Journal of Classification 18:159-183 (2001)
DOI: 10.1007/s00357-001-0014-1

NP-hard Approximation Problems in Overlapping Clustering

Jean-Pierre Barthélemy

ENST Bretagne, France

François Brucker

ENST Bretagne, France

Abstract: In this paper we prove that the approximation of a dissimilarity by an indexed pseudo-hierarchy (also called a pyramid) or an indexed quasi-hierarchy (also called an indexed weak hierarchy) is an NP-hard problem for any L_p-norm ($p < \infty$). These problems also correspond to the approximation by a strongly Robinson dissimilarity or by a dissimilarity fulfilling the four-point inequality (Bandelt 1992; Diatta and Fichet 1994). The results are extended to circular strongly Robinson dissimilarities, indexed k-hierarchies (Jardine and Sibson 1971, pp. 65–71), and to proper dissimilarities satisfying the Bertrand and Janowitz $(k + 2)$-point inequality (Bertrand and Janowitz 1999). Unidimensional scaling (linear or circular) is reinterpreted as a clustering problem and its hardness is established, but only for the L_1 norm.

Résumé: Nous montrons dans cet article que l'approximation d'une dissimilarité par une pseudo-hiérarchie indicée (également appelée pyramide) ou une quasi-hiérarchie indicée (également appelée hiérarchie faible indicée), est NP-difficile pour toutes les normes L_p ($p < \infty$). Ces problèmes correspondent également aux approximations par une dissimilarité fortement robinsonienne ou par une dissimilarité vérifiant l'inégalité des quatre points (Bandelt 1992, Diatta et Fichet 1994). Ces résultats sont généralisés aux dissimilarités circulaires fortement robinsoniennes, aux k-hiérarchies indicées (Jardine et Sibson 1971, pp. 65–71) ainsi qu'aux dissimilarité propres satisfaisant l'inégalité des $k + 2$ points de Bertrand et Janowitz (1999). L'approximation par une dissimilarité linéaire (ou circulaire) est réinterprétée comme un problème de classification et sa NP-difficulé est établie, mais seulement pour la norme L_1.

Keywords: Approximation problems; Complexity; Pseudo-hierarchies; k-hierarchies; k-weak hierarchies; Robinson dissimilarities.

The authors thank two anonymous referees and Phipps Arabie for many suggestions, improvements, and corrections.

Authors' Address: J.-P. Barthélemy and F. Brucker, École Nationale Supérieure des Télécommunications de Bretagne, département Intelligence artificielle et Sciences Cognitives, B.P. 832, 29285 Brest (France).

1. Introduction

This paper focuses on clustering methods that sort a set X of objects described by a pairwise dissimilarity d into clusters, thus building a clustering system. Many problems related to the search for cliques in graphs are known to be NP-hard: a clique of maximum size, minimal covering, or partitioning by cliques (*i.e.*, complete subgraphs of a graph but necessarily maximal), exact covering by triangles, and so on (for more detail, consult Garey and Johnson 1979, pp. 190–205). In many clustering models the clusters can be interpreted as families of cliques in graphs. So the above problems may be viewed as the first examples of NP-hard problems in clustering. They are followed by many others that can be sorted into three clustering models: the search for a single cluster (sequential clustering; Hansen, Jaumard, and Mladenovic 1995), partitioning, and the search for indexed hierarchies.

Search for a single cluster $C \subseteq X$ with a fixed size k (Hansen, Jaumard, and Mladenovic 1995). The problem is NP-hard in the following cases:

C has a minimum *diameter*, diam(C), with diam$(C) = \max_{x,y,\in C} d(x,y)$;
C minimizes $\sum_{x,y \in C} d(x,y)$;
C minimizes $\frac{1}{k}\sum_{x,y \in C} d^2(x,y)$.

Search for a partition $\mathcal{P}$ of X with a fixed number q of classes minimizing the objective function $f(\mathcal{P})$ (P. Brucker 1978). The problem is NP-hard for:

$f(\mathcal{P}) = \max_{C \in \mathcal{P}} \text{diam}(C)$ and $q \geq 3$; (this result was also obtained by Hansen and Delattre 1978)
$f(\mathcal{P}) = \sum_{C \in \mathcal{P}} \text{diam}(C)$ and $q \geq 3$;
$f(\mathcal{P}) = \sum_{C \in \mathcal{P}} \sum_{x,y \in C} d(x,y)$ and $q \geq 2$;
$f(\mathcal{P}) = \sum_{C \in \mathcal{P}} \frac{\sum_{x,y \in C} d(x,y)}{|C|}$ and $q \geq 2$;
$f(\mathcal{P}) = \max_{C \in \mathcal{P}} \sum_{x,y \in C} d(x,y)$ and $q \geq 2$.

It is worth noting that for $q = 2$ the first two problems can be solved in polynomial time (see Hansen, Jaumard, and Mladenovic 1995, and Hansen and Delattre 1978 for bibliographical references and a description of some objective functions whose minimization can be performed in polynomial time).

When the number of classes is not fixed, the dissimilarity d has to be replaced by a *cost function* c with possible negative values. In this framework the so-called *clique partitioning problem*

$$\min \sum_{C \in \mathcal{P}} \sum_{x,y \in C} d(x,y)$$

has been shown to be NP-hard (Grötschel and Wakabayashi 1990). The clique partitioning problem is a generalization of the Zahn (1964) problem which concerns the equivalence relations E closest to a symmetric relation S for the symmetric difference distance $\triangle(E,S)$ (see Section 3.2 for the definition). The Zahn problem is also NP-hard (Křivánek and Morávek 1986; Grötschel and Wakabayashi 1990).

Search for an indexed hierarchy. An indexed hierarchy on X is equivalent to an ultrametric on X, and the approximation of a dissimilarity by an ultrametric has been shown to be NP-hard by Křivánek and Morávek (1986) for the L_1 metric. This result was extended by Day (1987) both to the L_2-norm approximation and to additive tree metrics. The NP-hardness of hierarchical tree clustering, with hierarchical trees of a fixed height was established by Křivánek and Morávek (1986). These results extend straightforwardly to any L_p metric (with $p < \infty$), and mainly concern (apart from the exceptions of additive trees) nonoverlapping clustering (either two clusters are disjoint or one is a subset of the other). During the last fifteen years, alternative clustering models allowing overlap have appeared: pyramids, pseudo-hierarchies (Diday 1984; Fichet 1984); weak hierarchies, quasi-hierarchies (Bandelt and Dress 1989; Diatta and Fichet 1994); the Bandelt and Dress ordinal model for overlapping clustering (1993); k-weak hierarchies (Bertrand 1998).

In view of the complexity results in "classical" clustering, a question arises: does the relaxation of the nonoverlapping property reduce the complexity of approximation problems in classification?

This paper provides a negative answer to this question in the particular cases of pyramidal clustering (indexed pseudo-hierarchies) and weak hierarchical clustering (indexed quasi-hierarchies). Our work is divided into three sections. The first describes some basic material on clustering models that will be used throughout the paper. The second focuses on the two main results of the paper: the NP-completeness of the L_p-approximation in pyramidal clustering and weak clustering. The final section gives some extensions of these results (linear clustering, k-weak hierarchies, circularities, among others).

In considering NP-completeness we shall use the terminology of Garey and Johnson (1979, pp. 46–76). In particular, the problems are always stated as decision problems, $D \prec D'$ means that problem D reduces to problem D', and $D \simeq D'$ means that D and D' are polynomially equivalent.

2. Clustering Models

We shall discuss three kinds of clustering models: class models, distance models, and relational models. Bijection theorems between indexed class models, distance models, and indexed nested relational models show that these various points of view are equivalent. Only the bijections between indexed class models and distance models will be used in this paper.

2.1 Class Models

A *clustering system* (C.S.) on a finite set X is the set $\mathcal{K}$ of subsets of X such that:
C_1: $X \in \mathcal{K}$ and $\phi \notin \mathcal{K}$;
C_2: for each $x \in X, \{x\} \in \mathcal{K}$;
C_3: $A \in \mathcal{K}, B \in \mathcal{K}$ and $A \cap B \neq \phi \Rightarrow A \cap B \in \mathcal{K}$.

X is the *ground set* of $\mathcal{K}$; the elements of $\mathcal{K}$ are called the *clusters* of $\mathcal{K}$. The singletons $\{x\}$ and the ground set X are called *trivial clusters*. A cluster A is *minimal* if it is nontrivial and does not contain any other nontrivial cluster. Note that if A and B are minimal, then $|A \cap B| \leq 1$. Two clusters A and B are *compatible* whenever $A \subseteq B$ or $B \subseteq A$. A *chain* $\mathcal{C}$ of $\mathcal{K}$ is a set of pairwise compatible clusters. The *length* of the chain $\mathcal{C}$ is the number of its nontrivial clusters. The *height* of $\mathcal{K}$ is the maximum length of its chains.

