the likelihood function. The error term will be $o(x^2)$, which is a smaller order term when x is sufficiently small. This will then imply that for x sufficiently small, $\mathcal{L}_S(\mu_x)$ is close to $\mathcal{L}_S(\mu_0)$ and we simply need to compute the Hessian and Jacobian functions to bound $\mathcal{L}_S(\mu_x)$.

Before stating the general lemma, let us preview the terminology in our setting. We are interested in computing the likelihood for distribution μ_x where x > 0. We will use the distribution $\mu = \mu_0$ which corresponds to the tree with internal edge weight zero. (Note under μ each of the 4^4 assignments

to the leaves has positive probability.) To bound the maximum likelihood of μ_x on some tree topology S (which can be any topology), we need that the distribution is achievable on S. In our case this is clearly true as discussed above.

If x is small then

$$\mu_x = \mu + x\Delta\mu_x$$

where $\Delta \mu_x$ is a vector whose sum of coordinates is zero (thus, $\Delta \mu_x$ is a vector representing the change in μ_x from μ). As $x \to 0$ we will have $\Delta \mu_x \to \Delta \mu_0$. We will use $\Delta \mu$ for bounding the maximum likelihood of μ_x for x small. The maximum expected log-likelihood of μ_x on S will be expressed in terms of the expected log-likelihood of μ under the tree corresponding to μ_x , which is

$$(\mu_x)^T \ln(\mu) = (\mu^T + x(\Delta \mu_x)^T) \ln(\mu),$$

and the first few terms from the Taylor expansion.

Here is the formal statement of the lemma.

Lemma 6 (Štefankovič and Vigoda (2006)). Let μ be a probability distribution on Ω^n such that every element has non-zero probability. Let S be a leaf-labeled binary tree on n nodes. Suppose that there exists v in the closure of the model such that $\mu(S,v) = \mu$ and that v is the unique such weight. Let $\Delta \mu_x$ be such that $\Delta \mu_x^T 1 = 0$, and $\Delta \mu_x \to \Delta \mu_0$ as $x \to 0$.

Let $g(w) = \mathcal{L}_{S,w}(\mu)$, and $h_x(w) = (\Delta \mu_x)^T \ln \mu_{S,w}$. Let H be the Hessian of g at v and J_x be the Jacobian of h_x at v. Assume that H has full rank.

Then

$$\mathcal{L}_S(\mu + x\Delta\mu_x) \le \mu^T \ln \mu + xh_x(v) - \frac{x^2}{2}J_0H^{-1}J_0^T + o(x^2).$$
 (5)

If $(H^{-1}J^T)_i \leq 0$ for all i such that $v_i = 0$ then the inequality in (5) can be replaced by equality.

Remark 7. When $(H^{-1}J^T)_i < 0$ for all i such that $v_i = 0$ then the likelihood is maximized at non-trivial branch lengths. In particular, for the CFN model, the branch lengths are in the interval (0, 1/2), and there are no branches of length 0 or 1/2. Similarly for the Jukes-Cantor the lengths are in (0, 1/4).

We will need the following extension of the above result.

Lemma 8. Let μ be a probability distribution on Ω^n such that every element has non-zero probability. Let S be a leaf-labeled binary tree on n nodes. Suppose that there exists v in the closure of the model such that $\mu(S, v) = \mu$ and that v is the unique such weight. Let $\Delta \mu_x$ be such that $\Delta \mu_x^T 1 = 0$, and $\Delta \mu_x \to \Delta \mu_0$ as $x \to 0$. Let $g(w) = \mathcal{L}_{S,w}(\mu)$ and $h_x(w) = (\Delta \mu_x)^T \ln \mu_{S,w}$. Let H be the Hessian of g at v and J_x be the Jacobian of h_x at v. Assume that H has full rank. Then

$$\mathcal{L}_S(\mu + x\Delta\mu_x)$$

$$= \mu^T \ln \mu + xh_x(v) + \max_{\Delta w} \left(\frac{1}{2}(\Delta w)^T H(\Delta w) + xJ_0(\Delta w)\right) + o(x^2). \quad (6)$$

where the maximization is restricted to Δw such that $(\Delta w)_i \geq 0$ whenever $v_i = 0$.

