Hierarchical Representations of Network Data with Optimal Distortion Bounds

Zane Smith, Samir Chowdhury, and Facundo Mémoli

***Abstract*—Single linkage hierarchical clustering is a tool in unsupervised learning which has been fully characterized for finite metric spaces, but not for the unrestricted setting of general networks. We follow a recent line of work to complete the characterization for general networks, and moreover, we provide quantitative bounds on how much information is lost when applying our method to network data. These bounds are novel even in the setting of finite metric spaces. Finally, we propose a construction called a treegram that provides a visual summary of the result of applying our method to a network data set.**

I. Introduction

When faced with the difficult computational task of analyzing a complex network, a first approach is to perform some sort of exploratory data analysis. Ideally, this analysis would also lead to a reasonably faithful representation of the network that is easy to visualize. Networks are often most appropriately represented by adjacency matrices, and depending on the method of acquiring data, these matrices may initially be weighted and/or asymmetric. In practice, many methods of analyzing these matrices require a preprocessing step where additional structure is imposed on the matrices. For example, directed networks are often symmetrized to obtain symmetric $n \times n$ matrices with real valued entries, for which the spectral theorem guarantees a full set of eigenvalues that can inform the properties of dynamic processes running on the network. However, imposing any condition in the preprocessing step leads to a loss of data, which is undesirable. We propose a method that accepts *any* square matrix (possibly asymmetric) with real-valued entries as input, thus avoiding this data loss.

We adhere to the viewpoint [1]–[3] of looking at an n-point network as a (suitably generalized) n-point metric space. One motivation for this viewpoint is that in the simple case of networks endowed with the shortest path distance, we have a bona fide metric space, and are free to use data simplification techniques applicable to metric spaces. The other justification, which we know a posteriori, is that some of the data analysis methods that hold for metric spaces can be adequately extended to the most general networks, i.e. $n \times n$ real valued matrices. More specifically, in this paper we study an extension of the single linkage hierarchical clustering method (SLHC). In its classical form, SLHC takes a finite metric space as input and returns an ultrametric space on the same set of points—a space for which the strong triangle inequality holds. Such an ultrametric space has the highly desirable property of having a faithful visual representation—specifically, it can be represented by a rooted tree called a *dendrogram*.

The Network SLHC (nSLHC) method that we study is an extension of a *directed SLHC* method that was established in [2] to study dissimilarity networks. The nSLHC method returns networks satisfying the strong triangle inequality, which we call *ultranetworks*. By applying a symmetrization step to an ultranetwork, we are able to recover a certain generalization of a dendrogram, which we call a *treegram*.

We emphasize the following caveat: A treegram is only as good as the ultranetwork it represents. To make a treegram practically useful, one needs quantitative guarantees on how much data is lost when obtaining an ultranetwork from a given network. We interpret this "data loss" as the ℓ^∞ distortion between a network and its ultranetwork representation. The main result in our work is a bound on this distortion that depends only on the number of points in the network and a network dependent quantity that we call the *ultranetwork constant* of a network. By controlling this quantity, it is possible to control the distortion induced by nSLHC.

While searching the literature to see how our bound compared with the ones known for classical SLHC, we were surprised to find that no such bound appears to exist even in the case of metric spaces. Because our nSLHC method reduces to standard SLHC on metric spaces, we thus obtain a novel estimation of the "goodness-of-fit" of a dendrogram produced by single linkage to the underlying metric space.

Proofs, additional figures, and a movie can be found in [4].

II. Preliminaries

Recall that a finite metric space (X, d_X) is a finite set X together with a function $d_X : X \times X \to \mathbb{R}_+$ such that: (1) $d_X(x, x') = 0 \iff x = x'$, (2) $d_X(x, x') = d_X(x', x)$, and (3) $d_X(x, x') \leq d_X(x, x'') + d_X(x'', x')$ for any $x, x', x'' \in X$.

An ultrametric space (X, u_X) is a metric space satisfying the additional condition called *strong triangle inequality*:

$$u_X(x, x') \leq \max(u_X(x, x''), u_X(x'', x')),\ \forall x, x', x'' \in X. \tag{1}$$

The benefit of working with ultrametric spaces is that they can naturally be visualized as *dendrograms*. Given a finite set X, a dendrogram over X is a nested set of partitions $D(t)$ indexed by a resolution parameter $t \geq 0$, such that all points of X are clustered into singletons at $t = 0$, and into a single cluster for all t greater than some t_F. Because the partitions

Zane Smith is an undergraduate student with the Department of Computer Science and Engineering at The Ohio State University. Samir Chowdhury is a Ph.D. student with the Department of Mathematics at The Ohio State University. Facundo Mémoli is with the Departments of Mathematics and Computer Science and Engineering at The Ohio State University. Email at smith.9911@osu.edu, chowdhury.57@osu.edu, memoli@math.osu.edu.

are *nested*, once any two points merge into a cluster, they stay clustered together for all larger values of t. A useful result is that one may induce a dendrogram from an ultrametric, and vice versa, without losing any data [5], [6].

We define a network (X, ω_X) to be a finite set X together with a *weight function* $\omega_X : X \times X \to \mathbb{R}$. We will refer to the points of X as *nodes* and the images of ω_X as weights. Note that *none of the metric properties are assumed*; a priori, a network has no more structure than a generic $n \times n$ matrix with real entries. The collection of all networks will be denoted $\mathcal{N}$.

We will eventually be interested in the subclass of networks that satisfy the strong triangle inequality. We call these the family of *ultranetworks*, denoted $\mathcal{N}_{\text{ult}}$. Another interesting family is that of symmetric networks, denoted $\mathcal{N}^{\text{sym}}$. The family of symmetric ultranetworks will be denoted $\mathcal{N}^{\text{sym}}_{\text{ult}}$. Finally, there is the class of dissimilarity networks, denoted $\mathcal{N}^{\text{dis}}$, consisting of networks (X, ω_X) where ω_X takes values in the nonnegative reals and $\omega_X(x, x') = 0$ for any $x, x' \in X$ if and only if $x = x'$.

A. Hierarchical clustering methods and ultrametrics

In the setting of finite metric spaces, hierarchical clustering (HC) is a well-established method for representing complex data as a dendrogram that is easy to visualize and interpret. In particular, HC methods are heavily used in both data preprocessing and exploratory data analysis [7]. In recent years, numerous advances have been made towards formalizing the theory behind these methods. An axiomatic approach to clustering finite metric spaces has been put forward in [8], and further explored in [3], [6], [9], [10]. The stability of HC methods under small perturbations to the input data has been studied in [6], [11].