We denote by $\mathcal{K}_X^0$ the C.S. with ground set X whose clusters are all trivial. The height of $\mathcal{K}_X^0$ is 0.

A *hierarchy* is a C.S. in which two clusters are either disjoint or compatible. Thus, in a hierarchy noncompatible clusters never overlap. A number of combinatorial models allowing overlapping clusters have been designed in the last fifteen years. There are essentially two ways of constructing them:
(i) Replacing the axiom of hierarchies ($A \cap B \in \{A, B, \phi\}$) by weaker conditions;
(ii) Considering clusters as distinguished subsets of a given structured set.
k-weak hierarchies fall into the first category and pseudo-hierarchies into the second.

A *k-weak hierarchy* is a C.S. $\mathcal{K}$ such that the meet of any $k + 1$ clusters is equal to the meet of k clusters among them. It then satisfies:

H_k: For each family $\mathcal{A}$ of $k + 1$ clusters of $\mathcal{K}$, there exists a sub-family $\mathcal{B}$ of $\mathcal{A}$ such that $|\mathcal{B}| = k$ and $\cap\{A|A \in \mathcal{A}\} = \cap\{B|B \in \mathcal{B}\}$. For instance, H_2 can be written as for each $A,B,C \in \mathcal{K}$, $A \cap B \cap C \in \{A \cap B, A \cap C, B \cap C\}$.

2-weak hierarchies, are called *quasi-hierarchies*, following Diatta and Fichet (1994). Hypergraphs fulfilling the "triangle condition" H_2 were introduced for clustering purposes by Bandelt and Dress (1989) under the name of *weak hierarchies* and by Batbedat (1988, 1989) under the name of *maximum medinclus*.

A *pseudo-hierarchy* on X is a C.S. $\mathcal{K}$ with ground set X for which there exists a linear order L on X such that each cluster of $\mathcal{K}$ is an interval of L. The order L is said to be *compatible* with $\mathcal{K}$. Pseudo-hierarchies were introduced by Diday (1984, 1987) and Fichet (1984, 1986). Diday (1984) introduced the term *pyramid* that remains more popular, but could be reserved for its drawing, just as a dendrogram is for ultrametric distances.

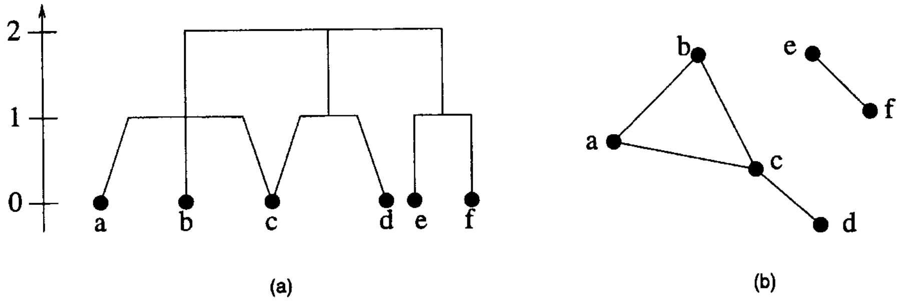


Figure 1. A pyramidal representation of a graphically indexed pseudo-hierarchy (a) and its associated graph (b)

The number of clusters in a quasi-hierarchy on X is $\mathcal{O}(|X|^2)$ (Bandelt and Dress 1989). It is easy to verify that the maximum number of clusters in a pseudo-hierarchy is $\frac{1}{2}|X|(|X|-1)$. It is well known that the number of clusters in a hierarchy is $\mathcal{O}(|X|)$. More generally, we can check that the number of clusters in a k-weak hierarchy is $\mathcal{O}(|X|^k)$.

We denote by $\mathcal{P}_X$,$\mathcal{H}_X$, and $\mathcal{H}_X^k$ the set of pseudo-hierarchies, hierarchies, and k-weak hierarchies on X, respectively. We have $\mathcal{H}_X \subset \mathcal{P}_X \subset \mathcal{H}_X^2 \subset \dots \subset \mathcal{H}_X^k \subset \mathcal{H}_X^{k+1} \subset \dots$

An *index* on a C.S. $\mathcal{K}$ is an integer-valued function f such that $f(\{x\}) = 0$ for each $x \in X$, and $A \subset B$ implies $f(A) < f(B)$. The pair $(\mathcal{K}, f)$ is called an *indexed C.S.* An index f on $\mathcal{K}$ is said to be *graphical* if and only if $f(X) = 2$. In this case we have $f(A) = 1$ for each minimal cluster, and $\mathcal{K}$ is of height 0 (if $\mathcal{K} = \mathcal{K}_X^0$) or 1. When f is graphical, the pair $(\mathcal{K}, f)$ is called a *graphically indexed* C.S.

2.2 Distance Models

A *dissimilarity* on X is a function d from $X \times X$ to the set of nonnegative integers such that: $d(x,y) = d(y,x)$ for $x, y \in X$ and $d(x,x) = 0$ for $x \in X$. The dissimilarity d is said to be *proper* whenever $d(x,y) = 0 \Rightarrow x = y$.

For a dissimilarity d on X we define:

the *diameter* of $A \subseteq X$ as the value $\text{diam}_d(A) = \max\{d(x,y)|x,y \in A\}$;
the *ball* of center x and radius $\rho \in \mathbb{N}$ as the set $B(x,\rho) = \{y|d(x,y) \leq \rho\}$;
the *2-ball* induced by $x, y \in X$ as the set $B_{xy} = B(x, d(x,y)) \cap B(y, d(x,y))$ (Durand and Fichet 1988).

Each indexed C.S. $(\mathcal{K}, f)$ is associated with the proper dissimilarity $\delta[\mathcal{K}, f]$ defined by:

$$\delta[\mathcal{K}, f](x,y) = \min\{f(C)|C \in \mathcal{K}, x \in C, y \in C\}.$$

Distance models for clustering are types of dissimilarities. There are essentially two issues concerning distance models:

(i) Bijection theorems make the search for an indexed C.S. of a given type equivalent to the search for a dissimilarity of a given type (classical bijection theorems between indexed hierarchies and ultrametrics have been established by Jardine, Jardine, and Sibson (1967), Johnson (1967), and Benzcri (1973, pp. 142–144)).

(ii) Usually, relevant data are described by a pairwise dissimilarity measure and – taking (i) into account – a clustering method becomes a transformation of a given dissimilarity into a dissimilarity of a given type.

Let d be a proper dissimilarity on X. We say that d is:

An *ultrametric* if and only if for each $x, y, z \in X$,
$d(x,y) \leq \max\{d(x,z), d(y,z)\}$;

A *strongly Robinson dissimilarity* if and only if there exists a linear order L on X such that:

(R_1) $xLyLz$ implies $\max\{d(x,y), d(y,z)\} \leq d(x,z)$;
(R_2) $xLyLzLt$ and $d(x,z) = d(y,z)$ implies $d(x,t) = d(y,t)$;
(R_2') $xLyLzLt$ and $d(y,t) = d(y,z)$ implies $d(x,z) = d(x,t)$;

A *quasi-ultrametric* if and only if it satisfies:

(Q_1) $z, t \in B_{xy}$ implies $B_{zt} \subseteq B_{xy}$ for all x, y, z, t;
(Q_2) $\mathrm{diam}(B_{xy}) = d(x,y)$ for all x, y.

A proper dissimilarity fulfilling only (R_1) is called a *Robinson dissimilarity* (Robinson 1951), and the order L is said to be *compatible* with d. In his pioneering work, Robinson (1951) considered similarities instead of dissimilarities, with linear order fulfilling the condition dual to (R_1). In that respect the term *anti-Robinson* is sometimes used (see, for instance, Hubert, Arabie, and Meulman 1998).

Conditions (Q_1) and (Q_2) (Diatta and Fichet 1994) are called the *inclusion condition* and the *diameter condition* respectively. They are equivalent to the *four-point inequality* that was independently found by Bandelt (1992) and Diatta and Fichet (1994):

$$\max\{d(z,x), d(z,y)\} \leq d(x,y) \Rightarrow \forall t, d(z,t) \leq \max\{d(t,x), d(t,y), d(x,y)\}.$$

A Robinson dissimilarity is a quasi-ultrametric if and only if it is a strongly Robinson dissimilarity.

We denote by $\mathcal{U}_X$, $\mathcal{R}_X$, and $\mathcal{QU}_X$ the set of all ultrametrics, all strongly Robinson dissimilarities, and all quasi-ultrametrics on X: $\mathcal{U}_X \subset \mathcal{R}_X \subset \mathcal{QU}_X$ respectively.