Remark 9. Now we remark on how the optimization problem in (6) can be solved. Let I be the set of indices with $v_i = 0$. Let u be an optimal solution of $(1/2)u^THu + xJ_0u$ subject to the restriction $u_i \geq 0$ for $i \in I$. Note that u exists because the optimization happens on a continuous function over a compact set. Let $I' \subseteq I$ be a set of indices that are zero in u. Let J' be J_0 restricted to indices that are not in I' and let H' be H restricted to indices not in I'. Then, by the optimality of u, we have that u restricted to indices not in I' is $(H')^{-1}(xJ')^T$.

Thus the maximization problem in (6) can be solved as follows. For each $I' \subseteq I$ do the following. Let J' be J_0 with entries at indices in I' zeroed out and let H' be H with entries at indices in I' zeroed out. Compute $(H')^{-1}(xJ')^T$ and check whether all entries at positions in I are non-negative. If they are then $xJ'(H')^{-1}(xJ')^T$ is a candidate solution. Pick the largest candidate solution. Note that only solutions that satisfy the restriction $u_i \geq 0$ for $i \in I$ are candidate, and by the previous paragraph an optimal solution is candidate. Thus using the procedure described we obtain an optimal solution of (6).

Proof of Lemma 8. In the proof of Lemma 6 from Štefankovič and Vigoda

(2006), it is proved that

$$\mathcal{L}_{T,v+\Delta v}(\mu + x\Delta\mu_x) = \mu^T \ln\mu + xh_x(v) + \frac{1}{2}(\Delta v)^T H(\Delta v) + xJ_x(\Delta v) + O\left(||\Delta v||^3 + x||\Delta v||^2\right), \quad (7)$$

and $\mathcal{L}_{v+\Delta v}(\mu + x\Delta\mu_x) < \mathcal{L}_v(\mu + x\Delta\mu_x)$ for $||\Delta v|| = \omega(x)$. Thus (7) is maximized for $||\Delta v|| = O(x)$.

The edge weights $v + \Delta v$ are invalid if, for some i, $v_i = 0$ and $(\Delta v)_i = 0$. On the other hand if for all i where $v_i = 0$ we have $(\Delta v)_i \geq 0$ and Δv are sufficiently small, then then $v + \Delta v$ are edge weights. Note that we are considering $||\Delta v|| = O(x)$. Hence, we have Δv is sufficiently small by choosing x sufficiently small. Therefore,

$$\mathcal{L}_S(\mu + x\Delta\mu_x) = \mu^T \ln \mu + xh(v) + \max_{\Delta w} \left(\frac{1}{2} (\Delta w)^T H(\Delta w) + xJ_x(\Delta w) \right) + O(x^3),$$

where the maximum is taken over Δw such that $(\Delta w)_i \geq 0$ whenever $v_i = 0$. From $\Delta \mu_x \to \Delta \mu_0$ it follows that

$$\max_{\Delta w} \left(\frac{1}{2} (\Delta w)^T H(\Delta w) + x J_x(\Delta w) \right)$$

$$= \max_{\Delta w} \left(\frac{1}{2} (\Delta w)^T H(\Delta w) + x J_0(\Delta w) \right) (1 + o(1)),$$

where the maxima are taken over Δw such that $(\Delta w)_i \geq 0$ whenever $v_i = 0$.

2.3 Proof of Theorem 2 for the CFN model

We begin with the simpler proof for the CFN model.

Let $\alpha = \beta \cdot x^2$ be the weight of the middle edge.

We will need the following polynomials in C:

$$Z := (1 - 4C + 4C^{2})(1 - 2C + 2C^{2})^{2}$$

$$U1 := 24C^{6} - 72C^{5} + 86C^{4} - 52C^{3} + 16C^{2} - 2C$$

$$U2 := -16C^{4} + 32C^{3} - 16C^{2} + 2$$

$$U3 := -8C^{4} + 16C^{3} - 16C^{2} + 8C - 2$$

$$U4 := -2(2C^{2} - 2C + 1)^{2}$$

We will need the following property of the polynomials.