Within the axiomatic frameworks of [3], [6], [9], it has been established that in the context of finite metric spaces and dissimilarity networks, single linkage is the unique "appropriate" method for hierarchical clustering. Thus we limit our attention to SLHC in this paper.

III. The Network SLHC Method

Given a network (X, ω_X), we define a new weight function $\overline{\omega}_X : X \times X \to \mathbb{R}$ as follows:

$$\overline{\omega}_X(x, x') := \max\left(\omega_X(x, x), \omega_X(x, x'), \omega_X(x', x')\right),$$

for $x, x' \in X$. We define a *chain* from x to x' as an ordered set of points starting at x and reaching x':

$$c = \{x_0, x_1, x_2, \ldots, x_n : x_0 = x, x_1 = x', x_i \in X \text{ for all } i\}.$$

The collection of all chains joining x and x' will be denoted $C_X(x, x')$. We define the *cost* of a chain $c \in C_X(x, x')$ as follows: $\text{cost}_X(c) := \max_{x_i, x_{i+1} \in c} \overline{\omega}_X(x_i, x_{i+1})$. The *minimum chain cost* $u^{\mathcal{H}}_X$ on $X \times X$ is defined by:

$$u^{\mathcal{H}}_X(x, x') := \min_{c \in C_X(x, x')} \text{cost}_X(c).$$

By directly appealing to the definition we obtain:

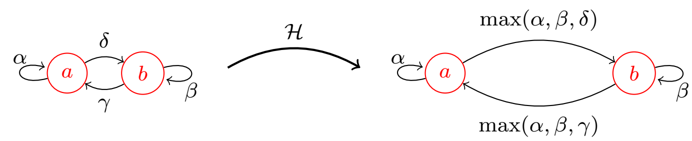


Fig. 1: Effect of applying $\mathcal{H}$ to a two-node network. Notice that the resulting network retains its asymmetry, as well as the weights of the self-loops.

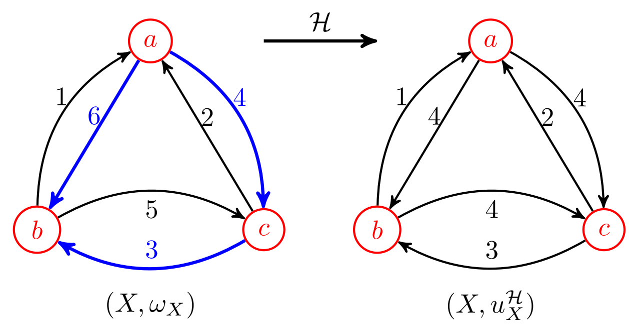


Fig. 2: The figure shows a network (X, ω_X) and its nSLHC output $(X, u^{\mathcal{H}}_X)$. Notice that $\text{ult}(X, \omega_X) = \Psi_X(a, c, b) = 2$, and $\text{ult}(X, u^{\mathcal{H}}_X) = 0$. Here ult refers to the ultranetwork constant defined in §III-B.

Proposition 1. *For any network (X, ω_X), the minimum chain cost $u^{\mathcal{H}}_X$ satisfies the strong triangle inequality.*

We now define the *network single linkage hierarchical clustering method* (nSLHC) $\mathcal{H} : \mathcal{N} \to \mathcal{N}_{\text{ult}}$ by:

$$\mathcal{H}(X, \omega_X) := (X, u^{\mathcal{H}}_X).$$

The effect of $\mathcal{H}$ on a simple two-node network is illustrated in Figure 1. The interaction of the method $\mathcal{H}$ with different input data is illustrated in Figure 2.

A. Stability and Characterization

For a data simplification method such as nSLHC to be practically useful, it needs to be *stable* in the following sense: small perturbations in the input data should result in small changes in the output. nSLHC enjoys the following quantitative stability property:

Theorem 2. *Let X be a finite set and let ω_1 and ω_2 be two different weight functions defined on $X \times X$. Write $(X, u^{\mathcal{H}}_1) := \mathcal{H}(X, \omega_1)$ and $(X, u^{\mathcal{H}}_2) := \mathcal{H}(X, \omega_2)$. Then we have:*

$$\|u^{\mathcal{H}}_1 - u^{\mathcal{H}}_2\|_{\ell^\infty(X \times X)} \le \|\omega_1 - \omega_2\|_{\ell^\infty(X \times X)}.$$

In analogy with dissimilarity networks [2], we are able to prove the following properties of nSLHC:

Proposition 3 (Property A1)**.** *For the two-point network $(X, \omega_X) = (\{p, q\}, \left(\begin{smallmatrix}\alpha & \gamma \\ \delta & \beta\end{smallmatrix}\right))$, we have $\mathcal{H}(X, \omega_X) = (\{p, q\}, \left(\begin{smallmatrix}\alpha & \Gamma \\ \Delta & \beta\end{smallmatrix}\right))$, where $\Gamma = \max\{\alpha, \beta, \gamma\}$ and $\Delta = \max\{\alpha, \beta, \delta\}$. This situation is illustrated in Figure 1.*

Proposition 4 (Property A2). *If $\phi : X \to Y$ satisfies $\omega_X(x,x') \geq \omega_Y(\phi(x),\phi(x'))$ for all $x,x' \in X$, then we also have $u^{\mathcal{H}}_X(x,x') \geq u^{\mathcal{H}}_Y(\phi(x),\phi(x'))$ for all $x,x' \in X$.*

In particular, we are able to prove the following *characterization* result for nSLHC: if a method of producing ultranetworks satisfies Properties A1-2, then its output ultranetworks are exactly the output ultranetworks of nSLHC. This result was previously known only for dissimilarity networks [2]. The properties of nSLHC that we establish in the next few sections, namely error bounds and visualization via treegrams, are our main contributions to the existing literature.

B. The ultranetwork constant and Error Bounds

Given a network (X,ω_X), consider the following function:

$$\Psi_X(x_1,x_2,x_3) := \omega_X(x_1,x_3) - \max\left(\omega_X(x_1,x_2),\omega_X(x_2,x_3)\right),$$

which measures for the three points $x_1,x_2,x_3 \in X$ the failure to satisfy the strong triangle inequality (1) in the triangle they define. We now define:

$$\mathrm{ult}(X) := \max_{x_1,x_2,x_3 \in X} \Psi_X(x_1,x_2,x_3),$$

and refer to this quantity as the *ultranetwork constant* of the network X. It measures the deviation of a network from satisfying the strong triangle inequality, and is a crucial quantity that we propose and study in this paper. A simple but important observation is that any ultranetwork X has $\mathrm{ult}(X) = 0$, and furthermore, if a network X has $\mathrm{ult}(X) = 0$ then X is actually an ultranetwork. In Figure 2, we illustrate a network with positive ultranetwork constant and its nSLHC output.

