Neighbor Joining (NJ) review:

We consider a dissimilarity map $d : X \times X \to \mathbb{R}$. Now for each $x \in X$ compute

$$u(x) = \frac{1}{|X| - 2} \sum_{y \in X} d(x, y).$$

Let $Q(x, y) = d(x, y) - u(x) - u(y)$. Find a pair $a, b$ with $Q(a, b)$ minimized. Join them, and call the new vertex $a \cdot b$. Set $d(a \cdot b, a) = \frac{1}{2}(d(a, b) + u(a) - u(b))$ and $d(a \cdot b, b) = \frac{1}{2}(d(a, b) + u(b) - u(a))$. Let $d'(y, x) = d(y, x)$ for all $y, x \neq a, b$ and $d'(a \cdot b, x) = \frac{1}{2}(d(a, x) + d(b, x) - d(a, b))$, and let $X' = X \setminus \{a, b\} \cup \{a \cdot b\}$. Iterate the above procedure using the pair $(X', d')$.

**Theorem:** Suppose $d$ is a tree metric $d = d(T, w)$, then the $Q$-criterion picks a cherry of $T$, and $d'$ is the induced tree metric on $T'$.

**Proof.** Let $d$ be a tree metric. We first observe that if $d$ is a tree metric and $c : X \to \mathbb{R}$ is a function, then $d'(x, y) := d(x, y) + c(x) + c(y)$ is a tree metric, provided it is a metric. (We will require $-c(x) \leq w$ (pendant edge going to $x$).

Furthermore, this scaling does not affect $NJ$ selection criterion. We will have

$$Q'(x, y) = Q'(x, y) + \frac{2}{|X| - 2} \sum_{z \in X} c(z).$$

So the pair $a, b$ minimizing $Q$ is the same pair $a, b$ minimizing $Q'$. Thus, we may assume that all pendant edges have weight 0.

Claim (Bryant 2005): $x' \in X$ is a possible maximizer of the function $u(x)$ if and only if $x'$ is part of some cherry. Because we don’t know how to prove this claim, we consider the example on the next page, which makes the claim fell at least plausible:
Then

\[
\begin{align*}
  u(1) &= u(2) = 6a + 5b + 2c + 3d + 2e \\
  u(3) &= u(4) = 2a + 3b + 6c + 3d + 2e \\
  u(5) &= 2a + 3b + 2c + 5d + 2e \\
  u(6) &= u(7) = 2a + 3b + 2c + 5d + 6e \\
  u(8) &= 2a + 5b + 2c + 3d + 2e.
\end{align*}
\]

Assuming the claim, we suppose that a pair \(c, d\) is not a cherry. Then \(d(c, d) > 0\). On the other hand, if \(a, b\) is a cherry such that \(u(a) = u(b)\) is the maximum value for \(u\), then \(Q(c, d) = d(c, d) - u(c) - u(d)\) and \(Q(a, b) = d(a, b) - u(a) - u(b)\). Now \(d(c, d) > 0 = d(a, b)\) and \(u(c) \leq u(a)\) and \(u(d) \leq u(b)\), from which it follows that \(Q(c, d) > Q(a, b)\). Hence, the minimizer must be a cherry.

**Theorem (Attesson):** Let \(d(T, w)\) be a tree metric. Let \(\varepsilon = \min_{e \in T} w(e)\). Suppose \(d\) is a tree a metric with \(||d(T, w) - d||_\infty < \frac{\varepsilon}{2}\). Then \(NJ\) applied to \(d\) returns a tree metric \(\hat{d}(T', w')\) such that \(T = T'\).

**Proof.** Omitted. 

---

2
There exists a philogenetic problem that is not NP-Hard!! Here it is: Minimizing $||\cdot||_\infty$ for $\epsilon T(X)$ (though for $T(X)$ it is NP-hard). Given a dissimilarity map $\delta : X \times X \rightarrow \mathbb{R}$, we will find an ultrametric $d$ such that $||\delta - d||_\infty$ is minimized.

Let $K_X$ be the complete graph with vertex set $X$, and let $w(x, y) = \delta(x, y)$. For a path $P$ in $K_X$ define $w(P) = \max_{e \in P} w(e)$, and define $\delta_u : X \times X \rightarrow \mathbb{R}$ by

$$
\delta_u(x, y) = \min_{\text{path } p \text{ connecting } x, y \text{ in } K_X} x(P).
$$

Then $\delta_u$ is called the subdominant ultrametric defined by $\delta$.