The well-known bijection theorem between $\mathcal{U}_X$ and the set of all indexed hierarchies on X has been generalized to other class models by Batbedat (1988, 1989, 1990), Bandelt and Dress (1989), Fichet (1986), Diday (1987), Durand and Fichet (1988), Diatta and Fichet (1994), and Bertrand and Janowitz (1999).

We denote by $\mathcal{B}_d$ the set of all the 2-balls of the dissimilarity d.

Theorem 1 (Diatta and Fichet 1994). *If the proper dissimilarity d on X satisfies both the inclusion and the diameter conditions, then $(\mathcal{B}_d, diam_d)$ is an indexed quasi-hierarchy. Moreover, $\mathcal{B}_d$ is a pseudo-hierarchy (respectively a hierarchy) if and only if d is a strong Robinson dissimilarity (resp. an ultrametric).*

Conversely if $(\mathcal{K}, f)$ is an indexed quasi-hierarchy on X, then $\delta = \delta[\mathcal{K}, f]$ is the unique quasi-ultrametric such that $(\mathcal{K}, f) = (\mathcal{B}_\delta, diam_\delta)$.

We denote by φ_X the bijection from the set of all indexed quasi-hierarchies to $\mathcal{QU}_X$ ($\varphi_X[\mathcal{K}, f] = \delta[\mathcal{K}, f]$). Clearly, for an indexed quasi-hierarchy $(\mathcal{K}, f)$ on X (resp. a quasi-ultrametric on X), $\varphi_X[\mathcal{K}, f]$ (resp. $\varphi_X^{-1}(d)$) can be constructed in polynomial time.

2.3 Graph Models and Relational Models

As well as distance models and class models, graph models are useful in classification. They correspond to the case where the nontrivial clusters can be seen as cliques of some (simple, loopless, nonoriented) graphs. The rewriting of a graph model as a relational model sometimes facilitates better formal statements. Standard problems in classification like the *Zahn* problem (1964) or the *Régnier* problem (1965) were initially formulated as relational problems.

A graphical dissimilarity is a proper dissimilarity d such that for $x, y \in X$, $d(x, y) \in \{0, 1, 2\}$. With such a dissimilarity is associated the graph $G_d = (X, E_d)$ with $\{x, y\} \in E_d$ if and only if $d(x, y) = 1$. Conversely, with each graph $G = (X, E)$ is associated the graphical dissimilarity d^G defined for $x \neq y$, $d^G(x, y) = 1$ whenever $\{x, y\} \in E$, $d^G(x, y) = 2$ otherwise.

Lemma 1. *For a graphical dissimilarity d the following assertions are equivalent:*

(i) *d is a graphical quasi-ultrametric.*
(ii) *The 2-balls of d are exactly the maximal cliques of G_d and X.*
(iii) *d satisfies the inclusion condition.*
(iv) *d satisfies the diameter condition.*
(v) *The configuration of Figure 2 is forbidden as a subgraph of G_d.*

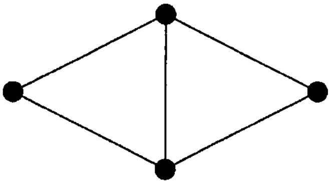


Figure 2. The forbidden configuration

Proof: The four-point inequality is equivalent to **(v)** and **(ii)** if d is a graphical dissimilarity. We then have **(i)** $\Leftrightarrow$ **(ii)** $\Leftrightarrow$ **(v)**. For a graphical dissimilarity, we also have **(iii)** $\Rightarrow$ **(iv)** because if $d(x,y) = 2$, then $B_{xy} = X$ and if $d(x,y) = 1$, the inclusion condition ensures us that $z,t \in B_{xy}$ implies $d(z,t) = 1$. Since **(iv)** obviously implies **(v)** because d is a graphical dissimilarity, we have: **(i)** $\Leftrightarrow$ **(ii)** $\Leftrightarrow$ **(v)** and **(iii)** $\Rightarrow$ **(iv)** $\Rightarrow$ **(v)**.

Noting that **(i)** implies both **(iii)** and **(iv)** concludes the proof □

Lemma 2. *A graphical dissimilarity d is strongly Robinson if and only if the maximal cliques of G_d may be labeled as $C_1, \dots, C_p$ in such a way that $|C_i \cap C_{i+1}| \leq 1$ for $1 \leq i < p$ and $C_i \cap C_j = \phi$ for $1 \leq i < j - 1 < p$.*

Proof: A strongly Robinson dissimilarity d is a quasi-ultrametric with a compatible order L (Diatta 1996). Thus, to avoid the forbidden configuration of Figure 2, the intersection between two maximal cliques of G_d is reduced to at most one element, and the linear order L allows us to label the maximal cliques as stated in Lemma 2.

Conversely, a dissimilarity d such that the maximal cliques of G_d can be labeled as stated above induces a linear order that makes d a strongly Robinson dissimilarity □

A graph fulfilling the condition of Lemma 2 is called a *strong Robinson Graph*, and a family of maximal cliques fulfilling the conditions of Lemma 2 is called a *clique chaining*.

The following condition generalizes the usual transitivity property. We say that G satisfies to (C^k) if and only if G allows no induced subgraphs with $(k+2)$ vertices and $\frac{(k+2)(k+1)}{2} - 1$ edges (*i.e.*, a complete subgraph with $k+2$ vertices minus a single edge).

Let R be a symmetric and reflexive relation. Denote by $G(R)$ the (nonoriented, loopless, and simple) graph associated with R. We say that R is a

Table 1. Summary table

Class Model	Distance Model	Drawing	Graph Model
hierarchy	ultrametric	dendrogram	partitioning into cliques
pseudo-hierarchy	Strong Robinson dissimilarity	Pyramidal representation	Strong Robinson graphs
quasi-hierarchy	quasi-ultrametric		graph satifying (C^2)

k-equivalence (resp. a *strongly Robinson relation*) if and only if $G(R)$ satisfies (C^k) (resp. $G(R)$ is strongly Robinson).

The 1-equivalences are the usual equivalence relations. Jardine and Sibson (1971, pp. 66–69) define the k-transitivity for a relation R such that if $S \subseteq X, |S| = k, a, b \in X$, then $[\{a\} \times S \cup S \times S \cup S \times \{b\}] \subseteq R \Rightarrow (a, b) \in R$. The k-equivalence relations are then the symmetric and reflexive k-transitive relation.

Table 1 summarizes the various types of models studied in this section.

3. The NP-hardness of the Approximation by a Strongly Robinson Dissimilarity or a Quasi-ultrametric

3.1 Preliminary Remark

Let d and d' be two dissimilarities, and set

$$(||d - d'||_p)^p = \sum_{x<y} |d(x, y) - d'(x, y)|^p.$$

(In this and subsequent expressions, the subscript p denotes the L_p norm, and the superscript p indicates a power.) Consider the problem (L_p-approximation) $\min_{\delta \in \mathcal{D}} (||d - \delta||_p)^p$ where $\mathcal{D}$ is a given set of dissimilarities.

If we relax the integrality condition on the values of the dissimilarities, the L_p approximation problem by a strongly Robinson dissimilarity or a quasi-ultrametric generally has no solution. The reason is that the cones of strongly Robinson dissimilarities and of quasi-ultrametrics are not closed (Diatta 1998).

The following example illustrates this drawback: the dissimilarity d depicted in Table 2 (a) is not a quasi-ultrametric; however, the dissimilarity δ_n of Table 2 (b) is a strong Robinson dissimilarity (with compatible order $xLyLzLt$) and

$$(||d - \delta||_p)^p = \frac{2}{n^p} \cdot$$

3.2 Problems and Basic Reductions

The symmetric difference distance between two binary relations R and S is defined by $\Delta(R, S) = |R \cup S| - |R \cap S|$. If we code R and S by their characteristic vectors, Δ becomes the so-called Hamming distance.

The decision problems associated with relational approximation problems can be stated as follows:

NAME : *k-Zahn.*
INSTANCE : A set X, a symmetric and reflexive relation R on X, an integer q.
QUESTION : Does there exist a k-equivalence relation S on X such that $\Delta(R, S) \leq q$?

NAME : *Robbin* (Strongly Robinson, binary case).
INSTANCE : A set X, a symmetric and reflexive relation R on X, an integer q.
QUESTION : Does there exist a strong Robinson relation S on X such that $\Delta(R, S) \leq q$?

1-Zahn is the standard (1964) *Zahn* problem. If we replace the instance of *k-Zahn* (resp. *Robbin*) by: "an undirected, simple, loopless graph G, an integer q", the question becomes: Is it possible to transform G into a graph fulfilling C^k (resp. into a strongly Robinson graph) by adding or deleting fewer than q edges?

The following four problems are called QH (quasi-hierarchies), PH (pseudo-hierarchies), QU (quasi-ultrametrics) and *Rob* (strongly Robinson dissimilarities). Let p be an integer.