The main theorem of this paper is the following bound on the distortion to ω_X caused by the nSLHC method:

Theorem 5. *For any n-point network $(X,\omega_X) \in \mathcal{N}$, we have:*

$$\|\omega_X - u^{\mathcal{H}}_X\|_{\ell^\infty(X\times X)} \leq \log_2(2n)\,\mathrm{ult}(X).$$

Moreover, this bound is asymptotically tight.

We first observe that the ultranetwork constant of any network can be easily computed by just considering all triples of points in the space. Intuitively, spaces which "almost" satisfy the strong triangle inequality are already close to being ultranetworks, and an application of Theorem 5 shows that applying nSLHC to such networks does, in fact, cause only small distortion. A particularly useful application is in the setting of metric spaces, where one now has a simple answer to the question "How much loss does my metric dataset incur when represented by a dendrogram"—by Theorem 5, this quantity can be estimated by a function on just the set of triangles in the dataset.

Sketch of proof. Let $x,x'' \in X$ and suppose we have a chain $c = \{x,x',x''\}$ joining x and x'' with minimal chain cost. Then by unpacking the definition of ultrametricity, we have $|\omega_X(x,x'') - u^{\mathcal{H}}_X(x,x'')| \leq \mathrm{ult}(X)$. By induction, one proves that a minimal cost chain of length 2^k+1 would admit an inequality of the form $|\omega_X(x,x'') - u^{\mathcal{H}}_X(x,x'')| \leq k\,\mathrm{ult}(X)$. But the maximal length of any chain (possibly with some repetition) can be bounded by $2^{\log_2(2n)}+1$, leading to the inequality $|\omega_X(x,x'') - u^{\mathcal{H}}_X(x,x'')| \leq \log_2(2n)\,\mathrm{ult}(X)$. Tightness can be proved even in the setting of metric spaces: we are able to construct a sequence of finite metric spaces that realizes the logarithmic error rate. □

IV. Treegrams and Related Methods

Because exploratory data analysis is one of the main applications of nSLHC, one desirable feature would be a visualization that summarizes the output and is easy to interpret. In classical SLHC, the output is a dendrogram, and because dendrograms are easy to visualize and interpret, one would hope for a similar construction in the setting of networks. However, there is a fundamental inconsistency in this expectation: a network is assumed to be asymmetric, whereas a dendrogram is symmetric, so any method that produces a dendrogram-like structure from a network must pass through a symmetrization step. Numerous choices are possible for this step. In this paper we proceed as follows: Define the $\max$-symmetrization map $\mathcal{S} : \mathcal{N} \to \mathcal{N}^{\mathrm{sym}}$ by: $\mathcal{S}(X,\omega_X) = (X,\widetilde{\omega}_X)$, where

$$\widetilde{\omega}_X(x,x') = \max(\omega_X(x,x'),\omega_X(x',x)) \text{ for all } x,x' \in X.$$

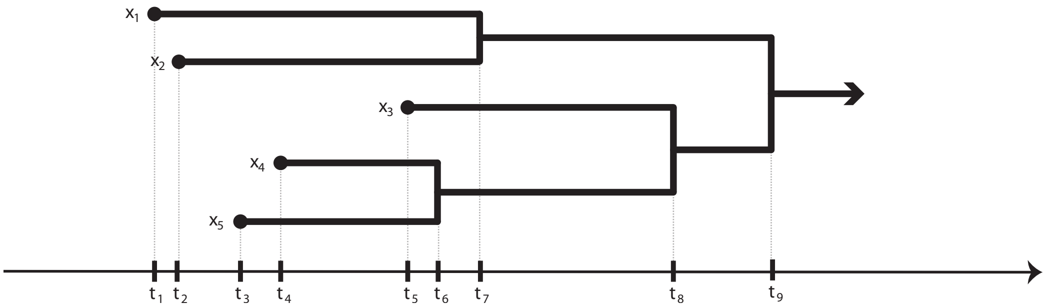
Notice that one may decide to symmetrize the network first and then apply nSLHC, or apply nSLHC first and then symmetrize the output. This apparent dichotomy leads to methods analogous to the *reciprocal* and *nonreciprocal* clustering methods described in [3]; further choices can be made to obtain methods that interpolate between these extremes. For the purposes of this paper we restrict ourselves to the case where symmetrization is applied *after* applying nSLHC. The motivation behind using this method is that it is truly sensitive to asymmetry, in contrast to the alternative of applying the symmetrization step first. To be more precise, consider (X,ω_X) in Figure 2, and suppose the edge weights between nodes a and c were swapped. Applying the symmetrization first would nullify the effect of this swap, whereas applying nSLHC first would fully capture this effect.

Let $\mathcal{T} = \mathcal{S}\circ\mathcal{H} : \mathcal{N} \to \mathcal{N}^{\mathrm{sym}}_{\mathrm{ult}}$ denote the method obtained by first applying nSLHC and then symmetrizing the output.

A. Treegrams

We now construct a tree structure that faithfully represents the symmetric ultranetworks that occur as an output of the method $\mathcal{T} = \mathcal{S}\circ\mathcal{H}$. We call this construction a *treegram*, and illustrate its appearance in Figure 3. We urge the reader to view the figure before the formal definition.

