Encoding X-trees with tree preorders

Yaokun Wu
Shanghai Jiao Tong University

February 2, 2015

New Zealand Phylogenomics Conference

Coauthors of this work

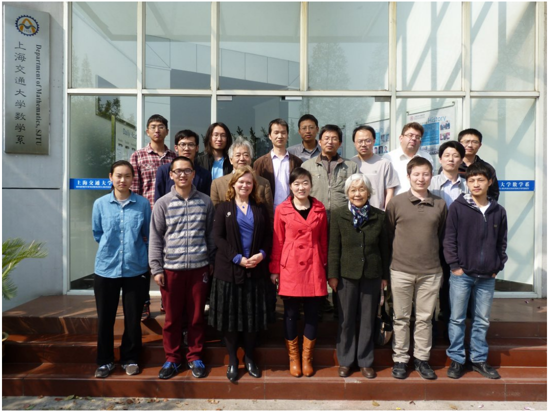


Figure: Ziqing Xiang, Zeying Xu, Yinfeng Zhu, April 5, 2014.

Metric and its associated total preorder

- A metric on a set X is a map D from $X \times X$ to $\mathbb{R} \cup \{+\infty\}$ such that $\mathrm{D}(x,x) = 0$ and $\mathrm{D}(x,y) \le \mathrm{D}(x,z) + \mathrm{D}(y,z)$ for all $x,y,z \in X$.
- A total preorder p on a set S is an ordered partition of S into a sequence of nonempty sets $S_1, \ldots, S_d$.
- For a metric D on a finite set X, its associated total preorder on $\binom{X}{2}$, denoted by p_{D}, lists the pairs of X according to their distances in D from smallest to largest, that is, p_{D} is the ordered partition of $\binom{X}{2}$ into nonempty sets, say $L_1, \ldots, L_d$, such that for any $i,j \in [d]$, $\{x,y\} \in L_i$ and $\{u,v\} \in L_j$, it holds $\mathrm{D}(x,y) < \mathrm{D}(u,v)$ if and only if $i < j$.

Reconstruct the metric

For any metric D on X and any positive number t, it is clear that D and $t\,\mathrm{D}$ generate the same total preorder on $\binom{X}{2}$. Also note that every total preorder on $\binom{X}{2}$ can come from a metric on X.

Reconstruct the metric

For any metric D on X and any positive number t, it is clear that D and $t\,\mathrm{D}$ generate the same total preorder on $\binom{X}{2}$. Also note that every total preorder on $\binom{X}{2}$ can come from a metric on X.

Which kind of information about the metric D is encoded by p_{D}? Can we reconstruct a metric from its associated total preorder, on the condition that we know some information about the shape of the metric?

Reconstruct the metric

For any metric D on X and any positive number t, it is clear that D and $t\,\mathrm{D}$ generate the same total preorder on $\binom{X}{2}$. Also note that every total preorder on $\binom{X}{2}$ can come from a metric on X.

Which kind of information about the metric D is encoded by p_{D}? Can we reconstruct a metric from its associated total preorder, on the condition that we know some information about the shape of the metric?

When the metric comes from an X-tree, we expect that such a metric is encoded by its associated preorder in some sense.

X-graphs

A graph G is an ordered pair $(V(G), E(G))$ comprising a set $V(G)$ of vertices and a set $E(G)$ of edges which is a subset of $\binom{V(G)}{2}$. Given a label set X, a graph G and a labelling map ϕ from X to $V(G)$ form an X-labelled graph (G, ϕ). If the image of the labelling map ϕ contains all vertices of G with degree no greater than 2, the X-labelled graph (G, ϕ) is called an X-graph. We call an X-graph $\mathcal{G} = (G, \phi)$ simple provided $X \subseteq V(G)$ and the labeling map ϕ is the inclusion map.