INSTANCE: A set X, a dissimilarity d on X, an integer q.
QUESTIONS: $[QH]_p$ Does there exist a quasi-hierarchy $\mathcal{K}$ on X and an index f on $\mathcal{K}$ such that $(||d - \delta[\mathcal{K}, f]||_p)^p \leq q$?
$[PH]_p$ Does there exist a pseudo-hierarchy $\mathcal{K}$ on X and an index f on $\mathcal{K}$ such that $(||d - \delta[\mathcal{K}, f]||_p)^p \leq q$?
$[QU]_p$ Does there exist a proper quasi-ultrametric δ on X such that $(||d - \delta||_p)^p \leq q$?
$[Rob]_p$ Does there exist a proper strong Robinson dissimilarity δ on X such that $(||d - \delta||_p)^p \leq q$?

Note that we have assumed (except in Section 3.1) that dissimilarities and indices have only integer values.

The *k-Zahn*, *Robbin*, QH, PH, QU, and *Rob* problems are clearly in NP.

Table 2. $\forall_n \delta_n$ is a quasi-ultrametric but d is not.

d	x	y	z	t
x	0	1	1	2
y		0	1	1
z			0	1
t				0

(a)

δ_n	x	y	z	t
x	0	1	$1+\frac{1}{n}$	2
y		0	1	$1+\frac{1}{n}$
z			0	1
t				0

(b)

Lemma 3. *$2-$Zahn $\prec [QU]_p \simeq [QH]_p$, and Robbin $\prec [Rob]_p \simeq [PH]_p$.*

Proof: The polynomial equivalence are consequences of the polynomiality of the constructions involved in the bijection theorem (Theorem 1).

Consider now an instance of both *2-Zahn* and *Robbin* (set X, symmetric and reflexive relation R on X, integer q). Let d_R be the graphical dissimilarity induced by R: $d_R(x,y) = 1$ if and only if xRy, and $x \neq y$. Consider the instance X, d_R, q of both $[QU]_p$ and $[Rob]_p$. Note that if δ is a strongly Robinson dissimilarity (resp. a quasi-ultrametric), then the graphical dissimilarity δ^* defined by, for $x \neq y$: $\delta^*(x,y) = 1$ whenever $\delta(x,y) = 1$ and $\delta(x,y) = 2$ otherwise is strongly Robinson (resp. quasi-ultrametric). Moreover, $(||d_R - \delta^*||_p)^p \leq ||d_R - \delta||_p^p$. Hence the instance (X, d_R, q) of $[Rob]_p$ (resp. of $[QU]_p$) has a solution if and only if (X, R, q) does □

3.3 Robinson Approximations

Theorem 2. *Robbin, $[Rob]_p$, and $[PH]_p$ are NP-complete.*

Proof: The problems $[Rob]_p$ and $[PH]_p$ are clearly in NP.

From Lemma 3 it suffices to show that *Robbin* is NP-complete. This result is obtained with a reduction from the Hamiltonian path problem (Garey and Johnson 1979, pp. 199–200).

NAME : *HAMP* (Hamiltonian path).
INSTANCE : A graph G
QUESTION : Does there exist a Hamiltonian path in G?

Let $G = (X, E)$ be a connected graph. Set $|X| = n$ and $|E| = m$. In the first step we construct a graph $G_1 = (X_1, E_1)$ by replacing each vertex x by a clique whose cardinality is of degree $\delta(x)$ of x, and by replacing an edge between x and y by an edge between y and only one vertex of the clique associated with

x. This procedure can be implemented iteratively by the following algorithm:

```
Begin:
$G_1 \leftarrow G$
$L \leftarrow X$
While $L \neq \{\phi\}$
    /* Take the first vertex $x$ in L. Replace it by a clique $C(x) = \{x^1, \dots, x^{\delta(x)}\}$. If $y_1, \dots, y_{\delta(x)}$ are the neighbors of $x$, make $x^j$ adjacent to $y_j$. */
        $X_1 \leftarrow (X_1 - \{x\}) \cup C(x)$;
        $E_1 \leftarrow (E_1 - \{\{x, y_j\}, 1 \leq j \leq \delta(x)\})$
        $\cup\{\{x^i, x^j\}, 1 \leq i < j \leq \delta(x)\} \cup \{\{x^j, y_j\}, 1 \leq j \leq \delta(x)\}$;
        $L \leftarrow L - \{x\}$.
End.
```

Figure 3 illustrates this construction.

The algorithm above performs in polynomial time ($|X_1| = 2m$, and $|E_1| = m + \sum_{x \in X} \frac{\delta(x)(\delta(x)-1)}{2} = \sum_{x \in X} \frac{\delta(x)^2}{2} = \mathcal{O}(n^3)$).

In a second step, from G_1 we get the graph $G^* = (X^*, E^*)$ by adding, for each $x \in X$, a clique $C'(x)$ with $2n^2 + 2$ vertices and making each vertex of $C'(x)$ adjacent to each vertex of $C(x)$. We observe that $|X^*| = 2m + 2n^3 + 2n = \mathcal{O}(n^3)$. We consider now the instance of *Robbin* with the relation R associated with the graph G^* and $q = m - n + 1$ ($q \geq 0$ because G is connected). Assume that *Robbin* allows a solution in this instance. Let G' be the graph associated with this solution. Then:

(i) For $x \in X$ the cliques $K(x) = C(x) \cup C'(x)$ (labeled such that $K(x) = \{x_1, \dots, x_{2n^2+2}, \dots\}$) are in G', because suppressing an edge $\{x_1, x_2\}$ in such a clique creates at least n^2 of the forbidden configurations depicted in Figure 2 (Lemma 1): $\{x_1, x_2, x_{2i+1}, x_{2i+2}\}$ for $1 \leq i \leq n^2$. We must then delete at least n^2 ($> m - n + 1$) edges to avoid these forbidden configurations.

(ii) Such a solution may be obtained without adding edges. The addition of edges either merges two cliques $K(x)$ or links another element of a clique $K(x)$ and one element of a clique $K(y)$ (Lemma 2). But because for each $x \in X$, $|C'(x)| > n^2 > m - n + 1$, the added edges only link one element of a clique $K(x)$ and one element of a clique $K(y)$, and we have $|K(x) \cap K(y)| \leq 1$ because G' is a strongly Robinson graph (Lemma 2). Deleting the added edges disconnects $K(x)$ and $K(y)$, and the graph remains a strongly Robinson Graph.

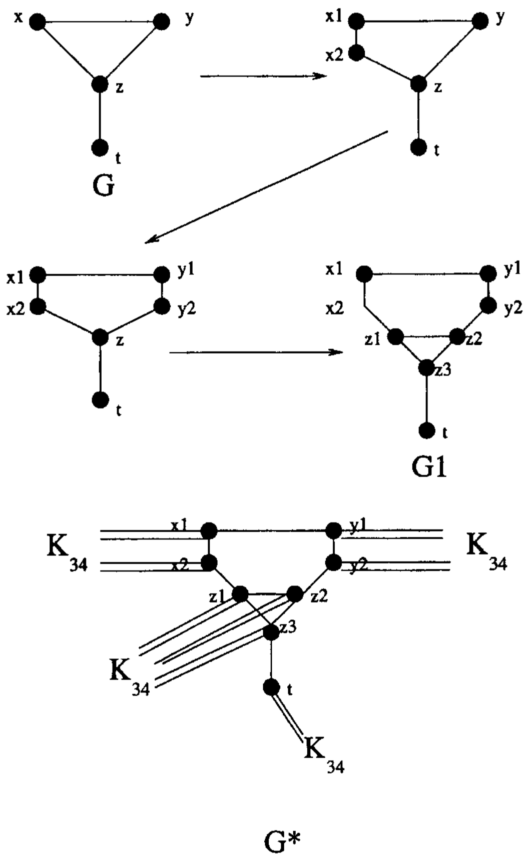


Figure 3. From *G to G**

Assume that G allows a Hamiltonian path $x_1, x_2, \dots, x_n$. Then we get a sequence $K(x_1), \{x_1^i, x_2^j\}, K(x_2), \{x_2^k, x_3^l\}, \dots, K(x_n)$, which is a clique chaining of G^* obtained by deleting $m - n + 1$ edges (the x_i are in X and the x_i^j in $C(x_i)$).

Conversely, assume that the *Robbin* problem with graph instance G^*, and $q = m - n + 1$ admits a solution G'. It follows from the above Remarks (i) and (ii) that only edges $\{x_j^i x_{j'}^{i'}\}$ with $j \neq j'$ have been removed. Hence, the clique chaining corresponding to G' can be written as $K(x_1)$, $\{x_1^i, x_2^j\}$, $K(x_2)$, $\{x_2^k, x_3^l\}$, ..., $K(x_n)$, and the $x_1, x_2, \dots, x_n$ constitute a Hamiltonian path of G □

3.4 Quasi-ultrametric Approximations

Theorem 3. *2-Zahn, $[QU]_p$, and $[QH]_p$ are NP-complete.*

Proof: *2-Zahn*, $[QU]_p$, and $[QH]_p$ are clearly in NP.