Observation 10. The polynomials Z, U_2 and $-U_1, -U_3, -U_4$ are positive for $C \in (0, 1/2)$.

The observation is easily proved by plugging in C = 1/4, and then checking that the polynomials do not have roots on (0, 1/2), e.g., using the method of Sturm sequences.

Let H_i be the Hessian and J_i the Jacobian for T_i . Because of symmetry we have that the Hessians are the same. We let $H := H_1 = H_2 = H_3$ and let H' be H with the last (fifth) row and column replaced by zeros. The Jacobians differ only in the last coordinate. Let J' be the vector which agrees with J_1, J_2, J_3 on the first four coordinates and is zero on the last (fifth) coordinate.

Let ℓ_i be the last coordinate of $-H^{-1}J_i$. Recall that if ℓ_i is positive then, by Lemma 6,

$$\mathcal{L}_{T_i}(\mu + x\Delta\mu_x) = \mu^T \ln \mu + xh_x(v) - \frac{x^2}{2}(J_i)_0 H^{-1}(J_i)_0^T + o(x^2).$$
 (8)

On the other hand if ℓ_i is not positive then, by Remark 9,

$$\mathcal{L}_{T_i}(\mu + x\Delta\mu_x) = \mu^T \ln \mu + xh_x(v) - \frac{x^2}{2}(J')(H')^{-1}(J')^T + o(x^2).$$
 (9)

Note that when ℓ_i is positive then the right hand side of (8) is strictly bigger than the right hand side of (9).

A tedious computation yields

$$\ell_1 = \beta + U_4/Z,$$

 $\ell_2 = (U_1\beta + U_3)/Z,$

 $\ell_3 = (U_1\beta + U_2)/Z.$

Since U_1 and U_3 are negative and Z is positive, then ℓ_2 is always negative.

For small β we have that ℓ_3 is positive and ℓ_1 is negative. Thus tree T_3 has higher likelihood than T_2 and T_1 (for sufficiently small x). In addition, by Remark 7, since $\ell_3 > 0$, the maximum likelihood for T_3 is achieved with non-zero branch lengths. For T_1 and T_2 the optimization procedure outlined

in Remark 9 may give zero length edges. But, since we only want to upper bound the likelihood for these trees, we can allow zero length edges. This proves Part (b) of Theorem 2.

We know from Theorem 5 that there exists an α'_c where the likelihood of T_1 and T_3 are equal. We want to show there exists α_c where the likelihoods of T_1 and T_3 are the same, and, for $\alpha < \alpha_c$, T_3 is the maximum likelihood tree. This requires showing there exists α_c where the following hold: at α_c we have $\ell_1 = \ell_3$ (and both are positive so that the maximum likelihoods on these trees are achieved at non-zero branch lengths); and for $\alpha < \alpha_c$ we have $\ell_1 > \ell_3$ (and ℓ_1 is positive).

For large β (i.e., large α) we have that ℓ_3 is negative and ℓ_1 is positive. Thus for large β , tree T_1 has higher likelihood than T_2 and T_3 .

We first argue that for any α , $\ell_1 > 0$ and/or $\ell_3 > 0$. Then the result follows easily by considering the smallest α where $\ell_1 = \ell_3$.

Multiply ℓ_1 by the positive polynomial $(-U_1)$ and add it to ℓ_3 . Since $-U_1$ is always positive, then if $-\ell_1U_1 + \ell_3 > 0$ then at least one of ℓ_1 and ℓ_3 must be positive. Note

$$-\ell_1 U_1 + \ell_3 = \frac{1}{Z} (U_2 - U_3 U_1).$$

We obtain

$$\frac{1}{Z}(U_2 - U_3 U_1) = \frac{1}{Z}(192C^{10} - 960C^9 + 2224C^8 - 3136C^7 + 2960C^6 - 1936C^5 + 890C^4 - 232C^3 + 32C^2 - 4C + 2),$$

which is positive for $C \in [0, 1/2]$ (this is proved using Sturm sequences as before). Thus at least one of ℓ_1 or ℓ_3 is always positive.

