Mathematical hierarchies and Biology (B. Mirkin, F.R. McMorris, F. Roberts, A. Rzhetsky, eds.), *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, Amer. Math. Soc., Providence, RI, 183-208.

Circular orders of tree metrics, and their uses for the reconstruction and fitting of phylogenetic trees

Vladimir Makarenkov and Bruno Leclerc

ABSTRACT. The circular orders associated with the planar drawings of an X-tree (or phylogenetic tree) have been studied by several authors. They allow an encoding of an X-tree by $2n$-3 numbers (where n is the number of elements of X), the lengths of some paths between leaves of the tree. It is shown here that circular orders are the same as those obtained from the table of a tree metric by a construction due to Yushmanov [28]. It is also observed that this construction applies to any dissimilarity, tree metric or not. Several fast algorithms (of complexity $O(n^2)$) are derived from these results: for the determination of a Yushmanov order; for the reconstruction of the valued X-tree represention of a tree metric; for the recognition of a tree metric; and for the fitting of a tree metric to a given dissimilarity; the fitting method is based on successive local least squares approximations. Tested on various experimental and real data, it gives satisfactory results.

RÉSUMÉ. Plusieurs auteurs ont étudié les ordres circulaires associés aux représentations planaires des X-arbres (ou arbres phylogénétiques), en particulier pour le codage d'un X-arbre valué par $2n$-3 longueurs de chemins d'une feuille à une autre (n étant le nombre d'éléments de X). On montre dans cet article que ces ordres circulaires sont les mêmes que ceux obtenus à partir de la table d'une distance d'arbre par une construction due à Yushmanov [28]. De plus, cette construction s'applique à toute dissimilarité, distance d'arbre ou non. Ces résultats permettent d'obtenir plusieurs algorithmes rapides (de complexité $O(n^2)$) : pour la détermination d'un ordre de Yushmanov ; pour la reconstruction du X-arbre valué représentant une distance d'arbre ; pour la reconnaissance d'une distance d'arbre ; et pour l'ajustement d'une distance d'arbre à une dissimilarité d donnée. Ce dernier comporte des approximations locales successives par les moindres carrés. Il conduit à une procédure d'ajustement qui donne des résultats intéressants sur les exemples sur lesquels elle a été testée.

This research was partly supported by Esprit LTR Project n° 20244-ALCOM-IT. The authors are indebted to the Département Informatique and the Centre Documentaire of the École Nationale Supérieure des Télécommunications de Paris for support in bibliographic research and programming work. They also wish to thank Alain Guénoche for helpful advice and comments.

1. Introduction

Let X be a finite set with n elements. We consider a tree T with leaves labelled according to X. When T is endowed with a non-negative edge length function, a *tree metric* d on X is defined as follows: for all $x, y \in X$, $d(x,y)$ is the length of the unique path of T between the leaves x and y.

The classical representation of a metric on X by the matrix of its values for all pairs of elements of X is fairly redundant in the case of a tree metric. Two ways of summarizing a tree metric by $2n$-3 entries, the minimum possible number, are presented in Leclerc [17]. Thay are both based on minimum spanning trees; previously, Barthélemy and Guénoche [2], generalizing a result of Chaiken, Dewdney and Slater [7], have shown the entire matrix of a tree metric d to be defined by its restriction to another set R of $2n$-3 entries associated with a so-called *diagonal plane order* on X. Diagonal plane orders are defined in geometrical terms, in relation with the planar drawings of T. On the contrary, the knowledge of such a graphical representation is not needed for the determination of *the linear orders defined by Yushmanov* [28], together with another way of summarizing a tree metric by $2n$-3 entries; Yushmanov orders are directly obtained from the matrix of d.

It is shown here in Section 3 that circular and Yushmanov orders are in fact exactly the same. This paper is devoted to the study of these orders, and their use in an approach of the problems of recognition, reconstruction and fitting of tree metrics.

The paper is organized as follows. Basic definitions are recalled in Section 2.1; Section 2.2 includes a synthetic presentation of the diagonal and circular orders associated with a valued X-tree, with some of their properties. Although the results of this section are not new, some of them are given in new statements, more complete than the previous ones, and with simplified proofs. Yushmanov's results, with the presentation of his orders, are recalled in Section 3.1. It is emphasized that the definition of Yushmanov orders extends to all dissimilarities. Two algorithms are given, the first one for the determination of a Yushmanov order and the second for the reconstruction of the valued X-tree associated with an initial tree dissimilarity. After an illustration of these algorithms in Section 3.2, it is shown in Section 3.3 that Yushmanov and circular orders are identical. Yushmanov orders are used in Section 4 in a new approach of the problem of fitting of a tree metric to a given observed dissimilarity. Both Algorithms 4 and 5 of Section 4.1 involve the solution of a least squares approximation problem at each step. The property that Yushmanov orders are circular is extensively used to obtain the low time complexities of $O(n)$ in the reconstruction algorithm 2 of Section 3.1 and $O(n^2)$ in the fitting algorithm 5. The presence of arbitrary initial choices in the last one deserves further studies. Presently, a procedure based on repeated uses of Algorithm 5 is proposed, involving all the possible first choices. In Section 4.2, some examples are presented. The procedure previously defined is compared with several good algorithms of the literature, and appears to be a competitive one. We conclude with some questions appealing for further developments.

2. The encoding of a valued tree with n leaves by $2n$-3 path lengths

2.1. Dissimilarities and trees. Let X be a set with n elements. A *dissimilarity* on X is a non-negative real function d on $X \infty X$ satisfying the following two conditions:

- for all $x, y \in X$, $d(x,y) = d(y,x)$;
- for all $x, y \in X$, $d(x,y) \geq d(x,x) = 0$.

The dissimilarity d is a *metric* if it satisfies the classical metric triangular inequality:

for all $x, y, z \in X$, $d(x,z) \leq d(x,y) + d(y,z)$.

A graph G is also denoted as $(V(G),E(G))$, where $V(G)$ is the vertex set and $E(G)$ is a set of unordered pairs of distinct elements of $V(G)$, the edge set of G; for sake of brevity, an edge is denoted vv' instead of $\{v,v'\}$. The degree $\partial(v)$ of a vertex v is the number of edges $e \in E(G)$ such that $v \in e$. A *leaf* is a vertex of degree one. In a graph G, a *path* P between two vertices v and v' is a sequence of edges vv_1, v_1v_2, …, $v_{k-1}v_k$, v_kv'. In fact,

since only paths with all edges distinct will be considered here, a path will be generally identified with the set of its edges. A tree T is a graph with a unique path, denoted $T(uv)$, between any two distinct vertices u and v. A tree T has exactly $|V(T)|$-1 edges.

Let u be an inner vertex (that is, not a leaf) of a tree T, and an edge uv. Consider the set Y containing u and all the vertices v' of T such that $uv \in T(uv')$. The induced subtree T_Y is said to be a *branch of T at the vertex u*. All the leaves of a branch, except u, are leaves of T.

We mainly consider here the so-called X-trees, related with the set X by two properties: (i) the set of leaves of T is X; (ii) for any $v \in V(T)$-X, $\partial(v) \geq 3$. According to the terminology in the Barthélemy and Guénoche book [2], this means that the X-trees considered here are *separated* and *free*. In such an X-tree, the role of the inner vertices, called also *latent vertices* is just to determine the shape of the tree; they are often indicated without labels in figures. An X-tree has at most $2n$-2 vertices and thus $2n$-3 edges, these numbers corresponding to the non-degenerate case where all the latent vertices have degree 3. The *articulation vertex* $a(x)$ of $x \in X$ is the unique latent vertex adjacent to x (that is such that $xa(x) \in E(T)$). The number of edges of T is denoted as $\mu(T)$ (or simply μ).

We make correspond an X-tree T_ℓ to any valued tree $T'_{\ell'}$ with X as set of leaves by the repetition of the following operation: choose a vertex u of degree two in $V(T')$-X, delete it and replace the edges vu and uv' incident to u by a unique edge vv'; set $\ell(vv') = \ell'(vu) + \ell'(uv')$. When no such vertex remains, an X-tree T is obtained; complete the end length function on T by setting $\ell(vv') = \ell'(vv')$ for all the edges common to T and T'. The trees T and T' have the same leaves and all the distances between leaves in $T'_{\ell'}$ are preserved in T_ℓ. We say that T_ℓ is the *reduced X-tree* corresponding with $T'_{\ell'}$.

We consider *valued X-trees* T_ℓ, where T is a tree and ℓ is a real *length* function on $E(T)$. The functions ℓ considered here are non-negative, with null values possible only on the edges adjacent to leaves. The *distance* $t(v,v')$ between two vertices v and v' is equal to $\sum_{e \in T(vv')} \ell(e)$; it satisfies the metric triangular inequality.

Let $\mathcal{X}$ be the class of all the X-trees T_ℓ, valued as above. A dissimilarity d on X is said to be a *tree metric* if it is representable by the length of the paths between the leaves of an element of $\mathcal{X}$. A tree metric is a metric. A dissimilarity on X satisfies the *four-point condition* if, for all $x, y, z, w \in X$,

$$d(x,y) + d(z,w) \leq \max\{\, d(x,z) + d(y,w)\ ,\ d(x,w) + d(y,z)\ \}.$$

THEOREM 2.1 (Zaretskii [29], Buneman [5], Patrinos and Hakimi [22], Dobson [15]). *Let d be a dissimilarity on X. The following two conditions are equivalent:*

(i) *d is a tree metric;*

(ii) *d satisfies the four-point condition.*

Moreover, a tree metric admits a unique tree representation.

As recalled in Leclerc [17], a dissimilarity satisfying the four-point condition for all *distinct* $x, y, z, w \in X$ is not necessarily a metric, but is still uniquely representable by a valued X-tree, possibly with negative lengths on the edges adjacent to the leaves. Such a dissimilarity will be said a *tree dissimilarity*.