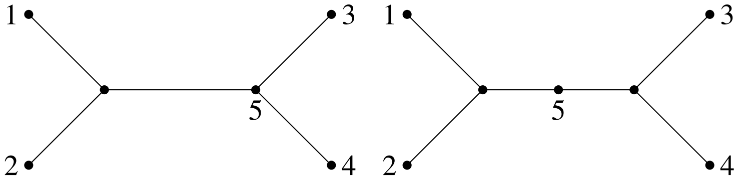


Figure: Two simple X-trees for $X = \{1, 2, 3, 4, 5\}$

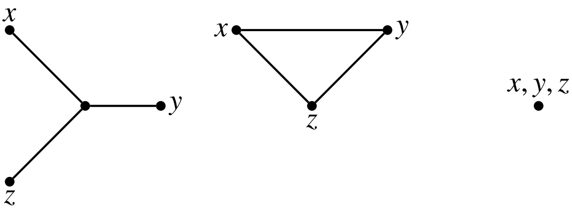
Metric/preorder from labelled graph

For each finite graph G, we use dist_G for the shortest path metric on $V(G)$, namely $\mathrm{dist}_G(u, v)$ is the length of a shortest path in G between u and v if such a path exists and $\mathrm{dist}_G(u, v) = +\infty$ otherwise. Unless we emphasize that the graph G is weighted, meaning that each edge of G carries a real number as its weight, the length of a path is the number of edges passed by it.

Let $\mathcal{G} = (G, \phi)$ be an X-labelled graph. The metric on X induced by the X-labeled graph $\mathcal{G}$, denoted $\mathrm{dist}_{\mathcal{G}}$, is given by $\mathrm{dist}_{\mathcal{G}}(x, y) = \mathrm{dist}_G(\phi(x), \phi(y))$ for all $x, y \in X$.

The total preorder on $\binom{X}{2}$ associated with $\mathrm{dist}_{\mathcal{G}}$ will often be recorded as $\mathrm{p}_{\mathcal{G}}$ and called the preorder associated with the X-labelled graph $\mathcal{G}$.

Knowing preorder is not enough: I



Let $X = \{x, y, z\}$. The above three X-graphs generate the same preorder on $\binom{X}{2}$. The first graph is a simple X-tree and the last graph is a non-simple X-tree.

Knowing preorder is not enough: II

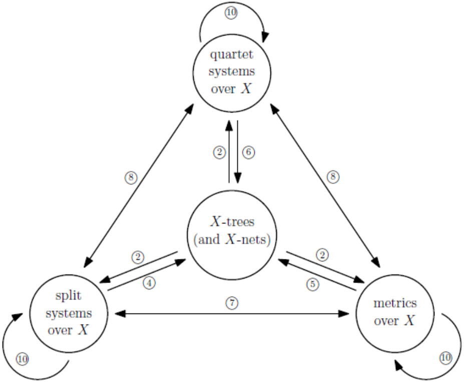
$\mathcal{T}_1$ $\mathcal{T}_2$

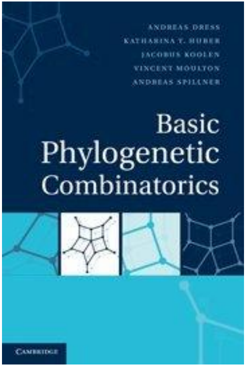
[Mike Steel] Let $X = \{a, b, c, d\}$. Consider the above two edge-weighted quartet phylogenetic X-trees $\mathcal{T}_1$ and $\mathcal{T}_2$. The two distinct distance functions on X, $\mathrm{dist}_{\mathcal{T}_1}$ and $\mathrm{dist}_{\mathcal{T}_2}$, have equal associated total preorder on $\binom{X}{2}$:

$$\mathcal{T}_1 : \; ab = 6 < ac = 11 < bc = 13 < ad = 19 < bd = 21 < cd = 24$$

$$\mathcal{T}_2 : \; ab = 7 < ac = 10 < bc = 13 < ad = 19 < bd = 20 < cd = 25$$

Success of encoding X-trees





It is a classical result in phylogenetics that two (weighted) X-trees are isomorphic if and only if their associated metrics on X coincide (K. A. Zaretskii, 1965; P. Buneman, 1974).

Labelled graph isomorphism

Let (G, ϕ) and (G', ϕ') be two X-labelled graphs. We say that (G, ϕ) and (G', ϕ') are isomorphic if there is a graph isomorphism f from G to G' such that $\phi' = f \circ \phi$.

