Seriation in the presence of errors: NP-hardness of l_∞-fitting Robinson structures to dissimilarity matrices

Victor Chepoi, Bernard Fichet, Morgan Seston

Laboratoire d'Informatique Fondamentale de Marseille,
Université de la Méditerranée, Faculté des Sciences de
Luminy, F-13288 Marseille Cedex 9, France,
{*chepoi,fichet,seston*}*@lif.univ-mrs.fr*

Abstract

In this paper, we establish that the following fitting problem is NP-hard: given a finite set X and a dissimilarity measure d on X (d is a symmetric function d from X^2 to the non-negative real numbers and vanishing on the diagonal), we wish to find a Robinsonian dissimilarity d_R on X minimizing the l_∞-error $||d - d_R||_\infty = \max_{x,y \in X}\{|d(x,y) - d_R(x,y)|\}$ between d and d_R. Recall that a dissimilarity d_R on X is called *monotone* (or *Robinsonian*) if there exists a total order $\prec$ on X such that $x \prec z \prec y$ implies that $d(x,y) \geq \max\{d(x,z), d(z,y)\}$. The Robinsonian dissimilarities appear in seriation and clustering problems, in sparse matrix ordering and DNA sequencing.

1 Introduction

A major issue in classification and data analysis is to visualize simple geometrical and relational structures between objects. Necessary for such an analysis is a similarity or a dissimilarity measure on a set of objects, which is either measured directly or computed from a data matrix. Many applied algorithmic problems ranging from archeological dating through DNA sequencing and numerical ecology to sparse matrix reordering and overlapping clustering involve ordering a set of objects so that closely coupled elements are placed near each other. The rearranged data may then speak for themselves. For example, the classical *seriation problem* [23, 25, 27] is to find a simultaneous ordering (or permutation) of the rows and the columns of the dissimilarity matrix with the objective of revealing an underlying one-dimensional structure. The basic idea is that small values should be concentrated around the main diagonal as closely as possible, whereas large values should fall as far from it as possible. This goal is best achieved by considering the so-called *Robinson property* [30]. Seriation is of importance in hypertext ordering [7], ecology [28], sparse matrix ordering [2], musicology [21], and DNA sequencing [6, 10, 29]. The most common methods for clustering provide a visual display of data in the form of hierarchical structures (or *dendrograms*). Dissimilarity matrices which are in perfect agreement with dendrograms satisfy the following Robinson

property: the distances from a given object increase when moving from this object to the left or to the right in the underlying dendrogram. These dissimilarities are best known under the name of *ultrametrics*. Generalizing the correspondence between ultrametrics and dendrograms, it has been shown in [8, 16, 17] that the Robinson dissimilarities can be visualized by pseudo-hierarchical structures also called pyramids.

Real experimental or archeological data always contain errors, therefore the dissimilarity between the objets can be measured only approximatively. As a consequence, any simultaneous permutation of the rows and the columns of the dissimilarity matrix gives a matrix which fails to satisfy the Robinson property, and we are lead to the problem of finding a matrix reordering which is as close as possible to a Robinson matrix. As an error measure one can use the usual l_p-distance between two matrices of equal size. In this paper, we show that the problem of optimal fitting a dissimilarity matrix by a Robinson matrix under the l_∞-error is NP-hard, thus answering an open question from [4]. The companion paper [13] provides a factor 16 polynomial time algorithm for solving this problem.

2 Definitions and Problems

Let X be a set of n elements to sequence, endowed with a dissimilarity function that reflects the desire for two elements to be near or far from each other in the sequence. Recall that a *dissimilarity* is a symmetric function d from X^2 to the nonnegative real numbers and vanishing on the diagonal, i.e. $d(x,y) = d(y,x) \geq 0$ and $d(x,x) = 0$. We call $d(x,y)$ the *distance* between the objects $x, y \in X$. Denote by d_0 the standard distance of the complete graph on X, i.e., $d_0(x,y) = 1$ if $x \neq y$ and $d_0(x,y) = 0$, otherwise. Denote by $\mathcal{D}$ the set of all dissimilarities on X. Notice that $\mathcal{D}$ is a convex cone, because the convex combination $d := \alpha \cdot d' + (1-\alpha) \cdot d''$ of two dissimilarities $d', d'' \in \mathcal{D}$ also belongs to $\mathcal{D}$; for $x, y \in X$, $d(x,y) = \alpha \cdot d'(x,y) + (1-\alpha) \cdot d''(x,y)$. Additionally, if $d \in \mathcal{D}$ and $\alpha > 0$ is a constant such that $d - \alpha \cdot d_0$ is non-negative, then $d - \alpha \cdot d_0 \in \mathcal{D}$.

A dissimilarity d and a total order $\prec$ on a set X are said to be *compatible* if $x \prec z \prec y$ implies that $d(x,y) \geq \max\{d(x,z), d(z,y)\}$. A dissimilarity d on X is said to be *Robinsonian* if it admits a compatible order. Equivalently, d is Robinsonian if its matrix can be symmetrically permuted so that its elements do not decrease when moving away from the main diagonal along any row or column. Such a matrix is called *Robinson* [14, 16, 20, 23] or *linear* in the terminology of Mirkin and Rodin [29]. Notice that if d is Robinsonian, then every increasing monotone transform φ of d is also Robinsonian (i.e., $d(x,y) < d(x',y')$ implies that $\varphi(d(x,y)) < \varphi(d(x',y'))$). In particular, for any Robinsonian dissimilarity d and any constant c, if $d + \alpha \cdot d_0$ is a dissimilarity, then $d + \alpha \cdot d_0 \in \mathcal{R}$. Basic examples of Robinson dissimilarities are the ultrametrics and the standard *line-distance* between n points on the line. Recall, that d is an *ultrametric* if $d(x,y) \leq \max\{d(x,z), d(y,z)\}$ for all $x, y, z \in X$.