Given three distinct vertices u, v, w of a tree T, there is a unique vertex which is common to the three paths $T(uv)$, $T(vw)$ and $T(uw)$. This vertex is called the *median vertex* of the triple (u,v,w) and denoted as $m(u,v,w)$. If x, y, z are three leaves of T, then $m(x,y,z) \neq x, y, z$. In a valued tree T_ℓ, the distance between x and the path $T(yz)$ is equal to $t(x,m(x,y,z)) = (t(x,y)+t(x,z)-t(y,z))/2$; this quantity is denoted as $dist(x,T(yz))$.

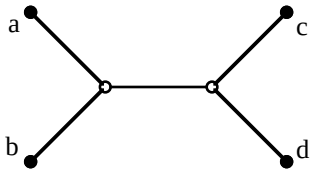

Figure 1

A triple (x,y,z) of leaves of T is said *well-formed* if $m(x,y,z) = a(y)$. Equivalently, the vertex $a(y)$ belongs to the path $T(xz)$. For instance, in the tree of Figure 1, the triple (a,c,d) is well-formed, whereas (a,c,b) is not. Note that, if x and y are two leaves such that $a(x) = a(y)$, then every triple (x,y,z) is well-formed. Here is a simple characterization of well-formed triples in terms of distances:

PROPOSITION 2.2. *A triple* (x,y,z) *of leaves of* T *is well-formed if and only if the equality* $dist(y,T(xz)) = \min_{w\in X,\, w\neq x,y} dist(y,T(xw))$ *holds.*

Proof. If (x,y,z) is a well-formed triple, then, $dist(y,T(xz)) = \ell(ya(y))$. For every $w \in X$, $w \neq x$, y, consider the path $T(yv)$ from y to the vertex v which is the closest to y on the path $T(xw)$; the path $T(yv)$ includes the edge $ya(y)$, and, so, has a length at least equal to $\ell(ya(y))$. Conversely, the quantity $dist(y,T(xw))$ is minimized by those leaves w such that $a(y)$ belongs to the path $T(xw)$. ◊

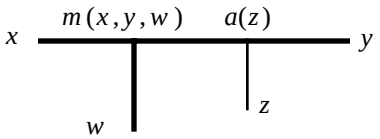

Figure 2

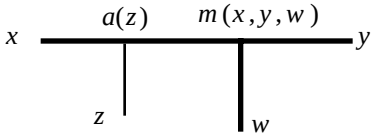

Figure 3

Bold lines represent paths ; thin ones represent single edges.

PROPOSITION 2.3. *Let* d *be a tree metric on the set* X *and* x, y, z *be three elements of* X *such that, in the* X*-tree representation* T_ℓ *of* d, *the triple* (x,z,y) *is well-formed. Then, the following equality holds for any* $w \in X-\{x,y,z\}$*:*

$$d(z,w) = \max\{\, d(x,w) + d(y,z)\ ,\ d(y,w) + d(x,z)\,\} - d(x,y).$$

Proof. For each $w \in X-\{x,y,z\}$, the median vertex $m(x,y,w)$ *lies either on the path* $T(xa(z))$ (Figure 2), or on the path $T(ya(z))$ (Figure 3), or is equal to $a(z)$. The four-point condition expresses as $d(x,y) + d(z,w) = d(x,z) + d(y,w) \geq d(x,w) + d(y,z)$ in the first case, as $d(x,y) + d(z,w) = d(x,w) + d(y,z) \geq d(x,z) + d(y,w)$ in the second, and as $d(x,y) + d(z,w) = d(x,z) + d(y,w) = d(x,w) + d(y,z)$ in the degenerate third case. The result follows.

So, when the triple (x,z,y) is well-formed, the distance between z and any element w of X may be computed from the values $d(x,y)$, $d(x,z)$, $d(y,z)$, $d(x,w)$ and $d(y,w)$.

The following notations are used when an indexing $x_1, x_2,\ldots, x_n$ of the set X is given. For three distinct leaves $x_i, x_j, x_k \in X$, we set $\Delta^k_{i,l} = dist(x_k,T(x_ix_j)) = \frac{1}{2}\,(d(x_i,x_k)+d(x_j,x_k)-d(x_i,x_j))$. Since d satisfies the triangular metric inequality, the quantities $\Delta^k_{i,l}$ are non negative: moreover, $d(x_i,x_j) = \Delta^i_{j,k} + \Delta^j_{i,k}$ and $\Delta^k_{i,l} > 0$ when the edge $x_ka(x_k)$ has a non-null length. We will also write a_k instead of $a(x_k)$.

2.2. Circular and diagonal orders. We now recall an encoding method, proposed in the Barthélemy and Guénoche [2] book as a generalization of a method given by Chaiken, Dewdney and Slater [7] in the special case of an unvalued tree. An order $x_1, x_2,\ldots, x_n$ on X is called a *diagonal order* of the X-tree T (Dewdney [14]) if, for any integer k (modulo n), the triple (x_{k-1}, x_k, x_{k+1}) is well-formed.

Consider a graphic planar representation of T (where two edges have no other common points than a vertex) and an order obtained as follows: first, the leaf x_1 is arbitrarily chosen; then, the leaves are indexed as x_1, $x_2,\ldots, x_n$ according to a circular (say, clockwise) scanning of the subset X of vertices of T. Such an order, frequently called *diagonal plane* in the literature, will be said here *circular*. It has the following property, for any integer k modulo n: when moving on the path $T(x_kx_{k+1})$ from x_k to x_{k+1}, all the branches at the encountered vertices are located on the right. A circular order is diagonal; otherwise, assume $m(x_{k-1},x_k,x_{k+1}) \neq a(x_k)$: there is a branch of the tree at the vertex $a(x_k)$, the leaves of which being either between x_{k-1} and x_k or between x_k and x_{k+1} in the clockwise scanning, a contradiction with the hypothesis that the order is circular. Hence the

following result holds:

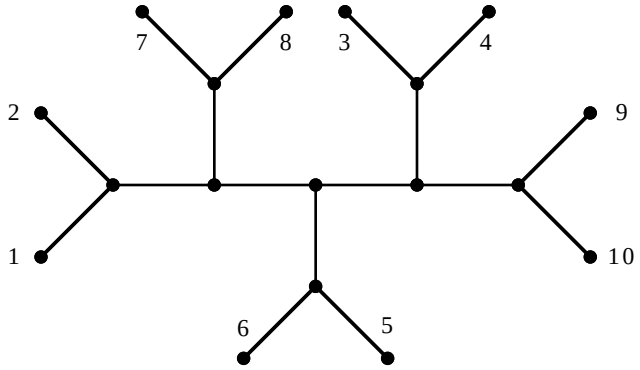

Figure 4

LEMMA 2.4. *Any X-tree T admits a diagonal order.*

As the following example shows, circular orders constitute in the general case a proper subclass of the diagonal ones. According to a remark above, if x and y are two leaves such that $a(x) = a(y)$, then the only condition on their places in a diagonal order is that they are consecutive (modulo n). So, in the example of Figure 4, the order (1 2 3 4 5 6 7 8 9 10) is diagonal. A simple investigation leads to the conclusion that it is not circular.

THEOREM 2.5 (Barthélemy and Guénoche [2]; see Leclerc and Makarenkov [18]). *Let* $X = \{ x_1, x_2,\ldots, x_n \}$ *be a linearly ordered set. For any sequence* $d_{12}, d_{13}, d_{23},\ldots, d_{1i}, d_{i,i+1},\ldots, d_{1n}, d_{n-1,n}$ *of* $2n$-3 *strictly positive real numbers, there exists a unique valued X-tree* T_ℓ *such that* $d_{ij} = \sum_{e\in T(x_i,x_j)} \ell(e)$ *and* $x_1, x_2,\ldots, x_n$ *is a circular order of* T.

As stated in Leclerc and Makarenkov [18], the function ℓ is non-negative if and only if the sequence d_{12}, d_{13}, d_{23},…, d_{1n}, $d_{n-1,n}$ is extracted from a metric array. For a counter-example, consider the X-tree of Figure 5, where $X = \{x_1, x_2, x_3, x_4\}$. For such a tree, all the linear orders on X are circular. The sequence of five positive real numbers given by $d_{12} = d_{13} = d_{14} = d_{23} = 2$ and $d_{34} = 10$ leads to a system of five linear equations which has no solution for this tree: $\alpha+\beta = 2$; $\alpha+\gamma = 2$; $\alpha+\delta = 2$; $\beta+\gamma = 2$; $\gamma+\delta = 10$. In fact, the sequence corresponds to the valued tree of Figure 6. Since the sequence was not extracted from a metric array, this tree has a negatively valued edge and is the representation of a tree dissimilarity.

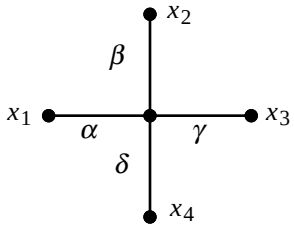

Figure 5

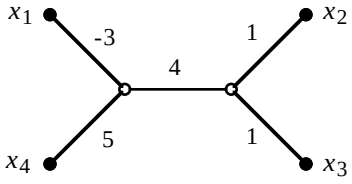

Figure 6

3. A combinatorial obtainment of circular orders

3.1. Yushmanov orders. In his 1984 paper [28], Yushmanov shows it possible to encode a positively valued X-tree T_ℓ by $2n$-3 lengths of paths between leaves. His work is independent of the Chaiken *et al.* one, even thought that, as these authors, he uses his results in a study of unvalued trees. The main results in Yushmanov's paper are recalled in this subsection with extensions to the cases of valued trees and dissimilarities. We also provide algorithms corresponding with Yushmanov's approach.