Recall that given a finite set X a partition of X is any collection $P = \{B_1,\ldots,B_k\}$ where each B_i is a subset of X referred to as a *block* or cluster of the partition P. Different blocks of P also need to be disjoint: for $i \neq j$ one has $B_i \cap B_j = \varnothing$, and the totality of the blocks must cover X completely: $\cup_{i=1}^k B_k = X$. From now on, for a finite set X we denote by $\mathrm{Part}(X)$ the set of all partitions of X. In order to keep track of partial clustering information, we also



$$\begin{array}{|c|c|c|c|c|}\hline t_1 & t_7 & t_9 & t_9 & t_9 \\ \hline & t_2 & t_9 & t_9 & t_9 \\ \hline & & t_5 & t_8 & t_8 \\ \hline & & & t_4 & t_6 \\ \hline & & & & t_3 \\ \hline \end{array}$$

$$u_{T_X}$$

Fig. 3: The figure on the left is a graphical representation of a treegram T_X. Notice that for $t < t_1$, $X_t = \varnothing$. For example, for $t \in [t_2, t_3)$, $X_t = \{x_1, x_2\}$ and $P_t = \{\{x_1\}, \{x_2\}\}$. Also, for example for $t \in [t_7, t_8)$, $X_t = \{x_1, x_2, x_3, x_4, x_5\}$ and $P_t = \{\{x_1, x_2\}, \{x_3\}, \{x_4, x_5\}\}$. The figure on the right is the corresponding symmetric ultranetwork u_{T_X}. Note that u_{T_X} can be read off of T_X, and T_X can be reconstructed from u_{T_X}.

consider *sub-partitions* of a finite set X: these will be pairs (X', P') where $X' \subseteq X$ and $P' \in \mathrm{Part}(X')$. We denote by $\mathrm{SubPart}(X)$ the set of all sub-partitions of X.

We will use the following simple fact: if $A' \subset A$ are sets and $P \in \mathrm{Part}(A)$, then the restricted partition $P|_{A'} := \bigcup_{B \in P} B \cap A'$ is a partition of A'.

Definition 1. *Let X be a finite set. A* treegram *over X is a function $T_X : \mathbb{R} \longrightarrow \mathrm{SubPart}(X)$ such that if for each $t \in \mathbb{R}$ we write $T_X(t) = (X_t, P_t)$ then*

1) *(hierarchy) For $t' \geq t$, $X_t \subseteq X_{t'}$ and $P_{t'}|_{X_t}$ is* coarser *than P_t.*
2) *$\exists t_F \in \mathbb{R}$ such that for all $t \geq t_F$, $X_t = X$ and P_t is the one block partition of X.*
3) *$\exists t_I \in \mathbb{R}$ such that X_t is empty for all $t < t_I$.*
4) *(right continuity) For all $t \in \mathbb{R}$ there exists $\varepsilon > 0$ such that $T_X(t') = T_X(t)$ for all $t' \in [t, t + \varepsilon]$.*

The definition is analogous but strictly more general than that of dendrograms [6]. The parameter t is referred to as *resolution*. Conditions 2 and 3 are called *boundary conditions*, and they specify the resolutions at which all the nodes of X are clustered together, and at which we only have the empty cluster. Condition 1 (*hierarchy*) emphasizes that as the resolution t increases, clusters can only be combined, not separated. Finally, we remark that the condition of *right continuity* is used to satisfy a technical condition in the proof of Theorem 6 below.

Consider as an example the treegram illustrated in Figure 3. In this case, the boundary conditions are $t_I = t_1$ and $t_F = t_9$. In the case of a standard dendrogram [5], all the points in the set appear simultaneously as singletons at the initial time $t_I = 0$. Treegrams are more general and in the example in Figure 3 we see the appearance of new nodes as far as t_5. The hierarchical structure is particularly easy to see from the figure; also note that nodes can only be combined (and not separated) as t increases.

In what follows, we explain how to obtain treegrams from symmetric ultranetworks, and vice versa.

From symmetric ultranetworks to treegrams. Let (X, u_X) be a symmetric ultranetwork. For each $t \in \mathbb{R}$ let $R_t = \{(x, x') \in X \times X : u_X(x, x') \leq t\}$. Then let $X_t = \pi_1(R_t) = \pi_2(R_t)$, where π_1 and π_2 are projections onto the first and second coordinates. Note that the last inequality follows because u_X is symmetric. If $X_t \neq \varnothing$, consider the relation $\sim_t$ on X_t defined as follows: $x \sim_t x' \iff (x, x') \in R_t$. One can verify that $\sim_t$ is a valid equivalence relation on X_t.

Now we have for each $t \in \mathbb{R}$ a possibly empty set X_t together with (a possibly empty) equivalence relation $\sim_t$ on the set. This is equivalent to saying that for each $t \in \mathbb{R}$ we have a pair (X_t, P_t) where $P_t \in \mathrm{Part}(X_t)$ is the partition induced by $\sim_t$. So we set $T_X(t) = (X_t, P_t)$.

Notice that if $t' \geq t$, then $R_{t'} \supseteq R_t$ by definition, and so $X_{t'} \supseteq X_t$ as well. Then it follows that $P_{t'}|_{X_t}$ is coarser than P_t. Thus the process described above defines a map from symmetric ultranetworks to treegrams, given by $u_X \mapsto T_X$.

It is also possible to define a lossless map from treegrams into ultranetworks. Details are posted in [4].

Theorem 6. *Any symmetric ultranetwork has a lossless realization as a treegram, and any treegram has a lossless realization as a symmetric ultranetwork.*

By virtue of this theorem, we have a completely faithful visual representation of symmetric ultranetworks.

V. An Application to a Social Network

Scenario: Assume n new teachers $A_1, \cdots A_n$ move to a city $\Omega \subset \mathbb{R}^2$ at different locations at different times $t_1, \cdots, t_n$. We model the initial locations as independent random variables uniformly distributed in Ω, and model the t_is as independent random variables with exponential distribution and common mean $T_I > 0$. The joint movement of the different teachers inside the city Ω is modeled as n independent random walks each respecting the initial conditions above. When two teachers A_i and A_j find themselves within a distance $R > 0$ of each other, they will attempt to exchange contact information, which we model as two independent processes with probability $\alpha \in [0, 1]$ for taking place: A_i will attempt to establish a one-directional link with A_j, and A_j will attempt to establish a one-directional link with A_i. The process runs from time 0 to a final time $T > 0$. Parameter α should be interpreted as an average "sociability" measure for the cohort of teachers.

Let $\mathbf{A} = \{1, \ldots, n\}$. Then, by keeping track of the history of pairwise of interactions between teachers, a network $(\mathbf{A}, \omega_{\mathbf{A}})$ consisting of exactly one node per teacher is defined where for $i, j \in \mathbf{A}$ with $i \neq j$ the weight $\omega_{\mathbf{A}}(i,j)$ is set to be as the *first* time the one directional link $i \to j$ alluded to above is established. Diagonal weights are defined as $\omega_{\mathbf{A}}(i,i) = t_i$. Informally, this network represents the *grapevine* through which colleagues can talk about their jobs, relay news, express successes and failures, and gain social recognition.