Let X be a set on n elements. Denote by $\mathcal{R}$, $\mathcal{R}^\prec$, and $\mathcal{U}$ the cone of all Robinson dissimilarities on X, the convex cone of all Robinson dissimilarities compatible with a total order $\prec$ on X, and the cone of all ultrametrics on X, respectively. For two dissimilarities

$d, d' \in \mathcal{D}$, and $p \geq 1$, define the l_p-error or l_p-norm between d and d' by

$$||d - d'||_p = (\sum_{x,y \in X} |d(x,y) - d'(x,y)|^p)^{\frac{1}{p}},$$

$$||d - d'||_\infty = \max_{x,y \in X}\{|d(x,y) - d'(x,y)|\}.$$

Endow the set of all dissimilarities $\mathcal{D}$ with a partial order $\leq$, where $d \leq d'$ if and only if $d(x,y) \leq d'(x,y)$ for all $x, y \in X$. For a dissimilarity d and a subset $\mathcal{D}'$ of $\mathcal{D}$, a dissimilarity $\hat{d} \in \mathcal{D}'$ is called a *sub-dominant* of d in $\mathcal{D}'$ if $\hat{d}$ is the (necessarily unique) maximum of the set $\{d' \in \mathcal{D}' : d' \leq d\}$. Analogously, a dissimilarity $\check{d} \in \mathcal{D}'$ is called a *super-dominated* of d in $\mathcal{D}'$ if $\check{d}$ is the (necessarily unique) minimum of the set $\{d' \in \mathcal{D}' : d \leq d'\}$. The dissimilarities $\hat{d}$ and/or $\check{d}$ not always exist, nevertheless they exist in some important cases, in particular, $\hat{d}_u$ exists if $\mathcal{D}'$ is the set $\mathcal{U}$ of all ultrametrics [33]. In this case, the ultrametric sub-dominant $\hat{d}_u$ can be defined in the following way: construct the minimum spanning tree T in the complete graph on X in which the length of the edge xy is $d(x,y)$, then $\hat{d}_u(x,y)$ is the length of the longest edge on the unique path of T connecting the vertices x and y.

In this paper, we study the complexity status of the following optimization problem, which can be viewed as the seriation problem in the presence of errors:

Problem l_∞-FITTING-BY-ROBINSON: *Given a dissimilarity $d \in \mathcal{D}$ find a Robinson dissimilarity $d_R \in \mathcal{R}$ minimizing the l_∞-error $||d - d_R||_\infty$.*

In other words, we are searching for a minimal value of ϵ such that for each pair x, y of different elements of X one can pick a value $d_R(x,y) \in [d(x,y) - \epsilon, d(x,y) + \epsilon]$ so that the resulting dissimilarity d_R is Robinsonian. To formulate the underlying decision problem, we relax the notions of compatible order and Robinson dissimilarity. Given $\epsilon \geq 0$, a total order $\prec$ on X is called *ϵ-compatible* if $u \prec x \prec y \prec v$ implies $d(u,v) + 2\epsilon \geq d(x,y)$; here u and x may coincide as well as y and v. (To show that a total order $\prec$ is ϵ-compatible it suffices to show that $d(x,z) \geq \max\{d(x,y), d(y,z)\} + \epsilon$ for any $x, y, z \in X$ such that $x \prec y \prec z$.) An *ϵ-Robinsonian dissimilarity* is a dissimilarity admitting an ϵ-compatible order. We are lead to the following recognition problem:

Problem ϵ-ROBINSON: *Given a dissimilarity d and a real number $\epsilon > 0$, is d ϵ-Robinsonian?*

We will show below, that the problem ϵ-ROBINSON is NP-complete and that, unless P=NP, it is NP-hard to approximate l_∞-FITTING-BY-ROBINSON within a factor smaller than 3/2.

3 Known results

Now, we briefly recall here what is known about Robinsonian structures and about l_∞-fitting of distances by simpler distances. In the original Robinson's paper [30] and in some other papers [2, 24, 28], compatible orders and Robinson matrices are defined for similarities; the elements of a Robinson similarity matrix not increase when moving away from the main

diagonal. Atkins et al. [2] showed that if S is a Robinson similarity matrix, then the coordinates of the eigenvector of its smallest nonzero eigenvalue of the Laplacian of S (the so-called Fiedler vector of S) constitute a monotone sequence of numbers. They use this result to design an algorithm of complexity $O(nT(n) + n^2 \log n)$ to recognize if a similarity matrix of size $n \times n$ is pre-Robinson, where $T(n)$ is the complexity of computing the Fiedler vector of a matrix. Mirkin and Rodin [29] describe an $O(n^4)$ algorithm for testing if a dissimilarity d on n points is Robinsonian. For this, they build up the hypergraph of all balls of d and test if this hypergraph is an interval hypergraph. A simple divide-and-conquer $O(n^3)$-time algorithm for the same recognition problem has been designed in [11]. Barthelemy and Brucker [4, 5] established that the problem of an optimal l_p-approximation of a dissimilarity by several types of simpler dissimilarities, in particular by a strongly Robinsonian dissimilarity, is NP-hard for $p < \infty$, leaving the case $p = \infty$ open.

The problem of fitting distances by simpler distances is a classical problem in data analysis, phylogeny, and, more recently, in computer science. We review here only the algorithmic results about l_∞-fitting of distances. Farach, Kannan, and Warnow [19] showed that l_∞-fitting of a distance d by an ultrametric is polynomial. This result has been used by Agarwala et al. [1] to design a factor 3 approximation algorithm for l_∞-fitting of distances by tree-distances, a problem which has been shown to be NP-hard in the same paper [1]. A unified and simplified treatment of these results of [1, 19] using sub-dominants was given in [12]. The l_∞-fitting of a dissimilarity by a line-distance (the so-called Matrix-To-Line problem) has been proven NP-hard by Saxe [31] in 1979. More recently, Håstad, Ivansson, and Lagergren [22] showed that this problem can be approximated within factor of 2, but unless P=NP cannot be approximated within a factor smaller than 7/5 (notice that Agarwala et al. [1] establish a similar non-approximability result for tree-distances with 7/5 replaced by 9/8). Bădoiu [3] extended the results of [22] to l_∞-fitting of distances by rectilinear (l_1-) distances in the 2-dimensional space and proposed a constant-factor algorithm for this problem.

4 Preliminary results

In this section, we establish some auxiliary results used in the proof of NP-hardness (these results are also used in the approximation algorithm from [13]). In particular, Lemma 4.2 and Proposition 4.3 follow the approach presented in [12].