Let d be a tree metric on X, his matrix D, and consider the sets P of pairs of elements such that the

knowledge of the entries ($d(x,y)$), $xy \in P$, allows us to recover the entire matrix D, provided the tree T is already known; the sets defined by Chaiken *et al.* and used in Theorem 2.5 are of this type. Denote as $\rho(T_\ell)$ the minimum cardinality of such a set. In his paper, Yushmanov first observes that every quantity $d(x,y)$ is a sum of lengths of edges of T. So, a subset P corresponds to a set of linear equations with rank $\mu(T) \leq 2n\text{-}3$. The equality $\rho(T_\ell) = \mu(T)$ follows; note that this observation remains valid when the edge lengths are no longer assumed to be positive. Similarly, consider the sets Q such that the knowledge of the entries ($d(x,y)$), $xy \in Q$, allows us to recover D without any further requirements. Denote their minimum cardinality as $\delta(T_\ell)$. Obviously, $\rho(T_\ell) \leq \delta(T_\ell)$. Indeed, Leclerc [17] gives two ways of determining a set Q of cardinality $2n\text{-}3 = \rho(T_\ell)$, thus proving the equality $\rho(T_\ell) = \delta(T_\ell)$. Yushmanov was the first to exhibit such sets, related with linear orderings $x_1, x_2,\ldots, x_n$ of X such that, for all $k = n, n\text{-}1,\ldots, 2$, the triple (x_1,x_k,x_{k-1}) is well-formed in a current tree T^k with k leaves. Such an order, called hereafter a *Yushmanov order*, is obtained by the procedure 1 described below:

Procedure 1
Initialization. Choose arbitrarily two leaves of T_ℓ and index them as x_1 and x_n. Set $V^n = \varnothing$ and $T^n = T$.
Step 1. Choose a leaf x_{n-1} in T^n in such a way that the vertex a_n lies on the path $T^n(x_1x_{n-1})$. Delete the leaf x_n and the edge a_nx_n in T^n and reduce the resulting tree in order to obtain the tree T^{n-1}; set $V^{n-1} = \{x_n\}$.
Step k+1. Assume the first k steps have led to a tree T^{n-k} the set of leaves of which is $X\text{-}V^{n-k}$, where $V^{n-k} = \{x_n,\ldots,x_{n-k+1}\}$.
If $k = n\text{-}2$, then $V^2 = \{x_n,\ldots,x_3\}$ and T^2 is reduced to the unique edge x_1x_2; since x_1 is already fixed, the Yushmanov indexing $x_1, x_2,\ldots, x_n$ is completely determined.
Otherwise, choose a leaf x_{n-k-1} in T^{n-k} in such a way that the vertex a_{n-k} lies on the path $T^{n-k}(x_1,x_{n-k-1})$; delete the leaf x_{n-k} and the edge $a_{n-k}x_{n-k}$ in T^{n-k}; reduce the resulting tree in order to obtain the tree T^{n-k-1}; set $V^{n-k-1} = V^{n-k}\cup\{x_{n-k}\}$.

The choice of the leaf x_{n-k-1} is always possible: the vertex a_{n-k} adjacent to x_{n-k} in T^{n-k} has degree at least three, defining at least three branches. Any leaf which belongs to a branch which neither contains the leaf x_1 nor consists of the edge $a_{n-k}x_{n-k}$ may be chosen as x_{n-k-1}.

The remaining problem is the determination of a Yushmanov order without the knowledge of the tree T. A solution is provided by arguments already used in the proof of Proposition 2.2: the possible choices are the elements x_{n-k-1} (different from x_1 and x_{n-k}) such that $dist(x_{n-k},T(x_{n-k-1}x_1)) = \min_{w\in X-V^{n-k}} dist(x_{n-k},T(wx_1))$. With the expression of $dist(x_{n-k},T(wx_1))$ recalled in Section 2.1, x_{n-k-1} is an element $w \in X\text{-}V^{n-k}$ minimizing the difference $d(x_{n-k},w)\text{-}d(x_1,w)$. Such an element is chosen directly on the matrix of d in the formal statement of Algorithm 1 below (see the notations of Section 2.1). In fact, Algorithm 1 works with any dissimilarity matrix, tree metric or not, as input; so, the definition of a Yushmanov order of X extends to all dissimilarities.

Algorithm 1: construction of a Yushmanov indexing $x_1, x_2,\ldots, x_n$ of the set X:
Input: a finite set X with n elements ; a dissimilarity d on X.
Output: a Yushmanov order $(x_1, x_2,\ldots, x_n)$ on X associated with d.

Initialization Choose arbitrarily two leaves x_1 and x_n; $W := X\text{-}\{x_1,x_n\}$; $k = 0$
Repeat
  Find x_{n-k-1} in W such that:
  $d(x_{n-k},x_{n-k-1})\text{-}d(x_1,x_{n-k-1}) = \min_{w\in W} d(x_{n-k},w)\text{-}d(x_1,w)$;
  $W := W\text{-}\{x_{n-k}\}$;
  $k = k+1$
Until $W = \varnothing$

In the k-th step of Algorithm 1, $n\text{-}k\text{-}2$ elements of X are examined. So, this algorithm has time complexity $O(n^2)$. A converse procedure allows to reconstruct the valued X-tree T_ℓ from the tree metric d and a Yushmanov order $x_1, x_2,\ldots, x_n$ obtained by Algorithm 1 applied to d. Consider the sequence $(d(x_1,x_2), d(x_1,x_3), d(x_2,x_3),\ldots, d(x_1,x_i), d(x_i,x_{i+1}),\ldots, d(x_1,x_n), d(x_{n-1},x_n))$ with $2n\text{-}3$ terms. For $k = 2,\ldots, n\text{-}1$, we are able to compute the quantities $\Delta^1_{k,k+1} = dist(x_1,T(x_kx_{k+1})) = \frac{1}{2}(d(x_1,x_k) + d(x_1,x_{k+1}) - d(x_k,x_{k+1}))$ and $\Delta^{k+1}_{1,k} =$

$dist(x_{k+1},T(x_1x_k)) = \frac{1}{2}(d(x_1,x_{k+1}) + d(x_k,x_{k+1}) - d(x_1,x_k))$; they are assumed to be positive, as it is the case when the values d_{ij} are extracted from a distance matrix. A sequence of valued trees T_ℓ^k with k leaves, $k = 2,\ldots, n$, is constructed according to the following Procedure 2:

Procedure 2
Step 1 (initialization). T_ℓ^2 is the tree reduced to the unique edge x_1x_2 with length $\ell(x_1x_2) = d(x_1,x_2)$.
Step k ($k = 2,\ldots,n$-1). A tree T_ℓ^k *with the leaves* $x_1, x_2,\ldots, x_k$ has been built. Two cases may occur for the path $T_\ell^k(x_1x_k)$:
Case 1. There exists a vertex u on this path such that $t(x_1,u) = \Delta_{k,k+1}^1$. In this case, the leaf x_{k+1} is the only vertex to add to T_ℓ^k in order to obtain T_ℓ^{k+1}, with the new edge ux_{k+1} of length $\ell(ux_{k+1}) = \Delta_{1,k}^{k+1}$.
Case 2. There exists an edge vv' on the path $T_\ell^k(x_1,x_k)$ such that $t(x_1,v) < \Delta_{k,k+1}^1 < t(x_1,v')$. In this case, a new inner vertex u is added on the edge vv', now divided into two edges uv and uv', with lengths $\ell(uv) = \Delta_{k,k+1}^1 - t(x_1,v)$ and $\ell(uv') = t(x_1,v') - \Delta_{k,k+1}^1$; then, as before, the leaf x_{k+1} is added to T_ℓ^k in order to obtain T_ℓ^{k+1}, with the new edge ux_{k+1} of length $\ell(u,x_{k+1}) = \Delta_{1,k}^{k+1}$.
For $k = n$, the valued tree T_ℓ^k *is equal to* T_ℓ.

In the next Algorithm 2, the k-th step consists of the examination from edge to edge (starting from x_k) of the path $T(x_kx_1)$ until the good place for the articulation vertex a_{k+1}, depending on $\Delta_{1,k+1}^k$, is found. The new leaf x_{k+1} is then added to the tree T with a new edge $a_{k+1}x_{k+1}$ of length $\Delta_{1,k}^{k+1}$. Note that the number of edges examined at this step is no more than $|T^n(x_kx_{k+1})|$, the number of edges of the path between x_k and x_{k+1} in the final tree T^n. Some edges are recognized as not belonging to a current path $P(T)$ in the next step k+1 and in the sequel; such edges are included in the set $E(T)$ (and their extremities in $V(T)$). This is due to the observation that any edge excluded from the linked list $P(T)$ will never return in this list, since $P(T)$ is always completed with one or two new edges. The complexity of Algorithm 2 is presently estimated as $O(\sum_{1\le k\le n-1} |T^n(x_kx_{k+1})|) + O(n)$, and will be shown to be in fact $O(n)$ in Section 3.4 (Corollary 3.7).

Some further notations are used in the following formal statement of Algorithm 2: given two leaves x, y of an X-tree T, $\ell(i,T(xy))$, $w(i,T(xy))$ and $w'(i,T(xy)) = w(i+1,T(xy))$ are respectively the length, the initial vertex and the terminal vertex of the i–th edge (starting from x) of the path $T(xy)$; $P(T)$ is the linked list of the edges of the path $T(x_1x_k)$ (starting from x_1), and $E(P(T))$ and $V(P(T))$ are, respectively, the edge set and the vertex set of $P(T)$.

Algorithm 2: Reconstructing a valued tree T_ℓ from a tree metric d and a corresponding Yushmanov order on X.
Input: a finite set X with n elements; the $2n$-3 entries $d(x,y)$, $xy \in \{x_1x_2, x_1x_3, x_2x_3,\ldots,x_1x_i, x_ix_{i+1},\ldots, x_1x_n, x_{n-1}x_n\}$ corresponding with a Yushmanov order $(x_1, x_2,\ldots, x_n)$ for a tree metric d on X.
Output: the valued X-tree $T_\ell = (V(T),E(T),\ell)$ associated with d.