From Lemma 3 it suffices to show that *2-Zahn* is NP-complete. This result follows from the result that *1-Zahn* is NP-complete (Křivánek and Morávek 1986), and the following Lemma □

Lemma 4. *k-Zahn $\prec$ $(k+1)$-Zahn.*

Proof: Let $G = (X, E)$ and q be the graph associated with the symmetric relation R and the integer of a k-*Zahn* instance, respectively.

Then let $G' = (X', E')$ be a graph such that:

$$X' = X \cup \{\omega\} \cup \{\alpha_{i,y,z} | 1 \leq i \leq |X|, 1 \leq y \leq q+1, 1 \leq z \leq k+1\};$$
$$E' = E \cup Cl_1 \cup Cl_2 \cup \cdots \cup Cl_{|X|},$$

Where Cl_i is the clique whose vertex set is $\{x_i, \omega, \alpha_{i,y,z} | 1 \leq y \leq q+1, 1 \leq z \leq k+1\}$. Figure 4 shows an example of this transformation for $k = 1$ and $q = 2$.

Because we can assume that $q < \frac{|X|(|X|-1)}{2}$ (otherwise $G^0 = (X, \phi)$ is a solution for any instance G of k-*Zahn*), the transformation from G to G' is performed in polynomial time.

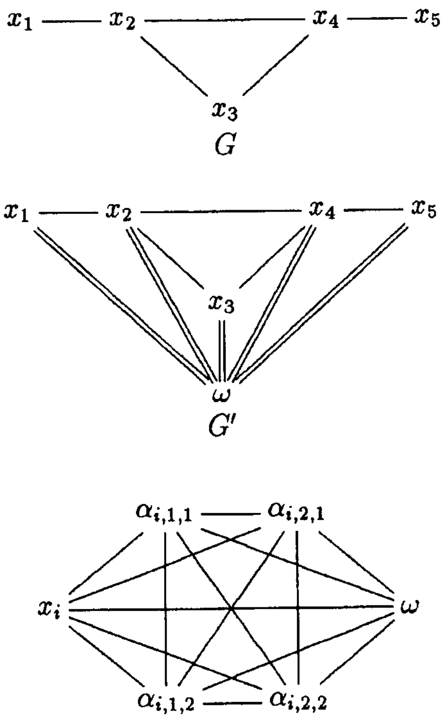
Now consider the instance R' (the symmetric relation associated with G'), q of $(k+1)$-*Zahn*, and a solution S' (associated with the graph $G[S'] = (X', E'_{S'})$) of this problem.

To solve this problem we can suppress at most q edges from G', but to satisfy the (C^{k+1}) condition we cannot delete any edge from a clique Cl_i, because $\forall x_i \in X$, $|Cl_i| = 2 + (q+1) * (k+1)$ (even if by deleting q edges of such a clique, one can always find $k+3$ vertices that violate (C^{k+1})).

Thus for each $x_i \in X$, $\{x_i, \omega\} \in E'_{S'}$. Hence, each subset Y of X with $k+2$ elements satisfies (C^k) because $Y \cup \{\omega\}$ satisfies (C^{k+1}). Therefore, $G[S] = (X, E_S)$, where E_S is $E'_{S'}$ restricted to $X \times X$, is a graph associated with a k-equivalence relation. Moreover, if S is the k-equivalence relation associated with $G[S]$, we have $|R \Delta S| \leq |R \Delta S'| \leq q$. Hence, S is a solution of our instance of k-*Zahn*.

Conversely, a solution S (associated with $G[S] = (X, E_S)$) of our k-*Zahn* instance can be transformed using the method above to a solution S' (associated with $G[S'] = (X', E'_{S'})$) of our $(k+1)$-*Zahn* instance.

Because all these transformations are performed in polynomial time, we finally have: k-*Zahn* $\prec$ $(k+1)$-*Zahn* □



Where $x_i = \omega$ means:

Figure 4. From G to G' with k =1 and q =2

4. Related Problems

4.1 Linear Distances, Circular Distances, Strongly Robinson Circular Dissimilarities

A *linear distance* is a strongly Robinson dissimilarity δ which is additive along a compatible order L. Thus, δ is such that $xLyLz \Rightarrow \delta(x,z) = \delta(x,y) + \delta(y,z)$. Despite the fact that the usual interest in linear distances resides in so-called unidimensional scaling or seriation, they correspond – as a strongly Robinson dissimilarity – to a specially indexed C.S.

Let L be a linear order on X; we assume that $x_1Lx_2L\ldots Lx_n$; a pseudo-hierarchy $\mathcal{K}$ is said to be *L-saturated* if and only if the clusters of $\mathcal{K}$ are exactly the intervals of L. In this case, $\mathcal{K}$ allows only two compatible orders: L and its reverse (for more detail about Robinson dissimilarities allowing only two compatible orders, consult Bertrand 1992). The following proposition specifies the bijection theorem (Theorem 1) for linear distances.

Proposition 1. *Let $(\mathcal{K}, f)$ be an indexed C.S. The following two assertions are equivalent:*

(i) $\delta[\mathcal{K}, f]$ *is a linear distance;*

(ii) $\mathcal{K}$ *is an L-saturated pseudo-hierarchy for some linear order L (we label X such that $x_1 L x_2 L \dots L x_n$), and f is such that for every cluster $[x_i, x_k]$ with $i < k$ we have:*

$$f([x_i, x_k]) = \sum_{i \le j < k} f([x_j, x_{j+1}]).$$

Proof: Assume that $\delta[\mathcal{K}, f]$ is a linear distance, thus requiring a linear order L on X. We have then for $i < j < k$: $\delta[\mathcal{K}, f](x_i, x_k) = \delta[\mathcal{K}, f](x_i, x_j) + \delta[\mathcal{K}, f](x_j, x_k)$. By induction we obtain that for $i < k$: $\delta[\mathcal{K}, f](x_i, x_k) = \sum_{i \le j < k} \delta[\mathcal{K}, f](x_j, x_{j+1})$. Because f is an index, $\delta[\mathcal{K}, f](x_j, x_{j+1}) > 0$ for all $1 \le j < n$, and then each interval $[x_i, x_j]$ is a 2-ball of $\delta[\mathcal{K}, f]$: $[x_i, x_j]$ is a cluster of $\mathcal{K}$, and $\delta[\mathcal{K}, f](x_j, x_{j+1}) = f([x_j, x_{j+1}])$.

Conversely, assume that $\mathcal{K}$ is L-saturated. We have, with $i < j$,

$$\delta[\mathcal{K}, f](x_i, x_j) = f([x_i, x_j]) = \sum_{k=i}^{j-1} f([x_k, x_{k+1}]);$$

hence for $i < k < j$ $(x_i L x_k L x_k)$,

$$\begin{aligned}\delta[\mathcal{K}, f](x_i, x_j) &= \sum_{l=i}^{k-1} f([x_l, x_{l+1}]) + \sum_{l=k}^{j-1} f([x_l, x_{l+1}]) \\ &= \delta[\mathcal{K}, f](x_i, x_k) + \delta[\mathcal{K}, f](x_k, x_j).\end{aligned}$$

$\delta[\mathcal{K}, f]$ is then linear □

Consider the following problem.

NAME: LIN_1 (L_1-approximation by a linear distance).
INSTANCE: A set X, a dissimilarity d with integer values on X, an integer q.
QUESTION: Does there exist a linear distance δ on X such that $||d - \delta||_1 \le q$?

Proposition 2. *LIN_1 is NP-complete.*

Proof: LIN_1 is obviously in NP. Let δ be a linear distance compatible with the order L and $|X| \geq 3$. Consider a graphical dissimilarity d on X. Set $d_{i,j} = d(x_i, x_j)$, $\delta_{i,j} = \delta(x_i, x_j)$ and $s_i = \delta_{i,i+1}$. Then $\delta_{i,j} = \sum_{1 \leq k < j} s_k$.