Lemma 4.1 *A dissimilarity d on X is Robinsonian if and only if there exists a total order $\prec$ on X, such that $d(x, y) \geq d(u, v)$ holds for any four (not necessarily distinct) elements $u, v, x, y \in X$ such that $x \prec u \prec v \prec y$.*

Proof. First suppose that the inequality $d(x, y) \geq d(u, v)$ holds for all $x \prec u \prec v \prec y$. To show that $\prec$ is compatible with d, pick the elements $x, y, z \in X$ such that $x \prec z \prec y$. Applying twice the 4-point inequality, first with $u = z, v = y$ and then with $u = x, v = z$, we deduce that $d(x, y) \geq d(z, y)$ and $d(x, y) \geq d(x, z)$, respectively. Thus $\prec$ is compatible with d, yielding that d is Robinsonian. Conversely, if d is Robinsonian, then there exists a total order

$\prec$ compatible with d. If $x \prec u \prec v \prec y$, then we have $d(x,y) \geq d(x,v)$ and $d(x,v) \geq d(u,v)$, therefore $d(x,y) \geq d(u,v)$, and the required 4-point inequality is proven. $\square$

Lemma 4.2 *For a total order $\prec$ on X and $d \in \mathcal{D}$, let $\check{d}_\prec$ and $\hat{d}_\prec$ be two dissimilarities defined in the following way: for $x, y \in X$ with $x \prec y, x \neq y$, set*

$$\check{d}_\prec(x,y) = \max\{d(u,v) : u, v \in X \text{ and } x \prec u \prec v \prec y\},$$

$$\hat{d}_\prec(x,y) = \min\{d(u,v) : u, v \in X \text{ and } u \prec x \prec y \prec v\}.$$

Then $\check{d}_\prec$ is the super-dominated of d in $\mathcal{R}_\prec$ and $\hat{d}_\prec$ is the sub-dominant of d in $\mathcal{R}_\prec$.

Proof. We will establish the result only for $\check{d}_\prec$, the proof of the second assertion is similar. From the definition of $\check{d}_\prec$ we infer that $d \leq \check{d}_\prec$. We assert that $\check{d}_\prec$ is Robinsonian, namely that $\prec$ is a compatible order for $\check{d}_\prec$. Let $E(u,v) = \{d(u',v') : u \prec u' \prec v' \prec v\}$. For a triplet $x \prec z \prec y$, the distance-sets $E(x,z)$ and $E(z,y)$ are obviously contained in $E(x,y)$. Taking the maximum in each of the sets $E(x,z), E(z,y)$ and $E(x,y)$, we deduce that $x \prec z \prec y$ implies $\check{d}_\prec(x,y) \geq \max\{\check{d}_\prec(x,z), \check{d}_\prec(z,y)\}$. It remains to show that if $d' \in \mathcal{R}_\prec$ and $d \leq d'$, then $\check{d}_\prec \leq d'$. Pick $x, y \in X$ with $x \prec y$. Let u, v be defined so that $x \prec u \prec v \prec y$ and $\check{d}_\prec(x,y) = d(u,v)$. Lemma 4.1 yields $d'(x,y) \geq d'(u,v)$. Since $d \leq d'$, we also have $d(u,v) \leq d'(u,v)$. Putting all together, we obtain $\check{d}_\prec(x,y) = d(u,v) \leq d'(u,v) \leq d'(x,y)$, thus $\check{d}_\prec \leq d'$. This shows that indeed $\check{d}_\prec$ is the super-dominated of d in $\mathcal{R}_\prec$. $\square$

The $\prec$-RESTRICTED l_∞-FITTING-BY-ROBINSON problem is obtained from the problem l_∞-FITTING-BY-ROBINSON by replacing $\mathcal{R}$ by $\mathcal{R}_\prec$. We show now that this restricted problem can be solved in polynomial time. Let $2\tilde{\epsilon}_\prec = ||d - \check{d}_\prec||_\infty$ and let $\tilde{d}_\prec$ be the (Robinsonian) dissimilarity obtained from $\check{d}_\prec$ by setting $\tilde{d}_\prec(x,y) = \max\{0, \check{d}_\prec(x,y) - \tilde{\epsilon}_\prec\}$ for all $x, y \in X, x \neq y$. Analogously, let $2\bar{\epsilon}_\prec = ||d - \hat{d}_\prec||_\infty$ and let $\bar{d}_\prec$ be the (Robinsonian) dissimilarity obtained from $\hat{d}_\prec$ by setting $\bar{d}_\prec(x,y) = \hat{d}_\prec(x,y) + \bar{\epsilon}_\prec$ for all $x, y \in X, x \neq y$; in other terms, $\bar{d}_\prec = \hat{d}_\prec + \bar{\epsilon}_\prec \cdot d_0$.

Proposition 4.3 *For a total order $\prec$ on X and $d \in \mathcal{D}$, $\tilde{d}_\prec$ and $\bar{d}_\prec$ are two Robinsonian dissimilarities that minimize $||d - d'||_\infty$ for $d' \in \mathcal{R}_\prec$. In particular, $\tilde{\epsilon}_\prec = \bar{\epsilon}_\prec$ holds.*

Proof. Again, we will establish the result only for $\tilde{d}_\prec$. First note that $\tilde{d}_\prec$ is a dissimilarity. Additionally, since $\check{d}_\prec \in \mathcal{R}_\prec$, we infer that $\tilde{d}_\prec \in \mathcal{R}_\prec$. Note that the l_∞-distance between d and $\tilde{d}_\prec$ is one-half of the l_∞-distance between d and $\check{d}_\prec$. Indeed, $\check{d}_\prec(x,y) \leq \tilde{d}_\prec(x,y) + \tilde{\epsilon}_\prec$, thus $0 \leq \tilde{d}_\prec(x,y) + \tilde{\epsilon}_\prec - d(x,y) \leq 2\tilde{\epsilon}_\prec$. Adding $-\tilde{\epsilon}_\prec$ to all parts of this inequality, we obtain $|\tilde{d}_\prec(x,y) - d(x,y)| \leq \tilde{\epsilon}_\prec$. On the other hand, there exist $u, v \in X$ such that $\check{d}_\prec(u,v) - d(u,v) = 2\tilde{\epsilon}_\prec$. This shows that $||d - \tilde{d}_\prec||_\infty = \tilde{\epsilon}_\prec$. To prove that $\tilde{d}_\prec$ is an optimal l_∞-approximation for d in $\mathcal{R}_\prec$, pick $d' \in \mathcal{R}_\prec$ and let $\epsilon' = ||d - d'||_\infty$. Then $d \leq d' + \epsilon' \cdot d_0$ and $d' + \epsilon' \cdot d_0 \in \mathcal{R}_\prec$. By Lemma 4.2, $d' + \epsilon' \cdot d_0 \geq \check{d}_\prec$. Since $\check{d}_\prec(u,v) - d(u,v) = 2\tilde{\epsilon}_\prec$, we obtain that $d'(u,v) + \epsilon' - d(u,v) \geq 2\tilde{\epsilon}_\prec$, showing that $2\tilde{\epsilon}_\prec - \epsilon' \leq d'(u,v) - d(u,v) \leq \epsilon'$. Hence $\tilde{\epsilon}_\prec \leq \epsilon'$. $\square$

We prove here that the optimal error ϵ^* in the problem l_∞-FITTING-BY-ROBINSON belongs to a compact list Δ of size $O(n^4)$, whose entries can be derived from the matrix of d.