Initialization $V(T) := \emptyset$; $E(T) := \emptyset$; $k := 1$; $P(T) := \{x_1x_2\}$;
$\ell(x_1x_2) := d(x_1,x_2)$
Repeat
$k := k+1$; $i := 1$; $S := 0$
if $\Delta_{1,k+1}^k > 0$ **then**
$S := \ell(1,T(x_kx_1))$
$u := w'(1,T(x_kx_1))$
$v := w(1,T(x_kx_1)) = x_k$
$P(T) := P(T) - \{uv\}$
if $S \ge \Delta_{1,k+1}^k$ **then**
$V(T) := V(T)+\{x_k\}$
else $u := x_k$

while $S < \Delta^k_{1,k+1}$ **do**

$V(T) := V(T)+\{u,v\}$

$E(T) := E(T)+\{uv\}$

$i := i+1$

$u := w'(i,T(x_k x_1))$

$v := w(i,T(x_k x_1))$

$S := S+\ell(i,T(x_k x_1))$

$P(T) := P(T) - \{uv\}$

if $S > \Delta^k_{1,k+1}$ **then**

$E(T) := E(T)+\{a_{k+1}v\}$

$P(T) := P(T)+\{ua_{k+1}\}+\{a_{k+1}x_{k+1}\}$

$\ell(a_{k+1}v) := \Delta^k_{1,k+1}-S+\ell(uv)$

$\ell(ua_{k+1}) := S-\Delta^k_{1,k+1}$

$\ell(a_{k+1}x_{k+1}) := \Delta^{k+1}_{1,k}$

else $P(T) := P(T)+\{ux_{k+1}\}$

$\ell(ux_{k+1}) := \Delta^{k+1}_{1,k}$

until ($k = n$-1)

$E(T) := E(T)+E(P(T))$; $V(T) := V(T)+V(P(T))$

THEOREM 3.1 (Yushmanov [28]). *The successive uses of Procedures 1 and 2 map any valued X-tree* T_ℓ *on itself.*

Proof. The result is true for $n = 3$: in this case, the X-tree T has one latent vertex a adjacent to its three leaves and every order on X is Yushmanov. For an arbitrary order x_1, x_2, x_3, the lengths $\Delta^1_{2,3}$, $\Delta^2_{1,3}$ and $\Delta^3_{1,2}$ of the edges ax_1, ax_2 and ax_3 are determined by Procedure 2.

Assume that the result is true for every X'-tree, where X' is a set of cardinality n-1. Then, let $x_1,\ldots, x_n$ be a Yushmanov order on X. By the induction hypothesis, Procedure 2 constructs, before its last iteration, a valued tree T^{n-1}_ℓ with n-1 leaves such that: (i) the leaves of T^{n-1}_ℓ are the elements $x_1, x_2,\ldots, x_{n-1}$; and (ii) the distances between these leaves are given by the restriction of d to X-$\{x_n\}$. By the rule used for choosing x_{n-1} at the first iteration of Procedure 1, one also knows that the latent vertex a_n lies on the path $T^{n-1}(x_1x_{n-1})$; then, the determination of the place of a_n on this path is made in the last step of Procedure 2 on such a way that the distance d is realized by the tree T^n_ℓ for the pairs x_1x_n and x_nx_{n-1}. By the proposition 2.3, the induction hypothesis, and the unicity of the tree representation recalled in Theorem 2.1, $T^n_\ell = T_\ell$ is the unique valued X-tree realizing the distance d on X. ◊

Many fitting algorithms of the literature transform a given dissimilarity matrix d_0 on X into a tree metric one d; this is the case of the decomposition algorithm of Brossier [4], the algorithm based on minimum spanning trees of Leclerc [17], or the reduction methods of Roux [25] and Gascuel and Levy [16]. Then, the problem of the reconstruction of the X-tree representation of d remains. Methods like ADDTREE (Sattah and Tversky [27]) or the scoring method of Luong [19] and Barthélemy and Guénoche [2] are sometimes proposed in the literature for determining the corresponding X-tree; as their time complexity $O(n^5)$ indicates, these methods have not been designed for this particular use. Starting from the distance matrix of d, the successive uses of Algorithms 1 and 2 provide the valued X-tree in $O(n^2)$ time, which does not exceed the complexity of any fitting method.

3.2. An example. For an illustration of Algorithms 1 and 2, consider the valued X-tree of Figure 7. Let us start from the corresponding tree metric array (Table 1).

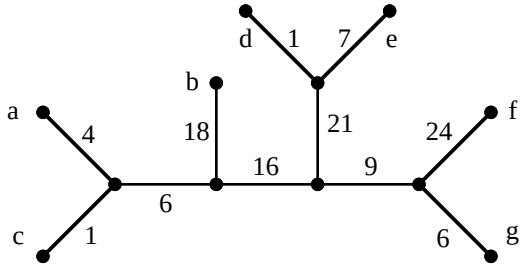

Figure 7

Set x_1 = a and x_7 = b ; then, Algorithm 1 computes the quantities *dist*(b,*T*(a*x*)) for *x* = c , d , e , f , g , which gives:

2*dist*(b,*T*(ac)) = 28+25-5 = 48 2*dist*(b,*T*(ad)) = 28+56-48 = 36

2*dist*(b,*T*(ae)) = 28+62-54 = 36 2*dist*(b,*T*(af)) = 28+67-59 = 36

2*dist*(b,*T*(ag)) = 28+49-41 = 36

So, x_6 may be chosen among d , e , f and g ; set, for instance, x_6 = d and compute*:*

2dist(d,*T*(ac)) = 48+45-5 = 88 2*dist*(d,*T*(ae)) = 48+8-54 = 2

2*dist*(d,*T*(af)) = 48+55-59 = 44 2*dist*(d,*T*(ag)) = 48+37-41 = 44

So, x_5 = e and, at the next step, we have: 2*dist*(e,*T*(ac)) = 54+51-5 = 100

2*dist*(e,*T*(af)) = 54+61-59 = 56 2*dist*(e,*T*(ag)) = 54+43-41 = 56

Choosing x_4 = f , we compute: 2*dist*(f,*T*(ac)) = 59+56-5 = 110

2*dist*(f,*T*(ag)) = 59+30-41 = 48, leading to x_3 = g and x_2 = c .

b	28					
c	5	25				
d	48	56	45			
e	54	62	51	8		
f	59	67	56	55	61	
g	41	49	38	37	43	30
	a	b	c	d	e	f

Table 1

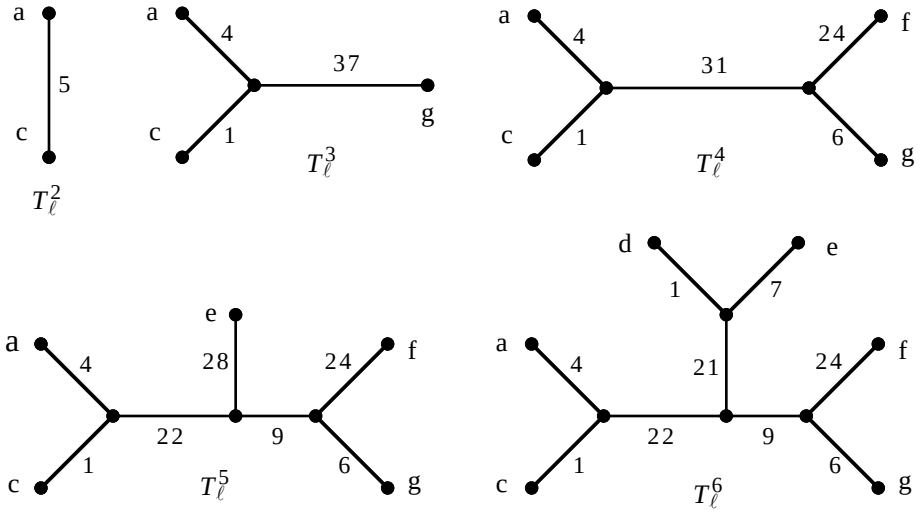

Figure 8

3.3. Yushmanov and circular orders are the same. As noticed just above, circular and Yushmanov orders are defined on two different ways. Both ways select well-formed triples. According to the results of this section, these two ways lead in fact to the same orders.

PROPOSITION 3.2. *Every circular order on the leaves of an X-tree T is Yushmanov.*

Proof. We prove the result by induction on n; for $n = 3$, all orders on X are both circular and Yushmanov.

Assume the result is true for all X-trees with at most n-1 leaves. Let T be an X-tree with n leaves x_1, x_2,…, x_n, indexed accordingly with a circular order. Consider the path $T(x_1x_{n-1})$ between x_1 and x_{n-1}. Since the triple (x_{n-1},x_n,x_1) is well-formed, we have the configuration of Figure 9, with $m(x_{n-1},x_n,x_1) = a_n$. So, the vertex a_n belongs to the path $T(x_1x_{n-1})$; it follows that Procedure 1 of Section 3.1 can be performed to obtain a Yushmanov indexing $y_1, y_2,…, y_n$ of X, with choices in Initialization and Step 1 leading to $y_1 = x_1$, $y_{n-1} = x_{n-1}$ and $y_n = x_n$.