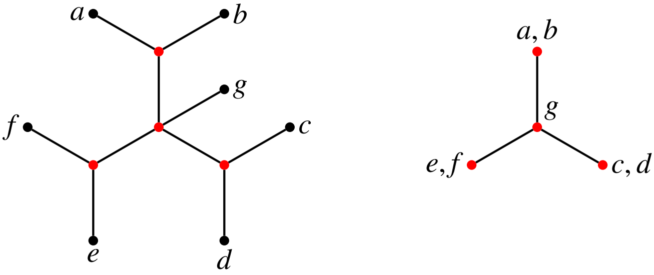
Main result

Theorem (W., Xiang, Xu, Zhu)

- *Two simple X-trees are isomorphic if and only if their associated total preorders on $\binom{X}{2}$ coincide.*
- *Two non-simple X-trees are isomorphic if and only if their associated total preorders on $\binom{X}{2}$ coincide.*
- *Given a total preorder on $\binom{X}{2}$, in $|X|^2 \log_2(|X|)$ time we can reconstruct the simple/non-simple X-tree or determine that no such tree exists.*

Simple vs non-simple

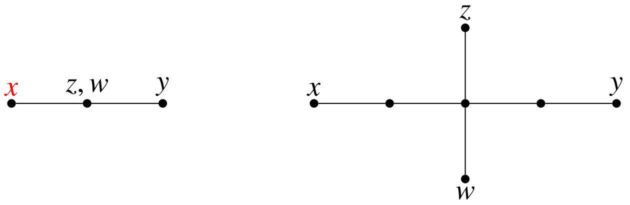
Let (T, ϕ) be a non-simple X-tree where every vertex of degree at most one is labelled via ϕ by at least two different elements of X. For every $x \in X$, we add a new vertex $\bar{x}$ to T and connect it to $\phi(x)$. We name the tree thus obtained as $\overline{T}$. We let $\overline{\phi}$ be the map which send $x \in X$ to $\bar{x}$. It is easy to see that (T, ϕ) and $(\overline{T}, \overline{\phi})$ share the same total preorder on $\binom{X}{2}$. Note that $(\overline{T}, \overline{\phi})$ has no labelled interior vertex and (T, ϕ) can be obtained from $(\overline{T}, \overline{\phi})$ by shrinking leaves.



Conjecture (W., Xiang, Xu, Zhu)

If $\overline{\mathcal{T}}$ is a simple X-tree which contains a labelled interior vertex, then $\overline{\mathcal{T}}$ is uniquely determined by its associated total preorder $\mathrm{p}_{\overline{\mathcal{T}}}$ among all X-trees. Equivalently, if $\mathcal{T}$ is a non-simple X-tree which contains a leaf vertex which is only labelled by one element of X, then $\mathcal{T}$ is uniquely determined by its associated total preorder $\mathrm{p}_{\mathcal{T}}$ among all X-trees.

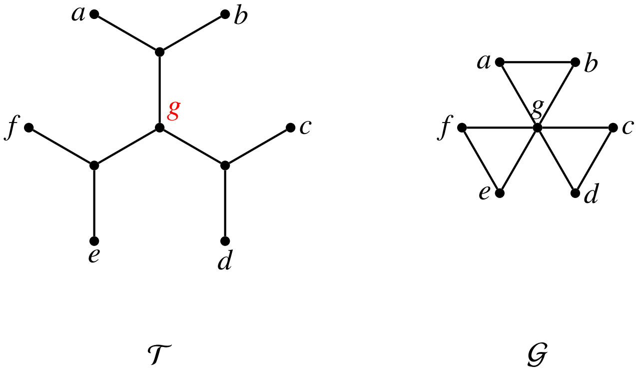
Non-simple X-tree



$\mathcal{T}_1$ $\mathcal{T}_2$

$\mathcal{T}_1$ is uniquely determined by $\mathrm{p}_{\mathcal{T}_1}$ among all X-trees but $\mathrm{p}_{\mathcal{T}_1}$ can come from graph metric other than $\mathrm{dist}_{\mathcal{T}_1}$.

Simple X-tree



$\mathcal{T}$ is uniquely determined by $\mathrm{p}_{\mathcal{T}}$ among all X-trees but but $\mathrm{p}_{\mathcal{T}}$ can come from graph metric other than $\mathrm{dist}_{\mathcal{T}}$.

Weak X-graphs

A weak X-graph is an X-labelled graph (G, ϕ) such that $\mathrm{Im}(\phi)$ contains every vertex of degree at most one in G. A weak X-graph (G, ϕ) is k-weak provided every connected component of G contains at most k unlabelled vertices of degree 2.

We can basically identify a weak X-graph with a weighted X-graph.