A linear dissimilarity that minimizes $||\delta - d||_1$ is such that $s_{i^*} = 1$, where $s_{i^*} = \max\{s_i | 1 \leq i < n\}$. To obtain this equality, suppose that $s_{i^*} \geq 2$; we may then define a linear distance δ° with compatible order L such that $\delta^\circ_{i^*,i^*+1} = 1$, and $\delta^\circ_{i,i+1} = \delta_{i,i+1}$ otherwise. We have

$$\begin{aligned} ||d - \delta||_1 = |d_{i^*,i^*+1} - s_{i^*}| &+ \sum_{i<j\leq i^*, i^*<i<j} |d_{i,j} - \delta_{i,j}| \\ &+ \sum_{i \leq i^* < j, (i,j) \neq (i^*, i^*+1)} |d_{i,j} - s_{i^*} - \sum_{i \leq k \neq i^* \leq j-1} s_k|. \end{aligned}$$

Using the fact that d is graphical and δ has integer values, we have

$$\begin{aligned} ||\delta - d||_1 - ||\delta^\circ - d||_1 = s_i^* &+ 1 - 2d_{i^*,i^*+1} \\ &+ \sum_{i \leq i^* < j, (i,j) \neq (i^*, i^*+1)} (s_{i^*} - 1) \geq 0; \end{aligned}$$

hence: $||\delta^\circ - d||_1 \leq ||\delta - d||_1$, and this inequality is strict for $n > 3$.

Thus, we can assume that a solution of $||d - \delta||_1 \leq q$ is such that $s_i = 1$ for $1 \leq i < n$. In this case we get

$$\begin{aligned} ||d - \delta||_1 &= \sum_{i=1}^{n-1} (d_{i,i+1} - 1) + \sum_{i<j-1} ((j - i) - d_{i,j}) \\ &= 2\sum_{i=1}^{n-1} (d_{i,i+1} - 1) + \sum_{i<j} ((j - i) - d_{i,j}). \end{aligned}$$

It is worth noting that the second part of this sum does not depend on the order L and is positive. Minimizing $||d - \delta||_1$ is then equivalent finding L such that $\sum_{i=1}^{n-1} (d_{i,i+1} - 1)$ is minimum.

If $\sum_{i=1}^{n-1} (d_{i,i+1} - 1) = 0$, $d_{i,i+1} = 1$ for all $1 \leq i < n$, and the set $\{\{x_1, x_2\}, \ldots, \{x_{n-1}, x_n\}\}$ is an Hamiltonian path of G_d. It is then equivalent to find an Hamiltonian Path in G_d and a linear distance δ such that $||d-\delta|| \leq \sum_{i<j}((j-i)-d_{i,j})$. Hence, the result, by reduction of the instance G of *HAMP* to the instance $d = d^G$, $q = \sum_{i<j}((j-i)-d_{i,j})$ of LIN_1 □

A *circular distance* δ is the shortest path metric of some weighted cycle, called a *circular order*. The approximation by a circular distance has been studied by Hubert, Arabie, and Meulman (1997). Consider the following problem:

NAME: CIR_1 (L_1 approximation by a circular distance).
INSTANCE: A set X, a dissimilarity d with integer values on X, an integer q.
QUESTION: Does there exist a circular distance δ such that $||d-\delta||_1 \leq q$?

Considering the proof of Proposition 2 and changing the linear order into a circular order and Hamiltonian paths into Hamiltonian cycles, by reduction from *HAM* (which is NP-complete, Garey and Johnson 1979, p. 199):

NAME: *HAM*.
INSTANCE: Graph G.
QUESTION: Is G Hamiltonian?

Hence:

Corollary 1. *CIR_1 is NP-complete.*

Circular Robinson dissimilarities allow a *compatible circular order*. The approximation by circular Robinson dissimilarities has been studied by Hubert, Arabie, and Meulman (1998) who also provided their own characterization, repeated here.

Let δ be a dissimilarity on X, and L a linear order on X and set $\delta_{i,j} = \delta(x_i, x_j)$. Note that δ is a *circular Robinson dissimilarity with compatible order* L if and only if, for $1 \leq i \leq n-2$ and $i+1 < j \leq n-1$:
CR_1 If $\delta_{i+1,j} \leq \delta_{i,j+1}$, then $\delta_{i+1,j} \leq \delta_{i,j}$ and $\delta_{i+1,j} \leq \delta_{i+1,j+1}$;
CR_2 If $\delta_{i+1,j} \geq \delta_{i,j+1}$, then $\delta_{i,j} \geq \delta_{i,j+1}$ and $\delta_{i+1,j+1} \geq \delta_{i,j+1}$;
CR_3 If $\delta_{i+1,n} \leq \delta_{i,1}$, then $\delta_{i+1,n} \leq \delta_{i,n}$ and $\delta_{i+1,n} \leq \delta_{i+1,1}$;
CR_4 If $\delta_{i+1,n} \geq \delta_{i,1}$, then $\delta_{i,n} \geq \delta_{i,1}$ and $\delta_{i+1,1} \geq \delta_{i,1}$.

Moreover, δ is a *circular strongly Robinson dissimilarity* if and only if it fulfills the above four conditions, and for $1 \leq i \leq n-2$, $i+1 < j \leq n-1$:
CSR_1 If $\delta_{i+1,j} \leq \delta_{i,j+1}$, then $\delta_{i+1,j} = \delta_{i,j}$ implies $\delta_{i+1,j+1} = \delta_{i,j+1}$ and $\delta_{i+1,j} = \delta_{i+1,j+1}$ implies $\delta_{i,j} = \delta_{i,j+1}$;
CSR_2 If $\delta_{i+1,j} \geq \delta_{i,j+1}$, then $\delta_{i,j+1} = \delta_{i+1,j+1}$ implies $\delta_{i,j} = \delta_{i+1,j}$ and $\delta_{i,j} = \delta_{i,j+1}$ implies $\delta_{i+1,j} = \delta_{i+1,j+1}$;
CSR_3 If $\delta_{i+1,n} \leq \delta_{i,1}$, then $\delta_{i+1,n} = \delta_{i,n}$ implies $\delta_{i+1,1} = \delta_{i,1}$ and $\delta_{i+1,n} = \delta_{i+1,1}$ implies $\delta_{i,n} = \delta_{i,1}$;

CSR_4 If $\delta_{i+1,n} \geq \delta_{i,1}$, then $\delta_{i,1} = \delta_{i+1,1}$ implies $\delta_{i,n} = \delta_{i+1,n}$ and $\delta_{i,n} = \delta_{i,1}$ implies $\delta_{i+1,n} = \delta_{i+1,1}$.

Lemma 5. *A graphical dissimilarity d is circular strongly Robinson if and only if the maximal cliques of G_d can be labeled as $C_1, \ldots, C_p$ in such a way that $|C_i \cap C_{i+1}| \leq 1$ for $1 \leq i \leq n$, $|C_p \cap C_1| \leq 1$, and $C_i \cap C_j = \phi$ otherwise.*

Proof: As for Lemma 2, this statement is just the reformulation of the definition taking into account that we have a graphical dissimilarity□

By applying both Lemma 5 and the proof of Theorem 2 with some changes: linear orders become circular orders, strongly Robinson dissimilarities become circular strongly dissimilarities, Hamilton paths become Hamilton cycles, and we therefore get the following result:

Corollary 2. *For any integer p, the following problem is NP-complete:*
NAME: *$[CRob]_p$ (circular strongly Robinson dissimilarity).*
INSTANCE: *A set X, a dissimilarity d with integer values on X, an integer q.*
QUESTION: *Does there exist a circular strongly Robinson dissimilarity δ with integer values such that $(||d - \delta||_p)^p \leq q$?*

4.2 The Case of k-weak Hierarchies

Theorem 3 does not extend *a priori* to k-weak hierarchies with $k \geq 3$ (remember that 2-weak hierarchies are the so-called quasi-hierarchies). The reason for this observation is that a set of two-element subsets of X complemented by X, together with the singletons constitute a 3-weak hierarchy. Thus, if d is any dissimilarity on X, we get an indexed 3-weak hierarchy whose nontrivial clusters are all the pairs $\{x, y\}$ indexed by $d(x, y)$.

The main problem here is that the clusters cannot be interpreted as maximal cliques of graphs. To avoid this drawback, Bertrand (1998) proposed the notion of a preindexed C.S. fulfilling the following condition (G). A *preindexed* C.S. is a pair $(\mathcal{K}, f)$ of a C.S., together with a function f from $\mathcal{K}$ to the set of integers such that $\forall\{x\}, f(\{x\}) = 0$, and $A \subseteq B$ implies $f(A) \leq f(B)$, and condition (G) can be formulated as follows:
(G): for all $C_1, C_2, C_3 \in \mathcal{K}$ there exists $B \in \mathcal{K}$ such that
$f(B) \leq \max\{f(C_1), f(C_2), f(C_3)\}$, and
$(C_1 \cap C_2) \cup (C_2 \cap C_3) \cup (C_1 \cap C_3) \subseteq B$.

Clearly, the index on the pairs of elements of X induced by a dissimilarity d does not satisfy the condition (G). Another useful condition is called (I_k) by Bertrand and Janowitz (1999): the preindexed C.S. $(\mathcal{K}, f)$ satisfies the condition (I_k) if and only if for all $A, B \in \mathcal{K}$ with $|A| \geq k$ and $A \subset B$ implies $f(A) < f(B)$.