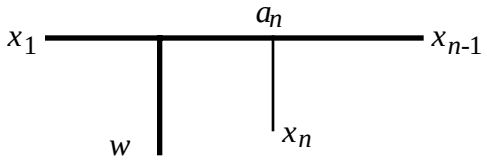

Figure 9 (with the same conventions as Figures 2-3)

Then, in the tree T, delete the vertex x_n and the edge a_nx_n and reduce the obtained tree. A new tree T' with the leaves x_1,…, x_{n-1} is obtained. From a planar representation of T admitting the previous circular indexing x_1, x_2,…,x_n, we derive a planar representation of T' with x_1,…, x_{n-1} as a circular indexing. So, by the induction hypothesis, this order is Yushmanov. It follows that it can be obtained by Procedure 1, with x_1 and x_{n-1} as vertices chosen in the initial step; that is, the Initialization and Step 1 mentioned just above can be continued according to Procedure 1 in order to obtain the order x_1,…, x_n on X. ◊

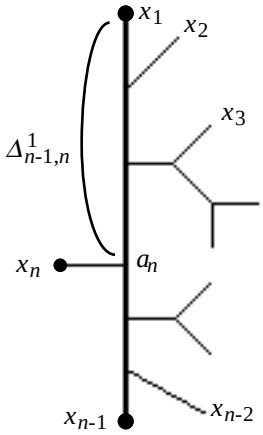

Figure 10

PROPOSITION 3.3. *Every Yushmanov order on the leaves of an X-tree T is circular.*

Proof. Let T be an X-tree with n leaves $x_1, x_2,…, x_n$, indexed according to a Yushmanov order. We must show that there exists a planar drawing of T for which this order is circular. This is obviously true for $n = 3$, and we proceed again by induction on n.

Assume the result is true for all X-trees with at most n-1 leaves, and let T be an X-tree with n leaves. After deletion of the edge a_nx_n and reduction of the obtained tree, a new tree T' is obtained, with leaves x_1,…, x_{n-1} still indexed accordingly with a Yushmanov order. By the induction hypothesis, there exists a planar drawing of T' such that this order is circular. As a consequence, there is no branch on the left when moving from x_{n-1} to x_1 on the path $T(x_{n-1}x_1)$. Add the vertex a_n on this path at the place specified by Algorithm 2; the drawing of the

edge $a_n x_n$ on the left of the path $T(x_{n-1}x_1)$ gives $x_1, x_2,\ldots, x_n$ as a circular order of T (Figure 10).

Propositions 3.2 and 3.3 assemble into the following

THEOREM 3.4. *An order $x_1, x_2, \ldots, x_n$ on the leaves of an X-tree T is circular if and only if it is Yushmanov.*

Now, Yushmanov's algorithm 1 appears to be a purely combinatorial, free of geometric representation, construction of circular orders. For a dissimilarity d on X, a linear order C on X derived from d by Algorithm 1 will be called "circular" if d is known to be a tree metric, and "Yushmanov" otherwise.

3.4. Further properties of circular orders and recognition of a tree metric. The following Algorithm 3 decides whether a given dissimilarity matrix d is a tree metric. The elements of X are examined in an order provided by Algorithm 1. When x_{k+1} is examined, Proposition 2.3 gives a formula leading to values $t(j,k+1)$ of the distance between the leaves x_j and x_{k+1} in the tree T, for all $j = 2,\ldots, k\text{-}1$, under the hypothesis that d is a tree metric. These computed values are compared with the actual ones and the conclusion follows; in the statement below, the "return" command means the exit from the algorithm.

Algorithm 3: recognition of a tree metric
Input: a finite set X with n elements; a dissimilarity d on X and a corresponding Yushmanov order $(x_1, x_2,\ldots, x_n)$ on X.
Output: the answer "Yes" if d is a tree metric; the answer "No" otherwise.

Initialization
Compute $\Delta^1_{2,3}$, $\Delta^2_{1,3}$ and $\Delta^3_{1,2}$.
if $(\Delta^1_{2,3} \geq 0,\ \Delta^2_{1,3} \geq 0,\ \Delta^3_{1,2} \geq 0)$
else print "No"
return
$k := 3$
repeat
Compute $\Delta^1_{k,k+1}$, $\Delta^k_{1,k+1}$ and $\Delta^{k+1}_{1,k}$
if $(\Delta^1_{k,k+1} \geq 0,\ \Delta^k_{1,k+1} \geq 0,\ \Delta^{k+1}_{1,k} \geq 0)$ **then**
for $j = 2,\ldots, k\text{-}1$ **do**
$t(j,k+1) := \max\{d(x_1,x_j)+ d(x_k,x_{k+1}), d(x_1,x_{k+1})+ d(x_k,x_k)\}- d(x_1,x_k)$
if $(t(j,k+1) \neq d(x_j,x_{k+1}))$ **then**
print "No"
return
else print "No"
return
$k := k+1$
until $(k = n+1)$
print "Yes"

Recall $\Delta^k_{i,j} = dist(x_k, T(x_i x_j)) = \frac{1}{2}(d(x_i,x_k)+d(x_j,x_k)-d(x_i,x_j))$. The time complexity of Algorithm 3 is $O(n^2)$, similar to previous algorithms addressing the same problem (final note in Bandelt [1], Leclerc [17]).

Theorem 3.4 suggests some remarks about circular orders. First, as illustrated in the example of Section 3.2 above, when d is a tree metric, several Yushmanov orders may be obtained with the same initial vertices x_1 and x_n; this is due to the fact that tree metrics are strongly constrained, although their regularities do not directly appear in their matrices. So, arbitrary choices may be needed at every step. This explains why the number $n(n\text{-}1)$ of possible initial choices differs from the number of circular orders (defined modulo n), for instance 2^{n-2} for an X-tree with the maximum number n-2 of latent vertices. On the other hand, in the general case of a dissimilarity

d without special properties, ties on the values of $d(x_{n-k},w)$-$d(x_1,w)$ rarely occur in Algorithm 1 and the number of different Yushmanov orders is close to $n(n\text{-}1)$. Note also that the choice of the initial vertex x_1 in Algorithm 1 finally does not matter when the purpose is just to obtain a circular order, because of the following consequence of Theorem 3.4:

COROLLARY 3.5. *If x_1, x_2,…, x_n is a circular order on the leaves of an X-tree T, then, for any $k \in \{1,\ldots,n\}$, the order x_k, x_{k+1},…, x_n, x_1, x_2,…, x_{k-1} is again circular.*

Another property of circular orders is related with the induction on the number n used in the proof of Proposition 3.3. With the remark that the difference between $d(x_{n-1},x_n)+d(x_n,x_1)$ and $d(x_{n-1},x_1)$ is two times the length of the new edge a_nx_n, the following Proposition 3.6, already stated by Yushmanov, holds:

PROPOSITION 3.6. *The sum $\ell(T)$ of the lengths of the edges of a valued X-tree T_ℓ is given by $2\ell(T) = d(x_1,x_2) + d(x_2,x_3) + d(x_3,x_4) +\ldots+ d(x_{n-1},x_n) + d(x_n,x_1)$, provided $x_1, x_2,\ldots, x_n$ is a circular order on X.*

Especially, in the unvalued case, the sum $d(x_1,x_2) + d(x_2,x_3) +\ldots+ d(x_n,x_1)$ is two times the number of edges of the tree.

COROLLARY 3.7. *Algorithm* 2 *reconstructs a valued X-tree T with n leaves in $O(n)$ operations.*

Proof. As a consequence of the previous result applied to the unvalued case, the number $\sum_{1\le k\le n-1} |T^n(x_kx_{k+1})|$, which is an upper bound of the number of the edges examined in the algorithm, is two times the number of edges of the final tree. ◊

4. A fitting method

4.1. Two fitting algorithms. This section is devoted to the problem, very often addressed in the literature, of fitting a tree metric t to a given dissimilarity d. Algorithms of various types have been given or recalled, for instance, by De Soete [12], Luong [19], Saitou and Nei [26], Barthélemy and Guénoche [2], Roux [25], Leclerc [17], De Soete and Caroll [13], Gascuel and Lévy [16]. The best time complexity of such algorithms is $O(n^2)$; the algorithms reaching such a complexity are rarely good for global criteria like the least squares one. In fact, the least square approximation of a dissimilarity d by a tree metric t is shown to be *NP*-hard in Day [10] (see also Day [11]). For this problem, the heuristics giving satisfactory results have usually a time complexity of $O(n^4)$ or $O(n^5)$, the Saitou and Nei NJ (nearest joining) method in $O(n^3)$ being a noticeable exception. For many problems of data analysis, where the purpose is to handle large data sets, a complexity order of $O(n^4)$ or $O(n^5)$ is too high. Since $O(n^5)$ algorithms are still currently proposed for the fitting of tree metrics, it seems that the situation is different in many applications of this problem. We describe here two algorithms based on Yushmanov orders. They proceed by successive local least squares approximations and are basically in $O(n^2)$. Global approximation and repetitive uses will increase this complexity up to $O(n^4)$ or $O(n^5)$ in Section 4.2, with seemingly good performances (Section 4.3).

The principle of the algorithm is as follows: at the step k, $2 \le k \le n\text{-}1$, a current valued tree T_ℓ^k has been determined, with the leaves $\{x_1,\ldots, x_k\}$. The vertex a_{k+1} is assumed to be on the path $T_\ell^k(x_1x_k)$ of this tree and a reconstruction procedure is introduced to obtain the tree T_ℓ^{k+1}, with the further problem of the determination of the best place of a_{k+1} on the path $T_\ell^k(x_1x_k)$. At the final step n-1, the valued X-tree T_ℓ corresponding to the tree metric t is obtained. In the determination of the best place of a_{k+1}, the lengths α, β and γ, of, respectively, the paths $T_\ell^{k+1}(x_1a_{k+1})$ and $T_\ell^{k+1}(a_{k+1}x_k)$ and the edge $a_{k+1}x_{k+1}$, are adjusted at each step according to a least squares criterion. Two methods are proposed here. In Algorithm 4, the computations at the Step k are based on the only two values $d(x_1,x_{k+1})$, $d(x_k,x_{k+1})$ of the initial dissimilarity, together with the value $t(x_1,x_k)$ determined at the previous step; this computation corresponds with Problem $P_{1,k}$. In Algorithm 5, the best place for a_{k+1} is determined for each edge uv of the path $T_\ell^k(x_1x_k)$, taking in account all the initial values $d(x_i,x_{k+1})$, $i = 1,\ldots,k$. This computation corresponds with Problem $P_{2,k}(uv)$. The edge leading to the best fitting is chosen.