Goal: To detect the *first time* $\tau = \tau(\Omega, n, T, T_I, R, \alpha)$ when the network of teachers is able to relay a "message through the grapevine" from *any node* i_0 *to any other node* j_0.

Procedure: For any given realization of the underlying stochastic process the goal can be accomplished by first applying the method $\mathcal{T}$ to $(\mathbf{A}, \omega_{\mathbf{A}})$ to obtain the symmetric ultranetwork $(\mathbf{A}, u^{\mathcal{T}}_{\mathbf{A}})$; then the value of τ is equal to $\tau^* := \max_{i,j} u^{\mathcal{T}}_{\mathbf{A}}(i,j)$.

Indeed, imagine that for some $\delta \geq 0$ teachers A_i and A_j are such that $u^{\mathcal{T}}_{\mathbf{A}}(i,j) \leq \delta$. Then, by the definition of $\mathcal{T}$, this means that both $u^{\mathcal{H}}_{\mathbf{A}}(i,j) \leq \delta$ and $u^{\mathcal{H}}_{\mathbf{A}}(j,i) \leq \delta$. Now, from the definition of $\mathcal{H}$ it follows that there exist chains $c \in C(i,j)$ and $c' \in C(j,i)$ with total chains costs not larger than δ. Consider what this means in the context of our application. Take c; the fact that $\mathrm{cost}_{\mathbf{A}}(c) \leq \delta$ means that any two consecutive points i_p and i_{p+1} in c (which are indices of teachers in $\mathbf{A}$) are such that the link $i_p \to i_{p+1}$ was established at time $\leq \delta$. Since this is true for *all consecutive pairs* in c, it means that by broadcasting a message at time δ, teacher A_i can reach teacher A_j. By analyzing the chain c' one can similarly conclude that by time δ teacher A_j can send a message to teacher A_i by relying on colleagues along the chain c'. Finally, it follows that when $\delta = \tau^*$, for any pair of teachers A_i and A_j it is possible to find two chains joining them with cost at most τ^*. By tracing definitions, one can see that τ^* is the *first* time this event can happen. Note that this may be a much smaller value than the first time when any pair of teachers can trade messages directly, which is what we would get by simply symmetrizing the original network.

Results: We considered Ω as a square grid-like discretization of $[0,1] \times [0,1]$ consisting of 21 equidistant points in each direction. Any point not on the boundary of the grid was connected to all 8 neighbors.. The random walk on the resulting graph was coded in matlab. We carried a simulation where $n \in \{5, 6, 7, \ldots, 50\}$, $R \in \{0.06, 0.07, \ldots, 0.5\}$, $T_I \in \{20, 100\}$, and $\alpha = \{0.1, 0.2, \ldots, 1\}$. For each of the 4 parameters the corresponding value of τ was averaged over 50 repetitions. An example treegram together with results and interpretation are shown in Figure 4. Results corresponding to other combinations of parameters, and a movie with the trajectory corresponding to the treegram in the figure can be viewed at [4].

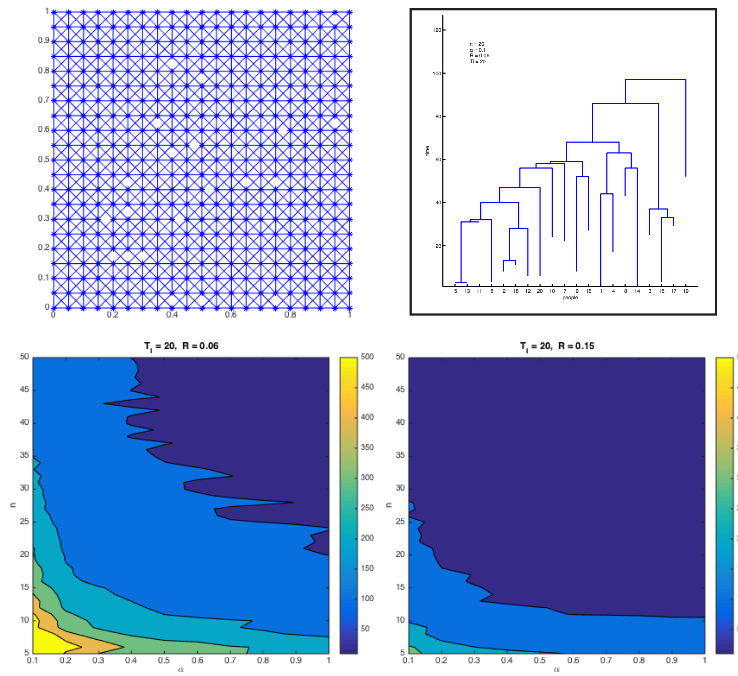


Fig. 4: **Top left:** Grid of discretization of $[0,1] \times [0,1]$. **Top right:** Treegram corresponding to parameters $n = 20, \alpha = 0.1, R = 0.06, T_I = 20$. **Bottom:** Plots of contour lines for τ as a function of n and α for two different values of R. The value of T_I was fixed at 20. Note that whereas for the smaller value $R = 0.06$ both an increase in n and α contribute to a decrease of τ, for $R = 0.15$ the dominant parameter is n.

References

[1] R. Rammal, G. Toulouse, and M. A. Virasoro, "Ultrametricity for physicists," *Reviews of Modern Physics*, vol. 58, no. 3, p. 765, 1986.

[2] G. E. Carlsson, F. Mémoli, A. Ribeiro, and S. Segarra, "Hierarchical quasi-clustering methods for asymmetric networks," in *ICML 2014*, 2014, pp. 352–360.

[3] G. Carlsson, F. Mémoli, A. Ribeiro, and S. Segarra, "Hierarchical clustering methods and algorithms for asymmetric networks," in *Signals, Systems and Computers, 2013 Asilomar Conference on*. IEEE, 2013, pp. 1773–1777.

[4] "Hierarchical representations of network data: Supplementary material," https://research.math.osu.edu/networks/asilo.html.

[5] N. Jardine and R. Sibson, *Mathematical Taxonomy*, ser. Wiley series in probability and mathematical statistics. Wiley, 1971. [Online]. Available: https://books.google.com/books?id=ka4KAQAAIAAJ

[6] G. Carlsson and F. Mémoli, "Characterization, stability and convergence of hierarchical clustering methods," *The Journal of Machine Learning Research*, vol. 11, pp. 1425–1470, 2010.

[7] I. Guyon, U. Von Luxburg, and R. C. Williamson, "Clustering: Science or art," in *NIPS 2009 workshop on clustering theory*, 2009, pp. 1–11.