An indexed C.S. just satisfies Condition (I_k) for $k = 1$.

We say that a dissimilarity d (proper or not) satisfies the k-point inequality (Bertrand and Janowitz 1999) if and only if for each $u \in X$ and each subset A of X, with $|A| = k - 2$, we have: $\max\{d(u,x)|x \in A\} \leq \mathrm{diam}_d(A)$ implies that for all $v \in X, d(u,v) \leq \mathrm{diam}_d(A \cup \{v\})$.

It is easy to verify that the four-point inequality (Section 2.2) corresponds to $k = 4$.

Let d be a dissimilarity on X and σ an integer. The *graph of d at the threshold* σ has X as vertex set and $\{x, y\}$ is an edge if and only if $d(x, y) \leq \sigma$. With each dissimilarity (proper or not) on X is associated a C.S. $\mathcal{K}[d]$ whose clusters are defined inductively as follows:

the singleton $\{x\}$, with $x \in X$ is a cluster;

for each σ: $0 \leq \sigma \leq \mathrm{diam}_d(X)$ the maximal cliques of the graph of d at the threshold σ are clusters;

the nonempty intersection of two clusters is a cluster.

The map $\psi_X : d \to (\mathcal{K}[d], \mathrm{diam}_d)$ extends the map φ_X^{-1} of Section 2.2 and the following result extends Theorem 1.

Theorem 4 (Bertrand and Janowitz 1999). *The mapping ψ_X defines a bijection from the set of dissimilarities satisfying the $(k + 2)$-point inequality onto the set of preindexed k-weak hierarchies fulfilling conditions (G) and (I_k).*

Note that $\psi_X(d)$ and $\psi_X^{-1}(\mathcal{K}, f)$ can be constructed in polynomial time.

We say that a preindex f on the k-weak hierarchy is *proper* if and only if it satisfies (G), (I_k), and is such that $\delta[\mathcal{K}, f](x, y) > 0$ for $x \neq y$. Thus, for $x, y \in X$ and $C \in \mathcal{K}$ such that $\{x, y\} \subseteq C$, we have $f(C) > 0$.

By restriction, ψ_X defines a bijection between the set of proper dissimilarities satisfying the $(k + 2)$-point inequality on the set of k-weak hierarchies with proper preindexes.

Note that the proper preindexes on a quasi-hierarchy $\mathcal{K}$ $(k = 2)$ are exactly the indices on $\mathcal{K}$.

Lemma 6 characterizes the graphical dissimilarities satisfying the $(k+2)$-point inequality. Note that the symmetric relations induced by these dissimilarities are exactly the k-equivalence relations.

Lemma 6. *Let d be a graphical dissimilarity; then d is a k-weak hierarchy if and only if $G(d)$ satisfies the condition (C^k) (Section 2.3).*

Proof: As for Lemma 1, it is the reformulation of the $(k + 2)$-point inequality for a graphical dissimilarity □

We can now consider the following problems:

NAME: $[(k+2)-PI]_p$ ($k+2$ points inequality in norm L_p).
INSTANCE: A set X, a dissimilarity d on X, an integer q.
QUESTION: Does there exist a proper dissimilarity δ fulfilling the $(k+2)$-point condition such that $(||d-\delta||_p)^p \leq q$?

NAME: $[k-H]_p$ (L_p approximation by a k-weak hierarchy).
INSTANCE: A set X, a dissimilarity d on X, an integer q.
QUESTION: Does there exist a k-weak hierarchy $\mathcal{K}$ and a proper preindex f on $\mathcal{K}$ such that $(||d-\delta[\mathcal{K},f]||_p)^p \leq q$?

Theorem 5. *Problems $[(k+2)-PI]_p$ and $[k-H]_p$ are NP-complete.*

Proof: First we note that $[(k+2)-PI]_p$ and $[k-H]_p$ are polynomially equivalent and in NP. We shall show that *k-Zahn* $\prec [(k+2)-PI]_p$. This demonstration will conclude the proof because *2-Zahn* is NP-complete (Theorem 3), and $(k-1)$-*Zahn* $\prec$ *k-Zahn* (Lemma 4) makes *k-Zahn* NP-complete. Adapting the arguments of Lemma 3 and Theorem 3 and using Lemma 4, the proof comes from the fact that if δ is a proper dissimilarity on X, with integer values, and fulfilling the $(k+2)$-point inequality, the dissimilarity δ_0 on X defined by:

$$\delta_0(x,y) = \delta(x,y) \text{ if } \delta(x,y) \leq 1, \text{ and}$$

$$\delta_0(x,y) = 2 \text{ if } \delta(x,y) > 1,$$

satisfies the $(k+2)$-point inequality. Moreover, $(||d-\delta_0||_p)^p \leq (||d-\delta||_p)^p$ if d is a proper graphical dissimilarity on X.

Hence, the result is obtained by reduction from *k-Zahn*, because $G(\delta_0)$ satisfies (G^k) (Lemma 6)□

4.3 The Jardine and Sibson k-ultrametrics

Jardine and Sibson (1971, pp. 65–71) proposed another generalization to ultrametric distances, named k-ultrametrics:

A dissimilarity d on X is k-ultrametric if whenever $S \subseteq X$; $|S| = k$; $a,b \in X$; then $d(a,b) \leq \max\{d(x,y)|x \in S \cup \{a,b\}, y \in S\}$.

This inequality means that in every $k+2$ elements subset Y of X, the two greatest dissimilarities between the elements of Y are always equal ($k=1$ corresponds to the usual ultrametrics).

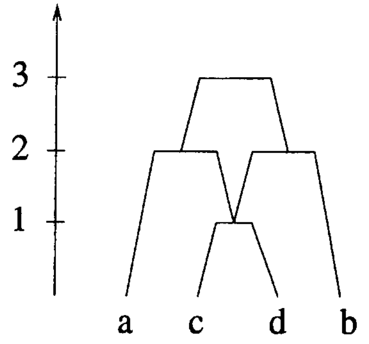


Figure 5. A quasi-ultrametric which is not a 2-ultrametric

Lemma 7 below shows that for graphical dissimilarities the k-ultrametrics and the dissimilarities satisfying the $(k+2)$-point inequality coincide. In general, k-ultrametrics are only a subset of dissimilarities satisfying the $(k+2)$-point inequality. Figure 5 shows an example: a quasi-ultrametric (which is also a strongly Robinson dissimilarity) which is not a 2-ultrametric because $d(a,b) > \max\{d(x,y)|x \in \{c,d\} \cup \{a,b\}, y \in \{c,d\}\}$.

Lemma 7. *Let d be a graphical dissimilarity; then d is a k-ultrametric if and only if $G(d)$ satisfies the condition (C^k) defined at Section 2.3 .*

Proof: As with Lemma 1, this is the reformulation of the definition for a graphical dissimilarity □

We can then consider the following problem which is a special case of $[(k+2) - PI]_p$.

NAME: $[(k - U]_p$ (k-ultrametrics in norm L_p).
INSTANCE: A set X, a dissimilarity d on X, an integer q.
QUESTION: Does there exist a proper k-ultrametric δ such that $(||d - \delta||_p)^p \leq q$?

Theorem 6. *Problem $[k - U]_p$ is NP-complete.*

Proof: $[k - U]_p$ is clearly in NP. As with Theorem 5 we shall show that *k-Zahn* $\prec [k - U]_p$. We adapt the arguments of Lemma 3 and Theorem 3 and use Lemma 4. If d is a proper graphical dissimilarity on X, and if δ is a proper k-ultrametric on X, with integer values, then the dissimilarity δ_0 defined by $\delta_0(x,y) = \delta(x,y)$ if $\delta(x,y) \leq 1$ and $\delta_0(x,y) = 2$ otherwise, is a graphical k-ultrametric.

Moreover, $(||d-\delta_0||_p)^p \leq (||d-\delta||_p)^p$. Thus, the result is obtained by reduction from *k-Zahn*, because $G(\delta_0)$ satisfies (G^k) (Lemma 7)□

4.4 Restriction on Height

Let h be an integer. Consider the problems $h-[k-H]_p$ and $h-[PH]_p$ whose common instance is a set X, a dissimilarity d on X, an integer q, and whose questions are:

$h-[k-H]_p$ Does there exist a k-weak hierarchy $\mathcal{K}$ with height h and a proper preindex f on $\mathcal{K}$ such that $(||d-\delta[\mathcal{K},f]||_p)^p \leq q$?

$h-[PH]_p$ Does there exist a pseudo-hierarchy $\mathcal{K}$ with height h and a proper preindex f on $\mathcal{K}$ such that $(||d-\delta[\mathcal{K},f]||_p)^p \leq q$?