Problem $P_{1,k}$ (see Figure 11)**:**

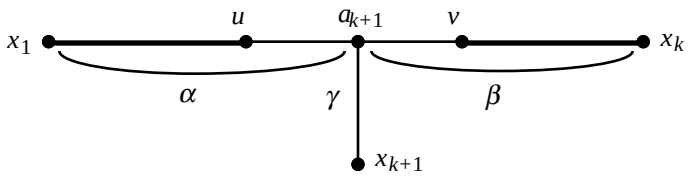

Figure 11

MINIMIZE $(\alpha + \gamma - d(x_1,x_{k+1}))^2 + (\beta + \gamma - d(x_k,x_{k+1}))^2$,
subject to: $\alpha+\beta = t(x_1,x_k)$ (determined at the previous step); $\alpha \geq 0$; $\beta \geq 0$; $\gamma \geq 0$.

Set $A = d(x_k,x_{k+1})-t(x_1,x_k)-d(x_1,x_{k+1})$ and $B = t(x_1,x_k)-d(x_k,x_{k+1})-d(x_1,x_{k+1})$; the problem reduces as:

MINIMIZE $\alpha^2 + \gamma^2 + A\alpha + B\gamma$,
subject to: $\alpha \geq 0$; $\gamma \geq 0$; $t(x_1,x_k) - \alpha \geq 0$.

With the use of the Lagrange function (see for instance Ciarlet [8] or Minoux [21]):

$$F(\lambda_1,\lambda_2,\lambda_3) = \alpha^2 + \gamma^2 + A\alpha + B\gamma - \lambda_1\alpha - \lambda_2\gamma + \lambda_3(\alpha - t(x_1,x_k)),$$

where $\lambda_i \geq 0$ for $i = 1, 2, 3$, we obtain a necessary condition on α and γ for reaching the minimum: $(\alpha,\gamma) \in \{(-\frac{A}{2},-\frac{B}{2}); (-\frac{A}{2},0); (t(x_1,x_k),-\frac{B}{2}); (t(x_1,x_k),0); (0,-\frac{B}{2}); (0,0)\}$.

Among these couples, choose the one satisfying the constraints and actually realizing the minimum.

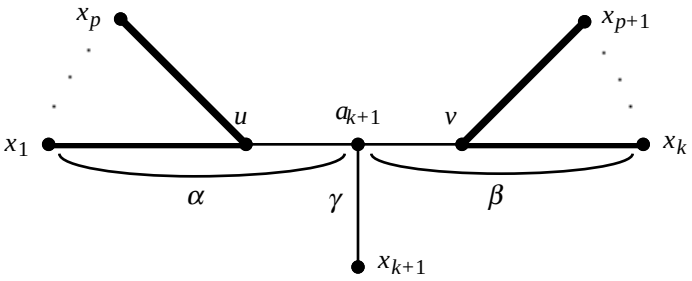

Figure 12

Problem $P_{2,k}$:
Let *uv* be an edge of the path $T_\ell^k(x_1x_k)$, *u* being its extremity closest to x_1. Since the order $x_1, x_2,\ldots, x_k$ on the leaves of T_ℓ^k is circular, there always exists an index *p* such that all the leaves $x_1,\ldots, x_p$ are on the same side of a_{k+1}, while $x_{p+1},\ldots,x_k$ are on the other side (Figure 12). Taking in account the distances $dist(x_i,T(x_1x_k))$ for all elements x_i, $2 \leq i \leq k$-1, we obtain the following quantity to minimize for the best place for a_{k+1} on the edge *uv*:

Problem $P_{2,k}(uv)$

MINIMIZE $(\alpha + \gamma - d(x_1,x_{k+1}))^2 + (\beta + \gamma - d(x_k,x_{k+1}))^2$
$+ \sum_{2\leq i\leq p} (d(x_i,x_{k+1}) - (\alpha - dist(x_1,T(x_ix_k)) + dist(x_i,T(x_1x_k)) + \gamma))^2$
$+ \sum_{p+1\leq i\leq k-1} (d(x_i,x_{k+1}) - \beta - dist(x_k,T(x_1x_i)) + dist(x_i,T(x_1x_k)) + \gamma)^2$,
subject to: $\alpha + \beta = t(x_1,x_k)$; $\beta \geq 0$; $\gamma \geq 0$; $t(x_1,u) \leq \alpha \leq t(x_1,v)$,

where $t(x_1,u)$ and $t(v,x_k)$ are the distances between the corresponding vertices in the valued tree T_ℓ^k.

Set $A_1 = d(x_1,x_{k+1})$, $A_k = d(x_k,x_{k+1}) - t(x_1,x_k)$,
$A_i = d(x_i,x_{k+1}) - t(x_i,x_k) + t(x_1,x_k)$ for $2 \leq i \leq p$, and
$A_i = d(x_i,x_{k+1}) - t(x_1,x_i)$ for $p+1 \leq i \leq k$-1. After reduction, the problem is now:

MINIMIZE $\sum_{1\leq i\leq p} (\alpha+\gamma-A_i)^2 + \sum_{p+1\leq i\leq k} (\alpha-\gamma+A_i)^2$,
subject to: $\gamma \geq 0$; $t(x_1,u) \leq \alpha \leq t(x_1,v)$.

Setting $B = 4p-2k$, $C = 2\sum_{p+1\le i\le k} A_i - 2\sum_{1\le i\le p} A_i$ and $D = -2\sum_{1\le i\le k} A_i$, one gets:

MINIMIZE $k\alpha^2 + k\gamma^2 + B\alpha\gamma + C\alpha + D\gamma$,
subject to the same constraints.

Consider the Lagrange function:

$F(\lambda_1,\lambda_2,\lambda_3) = k\alpha^2+k\gamma^2+B\alpha\gamma+C\alpha+D\gamma+\lambda_1(\alpha-t(x_1,v))-\lambda_2\gamma+\lambda_3(t(x_1,u)-\alpha)$.

The necessary conditions for minimum are:

$F'_\alpha = 2k\alpha + B\gamma + C + \lambda_1 - \lambda_3 = 0$;
$F'_\gamma = 2k\gamma + B\alpha + D - \lambda_2 = 0$;
$\lambda_1(\alpha-t(x_1,v)) = 0$; $\lambda_2\gamma = 0$; $\lambda_3(t(x_1,u)-\alpha) = 0$,

where $\lambda_j \ge 0$ for $j = 1, 2, 3$.

This system of equations leads to six possible solutions:

1. $\alpha = t(x_1,v)$, $\gamma = 0$;
2. $\alpha = t(x_1,v)$, $\gamma = -\dfrac{Bt(x_1,v)+D}{2k}$;
3. $\alpha = -\dfrac{C}{2k}$, $\gamma = 0$;
4. $\alpha = \dfrac{BD-2kC}{4k^2-B^2}$, $\gamma = \dfrac{BC-2kD}{4k^2-B^2}$;
5. $\alpha = t(x_1,u)$, $\gamma = 0$;
6. $\alpha = t(x_1,u)$, $\gamma = -\dfrac{Bt(x_1,u)+D}{2k}$.

Among the couples (α,γ) above, choose the one satisfying the constraints and actually realizing the minimum. End of Problem $P_{2,k}(uv)$.

Among the edges of the path $T_\ell^k(x_1x_k)$, choose the one realizing the minimum. End of Problem $P_{2,k}$.

In the following statement, the notations are the same as in Algorithm 2; moreover, $w(0,T(x_1x_k)) = w(1,T(x_1x_k)) = w'(0,T(x_1x_k)) = x_1$.

Algorithm 4: construction of a valued X-tree from a dissimilarity d
Input: a finite set X with n elements; a dissimilarity d on X.
Output: a valued X-tree $T_\ell = (V(T),E(T),\ell)$.

Initialization. Compute a Yushmanov order $(x_1, x_2,\ldots, x_n)$ on X;
$V(T) := \{x_1, x_2\}$; $E(T) := x_1x_2$; $\ell(x_1x_2) := d(x_1,x_2)$; $k := 1$
Repeat
$k := k+1$; $S := 0$; $i := 0$
The problem is to add the leaf x_{k+1} *to the current valued* tree T_ℓ^k *with leaves* $x_1,\ldots,x_k$.
Solve Problem $P_{1,k}(\alpha, \gamma)$ for the path $T_\ell^k(x_1x_k)$
while $S < \alpha$ **do**
$i := i+1$
$S = S+\ell(i,T(x_1x_k))$
$u := w(i,T(x_1x_k))$; $v := w'(i,T(x_1x_k))$
if $S = \alpha$ **then**
$V(T) := V(T)\cup\{x_{k+1}\}$
$E(T) := E(T)\cup\{ux_{k+1}\}$

$\ell(ux_{k+1}) := \gamma$

else $V(T) := V(T) \cup \{a_{k+1}, x_{k+1}\}$

$E(T) := (E(T) - \{uv\}) \cup \{ua_{k+1}, va_{k+1}, a_{k+1}x_{k+1}\}$

$\ell(a_{k+1}x_{k+1}) := \gamma$

$\ell(ua_{k+1}) := \alpha - S + \ell(i, T(x_1x_k))$

$\ell(va_{k+1}) := S - \alpha$

until ($k = n$)

Algorithm 5: construction of a valued X-tree from a dissimilarity d

This algorithm is identical to Algorithm 4, except the instruction "**Solve** Problem $P_{1,k}$", which is replaced with "**Solve** Problem $P_{2,k}$".