Lemma 4.4 *For a total order $\prec$ on X and $d \in \mathcal{D}$, if $x \prec u \prec v \prec y$, then $\hat{\epsilon}_\prec \geq \frac{d(u,v)-d(x,y)}{2}$.*

Proof. By Lemma 4.2, $\check{d}_\prec(x,y) \geq \check{d}_\prec(u,v) \geq d(u,v)$. Hence $d(u,v) - d(x,y) \leq \check{d}_\prec(x,y) - d(x,y) \leq ||d - \check{d}_\prec||_\infty = 2\hat{\epsilon}_\prec$, where the last equality follows from Proposition 4.3. □

Lemma 4.5 *For a dissimilarity $d \in \mathcal{D}$, the optimal error ϵ^* of l_∞-fitting of d by a Robinsonian dissimilarity belongs to the set $\Delta = \{|d(x,y) - d(x',y')|/2 : x, y, x', y' \in X\}$.*

Proof. Let d^* be an optimal Robinsonian dissimilarity for d and let $\prec$ be a total order compatible with d^*. From Proposition 4.3 we obtain

$$\epsilon^* = ||d - d^*||_\infty = ||d - \tilde{d}_\prec||_\infty = \frac{1}{2}||d - \check{d}_\prec||_\infty = \frac{1}{2}\max_{x,y \in X}|d(x,y) - \check{d}_\prec(x,y)|.$$

By Lemma 4.2, $\check{d}_\prec(x,y) = d(u,v)$ for some $u, v \in X$ such that $x \prec u \prec v \prec y$, whence ϵ^* has the form $\frac{1}{2}|d(x,y) - d(u,v)|$. This proves that $\epsilon^* \in \Delta$. □

5 NP-hardness

In this section, we establish the NP-hardness of the problems investigated in this paper.

Theorem 5.1 *The problem* ϵ-ROBINSON *is NP-complete. The problem* l_∞-FITTING-BY-ROBINSON *is NP-hard to approximate within* $3/2 - \delta$ *for any* $\delta > 0$, *unless P=NP.*

Proof. We will use a polynomial transformation from the NP-complete problem NOT-ALL-EQUAL 3-SAT. In this respect, our reduction follows the main lines of the proof by Saxe [31] and Håstad et al. [22] that the problem MATRIX-TO-LINE is NP-hard. Nevertheless, the technical details of both proofs are quite different. In particular, the distance matrices derived from the same 3-SAT formulae are different. Also, to establish the non-approximability result, Håstad et al. [22] use a linear program, while for our problem a simple reasoning suffices. Let X be a set of variables and let C be a collection of clauses over X, such that each clause $c \in C$ has three literals. Then (X, C) belongs to NOT-ALL-EQUAL 3-SAT (NAE 3-SAT) if there is a true assignment that for each clause $c \in C$ assigns at least one literal of c the value true and at least one literal of c the value false. Deciding NAE 3-SAT was shown to be NP-complete by Schaefer [32].

Given an NAE 3-SAT instance (X, C), we define a corresponding ϵ-ROBINSON and l_∞-FITTING-BY-ROBINSON instance (P, d) in the following way. As in [22], for each variable $x \in X$ and its complement $\bar{x}$ we define two points p_x and $p_{\bar{x}}$ of P and for each clause $c \in C$ we define three points c_1, c_2, and c_3 of P. Additionally to these points, P contains two other points t and f. Thus $P = \{p_x, p_{\bar{x}}\}_{x \in X} \cup \{c_1, c_2, c_3\}_{c \in C} \cup \{t, f\}$. The dissimilarity d takes four distinct values $0, 3, 6$, and 9, and has the following entries. First, let $d(t,f) = 6$. For each variable $x \in X$, set $d(p_x, p_{\bar{x}}) = 9, d(p_x, t) = d(p_x, f) = d(p_{\bar{x}}, t) = d(p_{\bar{x}}, f) = 0$. For each clause $c = (u \vee v \vee w)$ of C, we set

$d(c_1, p_u) = d(c_1, p_{\bar{v}}) = 6$, $d(c_1, p_v) = d(c_1, p_{\bar{u}}) = 0$,
$d(c_2, p_v) = d(c_2, p_{\bar{w}}) = 6$, $d(c_2, p_w) = d(c_2, p_{\bar{v}}) = 0$,
$d(c_3, p_w) = d(c_3, p_{\bar{u}}) = 6$, $d(c_3, p_u) = d(c_3, p_{\bar{w}}) = 0$,
$d(c_1, c_2) = d(c_2, c_3) = d(c_1, c_3) = 9$.
All remaining distances are set to 3. We assert that d is 3-Robinson if and only of (X, C) belongs to NOT ALL EQUAL 3-SAT. We also show that a polynomial approximation algorithm for l_∞-FITTING-BY-ROBINSON having factor $< 3/2$ can be used to decide if (X, C) belongs to NAE 3-SAT.

Define a partition of all total orders $\prec$ on P into two sets $\mathcal{O}_1$ and $\mathcal{O}_2$. In $\mathcal{O}_1$ we include all total orders $\prec$ such that both t, f are located between p_x and $p_{\bar{x}}$ for all $x \in X$. In $\mathcal{O}_2$ we include the remaining total orders on P, i.e., for which there exists $x \in X$ such that either t or f is located outside $[p_x, p_{\bar{x}}]$. Notice that for any order $\prec$ from $\mathcal{O}_1$, we have $\tilde{\epsilon}_\prec \geq 3$. Indeed, pick $x \in X$. The distances from t and f to p_x and $p_{\bar{x}}$ are all equal to 0. Since $d(t, f) = 6$ and the points t and f are both located between p_x and $p_{\bar{x}}$, from Proposition 4.3 we deduce that $\tilde{\epsilon}_\prec \geq 3$. On the other hand, for any order $\prec$ from $\mathcal{O}_2$ we have $\tilde{\epsilon}_\prec \geq 9/2$. Indeed, we can find a variable x such that one of the points t, f is outside $[p_x, p_{\bar{x}}]$, say $p_x \prec p_{\bar{x}} \prec t$. Since $d(p_x, p_{\bar{x}}) = 9$ and $d(p_x, t) = 0$, Proposition 4.3 yields $\tilde{\epsilon}_\prec \geq 9/2$. This shows that $\epsilon^* \geq 3$.

