• Recall that the **CFN model** is a Markov Chain on a tree \((T, P, \mu_p)\), where \(T\) is a directed binary tree rooted at \(p\) and with leaf set \([n]\), whose edges have transition matrices

\[
P_e = \begin{bmatrix}
1 - p_e & p_e \\
p_e & 1 - p_e
\end{bmatrix}
\]

over the character set \(\{0,1\}\), and \(\mu_p = \left(\frac{1}{2}, \frac{1}{2}\right)\).

[\(p_e\) is the mutation probability of the edge]

• The **tree reconstruction problem** is the following:

- given \(X = (X_{[n]}, X_{[n]}, ..., X_{[n]})\), that is \(k\) independent samples from the CFN model at the leaves of \(T\), the goal is to reconstruct the unrooted, undirected tree \(T^{-p}\)

- the strong reconstruction problem also asks for a CFN model over the leaf set \([n]\) whose leaf character distn' is within \(\epsilon\) total variation distance from the distn' of the actual model (which the samples \(X\) were sampled from)

• Our goal is to reconstruct \(T^{-p}\) using a tree metric

• Our tree metric is inspired by the following decomposition:

\[
P_e = (1 - 2p_e) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + 2p_e \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}
\]

interpretation: on an edge \(e = (u,v)\):

\[
\begin{align*}
P_{uv} &= \frac{3p_e}{3p_e + 2p_e} \quad \text{w/pr} \quad 1 - 2p_e \\
\frac{3p_e}{3p_e + 2p_e} + p_e \quad \text{w/pr} \quad 2p_e
\end{align*}
\]
Consider now a path in tree. \( \text{Path}(a, b) = \{e_1, \ldots, e_k\} \)

\[
\mathbb{F}_b = \left\{ \mathbb{F}_a \quad \text{w.p.} \quad \prod_{i=1}^{k} \theta_i \right\}
\]

Since \( \theta_i \)'s multiply, reasonable to define

\[
\omega = -\log \theta_i
\]

Then

\[
\sum_{i=1}^{k} \omega_i = -\log \prod \theta_i = -\log \theta(a, b)
\]

So CEN model defines tree metric \( \delta(a, b) = -\log \theta(a, b) \) minimizes \( \omega_i \).

If I could estimate \( \delta(a, b) \) to within \( \frac{1}{4} \omega(a, b) \), I could solve the reconstruction problem (from last lecture).

**Estimating \( \delta(a, b) \) from \( \Omega \)**

Define \( \hat{\delta}(a, b) := \frac{1}{k} \sum_{i=1}^{k} \mathbb{1}\{ \omega_i \neq \omega_j \} \); then

\[
\mathbb{E}[\hat{\delta}(a, b)] = \frac{1-\theta(a, b)}{2} = p(a, b)
\]

Set

\[
\hat{\delta}^2(a, b) = \begin{cases} 
-\log \left[1 - 2\hat{p}^2 \right] & \text{if } \hat{p} < \frac{1}{2} \\
\text{equal to } +\infty & \text{if } \hat{p} \geq \frac{1}{2}
\end{cases}
\]
Claim: Suppose that \( a, b \in [n] \)

\[
| p_{ab}^{\hat{\theta}} - p_{ab}^{\theta} | < \varepsilon \leq \frac{1}{2} (1 - e^{-W^*/4}) (1 - 2p_{ab}^{\theta})
\]

Then

\[
\max_{q \in \{a,b,c,d\}} | \delta(q) - \delta(q)_{\theta} | \leq 2 \max_{a,b} | \delta(a,b) - \delta(a,b)_{\theta} | < \frac{1}{2} W^*.
\]

where

\[
\delta(q) = \frac{1}{2} \left[ \delta(a,c) + \delta(b,d) - \delta(a,b) - \delta(c,d) \right].
\]

Observation:

\[
p_{ab}^{\hat{\theta}} + \varepsilon < p_{ab}^{\theta} + \frac{1}{2} - p_{ab}^{\hat{\theta}} < \frac{1}{2}
\]

Hence \( -\log(1 - 2(p_{ab}^{\theta} \pm \varepsilon)) \) well-defined (i.e. not \( +\infty \))