4.2. Time complexity and strategies for the use of Algorithms 4 and 5. Problem $P_{1,k}$ is solved in $O(1)$ and the obtainment of a Yushmanov order by Algorithm 1 is $O(n^2)$. Then, Algorithm 4 has the same time complexity $O(n^2)$. In Problem $P_{2,k}$, using the fact that the Yushmanov order is circular, we can proceed to a careful updating from edge to edge on the path $T_\ell^k(x_1x_k)$: when moving from the edge uv to the next edge vw, it is not necessary to compute again all the values A_i. This computation has to be done just for the A_i's such that $m(x_1,x_i,x_k) = v$ and $i \neq 1, k$; their number is the degree of v minus two. The total number of such operations related to the path $T_\ell^k(x_1x_k)$ is exactly k-2, for $k = 3,\ldots, n$. Finally, though the steps of Algorithm 5 seem more complicated than those of Algorithm 4, each step is at most $O(k)$ and this algorithm is $O(n^2)$ again.

As reported at the beginning of this section, such a time complexity is very good. Two $O(n^2)$ methods are proposed in Leclerc [17]. They are based on combinatorial properties of tree metrics, and not related with the least squares criterion.

Indeed, this time complexity $O(n^2)$ is not realistic (at least at the present state of this study): the initial choice of the elements x_1 and x_n is important in the use of algorithms 4 or 5, since it determines in fact a Yushmanov order among all the possible ones. Further studies or experimentations will be necessary to have an idea of the best strategy for this choice. In the experimental testings of the next subsection, we use the low complexity of Algorithms 4 and 5 to develop the alternative approach consisting of trying all the possible initial pairs (x_1,x_n). It gives two $O(n^4)$ methods, called Methods 1 and 2, based on Algorithms 4 and 5, respectively.

Both Methods 1 and 2 are completed with an adjustment of the lengths of the edges, once the topology of the obtained tree is fixed. Based on a least squares criterion on the differences between the obtained tree metric and the initial dissimilarity, this quadratic approximation is performed with the Gauss-Seidel method proposed in Barthélemy and Guénoche ([8], pp. 60-66; see for instance Ciarlet [8]). Since it requires an $O(n^4)$ time, it is mainly interesting for algorithms having at least this complexity. Even with this improvement, good results may hardly be expected from Method 1, because of the local character of Problem $P_{1,k}$. In Method 2, the approximation is done on the obtained trees that give the best result for the least squares criterion. The number of these trees is a small fixed one in Method M21 (in $O(n^4)$), and of order n in Method M22 (in $O(n^5)$).

4.3. Experimental results. Algorithms 1, 4 and 5 have been programmed in the C^{++} programming language and tested on a MS-DOS machine of IBM-PC type.

We use an evaluation method similar to that of Pruzanski, Tversky et Caroll [23], De Soete [12] and Gascuel et Levy [16]. Each data set is obtained as follows: first, an X-tree with n leaves and $2n$-3 arêtes is generated at random (for n = 12, 18, 24). The lengths of the edges are chosen at random from a uniform distribution on the real interval [0,1]. Then, the values of the corresponding tree metric are computed and normalized to have a unit variance, leading to a valued X-tree TT. A normal random noise of mean 0 and variance σ^2 = 0.1, 0.25, 0.5 is added to these values to obtain the distance d; in the rare cases where a negative value $d(x,y)$ results from these operations, this value is replaced with 0.01. A number of 100 data sets is generated for each pair of values of n and σ.

The results obtained from our methods M1 and M2 (with the variants M21 and M22 described above) are compared with those of the classical NJ method. We also consider the true tree TT which, contrary to the case of

observed data, is known in these experimental ones.

The quality of the adjustement is evaluated by the means, computed on all the tests corresponding to each pair (n,σ), of two quantities:

1. The proportion of explained variance, as given by a formula of Pruzanski,Tversky and Caroll [23], where $m(d)$ is the mean value of d and t is the fitted tree metric:

$$\%\mathrm{Var} = 100\left(1-\frac{\sum_{xy\in X^2}(d(xy)-t(xy))^2}{\sum_{xy\in X^2}(d(xy)-m(d))^2}\right)$$

This quantity is also determined for the tree metric obtained after quadratic approximation (column "%Var+"). With this approximation, the NJ method becomes an $O(n^4)$ one.

2. The topological distance of Robinson and Foulds [24] between the true tree TT and the X-tree representation of t. It is a least move distance, the elementary move between two X-trees being the deletion or the addition of the split corresponding with an edge; that is, it is the symmetric difference metric on X-trees defined as sets of splits (Buneman [5]; see Barthélemy and Guénoche [2], ch. V). The distance between two trees is expressed as a percentage of the maximum value $3n$-6.

		$\sigma^2 = 0.1$			$\sigma^2 = 0.25$			$\sigma^2 = 0.5$		
	M1	88.14	93.54	12.76	73.93	84.98	18.06	61.49	75.84	26.57
	M21	92.62	93.66	10.47	83.24	85.44	16.00	73.86	76.96	21.63
n = 12	M22	92.62	93.69	9.80	83.24	85.56	14.70	73.86	77.17	20.90
	NJ	92.70	93.63	10.47	83.32	85.44	16.27	73.49	76.75	22.63
	TT	90.13	93.49		78.08	84.94		64.67	75.64	
	M1	84.80	92.26	17.58	68.65	82.58	27.12	51.52	69.89	37.48
	M21	91.33	92.57	12.19	80.89	83.55	20.77	68.30	72.12	31.75
n = 18	M22	91.33	92.61	11.64	80.89	83.73	19.73	68.30	72.55	29.08
	NJ	91.42	92.55	12.81	81.06	83.51	21.75	67.96	72.33	31.04
	TT	90.17	92.46		78.23	83.28		63.72	71.80	
	M1	82.74	91.28	23.64	65.46	80.05	37.83	46.47	66.24	46.92
	M21	90.42	91.83	16.62	79.04	82.07	28.13	64.62	69.65	37.02
n = 24	M22	90.42	91.93	13.54	79.04	82.34	28.08	64.62	70.24	35.60
	NJ	90.67	91.89	15.91	79.41	82.27	26.82	66.68	70.56	33.79
	TT	90.00	91.86		78.40	82.20		64.41	70.35	
		% VAR	%VAR+	RF	% VAR	%VAR+	RF	% VAR	%VAR+	RF

Table 2

The analysis of the results leads to the following observations: compared with the others, Method M1 is too elementary to give satisfactory results. The quadratic approximation is a very efficient tool for the improvement of the variance percentage. When the rank n is 12 or 18, Method M21 gives globally better results than NJ, and Method M22 is globally the best one in these tests. Nevertheless, method NJ is the most robust when the size of the data and the variance of the noise both increase. For n = 24 and σ^2 = 0.5, it still gives the best results.

A further experiment is based on the data of Case [6] (immunological distances between nine species of frogs), frequently used for similar testing (see for instance Saitou and Nei [26] and Gascuel and Lévy [16]). Table 9 gives these distances.

	1	2	3	4	5	6	7	8
1: Aurora								
2: Boylii	**10**							
3: Cascadae	**13**	**7**						
4: Muscosa	**12**	**7**	**7**					
5: Temporaria	**57**	**50**	**40**	**45**				
6: Pretiosa	**22**	**9**	**11**	**15**	**48**			
7: Catesbiana	**86**	**65**	**54**	**48**	**85**	**54**		
8: Pipiens	**89**	**67**	**66**	**49**	**83**	**55**	**54**	
9: Tarahumarae	**97**	**72**	**79**	**67**	**107**	**60**	**59**	**48**

Table 3

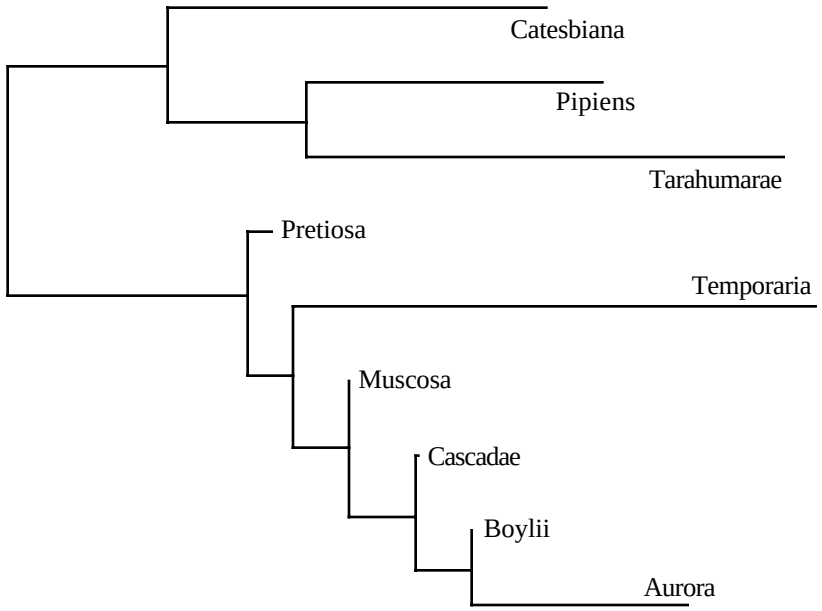

Figure 13

Five fitting methods are compared here: the methods M21 and M22 of this paper, the NJ method, and two other methods recognized to give generally good results: the method of scores (MS) of Luong [19] and Barthélemy and Guénoche [2], based on the grouping of pairs x, y such that $d(x,y) + d(z,w) < \max\{ d(x,z) + d(y,w) , d(x,w) + d(y,z) \}$ for a maximum number of pairs z, w (it may be considered as a refinement of the ADDTREE method of Sattah and Tversky [27]); and the reduction method of Gascuel and Lévy [16], denoted here GL (it iteratively modifies the values of the dissimilarity towards a dissimilarity satisfying the four point-condition). These last two methods are $O(n^5)$. The best tree obtained by Method 2.2 and the corresponding tree metric are given in Figure 13 and Table 4.