[8] J. Kleinberg, "An impossibility theorem for clustering," *Advances in neural information processing systems*, pp. 463–470, 2003.

[9] R. B. Zadeh and S. Ben-David, "A uniqueness theorem for clustering," in *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*. AUAI Press, 2009, pp. 639–646.

[10] S. Ben-David and M. Ackerman, "Measures of clustering quality: A working set of axioms for clustering," in *Advances in neural information processing systems*, 2009, pp. 121–128.

[11] A. Martínez-Pérez, "Gromov–Hausdorff stability of linkage-based hierarchical clustering methods," *Advances in Mathematics*, vol. 279, no. Complete, pp. 234–262, 2015.

VI. Appendix

Proposition 1. *For any network (X,ω_X), the minimum chain cost $u_X^{\mathcal{H}}$ satisfies the strong triangle inequality.*

Proof. Let $x,x',x''\in X$. We wish to show:

$$u_X^{\mathcal{H}}(x,x'')\leq \max\left(u_X^{\mathcal{H}}(x,x'),u_X^{\mathcal{H}}(x',x'')\right).$$

Let c',c'' be chains such that $u_X^{\mathcal{H}}(x,x')=\mathrm{cost}_X(c')$ and $u_X^{\mathcal{H}}(x',x'')=\mathrm{cost}_X(c'')$. Consider the composed chain $c:=c''\circ c'$ beginning at x and ending at x''. Then $\mathrm{cost}_X(c)\leq\max\left(\mathrm{cost}_X(c'),\mathrm{cost}_X(c'')\right)$. The result follows. □

Theorem 2. *Let X be a finite set and let ω_1 and ω_2 be two different weight functions defined on $X\times X$. Write $(X,u_1^{\mathcal{H}}):=\mathcal{H}(X,\omega_1)$ and $(X,u_2^{\mathcal{H}}):=\mathcal{H}(X,\omega_2)$. Then we have:*

$$\|u_1^{\mathcal{H}}-u_2^{\mathcal{H}}\|_{\ell^\infty(X\times X)}\leq\|\omega_1-\omega_2\|_{\ell^\infty(X\times X)}.$$

Proof. Let $x,x'\in X$ be such that $\|u_1^{\mathcal{H}}-u_2^{\mathcal{H}}\|_{\ell^\infty(X\times X)}=|u_1^{\mathcal{H}}(x,x')-u_2^{\mathcal{H}}(x,x')|$. Then we have

$$|\overline{\omega}_1(x,x')-\overline{\omega}_2(x,x')|\leq\|\omega_1-\omega_2\|_{\ell^\infty(X\times X)}.$$

Next let $c\in C_{(X,\omega_1)}(x,x')$ be an optimal chain, i.e. a chain such that

$$\max_{x_i,x_{i+1}\in c}\overline{\omega}_1(x_i,x_{i+1})=u_1^{\mathcal{H}}(x,x').$$

Here we are writing $c=\{x_0=x,x_1,\ldots,x_n=x'\}$. Then for any $1\leq i\leq n$, we obtain:

$$\begin{aligned}\overline{\omega}_2(x_i,x_{i+1})&\leq\overline{\omega}_1(x_i,x_{i+1})+\|\omega_1-\omega_2\|_{\ell^\infty(X\times X)}\\&\leq u_1^{\mathcal{H}}(x,x')+\|\omega_1-\omega_2\|_{\ell^\infty(X\times X)}.\end{aligned}$$

Since this holds for any $i\in\{1,\ldots,n\}$, we can minimize over all chains to obtain:

$$u_2^{\mathcal{H}}(x,x')-u_1^{\mathcal{H}}(x,x')\leq\|\omega_1-\omega_2\|_{\ell^\infty(X\times X)}.$$

A similar argument shows:

$$u_1^{\mathcal{H}}(x,x')-u_2^{\mathcal{H}}(x,x')\leq\|\omega_1-\omega_2\|_{\ell^\infty(X\times X)}.$$

This concludes the proof. □

Proposition 3 (Property A1)**.** *For the two-point network $(X,\omega_X)=\left(\{p,q\},\left(\begin{smallmatrix}\alpha&\gamma\\\delta&\beta\end{smallmatrix}\right)\right)$, we have $\mathcal{H}(X,\omega_X)=\left(\{p,q\},\left(\begin{smallmatrix}\alpha&\Gamma\\\Delta&\beta\end{smallmatrix}\right)\right)$, where $\Gamma=\max\{\alpha,\beta,\gamma\}$ and $\Delta=\max\{\alpha,\beta,\delta\}$.*

Proof. Note that the minimum cost chain from p to itself is the stationary chain, so $u_X^{\mathcal{H}}(p,p)=\alpha$. Similarly $u_X^{\mathcal{H}}(q,q)=\beta$. The values Γ and Δ appear from the definition of $\overline{\omega}_X$. □

Proposition 4 (Property A2)**.** *If $\phi:X\to Y$ satisfies $\omega_X(x,x')\geq\omega_Y(\phi(x),\phi(x'))$ for all $x,x'\in X$, then we also have $u_X^{\mathcal{H}}(x,x')\geq u_Y^{\mathcal{H}}(\phi(x),\phi(x'))$ for all $x,x'\in X$.*

Proof. It follows from the assumption that we have:

$$\overline{\omega}_X(x,x')\geq\overline{\omega}_Y(\phi(x),\phi(x'))\text{ for all }x,x'\in X.$$

Let $c\in C_X(x,x')$ be a chain such that $\mathrm{cost}_X(c)=u_X^{\mathcal{H}}(x,x')$. Write $c=\{x=x_0,\ldots,x_n=x'\}$. Consider the chain $\phi(c)=\{\phi(x_0),\ldots,\phi(x_n)\}$. Then $\phi(c)\in C_Y(\phi(x),\phi(x'))$, and $\mathrm{cost}_Y(\phi(c))\leq\mathrm{cost}_X(c)$ by the preceding observation. The result follows. □

Theorem 5. *For any n-point space $X\in\mathcal{N}$, write $(X,u_X)=\mathcal{H}(X,\omega_X)$. Then we have:*

$$\|\omega_X-u_X\|_{l^\infty(X\times X)}\leq\log_2(2n)\,\mathrm{ult}(X).$$

Moreover, this bound is asymptotically tight.