Now:

\[
| \delta(a,b) - \delta(a,b)_{\theta} | = \left| \log(1 - 2p_{ab}^{\theta}) - \log(1 - 2p_{ab}^{\hat{\theta}}) \right|
\]

\[
= \left| \log \frac{1 - 2p_{ab}^{\hat{\theta}} \pm 2\varepsilon}{1 - 2p_{ab}^{\hat{\theta}}} \right|
\]

\[
= \left| \log \left( 1 \pm \frac{2\varepsilon}{1 - 2p_{ab}^{\theta}} \right) \right| \leq \frac{1}{4} W^*
\]

Indeed:

\[
| \log (1 \pm \frac{2\varepsilon}{1 - 2p_{ab}^{\theta}}) | \leq \max \left\{ \log \left( 1 + \frac{2\varepsilon}{1 - 2p_{ab}^{\theta}} \right), -\log \left( 1 - \frac{2\varepsilon}{1 - 2p_{ab}^{\theta}} \right) \right\}
\]

\[
\leq -\log (1 - \frac{2\varepsilon}{1 - 2p_{ab}^{\theta}}) \leq \frac{1}{4} W^*
\]
**Theorem 1:** Let \( w_* = \min_{e} w(e) \) and \( W_* = \max_{a,b} \delta(a,b) \).

Then \( k = \Theta \left( \frac{e^{2W_*}}{(1 - e^{-W_*})^2 \cdot \log n} \right) \) samples suffice to get the correct tree \( T \) w/ prob \( (1 - o(1)) \) as \( n \to \infty \).

**Proof:** From Chernoff bounds:

\[
\Pr \left[ | \hat{p}_{ab} - p_{ab} | > \varepsilon \right] \leq 2 \exp \left( - \frac{2 \varepsilon^2 k}{n} \right) < \frac{1}{n^3} \quad (\text{a})
\]

choosing \( k = \Omega \left( \frac{1}{\varepsilon^2 \log n} \right) \).

- Use \( \varepsilon = \frac{1}{2} \left( 1 - e^{-w_*/4} \right) \cdot e^{-\sqrt{W_*}} \) \((\text{xx})\)

By probability \( \geq 1 - \frac{1}{n} \):

- For all \( a, b \in [n] \):

\[
| \hat{p}_{ab} - p_{ab} | < \frac{1}{2} \left( 1 - e^{-w_*/4} \right) (1 - 2p_{ab}).
\]

- Hence we'll get all quartets right w. pr. \( \geq 1 - \frac{1}{n} \), and therefore the correct tree w/ prob. \( \geq 1 - \frac{1}{n} \).

- From (x), (xx) it follows that \( k = \Theta \left( \frac{e^{2W_*}}{(1 - e^{-W_*})^2 \cdot \log n} \right) \) suffices for this purpose.

\( \Box \)
So assume that mutation probabilities are bounded away from 0 and 1/2, i.e. $0 < c_2 < P_e < c_2 < 1/2$, for all edges $e$.

Then
\[
\frac{D < \log(1/2c_2)}{f} \leq \frac{W_e \leq -\log(1 - 2c_2)}{g} < +\infty
\]
\[
(f, g \text{ are constants})
\]

In this case, $\min_e W_e \geq f$ and $W^* \leq (n+1)g$

So from previous theorem: $L = 2^{O(n)}$ suffices.

This is not tight; it turns out that Theorem 1 holds if we replace $W_e$ by the weighted depth of the tree.

**Def (Weighted Depth):** The depth of an edge $e$ is the length (under $\delta$) of the shortest path between two leaves crossing $e$. The depth of a tree is the largest depth of an edge in the tree.

**Ex (Expt):** If $w_e = 1/4$, then the depth of an edge $e$ in any binary tree is at most $2 \log_2 n + 2$. 
Hence, if $f \leq w \leq g$, and we use the modified version of Thm 1, it follows that w/ $k = \text{poly}(n)$ samples from the CFN model we can reconstruct the tree w/ prob. $1 - o(1)$ as $n \to \infty$.

Can we do better than $\text{poly}(n)$ sequence length? (recall that our counting lower bound was just $\Omega(\log n)$)

**Steel's Conjecture:**

Let $\Theta^* = \frac{1}{2}$. Then, if $\Theta(e) \leq \Theta^*$, $\forall e$, $\text{poly}(n)$ samples are necessary for the tree reconstruction problem, while if $\Theta(e) > \Theta^*$, $O(\log n)$ samples suffice.

More on Steel's Conjecture next time!