Five criteria are used for the comparison of the methods. The combinatorial criterion NI (number of inversions) is the number of quadruples $xyzw$ such that a unique minimum among the three sums $d(x,y) + d(z,w)$, $d(x,z) + d(y,w)$ and $d(x,w) + d(y,z)$ disagrees with the configuration of the obtained X-tree. The criteria AAD , MAD and MSD respectively correspond to the average absolute difference, the maximum absolute difference and the mean squared difference between the values of the initial dissimilarity matrix and the obtained tree metric one. The criterion L , not available here for the reduction method, is the total length of the obtained valued X-tree, a short length being in agreement with the "parsimony principle" of phylogenetics. The criterion values given in Table 5 are those obtained after the quadratic approximation of the edge lengths mentioned in the previous subsection. In Method M21, the approximation is just done one time, on the best tree for the least squares criterion. Here, Method M22, where n trees are used for the approximation, provides a significant improvement to the few number of trees (here, three) considered in Method M21. The obtained tree is topologically an intermediate between those obtained by Saitou and Nei (also by method M21), on the one hand,

and Gascuel and Lévy, on the other hand: it just differs from the former by the exchange of the places of *Muscosa* and *Cascadae*, and from the latter by the exchange of *Temporaria* and *Pretiosa*. Contrary to the Gascuel and Lévy method, the quadratic approximation is necessary to obtain a good fit.

	1	2	3	4	5	6	7	8
2	**13.22**							
3	**16.61**	**3.39**						
4	**20.88**	**7.66**	**4.35**					
5	**60.66**	**47.43**	**44.13**	**39.78**				
6	**28.71**	**15.49**	**12.18**	**7.83**	**40.78**			
7	**76.03**	**62.81**	**59.50**	**55.15**	**88.12**	**50.39**		
8	**79.37**	**66.15**	**62.85**	**59.00**	**91.47**	**53.73**	**50.93**	
9	**90.52**	**77.30**	**73.99**	**69.64**	**102.61**	**64.87**	**62.07**	**48.00**

Table 4

	ANI	AAD	MAD	MSD	L
Scores	**n.a.**	**4.76**	**12.25**	**32.80**	**172.0**
NJ	**26**	**4.71**	**11.21**	**30.12**	**171.9**
GL	**23**	**4.52**	**10.05**	**28.95**	**n.a.**
M21	**26**	**4.71**	**11.21**	**30.12**	**171.9**
M22	**23**	**4.61**	**9.97**	**28.27**	**170.7**

Table 5

5. Conclusion

The results and algorithms presented in this paper give evidence that Yushmanov orders are an interesting tool for the study of tree metrics. Among the questions arising about these orders, two of them seem especially interesting: the possible significance of Yushmanov orders for other types of dissimilarities; and their relations with the previously known combinatorial properties of tree metrics or X-trees (for instance the 4-ary relations characterized by Colonius and Schutze [9], the sets of splits of Buneman [5] and the relations with minimum spanning trees in Leclerc [17]).

Concerning the fitting algorithms of Section 4, the natural question is to device a way of preserving the low complexity of Algorithms 4 or 5. Here, this low complexity just allowed us to generate many good candidate trees, and to look for the best solutions among these candidates. Another direction of research is to generalize the method to other criteria, for instance the weighted least square one. An algorithm based on this criterion and sharing some features with those presented here, but without the use of circular orders, has been proposed by Makarenkov [20].

An important fact is the geometric signification, related to planar representations, of Yushmanov orders. Some uses of these orders for the drawing of trees, possibly directly from a dissimilarity array, may be expected: for instance, circular orders correspond to the so-called *hierarchical drawings* of an X-tree (Barthélemy and Guénoche [2], p.28). In such a drawing, inspired by the dendrograms of Numerical Taxonomy, the latent vertices are represented by horizontal lines, the upper one corresponding to the choice of a root. The edges of the tree are represented by vertical lines and no crossings are allowed. These orders, that can be obtained as right-left orders on the leaves in a hierarchical drawing, have been studied by Brossier [3] in the case of dendrograms; Figure 14 shows a hierarchical drawing of the tree of Figure 4, with the right-left order 10 9 4 3 6 5 8 7 2 1, which is circular. One may also expect that the algorithms described here could be modified in order to lead to a new family of low complexity methods of hierarchical classification, very different from the single linkage algorithm, which is, up to our knowledge, the only one in $O(n^2)$.

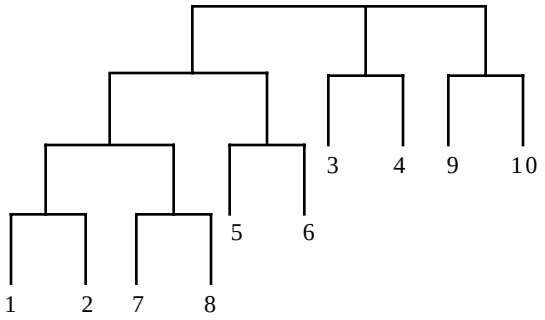

Figure 14

References

1. H.J. Bandelt, Recognition of tree metrics, *SIAM Journal on Discrete Mathematics* **3** (1990), 1-6.
2. J.P. Barthélemy, A. Guénoche, *Les arbres et les représentations des proximités*, Masson, Paris, 1988, transl. *Trees and proximity representations*,Wiley, New York, 1991.
3. G. Brossier, Représentation ordonnée des classifications hiérarchiques, *Statistique et Analyse des Données* **2** (1980), 31-44.
4. G. Brossier, Approximation des dissimilarités par des arbres additifs, *Mathématiques et Sciences humaines* **9 1** (1985), 5-21.
5. P. Buneman, The Recovery of Trees from Measures of Dissimilarity, in *Mathematics in Archaeological and Historical Sciences* (eds. F.R. Hodson, D.G. Kendall and P. Tautu), Edinburgh University Press, Edinburgh, 1971, 387-395.
6. S.M. Case, Biochemical systematics of members of the genus *Rana* native to wzstern North America, *Systematic Zoology* **27** (1978), 299-311.
7. S. Chaiken, A.K. Dewdney, P.J. Slater, An optimal diagonal tree-code, *SIAM Journal on Algebraic and Discrete Methods*, **4** (1983), 42-49.
8. P.G. Ciarlet, *Introduction à l'analyse numérique matricielle et à l'optimisation*, Masson, Paris, 1985.
9. H. Colonius, H.H. Schulze, Tree structure for proximity data, *British J. Math. Statist. Psychol.* **34** (1981), 167-180.
10. W.H.E. Day, Computational complexity of inferring phylogenies from dissimilarity matrices, *Bull. of mathematical Biology* **49** (1987), 461-467.
11. W.H.E. Day, Complexity theory: an introduction for practitioners of classification, in *Clustering and Classification* (eds P. Arabie, L.J. Hubert and G. De Soete), World Scientific Publ., River Edge, NJ, 1996, 199-233.
12. G. De Soete, A Least squares Algorithm for Fitting Additive Trees to Proximity Data, *Psychometrika* **48** (1983), 621-626.
13. G. De Soete, J.D. Caroll, Tree and other network models for representing proximity data, in *Clustering and Classification* (eds P. Arabie, L.J. Hubert and G. De Soete), World Scientific Publ., River Edge, NJ, 1996, 157-197.
14. A.K. Dewdney, Diagonal tree-codes, *Information and Control* **40** (1979), 234-239.
15. A.J. Dobson, Unrooted trees for numerical taxonomy, *J. Appl. Prob.* **11**(1974), 32-42.
16. O. Gascuel, D. Lévy, A reduction algorithm for approximating a (nonmetric) dissimilarity by a tree distance, *J. of Classification* **13** (1996), 129-155.
17. B. Leclerc, Minimum spanning trees for tree metrics: abridgements and adjustments, *J. of Classification* **12** (1995), 207-241.
18. B. Leclerc, V. Makarenkov, On some relations between 2-trees and tree metrics, Research Report n° 131, C.A.M.S., Paris, 1997, submitted.
19. X. Luong, Thesis, Université René Descartes, Paris, 1987.
20. V. Makarenkov, Deux algorithmes d'approximation d'une dissimilarité par une distance d'arbre au sens du critère des moindres carrés pondérés", *Les cahiers du C.A.M.S.*, n° 128, 1996.
21. M. Minoux, *Programmation mathématique, théorie et algorithmes*, Dunod, Paris, 1983.
22. A.N. Patrinos, S.L. Hakimi, The distance matrix of a graph and its tree realization, *Quart. Appl. Math.* **30** (1972), 255-269.
23. S. Pruzansky, A. Tversky, , J.D. Caroll, Spatial Versus Tree Representations of Proximity Data, *Psychometrika* **4 7** (1982), 3-19.

24. D.R. Robinson, L.R. Foulds, Comparison of phylogenetic trees, *Mathematical Biosciences* **53** (1981), 131-147.

25. M. Roux, Techniques of approximation for building two tree structures, in *Recent Devlopments in Clustering and Data Analysis* (eds. C. Hayashi, E. Diday, M. Jambu, N. Ohsumi), Academic Press, New York, 1988, 151-170.

26. N. Saitou, M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees, *Molecular Biology Evolution* **4** (1987), 406-425.

27. S. Sattah, A. Tversky, Additive similarity trees, *Psychometrika* **42** (1977), 319-345.

28. S.V. Yushmanov, Construction of a tree with p leaves from $2p$-3 elements of its distance matrix (russian), *Matematicheskie Zametki* **35** (1984), 877-887.

29. K. Zaretskii, Construction of a tree on the basis of a set of distances between its leaves (russian), *Uspekhi Mat. Nauk*. **20** (1965), 90-92.

CENTRE D'ANALYSE ET DE MATHÉMATIQUE SOCIALES, ÉCOLE DES HAUTES ÉTUDES EN SCIENCES SOCIALES, 54 Boulevard Raspail, F-75270 PARIS CEDEX 06, FRANCE.
E-mail Address: leclerc@ehess.fr

INSTITUTE OF CONTROL SCIENCES, 65 Profsoyuznaya, MOSCOW 117806, RUSSIA.