Claim 1: *If (X, C) belongs to* NAE 3-SAT*, then for the instance (P, d) of the related l_∞-*FITTING-BY-ROBINSON *problem we have $\epsilon^* = 3$; in particular d is 3-Robinson.*

Proof. Let A be a NAE-satisfying assignment for (X, C) (for compactness, we let A be a (0,1)-assignment rather than a (false,true)-assignment). We define three disjoint subsets Q_1, Q_2, Q_3 of P. Pick a clause $c = (u \vee v \vee w) \in C$. We insert the point c_1 in Q_1 if $A(u) = 1, A(v) = 0$, in Q_2 if $A(u) = A(v)$, and in Q_3 if $A(u) = 0, A(v) = 1$. We insert the point c_2 in Q_1 if $A(v) = 1, A(w) = 0$, in Q_2 if $A(v) = A(w)$, and in Q_3 if $A(v) = 0, A(w) = 1$. Finally, we insert the point c_3 in Q_1 if $A(w) = 1, A(u) = 0$, in Q_2 if $A(w) = A(u)$, and in Q_3 if $A(w) = 0, A(u) = 1$. Obviously, every point c_i is assigned to a unique set Q_j, thus the sets Q_1, Q_2, and Q_3 are disjoint. Define also the sets P_f and P_t : for a variable $x \in X$, we include p_x in P_f if $A(x) = 0$ and in P_t if $A(x) = 1$. Analogously, we include $p_{\bar{x}}$ in P_f if $A(x) = 1$ and in P_t if $A(x) = 0$. Notice that $p_x \in P_f$ if and only if $p_{\bar{x}} \in P_t$ and $p_x \in P_t$ if and only if $p_{\bar{x}} \in P_f$. Thus the sets Q_1, Q_2, Q_3, P_f, and P_t define a partition of the set $P \setminus \{f, t\}$.

Now, we will define a Robinson dissimilarity d' on P which is compatible with any linear extension of the order $Q_1 \prec P_f \prec f \prec Q_2 \prec t \prec P_t \prec Q_3$. We will also show that $||d - d'||_\infty = 3$, establishing thus that $\epsilon^* = 3$ and that d is 3-Robinson. The dissimilarity d' takes three distinct values $0, 3$, and 6. Inside of each of five blocks Q_1, Q_2, Q_3, P_f, P_t, the dissimilarity d' is 0. Moreover, d' is 0 on each of the sets $Q_1 \cup P_f \cup \{f\}$ and $\{t\} \cup P_t \cup Q_3$. The distance between two points from different sets Q_i and Q_j or P_f and P_t is 6. The value 6 is also taken by all distances between two points of P_f and Q_3, and of P_t and Q_1. Finally the distance from f to any point of Q_3 and from t to any point of Q_1 is 6. All remaining distances are equal to 3. Namely, the distance from any point of Q_2 to any point of $P_f \cup P_t \cup \{f, t\}$ is 3, as well as the distances from f to all points of $P_t \cup \{t\}$ and the distances from t to all points of $P_f \cup \{f\}$. Summarizing, we obtain the following distance matrix for d':

	Q_1	P_f	f	Q_2	t	P_t	Q_3
Q_1	0	0	0	6	6	6	6
P_f		0	0	3	3	6	6
f			0	3	3	3	6
Q_2				0	3	3	6
t					0	0	0
P_t						0	0
Q_3							0

It can be easily seen that the resulting dissimilarity d' is compatible with the total order $Q_1 \prec P_f \prec f \prec Q_2 \prec t \prec P_t \prec Q_3$ and d' remains compatible with respect to any linear extension of this order to P. Thus, d' is Robinson.

It remains to show that $||d - d'||_\infty = 3$. First notice that $d(t, f) = 6$ and $d'(t, f) = 3$. Next, the d-distance from f or t to any point of $P_f \cup P_t$ is 0, while the similar d'-distance is either 0 or 3. Analogously, the d-distance from any point of $Q_1 \cup Q_2 \cup Q_3$ to t or f is equal to 3, while the similar d'-distance takes one of the values $0, 3$, or 6. Now, $d(p_x, p_{\bar{x}}) = 9$ and $d'(p_x, p_{\bar{x}}) = 6$ because p_x and $p_{\bar{x}}$ belong to distinct sets P_f and P_t. On the other hand, the d-distance between any other points of $P_f \cup P_t$ is by definition 3, while the d'-distance is either 0 or 6. Notice also that the d-distance between two points of $Q_1 \cup Q_2 \cup Q_3$ is 9 unless these points come from different clauses, in which case this distance is 3. All points from the same set Q_i come from different clauses: for sets Q_1 and Q_3 this follows from the definitions of these sets and for the set Q_2 this is so because A is a NAE-satisfying assignment. Thus the d-distance between two points of the same set Q_i is 3. Now, the d'-distance between such points is 0. All other d'-distances between the points of $Q_1 \cup Q_2 \cup Q_3$ are equal to 6, showing that $|d(c, c') - d'(c, c')| \leq 3$ if $c, c' \in Q_1 \cup Q_2 \cup Q_3$. To complete the proof, it remain to compare the d- and d'-distances between a point c_i of $Q_1 \cup Q_2 \cup Q_3$ corresponding to a literal of some clause c and a point p of $P_f \cup P_t$ corresponding to some variable u. If neither u nor $\bar{u}$ do not belong to c, then $d(c_i, p) = 3$, while $d'(c_i, p)$ takes one of the values $0, 3$, or 6. Now, assume without loss of generality that u is a literal of c, namely its first literal (the case when $\bar{u}$ is a literal of c is analogous). Then $p = p_u$. By definition of d we conclude that $d(p_u, c_1) = 6, d(p_u, c_2) = 3$, and $d(p_u, c_3) = 0$. Since $d(p_u, c_2) = 3$ independently of the position of p_u and c_2, we deduce that $|d(p_u, c_2) - d'(p_u, c_2)| \leq 3$. If $p_u \in P_f$, then $A(u) = 0$, therefore c_1 belongs to $Q_2 \cup Q_3$. If $c_1 \in Q_2$, then $d'(p_u, c_1) = 3$ and if $c_1 \in Q_3$, then $d'(p_u, c_1) = 6$. On the other hand, since $A(u) = 0$, the point c_3 does not belong to Q_3. If $c_3 \in Q_1$, then $d'(p_u, c_3) = 0$ and if $c_3 \in Q_2$, then $d'(p_u, c_3) = 3$. Since $d(p_u, c_3) = 0$, we obtain the required inequality $|d(p_u, c_3) - d'(p_u, c_3)| \leq 3$. This settles the case $p_u \in P_f$. The case $p_u \in P_t$ is completely analogous: since $A(u) = 1$, the point c_1 is located in $Q_1 \cup Q_2$, while the point c_3 is located in $Q_2 \cup Q_3$. Thus $d'(p_u, c_1)$ is 6 or 3, while $d'(p_u, c_3)$ is 3 or 0. Since $d(p_u, c_1) = 6$ and $d(p_u, c_3) = 0$, we are done. This shows that $||d - d'||_\infty = 3$. Since d' is Robinson, we deduce that $\epsilon^* = 3$, thus the d is 3-Robinson. This establishes Claim 1.