A conjecture on 1-weak X-trees

Conjecture (W., Xiang, Xu, Zhu)

Two non-isomorphic (non-)simple 1-weak X-trees can share the same total preorder if and only if they are as depicted below:

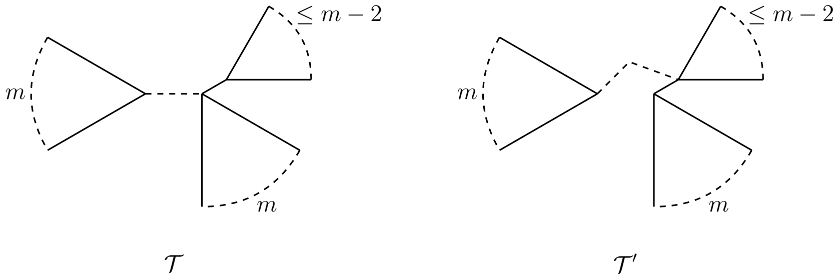


Figure: The distance from a labelled vertex to the relevant root vertex is indicated. The maximum distance on X read from $\mathcal{T}$ and $\mathcal{T}'$ are $2m+1$ and $2m+2$, resp..

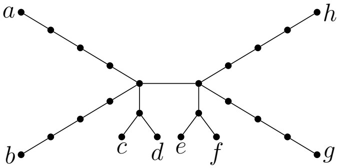
A tree T is an extended regular tree if, for any two vertices of degree larger than 2, the vertices on the unique path between them all have the same degree in T. If T is an extended regular tree and X is its set of leaves, we say that (T, X) forms an extended regular weak X-tree.

Theorem (W., Xiang, Xu, Zhu)

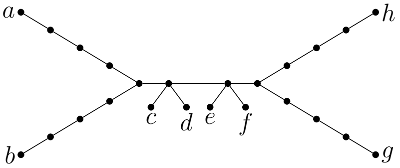
If $\mathcal{T}$ and $\mathcal{T}'$ are two extended regular weak X-trees, then they are isomorphic as X-graphs if and only if the following hold: 1) The total preorders $\mathrm{p}_{\mathcal{T}}$ and $\mathrm{p}_{\mathcal{T}'}$ coincide; 2) If the unique path in $\mathcal{T}$ connecting $x_1 \in X$ and $x_2 \in X$ contains at most one vertex of degree greater than 2, then $\mathrm{dist}_{\mathcal{T}}(x_1, x_2) = \mathrm{dist}_{\mathcal{T}'}(x_1, x_2)$; 3) For every $x \in X$, the longest path in $\mathcal{T}$ containing x but not any vertex of degree greater than 2 has the same length with the the longest path in $\mathcal{T}'$ containing x but not any vertex of degree greater than 2.

For simple X-trees, conditions 2) and 3) are always fulfilled.

Two non-isomorphic weak X-trees satisfying all three conditions in last theorem



T_1

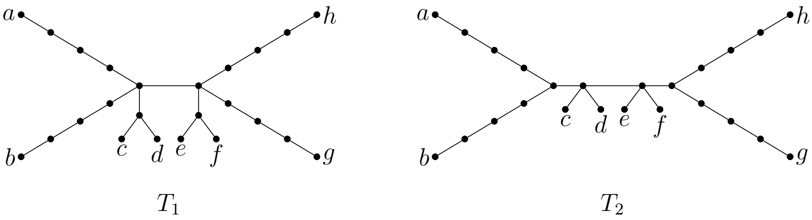


T_2

T_1 : $cd = ef = 2 < ce = cf = de = df = 5 < ac = ad = bc = bd = eg = eh = fg = fh = 6 < ae = af = be = bf = cg = ch = dg = dh = 7 < ab = gh = 8 < ag = ah = bg = bh = 9$

T_2 : $cd = ef = 2 < ce = cf = de = df = 3 < ac = ad = bc = bd = eg = eh = fg = fh = 6 < ae = af = be = bf = cg = ch = dg = dh = 7 < ab = gh = 8 < ag = ah = bg = bh = 11$

Two non-isomorphic weak X-trees satisfying all three conditions in last theorem

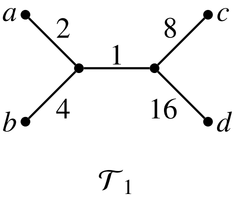


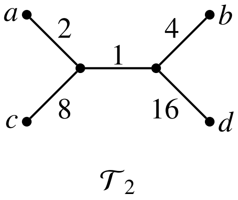
T_1 : $cd = ef = 2 < ce = cf = de = df = 5 < ac = ad = bc = bd = eg = eh = fg = fh = 6 < ae = af = be = bf = cg = ch = dg = dh = 7 < ab = gh = 8 < ag = ah = bg = bh = 9$

T_2 : $cd = ef = 2 < ce = cf = de = df = 3 < ac = ad = bc = bd = eg = eh = fg = fh = 6 < ae = af = be = bf = cg = ch = dg = dh = 7 < ab = gh = 8 < ag = ah = bg = bh = 11$

T_1 is not an extended regular tree!