Proof. Let $\delta=\mathrm{ult}(X,d_X)$. First we claim that for any sequence of 2^k+1 points, we have:

$$\max_{1\leq i\leq 2^k}\overline{\omega}_X(x_i,x_{i+1})\geq\overline{\omega}_X(x_1,x_{2^k+1})-k\delta.$$

To see this, we proceed by induction. Notice that for $k=1$, we have the following by the definition of the ultranetwork constant:

$$\begin{aligned}\omega_X(x_1,x_3)&\leq\max\left(\omega_X(x_1,x_2),\omega_X(x_2,x_3)\right)+\delta\\&\leq\max\left(\overline{\omega}_X(x_1,x_2),\overline{\omega}_X(x_2,x_3)\right)+\delta.\end{aligned}$$

Also, we have

$$\begin{aligned}&\omega_X(x_1,x_1)\leq\overline{\omega}_X(x_1,x_2),\\&\omega_X(x_3,x_3)\leq\overline{\omega}_X(x_2,x_3).\end{aligned}$$

Then it follows that:

$$\overline{\omega}_X(x_1,x_3)\leq\max\left(\overline{\omega}_X(x_1,x_2),\overline{\omega}_X(x_2,x_3)\right)+\delta.$$

This proves the base case. Next let $k\in\mathbb{N}$, and suppose the claim holds for k. By the base case, we obtain:

$$\overline{\omega}_X(x_1,x_{2^{k+1}+1})\leq\max\left(\overline{\omega}_X(x_1,x_{2^k+1}),\overline{\omega}_X(x_{2^k+1},x_{2^k+1+2^k})\right)+\delta.$$

But by the induction step, we have:

$$\begin{aligned}\overline{\omega}_X(x_1,x_{2^k+1})&\leq\max_{1\leq i\leq 2^k}\overline{\omega}_X(x_i,x_{i+1})+k\delta\\\overline{\omega}_X(x_{2^k+1},x_{2^k+1+2^k})&\leq\max_{2^k+1\leq i\leq 2^{k+1}}\overline{\omega}_X(x_i,x_{i+1})+k\delta\end{aligned}$$

Thus, taking the maximum of the two, we obtain:

$$\overline{\omega}_X(x_1,x_{2^{k+1}+1})\leq\max_{1\leq i\leq 2^{k+1}}\overline{\omega}_X(x_i,x_{i+1})+(k+1)\delta.$$

This proves the claim. Next, let $x,x'\in X$. Let $c\in C(x,x')$. Write $c=\{x=x_1,\ldots,x_p=x'\}$. Note that if c contains any repetition, i.e. if there exist $i<j\leq p$ with $x_i=x_j$, then we may replace c by $c'=\{x_1,\ldots,x_i,x_{j+1},\ldots,x_p\}$. Thus by reindexing if necessary, we obtain a chain of distinct elements $c'=\left\{x=x_1',\ldots,x_q'=x'\right\}$, with $q<p$. Also note that $\mathrm{cost}(c')\leq\mathrm{cost}(c)$. Next let k be the greatest integer such that $2^k\leq n$. Then we have $n\leq 2^{k+1}\leq 2n$. Since c' has length $q\leq n$, we can define:

$$\bar{c}=\left\{x_1',\ldots,x_q',x_q',\ldots,x_q'\right\},$$

where $\bar{c}$ is obtained from c' by padding copies of the endpoint x_q' until $\bar{c}$ has length $2^{k+1}+1$. Notice that $\mathrm{cost}(\bar{c})=\mathrm{cost}(c')$.

By applying the claim to $\bar{c}$, we obtain $\mathrm{cost}(c) \geq \mathrm{cost}(\bar{c}) \geq d_X(x, x') - (k+1)\delta$. Since c was arbitrary, we also have:

$$\min_{c \in C(x,x')} \mathrm{cost}(c) = u_X(x, x') \geq d_X(x, x') - (k+1)\delta.$$

Since x, x' were also arbitrary, we obtain:

$$\max_{x,x' \in X} \left(d_X(x, x') - u_X(x, x') \right) \leq (k+1)\delta \leq \log_2(2n)\,\mathrm{ult}(X).$$

This concludes the proof of the upper bound.

An example to show tightness. We formulate our example in terms of metric spaces. First we need to describe some constructions. Our main tool is a *metric transform*, which is a continuous, monotone increasing function $\Psi : \mathbb{R}_+ \to \mathbb{R}_+$ with $\Psi(0) = 0$. In particular, Ψ maps metrics to metrics. For any metric space (X, d_X), we let $\Psi(X)$ denote $(X, \Psi(d_X))$. For spaces X and transforms $\Psi(X)$ such that $\mathrm{ult}(\Psi(X)) \neq 0$, we define the following quantity:

$$R(\Psi) := \frac{\|\Psi(d_X) - \Psi(u_X)\|_\infty}{\mathrm{ult}(\Psi(X))}.$$

For any $x, x' \in X$, we also define:

$$d_X^{(1)}(x, x') := \min \left\{ \max \left(d_X(x, z), d_X(z, x') \right) : z \in X \right\}.$$

One can verify the following reformulation of $\mathrm{ult}(X)$:

$$\mathrm{ult}(X) = \|d_X - d_X^{(1)}\|_\infty. \tag{2}$$

Let $0 < \varepsilon \ll 1$. Consider the *snowflake* metric transform $\Psi_\varepsilon(\alpha) = \alpha^\varepsilon$. In the limit, when $\varepsilon \to 0$, all non-zero distances would become 1. That is, $\lim_{\varepsilon \downarrow 0} \Psi_\varepsilon(X)$ would be equal to the metric space with underlying set X and the *discrete metric* (i.e. the metric attaining only the values 0 and 1). Note that the discrete metric is actually an ultrametric.

Next let X be a finite set with $n > 1$ points, E a subset of $X \times X$, such that $G = (X, E)$ becomes a connected graph with edge weights 0 (for absence of an edge) or 1 (for presence of an edge). Let (X, d_X) represent the finite metric space with n points arising from computing the graph (or path length) distance on G. Specifically,

$$d_X(x, x') := \min\{|c| : c \in C(x, x')\},$$

where $C(x, x')$ is the set of all chains connecting x and x'. In this case, u_X, the SLHC output ultrametric, will be 1 between different points. Also note that d_X takes integer values, and for any two points x, x', we have $d_X^{(1)}(x, x') = \lceil \frac{d_X(x,x')}{2} \rceil$. Such a space will be called a *graph metric space*.