Claim 2: *If* (X, C) *does not belong to* NAE 3-SAT*, then for the instance* (P, d) *of the related* l_∞-FITTING-BY-ROBINSON *problem we have* $\epsilon^* \geq 9/2$*; in particular* d *is not* 3*-Robinson.*

Proof. Pick a total order $\prec$ on P. If $\prec$ belongs to $\mathcal{O}_2$, then, as we have noticed above, $\tilde{\epsilon}_\prec \geq 9/2$. Now, suppose that $\prec$ belongs to the set $\mathcal{O}_1$. Additionally, suppose without loss of generality that $f \prec t$. Since both f and t are located between p_x and $p_{\bar{x}}$ for all $x \in X$, with $\prec$ we can associate an assignment $A_\prec$ defined in the following way: if $p_x \prec f$, then set $A_\prec(x) = 0$ and if $p_{\bar{x}} \prec f$, then set $A_\prec(x) = 1$. Notice that for at least one clause $c = (u \vee v \vee w)$, the points p_u, p_v, p_w must be located on one side of f and t, either all to the left of f or all to the right of t, and all points $p_{\bar{u}}, p_{\bar{v}}, p_{\bar{w}}$ to another side of f and t. Indeed, if for each clause $c = (u \vee v \vee w)$, say p_u is to the left of f and p_v is to the right of t, then $A_\prec$ is a NAE-satisfying assignment for all c, and therefore for C. Since (X, C) does not belong to NAE 3-SAT, a clause $c = (u \vee v \vee w)$ with required property necessarily exists. Let c_1, c_2, c_3 be the points defined for c. Suppose without loss of generality that $p_{\bar{u}}, p_{\bar{v}}, p_{\bar{w}}$ are to the left of f and p_u, p_v, p_w are to the right of t.

We assert that for the restriction of d on the set $p_{\bar{u}}, p_{\bar{v}}, p_{\bar{w}}, p_u, p_v, p_w, f, t, c_1, c_2, c_3$ of 11 points we have $\tilde{\epsilon}_\prec \geq 9/2$. Suppose by way of contradiction that $\tilde{\epsilon}_\prec < 9/2$. Since the entries of the distance matrix take only the values $0, 3, 6$, and 9, from Lemma 4.2 we infer that a necessary and sufficient condition to have this inequality is to not locate two points at distance 9 between two points at distance 0. In particular, since any point c_i is at distance 0 to one point of the triplet $p_{\bar{u}}, p_{\bar{v}}, p_{\bar{w}}$ and c_i is at distance 9 to any other point c_j, we cannot locate c_j between c_i and the triplet $p_{\bar{u}}, p_{\bar{v}}, p_{\bar{w}}$. This shows that we cannot locate two of the points c_1, c_2, c_3 to the right of the points $p_{\bar{u}}, p_{\bar{v}}$, and $p_{\bar{w}}$. Analogously, we cannot locate two of the points c_1, c_2, c_3 to the left of the points p_u, p_v, and p_w. But then, no admissible location of all three points c_1, c_2, c_3 exists. This concludes the proof of Claim 2.

From Claims 1 and 2 we conclude that (X, C) belongs to NAE 3-SAT if and only if the dissimilarity d is 3-Robinson on P, thus the decision problem ϵ-ROBINSON is NP-complete. From Claims 1 and 2 we also infer that if (X, C) belongs to NAE 3-SAT, then the optimal solution for problem l_∞-FITTING-BY-ROBINSON has value $\epsilon^* = 3$, otherwise, if (X, C) does not belong to NAE 3-SAT then $\epsilon^* \geq 9/2$. Suppose that l_∞-FITTING-BY-ROBINSON admits a polynomial approximation algorithm with factor $< 3/2$. Running this algorithm on instances (P, d) derived from (X, C), for "yes"-instances it will return a solution of value $< 3 \times (3/2) = 9/2$ and on "no"-instances it will return solutions of value $> 9/2$. Therefore this algorithm will decide in polynomial type if (X, C) belongs to NAE 3-SAT. This is impossible, unless P=NP. This concludes the proof of the theorem. □

Remark: Consider a dissimilarity d' obtained from the dissimilarity d occurring in the proof of Theorem 5.1 by adding a positive constant α, i.e., $d' = d + \alpha \cdot d_0$ (recall that d_0 is the standard distance of the complete graph). Notice that if $\alpha = 9$, then the resulting dissimilarity d' is a metric, i.e., d' satisfies the triangle condition. On the other hand, it

is well-known [20, 26] that for any dissimilarity d (in particular, for our d) one can find a sufficiently large constant α such that $d' = d + \alpha \cdot d_0$ is of Euclidean type, of l_1-type, a k-hypermetric, a hypermetric, or a metric of negative type (for definitions see [14, 15]). From Lemma 4.5 we conclude that the optimal values for the problem l_∞-FITTING-BY-ROBINSON on instances d and d' are the same. Therefore d is 3-Robinson if and and only if d' is 3-Robinson. This shows that the problem ϵ-ROBINSON is NP-complete if the input dissimilarity is a metric of type indicated above. The non-approximability result also holds in all these cases. Finally, notice that the NP-completeness result of Theorem 5.1 can be also extended to any set $\mathcal{E}$ of dissimilarities such that $\varphi(d) \in \mathcal{E}$ for an increasing monotone transform φ.