Proposition 3. *$h-[k-H]_p$ and $h-[PH]_p$ are NP-complete*

Proof: $h-[k-H]_p$ and $h-[PH]_p$ are obviously in NP. Consider the problem $[PH]_p^h$ of the approximation of d by an indexed hierarchy with height h. Křivánek and Morávek (1986) proved that $[1-H]_p^h \prec [1-H]_p^{h+1}$. This proof works without any change in the case of k-weak hierarchies and in the case of pseudo-hierarchies. Hence, the results from the case $h=1$, which correspond to *k-Zahn* and *Robbin* respectively□

4.5 Some Open Questions

We have observed that the approximation of a dissimilarity by an indexed k-weak hierarchy is, for $k \geq 3$, trivial: d itself is the solution which is no longer related to the $(k+2)$-point inequality. We have shown that for a proper preindexed dissimilarity, the problem is NP-hard. But our proof, based on "graphical" features, does not extend to a possibly nonproper dissimilarity satisfying the $(k+2)$-point inequality. For the same reason, our proofs do not apply to $p=\infty$ (graphical problems are trivial in the L_∞ norm). It is worth noting that the L_∞ approximation of a dissimilarity by an ultrametric can be performed in polynomial time, as has been proved by Farach, Kannan, and Warnow (1995), and by Chepoi and Fichet (2000) who provide a clearer proof.

References

BANDELT, H.-J. (1992), "Four point characterization of the dissimilarity functions obtained from indexed closed weak hierarchies," *Mathematisches Semminar*, Mathematisches Forsuch Center, Universität Hamburg, Germany.

BANDELT, H.-J., and DRESS, A. W. M. (1989), "Weak Hierarchies Associated with Similarity Measures – an Additive Clustering Technique,"*Bulletin of Mathematical Biology*, *51*, 133–166.

BANDELT, H.-J., and DRESS, A. W. M. (1993), "An Order Theoretic Framework for Overlapping Clustering," *Discrete Mathematics, 136*, 21–37.

BATBEDAT, A. (1988), "Les isomorphismes HTS et HTE (après la bijection de Benzécri-Johnson)," *Metron, 46*, 47–59.

BATBEDAT, A. (1989), "Les dissimilarités médias et arbas," *Statistiques et analyse des données, 14*, 1–18.

BATBEDAT, A. (1990), *Les approches pyramidales dans la classification arborée*, Paris: Masson.

BENZÉCRI, J. P. (1973), *L'analyse des données (Volume 1: Taxonomie)*, Paris: Dunod.

BERTRAND, P. (1992), "Propriétés et caractérisations topologiques d'une représentation pyramidale," *Mathématiques, informatiques et sciences humaines, 117*, 5–28.

BERTRAND, P. (1998), *Set Systems and Dissimilarities cahiers du CEREMADE* 9853, Centre de Recherche en Mathématiques de la Décision, University of PARIS IX-Dauphine, France.

BERTRAND, P., and JANOWITZ, M. F. (1999), "The k-weak Hierarchies: an Extension of the Weak-hierarchies," Technical Report, Amherst: Department of Mathematics and Statistics, University of Massachusetts.

BRUCKER, P. (1978), "On the Complexity of Clustering Problems," In *Optimization and Operations Research*, Eds., M. Beckmann, and H. P. Kunzi, Heidelberg: Springer-Verlag, 45–54.

CHEPOI, V., and FICHET, B. (2000), "L_∞-Approximation via Subdominants," *Journal of Mathematical Psychology, 44*, 600–616.

DAY, W. H. E. (1987), "Computational Complexity of Inferring Phylogenies from Dissimilarity Matrices," *Bulletin of Mathematical Biology, 49*, 461–467.

DIATTA, J. (1996), *Une extension de la classification hiérarchique : les quasi-hiérarchies*, Ph.D. Thesis, Mathématiques appliquées, Université de Provence - Aix Marseille I, France.

DIATTA, J. (1998), "Approximating Dissimilarities by Quasi-ultrametrics," *Discrete Mathematics, 192*, 81–86.

DIATTA, J., and FICHET, B. (1994), "From Asprejan Hierarchies and Bandelt-Dress Weak-hierarchies to Quasi-hierarchies," in *New Approaches in Classification and Data Analysis*, Eds., E. Diday,Y. Lechevallier, M. Schader, P. Bertrand, and B. Burtschy, Berlin: Springer-Verlag, 111–118.

DIATTA, J., and FICHET, B. (1998), "Quasi-ultrametrics and Their 2-balls Hypergraphs," *Discrete Mathematics, 192*, 87–102.

DIDAY, E. (1984), *Une représentation visuelle des classes empiétantes: les Pyramides*, Research report 291, Institut National de Recherche en Informatique et en Automatique (INRIA), centre de Rocquencourt Domaine de Voluceau, B.P. 105, 78153 Le Chesnay Cedex, France.

DIDAY, E. (1986), "Orders and Overlapping Clusters in Pyramids," in *Multidimentional Data Analysis Proceedings*, Eds., J. De Leeuw, W. Heiser, J. Meulman, and F. Critchley. Leiden: DSWO Press, 201–234.

DURAND, C. (1989), *Ordre et graphes pseudo-hiérarchiques : théorie et optimisation algorithmique*. Doctoral Thesis, Mathématiques appliquées, Université de Provence - Aix Marseille I, France.

DURAND, C., and FICHET, B. (1988), "One to one Correspondances in Pyramidal Representation: an Unified Approach," in *Classification and Related Methods of Data Analysis*, Eds., H. H. Bock, Amsterdam: North-Holland, 85–90.

FARACH, M., KANNAN, S., and WARNOW, T. (1995), "A robust model for finding optimal evolutionary trees," *Algorithmica, 13*, 155–179.

FICHET, B. (1984), "Sur une extention de la notion de hiérarchie et son équivalence avec quelques matrices de robinson," Lecture at *Journée de statistique de la grande Motte*, France.

FICHET, B. (1986), "Data Analysis: Geometric and Algebric Structures," In *First World Congress of the Bernoulli Society Proceedings*, Eds., Y. A. Prohorov, and V. U. Sasonov V.U, Utrecht: V.N.U. Science Press, 123–132.

GAREY, M. R., and JOHNSON, D. S. (1979), *Computer and Intractability, a Guide in the Theory of NP-completeness*, New York: Freeman.

GRÖTSCHEL, M., and WAKABAYASHI, Y. (1990), "Facets of the Clique Partitioning Problem,"

HANSEN, P., and DELATTRE, M. (1978), "Complete-link Cluster Analysis by Graph Colouring," *Journal of the American Statistical Association, 73*, 397–403.

HANSEN, P., B., JAUMARD, B. , and MLADENOVIC, N. (1995), "How to Choose k Entities among N?," In *Partitioning Data Sets, DIMACS Series in Discrete Mathematics and Theoretical Computer Science (Volume 19)*, Eds., I. J. Cox, P. Hansen, and B. Julesz, Providence, RI: American Mathematical Society, 105–116.

HUBERT, L., ARABIE, P., and MEULMAN, J. (1997), "Linear and Circular Unidimensional Scaling for Symmetric Proximity Matrices," *British Journal of Mathematical and Statistical Psychology, 50*, 253–284.

HUBERT, L., ARABIE, P., and MEULMAN, J. (1998), "Graph-theoric Representations for Proximity Matrices through Strongly-anti-Robinson or Circular Strongly-anti-Robinson Matrices," *Psychometrika, 63*, 341–358.

JARDINE, J. P. J., JARDINE, N., and SIBSON, R. S. (1967), "The Structure and Construction of Taxonomic Hierarchies," *Mathematical Biosciences, 1*, 171–179.

JARDINE, N., and SIBSON, R. (1971), *Mathematical Taxonomy*, London: Wiley.

JOHNSON, S. C. (1967), "Hierarchical Clustering Schemes," *Psychometrika, 32*, 241–254.

KŘIVÁNEK, M., and MORÁVEK, J. (1986), "NP-Hard Problems in Hierarchical-tree Clustering," *Acta Informatica, 23*, 311–323.

RÉGNIER, S. (1965), "Sur quelques aspects mathématiques des problèmes de classification automatique," *Mathématiques et Sciences Humaines* (1983), 13-29. Reprinted from *ICC Bulletin* (1965), *4*, 175–191.

ROBINSON, W. S. (1951), "A Method for Chronologically Ordering Archeological Deposits," *American Antiquity, 16*, 295–301.

WAKABAYASHI, Y. (1986), *Aggregation of Binary Relations: Algorithmic and Polyhedral Investigations*. Doctoral thesis, der Naturwissenschaftlichen Fakultät der Universität Ausburg.

ZAHN, C. T. (1964), "Approximating Symmetric Relations by an Equivalence Relation," *SIAM Journal on Applied Mathematics, 12*, 840–847.