Example of Mike Steel again

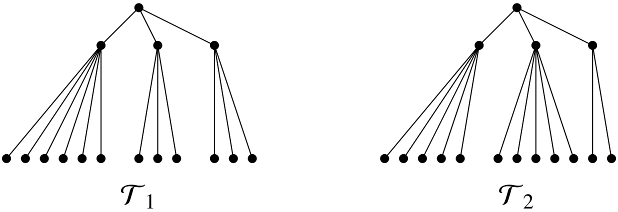




For the previous example of Mike Steel, we can view the two trees as extended regular weak X-trees. They fail condition 2) and so this example does not contradict with our theorem on extended regular trees.

Another way of relaxing the metric encoding method: Distance sequence?

A phylogenetic X-graph is a simple X-graph such that all vertices in X has degree at most one in the graph.



[John Goldwasser] Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be the two phylogenetic X-trees shown above.

Both of them have 12 leaves, with 21 pairs of leaves at distance 2 and 45 pairs of leaves at distance 4.

Consistency condition: Capture global aspect with local information

If a total preorder comes from the metric of a simple X-tree, then we call it a simple X-tree preorder, or just a simple tree preorder, and so on.

Total preorder is not enough to reconstruct weighted trees; but total preorder is enough to reconstruct simple/non-simple X-trees.

Many encoding aspects of X-trees have very simple local consistency conditions. For example, a metric arises from a weighted X-tree if and only if it satisfies the famous 4-point condition (Zaretskii 1965; Simoes-Pereira 1969; Buneman 1974; Dress 1984).

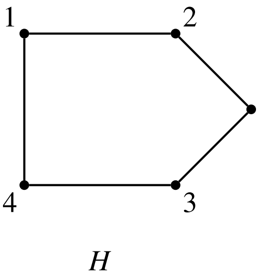
Is there any local characterization for the tree preorders or simple tree preorders? Is there any local characterization of weak tree preorder (weighted tree preorder)?

Optimal realization

An optimal realization of a total preorder σ on $\binom{X}{2}$ (among some graph classes) is an X-labelled graph (in that graph class) $\mathcal{G}$ with $\mathrm{p}_{\mathcal{G}} = \sigma$ and with fewest number of edges.

If the total preorder σ is a simple X-tree preorder, is it true that its optimal simple X-graph realization is just that unique simple X-tree?

The corresponding result for metric is known (Hakimi-Yao, 1964).



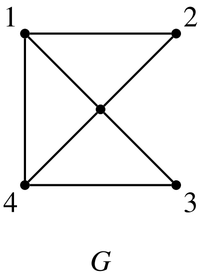
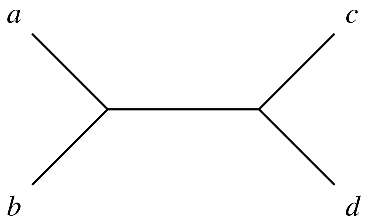


Figure: Consider the preorder $\sigma: \; 12 = 14 = 34 < 23 = 13 = 24$. It is not a weighted tree preorder. The graph H is an optimal X-labelled graph realization and the graph G is an optimal X-graph realization.

Lasso

Can we encode X-tree from partial information of its associated preorder?



Let $X = \{a, b, c, d\}$. For a simple X-tree $\mathcal{T}$, if we know $ab = cd < ac = ad = bc$ in the preorder $\mathrm{p}_{\mathcal{T}}$, we can determine $\mathrm{p}_{\mathcal{T}}$ and hence the simple X-tree.

Triple cover conjecture for total preorders on binary phylogenetic X-tree.

Further topics

- Association scheme and total preorder arising from tree metrics.
- Reconstruction from ball system.
- Shape of tree distance set and tree distance sequence.
- Total preorder in terms of diversity, quartet system, etc..
- Influence on tight span.

Figure: Piet Mondrian, Gray Tree, 1911.

https://en.wikipedia.org/wiki/Gray_Tree