Open question: Does the problem ϵ-ROBINSON remain NP-complete if the input dissimilarity d is a tree-distance?

6 Robinsonian dissimilarities with interval data

The fact that the decision problem ϵ-ROBINSON is NP-complete shows that the following ROBINSON-WITH-INTERVAL-DATA problem is NP-complete as well: given a finite set of objects X and for each pair $x, y \in X$ an interval $[\underline{d}(x,y), \overline{d}(x,y)]$ of possible values which the distance between x and y may take, we ask if there exists a Robinsonian dissimilarity d_R on X, such that $d_R(x,y) \in [\underline{d}(x,y), \overline{d}(x,y)]$ for all x, y? To see that this problem is NP-complete, consider an instance (X, d, ϵ) of ϵ-ROBINSON and for each pair $x, y \in X$ set $\underline{d}(x,y) = d(x,y) - \epsilon$ and $\overline{d}(x,y) = d(x,y) + \epsilon$ if $x \neq y$ and $\underline{d}(x,y) = 0 = \overline{d}(x,y)$ if $x = y$. Then ϵ-ROBINSON has an admissible solution if and only if there exists a Robinsonian dissimilarity d_R such that $\underline{d} \leq d_R \leq \overline{d}$.

On the other hand, if we consider the problem ROBINSON-WITH-INTERVAL-DATA for a given total order $\prec$, then it becomes polynomial. For this, we either can construct two Robinsonian dissimilarities $\check{d}$ and $\hat{d}$ compatible with $\prec$ and satisfying $\underline{d} \leq \hat{d} \leq \check{d} \leq \overline{d}$ or we conclude that the problem has answer "not". The definition of these dissimilarities is quite similar with the definition of the Robinsonian sub- and super-dominated of a given dissimilarity provided by Lemma 4.2. Given two pairs of points x, y and u, v such that $u \prec x \prec y \prec v$, to define $\hat{d}$ we compute $\hat{d}(x,y)$ before $\hat{d}(u,v)$, while to define $\check{d}$ we compute $\check{d}(u,v)$ before $\check{d}(x,y)$. Namely, let

$$\hat{d}(u,v) = \min\{d(u,v) \in [\underline{d}(u,v), \overline{d}(u,v)] : d(u,v) \geq \max\{\hat{d}(x,y) : u \prec x \prec y \prec v\}\}.$$

$$\check{d}(x,y) = \max\{d(x,y) \in [\underline{d}(x,y), \overline{d}(x,y)] : d(x,y) \leq \min\{\check{d}(u,v) : u \prec x \prec y \prec v\}\}.$$

If for some pair, one of these values does not exist, then the algorithm returns the answer "not". Obviously, both $\hat{d}$ and $\check{d}$ are Robinsonian dissimilarities compatible with $\prec$. Using induction one can easily show that for any admissible solution $d_\prec$ of the problem ROBINSON-WITH-INTERVAL-DATA we have $\hat{d} \leq d_\prec \leq \check{d}$. In other words, if we will replace each interval $[\underline{d}(x,y), \overline{d}(x,y)]$ with the interval $[\hat{d}(x,y), \check{d}(x,y)]$ $(x, y \in X)$, we will not loose any admissible solution of our problem and this is the sharpest reduction which has this property.

There is a strong similarity between our two versions of ROBINSON-WITH-INTERVAL-DATA and the problem of NARROWING OF A BLOCK OF SORTINGS investigated by Bleuzen-Guernalec and Colmerauer [9]. In this problem, given two sets of n intervals each $S_1 = \{[\underline{a}_1, \overline{a}_1], \ldots, [\underline{a}_n, \overline{a}_n]\}$ and $S_2 = \{[\underline{b}_1, \overline{b}_1], \ldots, [\underline{b}_n, \overline{b}_n]\}$ one ask to maximally reduce these intervals to obtain two sets of intervals $S_1' = \{[\hat{a}_1, \check{a}_1], \ldots, [\hat{a}_n, \check{a}_n]\}$ and $S_2' = \{[\hat{b}_1, \check{b}_1], \ldots, [\hat{b}_n, \check{b}_n]\}$ such that whenever there exists a sorted list $x_1 \prec \ldots x_n$ of n numbers and a permutation π on $\{1, \ldots, n\}$ satisfying $x_i \in [\underline{a}_i, \overline{a}_i] \cap [\underline{b}_{\pi(i)}, \overline{b}_{\pi(i)}]$ for all $i = 1, \ldots, n$, we also have $x_i \in [\hat{a}_i, \check{a}_i] \cap [\hat{b}_{\pi(i)}, \check{b}_{\pi(i)}]$. In other words, it is necessary to maximally reduce the $2n$ intervals without loosing any sorted list of n numbers contained in these intervals, where the order between the first set of n intervals is fixed while the order between the second set of n intervals is free. The authors present in [9] an optimal $O(n \log n)$ algorithm for solving this problem.

In analogy with NARROWING OF A BLOCK OF SORTINGS, one can formulate the following optimal narrowing problem for Robinsonian dissimilarities and compatible orders. Given a finite set X, a total order $\prec$ on X, and for each pair $x, y \in X$ two intervals $[\underline{a}(x,y), \overline{a}(x,y)]$ and $[\underline{b}(x,y), \overline{b}(x,y)]$ of possible values which the distance between x and y may take, one ask to maximally reduce these intervals to obtain the intervals $[\hat{a}(x,y), \check{a}(x,y)]$ and $[\hat{b}(x,y), \check{b}(x,y)]$ such that whenever there exists a Robinsonian dissimilarity d_R compatible with the total order $\prec$ and a permutation π on X satisfying $d_R(x,y) \in [\underline{a}(x,y), \overline{a}(x,y)] \cap [\underline{b}(\pi(x), \pi(y)), \overline{b}(\pi(x), \pi(y))]$ for all $x, y \in X$, we also have $d_R(x,y) \in [\hat{a}(x,y), \check{a}(x,y)] \cap [\hat{b}(\pi(x), \pi(y)), \check{b}(\pi(x), \pi(y))]$. From previous discussion we know that the optimal reduction of the intervals $[\underline{a}(x,y), \overline{a}(x,y)]$ $(x, y \in X)$ can be done in polynomial time, while the reduction of the intervals $[\underline{b}(x,y), \overline{b}(x,y)]$ $(x, y \in X)$ is NP-hard, because the problem ROBINSON-WITH-INTERVAL-DATA is NP-complete. This is a strong evidence (but not yet a proof) that the optimal narrowing problem for Robinsonian dissimilarities is NP-hard.