Take the smallest β_c where $\ell_1 = \ell_3$. We know there exists at least one such β_c since for small β we have $\ell_1 > 0$ and $\ell_3 < 0$, whereas for large β we have $\ell_1 < 0$ and $\ell_3 > 0$. Since at least one of ℓ_1 and ℓ_3 is positive, we have that both ℓ_1 and ℓ_3 are positive when $\ell_1 = \ell_3$. Thus the maximum likelihoods on T_1 and T_3 are the same at β_c (since $\ell_1 = \ell_3$), and are achieved on non-zero branch lengths (since $\ell_1, \ell_3 > 0$).

All the formulas are continuous in β and hence for all $\beta < \beta_c := \alpha_c/x^{-2}$ the likelihood of T_3 is higher than the likelihood of T_1 , and for all $\beta > \beta_c$ the likelihood of T_3 is smaller than the likelihood of T_1 .

This completes the proof of the theorem.

2.4 Proof of Theorem 1 for the Jukes-Cantor model

The proof for the Jukes-Cantor model will follow the same lines as the argument for the CFN model, however some of the quantities are more complicated.

Let $\alpha = \beta \cdot x^2$ be the weight of the middle edge.

We will need the following polynomials in C:

$$Z = (-1+2C) \left(2359296C^{15} + 11010048C^{14} - 42385408C^{13} + 55336960C^{12}\right)$$
$$-33972224C^{11} + 4602880C^{10} + 9055232C^{9} - 8473344C^{8} + 4156416C^{7}$$
$$-1372704C^{6} + 326728C^{5} - 57200C^{4} + 7308C^{3} - 654C^{2} + 37C - 1\right)$$

$$U_1 = 21233664C^{16} - 98697216C^{15} + 203702272C^{14} - 247480320C^{13}$$

$$+196929536C^{12} - 107718656C^{11} + 41333888C^{10} - 11267328C^{9}$$

$$+2317632C^{8} - 465360C^{7} + 123492C^{6} - 33482C^{5} + 6644C^{4}$$

$$-846C^{3} + 62C^{2} - 2C$$

$$U_2 = 393216C^{12} - 1515520C^{11} + 2001920C^{10} - 917504C^9 - 441728C^8$$
$$+830176C^7 - 517696C^6 + 183344C^5$$
$$-39128C^4 + 4604C^3 - 164C^2 - 22C + 2$$

$$U_3 = 49152C^{12} + 548864C^{11} - 1646080C^{10} + 2025984C^9 - 1440896C^8$$
$$+672800C^7 - 226064C^6 + 60968C^5 - 14600C^4 + 3044C^3$$
$$-476C^2 + 46C - 2$$

We will need the following property of the polynomials.

Observation 11. The polynomials Z, U_2 and $-U_1, -U_3$ are positive for $C \in (0, 1/4)$.

Once again the observation is proved by plugging in C = 1/8 and then using Sturm sequences to prove that the polynomials do not have roots on (0, 1/4).

Define H, J_1, J_2, J_3 and J as in the proof for the CFN model. More precisely, let H_i be the Hessian and J_i the Jacobian for T_i . Because of symmetry we have that the Hessians are the same, thus let $H := H_1 = H_2 =$ H_3 . The Jacobians differ only in the last coordinate. Let J be the vector which agrees with J_1, J_2, J_3 on the first four coordinates and is zero on the last (fifth) coordinate. Let ℓ_i be the last coordinate of $-H^{-1}J_i$. Recall that if ℓ_i is positive then (8) holds, and if ℓ_i is not positive then (9) holds.

Note that when ℓ_i is positive then the right hand side of (8) is strictly

bigger than the right hand side of (9). We now have

$$\ell_1 = \beta + U_3/Z,$$

 $\ell_2 = (U_1\beta + U_3)/Z,$

 $\ell_3 = (U_1\beta + U_2)/Z.$

Note that ℓ_2 is always negative. As in the CFN model, for small β we have that ℓ_3 is positive and ℓ_1 is negative. Thus tree T_3 has higher likelihood than T_2 and T_1 (for sufficiently small x). The proof of part (a) is along identical lines as for the CFN model. However the polynomial of interest, i.e., $(1/Z)(U_2 - U_3 \cdot U_1)$ is significantly more complicated.

This completes the proof of the theorem.