**Theorem:** Let $\delta$ be a dissimilarity map on $X$. Then

1. $\delta_u$ is an ultrametric
2. $\delta_u(x, y) \leq \delta(x, y) \forall x, y$
3. If $\delta'$ is an ultrametric s.t. $\delta'(x, y) \leq \delta(x, y) \forall x, y$ then $\delta'(x, y) \leq \delta_u(x, y) \forall x, y$.
4. If $\delta$ is an ultrametric, then $\delta = \delta_u$.

**Proof.**

1. Let $P_1$ be a path from $x$ to $y$ such that $w(P) = \delta_u(x,y)$. Let $P_2$ be a path from $y$ to $z$ such that $w(P_2) = \delta_u(y,z)$. Then $P_1P_2$ is a path from $x$ to $z$, $w(P_1P_2) = \max\{\delta_u(x, y), \delta_u(y, z)\} \geq \delta_u(x, z)$.

2. We observe that $w(x, y) = \delta(x, y)$ is a path from $x$ to $y$ of weight $\delta(x, y) = \min$ over all paths $\delta_u(x, y)$.

3. Let $\delta'$ be an ultrametric such that $\delta'(x, y) \leq \delta(x, y) \forall x, y$. Let $P$ be the path from $x$ to $y$ such that $\delta_u(x, y) = w(P)$. Write $P = V_0V_1 \cdots V_t$ where $x = V_0$ and $y = V_t$. Then

$$
\delta(x, y) \leq \max\{\delta'(x, v_{i-1}), \delta'(v_{t-1}, y)\} \leq \max\{\max\{\delta'(x, v_{i-2}), \delta'(v_{t-2}, v_{t-1})\}, (\delta'(v_{t-1}, y))\} \leq \max_{i \in [t-1]} \{\delta'(v_i, v_{i+1})\} = w(P) = \delta_u(x, y)
$$

4. By (2) $\delta_u(x, y) \leq \delta(x, y)$, and by (3) $\delta_u(x, y) \geq \delta(x, y)$. Then $\delta_u = \delta$.

We note that $\delta_u$ can be computed in polynomial time! Let $G$ be a graph, and $w : V \rightarrow \mathbb{R}$ a weight function. A spanning tree of $G$ is a subset of the edges that form a tree and contain all vertices. Let $T \subseteq G$ be a spanning tree. Then $w(T) = \sum_{e \in T} w(e)$. A minimal weight spanning tree is a spanning tree of minimal weight.

**Theorem** (Kruskal): The greedy algorithm finds a minimal weight spanning tree.

**Proof.** Omitted.
Example:

**Proposition:** Let $\delta$ be a dissimilarity map of $X$, $T$ a minimal weight spanning tree of $K_X$. Then $\forall x, y \delta_T(x, y) = \delta_u(x, y)$ and $\delta_T(x, y) = w$(unique path $x$ to $y$ in $T$).

**Proof.** Omitted.

**Definition:** For $\varepsilon \in \mathbb{R}$, let $\delta^+\varepsilon(x, y) = \delta(x, y) + \varepsilon \, \forall x, y$.

**Theorem:** Let $\delta$ be a dissimilarity map on $X$, and let $\varepsilon = \frac{||\delta - \delta_u||_\infty}{2}$. Then $\delta^+\varepsilon$ is an $L_\infty$ optimal ultrametric for $\delta$ and $||\delta^+\varepsilon - \delta||_\infty = \varepsilon$.

**Proof.** We note that $\delta_u < \delta$ by a previous theorem, allowing for $||\delta^+\varepsilon - \delta||_\infty = \varepsilon$. Now since $\delta_u(x, y) \leq \delta(x, y)$, there exists $x_0, y_0$ such that $\delta_u(x_0, y_0) = \delta(x, y) - 2\varepsilon$ and so $\delta^+\varepsilon = \delta_u(x_0, y_0) + \varepsilon = \delta(x, y) - \varepsilon$.

Suppose $\delta'$ is $L_\infty$ optimal. Then $||\delta' - \delta||_\infty \leq \varepsilon$. Then $\forall x, y \in X$, $\delta'(x, y) \leq \delta^+\varepsilon(x, y)$ (if not, then $\delta'(x, y) > \delta(x, y) + \varepsilon$ implying $\delta'$ not optimal). Hence $\delta'(x, y) \leq \delta^+\varepsilon(x, y)$.

Then we have $\delta'(x_0, y_0) \leq \delta^+\varepsilon(x_0, y_0) = \delta_u(x_0, y_0) + \varepsilon = \delta(x_0, y_0) - \varepsilon$, so $||\delta' - \delta||_\infty \geq \varepsilon$, and it follows that $\delta^+\varepsilon$ is optimal.

$\square$