References

[1] R. Agarwala, V. Bafna, M. Farach, B. Narayanan, M. Paterson, and M. Thorup, *On the approximability of numerical taxonomy (fitting distances by tree metrics)*, SIAM Journal on Computing **17** (1999), 1073–1085.

[2] J.E. Atkins, E.G. Boman, and B. Hendrickson, *A spectral algorithm for seriation and the consecutive ones problem,* SIAM Journal on Computing, **28** (1998), 297–310.

[3] M. Bădoiu, *Approximation algorithm for embedding metrics into a two-dimensional space,* in the Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), 2003.

[4] J.-P. Barthelemy and F. Brucker, *NP-hard approximation problems in overlapping clustering*, Journal of Classification **18** (2001), 159–183.

[5] F. Brucker, *Modèles de classification en classes empiétantes*, Thèse, EHESS (2001).

[6] S. Benzer, *The fine structure of the gene,* Scientific American, **206** (1962), 70–84.

[7] M.W. Berry, B. Hendrickson, and P. Raghavan, *Sparse matrix reordering schemes for browsing hypertext,* In J. Renegar, M. Shub, and S. Smale, editors, Lectures in Applied Mathematics, Vol. 32: The Mathematics of Numerical Analysis, American Mathematical Society, 1996.

[8] P. Bertrand and E. Diday, *A visual representation of the compatibility between an order and a dissimilarity index: the pyramids,* Computational Statistics Quarterly, **2** (1985), 31–44.

[9] N. Bleuzen-Guernalec and A. Colmerauer, *Optimal narrowing of a block of sortings in optimal time,* Constraints, **5** (2000), 85–118.

[10] G. Caraux and S. Pinloche, *PermutMatrix: a graphical environment to arrange gene expression profiles in optimal linear order,* Bioinformatics **21**(2005), 1280–1281.

[11] V. Chepoi and B. Fichet, *Recognition of Robinsonian dissimilarities*, Journal of Classification **14** (1997), 311–325.

[12] V. Chepoi and B. Fichet, l_∞*-approximation via subdominants*, Journal of Mathematical Psychology **44** (2000), 600–616.

[13] V. Chepoi, and M. Seston, *Seriation in the presence of errors: an approximation algorithm for fitting Robinson structures to dissimilarity matrices* (submitted).

[14] F. Critchley and B. Fichet, *The partial order by inclusion of the principal classes of dissimilarities on a finite set, and some of their basic properties*, In B. van Cutsen (Ed.) Classification and Dissimilarity Analysis. Leture Notes In Statistics (1994), 5–65.

[15] M. Deza and M. Laurent, *Geometry of Cuts and Metrics,* Springer, Berlin, 1997.

[16] E. Diday, *Orders and overlapping clusters by pyramids*, in Multidimensional Data Analysis, Eds., J. de Leeuw, W. Heiser, J. Meulman, and F. Critchley, Leiden: DSWO (1986), 201–234.

[17] C. Durand and B. Fichet, *One–to–one correspondences in pyramidal representation: a unified approach*, in Classification and Related Methods of Data Analysis, Ed., H.H. Bock, Amsterdam: North–Holland (1988), 85–90.

[18] C. Durand, *Ordre et graphes pseudo-hierachiques: théorie et optimisation algorithmique*, Thèse, Université de Provence (1988).

[19] M. Farach, S. Kannan, et T. Warnow, *A robust model for finding optimal evolutionary trees*, Algorithmica **13** (1995), 155–179.

[20] B. Fichet, *Data analysis: geometric and algebraic structures,* in First World Congress of Bernoulli Society Proceedings, vol.2, Tashkent, USSR, 1986, Eds., Yu.A. Prohorov and V.V. Sazonov, Utrecht: VNU Science Press, 123–132.

[21] D. Halperin, *Musical chronology by seriation,* Computers and the Humanities, **28** (1994), 13–18.

[22] J. Håstad, L. Ivansson, and J. Lagergren, *Fitting points on the real line and its application to RH mapping*, Journal of Algorithms **49** (2003), no. 1, 42–62.

[23] L.J. Hubert, *Some applications of graph theory and related nonmetric techniques to problems of approximate seriation: The case of symmetric proximity measures,* British Journal of Mathematical Statistics and Psychology, **27** (1974), 133–153.

[24] D.G. Kendall, *Incidence matrices, interval graphs and seriation in archaeology,* Pacific Journal of Mathematics, **28** (1969), 565–570.

[25] D.G. Kendall, *Seriation,* in Encyclopedia of Stastistical Sciences, 1982. Edited by S. Kotz and N.L. Johnson, pp. 417–424, New York, NY Wiley-Interscience Vol. 8.

[26] J.C. Lingoes, *S*ome boundary conditions for a monotone analysis of symmetric matrices, Psychometrika, **36** (1971), 195–203.

[27] W. Marquardt, *Advances in archaeological seriation,* Advances in Archaeological Method and Theory 1, 1978, Edited by M. Schiffer, pp. 257–314, Academic Press, New York.

[28] I. Miklos, I. Somodi, and I. Podani, *Rearrangement of ecological data matrices via Markov chain Monte Carlo simulation,* Ecology, **86** (2005), 3398–3410.

[29] B. Mirkin and S. Rodin, *Graphs and Genes*, Springer-Verlag, Berlin, 1984.

[30] W. S. Robinson, *A method for chronologically ordering archaeological deposits*, American Antiquity **16** (1951), 293–301.

[31] J. B. Saxe, *Embeddability of weighted graphs in k-space is strongly NP-hard*, Proceedings of the 17th Allerton Conference on Communications, Control, and Computing, 1979, pp. 480–489.

[32] T. J. Schaefer, *The complexity of satisfiability problems*, STOC '78: Proceedings of the tenth annual ACM symposium on Theory of computing (New York, NY, USA), ACM Press, 1978, pp. 216–226.

[33] C. Semple and M. Steel, *Phylogenetics*, Oxford University Press, 2003.