Proof of tightness. For $n \geq 2$, fix $\varepsilon_n = \frac{1}{\log^2 n}$ and let $\widehat{X_n}$ denote the graph metric space on $(n+1)$ points. Note that $\mathrm{diam}(\widehat{X_n}) = n$ for each n. Next let $X_n = \Psi_{\varepsilon_n}(\widehat{X_n})$. Notice that the numerator of $R(\Psi_{\varepsilon_n})$ is now:

$$\max_{\alpha \in [0,n]} (\alpha^{\varepsilon_n} - 1^{\varepsilon_n}) = n^{\varepsilon_n} - 1.$$

By applying Equation 2, the denominator of $R(\Psi_{\varepsilon_n})$ becomes:

$$\begin{aligned}
\max_{\alpha \in [0,n]} (\alpha^{\varepsilon_n} - \left\lceil \frac{\alpha}{2} \right\rceil^{\varepsilon_n}) &\approx \max_{\alpha \in [0,n]} \left(\alpha^{\varepsilon_n} - \left(\frac{\alpha}{2} \right)^{\varepsilon_n} \right) \\
&= n^{\varepsilon_n}(1 - 2^{-\varepsilon_n}) \\
&= \left(\frac{n}{2} \right)^{\varepsilon_n} (2^{\varepsilon_n} - 1).
\end{aligned}$$

Notice that equality holds above for even values of n. The expression for $R(\Psi_{\varepsilon_n})$ now becomes:

$$\begin{aligned}
R(\Psi_{\varepsilon_n}) = \frac{(n^{\varepsilon_n} - 1)2^{\varepsilon_n}}{n^{\varepsilon_n}(2^{\varepsilon_n} - 1)} &= \frac{n^{\varepsilon_n} - 1}{n^{\varepsilon_n}} \cdot \frac{2^{\varepsilon_n}}{2^{\varepsilon_n} - 1} \\
&= \frac{e^{\varepsilon_n \log n} - 1}{e^{\varepsilon_n \log n}} \cdot \frac{2^{\varepsilon_n}}{e^{\varepsilon_n \log 2} - 1} \\
&= \frac{e^{\frac{1}{\log n}} - 1}{e^{\frac{1}{\log n}}} \cdot \frac{2^{\varepsilon_n}}{e^{\frac{\log 2}{\log^2 n}} - 1}.
\end{aligned}$$

For large n, this becomes $\approx \dfrac{\frac{1}{\log n}}{\frac{\log 2}{\log^2 n}} = \log_2(n) \approx \log_2(2n)$.

This proves tightness, and we conclude our proof. ☐

Additional details of the treegram construction.

From treegrams to symmetric ultranetworks. Let T_X be a treegram over X. For each $x, x' \in X$, define:

$$u_{T_X}(x, x') := \min\{t \in \mathbb{R} |\, x, x' \in X_t \,\text{and}\, x \sim_t x'\}.$$

One can see that u_{T_X} defines a symmetric ultranetwork over X. Here symmetry follows because T_X is symmetric. We need to check the strong triangle inequality. Let $x, x', x'' \in X$. Let $t = \max(u_{T_X}(x, x''), u_{T_X}(x'', x'))$. Then $x, x', x'' \in X_t$ and $x \sim_t x'' \sim_t x'$. Thus $u_{T_X}(x, x') \leq t$.

This defines a map from treegrams to symmetric ultranetworks given by $T_X \mapsto u_{T_X}$.

Theorem 6. *Any symmetric ultranetwork has a lossless realization as a treegram, and any treegram has a lossless realization as a symmetric ultranetwork.*

More specifically, given a finite set X, let $\mathrm{Tree}(X)$ *denote the set of all treegrams on X and let* $\mathrm{Ultra}(X)$ *denote the set of all symmetric ultranetworks on X. Next consider the maps* $\Phi : \mathrm{Tree}(X) \to \mathrm{Ultra}(X)$ *and* $\Psi : \mathrm{Ultra}(X) \to \mathrm{Tree}(X)$ *given by $T_X \mapsto u_X$ and $u_X \mapsto T_X$. Then we have* $\Phi \circ \Psi = \mathrm{Id}_{\mathrm{Ultra}(X)}$ *and* $\Psi \circ \Phi = \mathrm{Id}_{\mathrm{Tree}(X)}$.

Proof. Let (X, ω_X) be a symmetric ultranetwork. Let $x, x' \in X$, and let $t = \omega_X(x, x')$. Then $(x, x') \in R_t$, where R_t is defined as before:

$$R_t := \{(x, x') \in X \times X : \omega_X(x, x') \leq t\}.$$

In particular, $(x, x') \notin R_s$ for $s < t$. Thus $x \sim_t x'$. So $x, x' \in X_t$ and $P_t(x) = P_t(x')$, where $(X_t, P_t) = T_X(t)$ and $T_X = \Psi(\omega_X)$. Since $(x, x') \notin R_s$ for $s < t$, it follows that $u_X(x, x') = t$, where $u_X = \Phi(T_X) = \Phi(\Psi(\omega_X))$. This holds for arbitrary $x, x' \in X$. Hence $u_X = \omega_X$, so $\Phi \circ \Psi = \mathrm{Id}_{\mathrm{Ultra}(X)}$.

Next let (X, T_X) be a treegram with subpartitions (X_t, P_t) for $t \in \mathbb{R}$, where the equivalence relation defining P_t is denoted $\sim_t$. Denote $\Phi(T_X)$ by u_X. For each $t \in \mathbb{R}$, let R_t be the relation on $X \times X$ induced by u_X. Let $X'_t = \pi_1(R_t) = \pi_2(R_t)$ and let $P'_t \in \mathrm{Part}(X'_t)$ be the partition induced by $\sim'_t$, where $x \sim'_t x'$ if and only if $(x, x') \in R_t$. Let $T'_X(t) = (X'_t, P'_t)$. Notice that $T'_X = \Psi(u_X)$.

But notice that $x \sim'_t x'$ if and only if $(x, x') \in R_t$ if and only if $x \sim_t x'$. Thus $\sim'_t$ and $\sim_t$ define the same equivalence relation for each $t \in \mathbb{R}$. It follows then that $P'_t = P_t$ for each $t \in \mathbb{R}$. We also have $X'_t = \pi_1(R_t) = X_t$. Thus $T'_X = T_X$ for all $t \in \mathbb{R}$. So $\Psi \circ \Phi = \mathrm{Id}_{\mathrm{Tree}(X)}$. This proves the result. $\square$