UPPSALA DISSERTATIONS IN MATHEMATICS

52

Optimal and Hereditarily Optimal Realizations of Metric Spaces

Alice Lesser

Department of Mathematics
Uppsala University
UPPSALA 2007

Dissertation presented at Uppsala University to be publicly examined in Polhemssalen, Ångströmlaboratoriet, Lägerhyddsvägen 1, Uppsala, Friday, November 30, 2007 at 10:15 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract
Lesser, A. 2007. Optimal and Hereditarily Optimal Realizations of Metric Spaces. (Optimala och ärftligt optimala realiseringar av metriker). *Uppsala Dissertations in Mathematics* 52. 70 pp. Uppsala. ISBN 978-91-506-1967-6.

This PhD thesis, consisting of an introduction, four papers, and some supplementary results, studies the problem of finding an *optimal realization* of a given finite metric space: a weighted graph which preserves the metric's distances and has minimal total edge weight. This problem is known to be NP-hard, and solutions are not necessarily unique.

It has been conjectured that *extremally weighted* optimal realizations may be found as subgraphs of the *hereditarily optimal realization* Γ_d, a graph which in general has a higher total edge weight than the optimal realization but has the advantages of being unique, and possible to construct explicitly via the *tight span* of the metric.

In Paper I, we prove that the graph Γ_d is equivalent to the 1-skeleton of the tight span precisely when the metric considered is *totally split-decomposable*. For the subset of totally split-decomposable metrics known as *consistent* metrics this implies that Γ_d is isomorphic to the easily constructed *Buneman graph*.

In Paper II, we show that for any metric on at most five points, any optimal realization can be found as a subgraph of Γ_d.

In Paper III we provide a series of counterexamples; metrics for which there exist extremally weighted optimal realizations which are not subgraphs of Γ_d. However, for these examples there also exists at least one optimal realization which is a subgraph.

Finally, Paper IV examines a weakened conjecture suggested by the above counterexamples: can we always find some optimal realization as a subgraph in Γ_d? Defining *extremal* optimal realizations as those having the maximum possible number of shortest paths, we prove that any embedding of the vertices of an extremal optimal realization into Γ_d is injective. Moreover, we prove that this weakened conjecture holds for the subset of consistent metrics which have a 2-dimensional tight span

Keywords: optimal realization, hereditarily optimal realization, tight span, phylogenetic network, Buneman graph, split decomposition, T-theory, finite metric space, topological graph theory, discrete geometry

Alice Lesser, Department of Mathematics, Box 480, Uppsala University, SE-75106 Uppsala, Sweden. The Linnaeus Centre for Bioinformatics, Box 598, Uppsala University, SE-75124 Uppsala, Sweden

ISSN 1401-2049
ISBN 978-91-506-1967-6
urn:nbn:se:uu:diva-8297 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-8297)

"Don't be too sure," said the child patiently, "for one of the nicest things about mathematics, or anything else you might care to learn, is that many of the things which can never be, often are. You see," he went on, "it's very much like your trying to reach Infinity. You know that it's there, but you just don't know where — but just because you can never reach it doesn't mean that it's not worth looking for."

Norton Juster, *The Phantom Tollbooth*

Cover image: An optimal realization of a metric on 16 elements, as a subgraph of its hereditarily optimal realization. The red graph is an optimal realization by results in section 4.2.2. The hereditarily optimal realization was calculated using `polymake` and `SplitsTree`, using results in Paper I. Image rendered in `POV-Ray`.

List of Papers

This PhD thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I — Andreas Dress, Katharina T. Huber, Alice Lesser, and Vincent Moulton, **Hereditarily optimal realizations of consistent metrics**
Annals of Combinatorics, 10(1):63–76 (2006)

II — Jack Koolen, Alice Lesser, and Vincent Moulton, **Optimal realizations of generic 5-point metrics**
Isaac Newton Institue preprint NI07070, submitted (2007)

IIa — Alice Lesser, **Optimal and h-optimal realizations for 5-point metrics**
appendix to II (2007)

III — Jack Koolen, Alice Lesser, and Vincent Moulton, **Concerning the relationship between realizations and tight spans of finite metrics**
Journal of Discrete and Computational Geometry, to appear (2007)

IV — Alice Lesser, **Extremal optimal realizations**
Uppsala University, Department of Mathematics preprint 2007:53 (2007)

Contents

1. Introduction and Historical Review

This PhD thesis is concerned with the problem of finding an *optimal realization of a given finite metric space*. Before we start our search for these objects, it might help to know what we are looking for.

A *metric space* is a set of objects and a specified distance between any two of them. More formally, a metric space consists of a pair (X, d), where X is a set, which we will consider to be finite unless stated otherwise, and d is a distance function or *metric*, i.e. a function $d : X \times X \to \mathbb{R}$ satisfying the following properties:

- All distances are *nonnegative*: $d(x, y) \geq 0$ and $d(x, x) = 0$ for all x and y in X.
- Every distance is *symmetric*: $d(x, y) = d(y, x)$ for all x and y in X.
- The distances satisfy the *triangle inequality*: $d(x, z) \leq d(x, y) + d(y, z)$ for all x, y, z in X.

Note that we allow $d(x, y) = 0$ for distinct x and y; thus our definition of a metric is equivalent to what is often called a *pseudometric*.

Searching for an optimal realization of a given metric d on a set X can be thought of as searching for a way to "connect the dots" as efficiently as possible. Formally, we consider a weighted graph $G = (V, E, w \geq 0)$ such that $X \subseteq V$, i.e. G contains all elements of the set X as vertices, and possibly some additional vertices which we will call *auxiliary vertices*. If, for any two elements $x, y \in X$ we have $d(x, y) = d_G(x, y)$, that is the shortest path joining x and y in the graph G has length precisely equal to their distance in the metric space, we call G a *realization* of the metric space (X, d).

$$
\begin{array}{c|ccc}
 & a & b & c \\
\hline
a & 0 & 1 & 1 \\
b & 1 & 0 & 1 \\
c & 1 & 1 & 0
\end{array}
$$

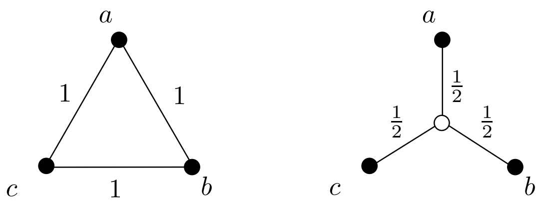


Figure 1.1: A simple metric space, consisting of three elements a, b and c such that the distance between any pair is 1; a realization of this metric, with total edge weight 3; and the optimal realization, with total edge weight $\frac{3}{2}$.

Every metric space has at least one realization, since we can let the vertices V be precisely the elements X, and construct the complete graph on these vertices with all edges $\{x, y\}$ having weight $d(x, y)$, although in most cases we can create a "more efficient" realization by letting some paths intersect. If the *total edge weight* $\|G\| = \sum_{e \in E} w(e)$ of a realization G is minimal among all realizations, we will say that G is an *optimal realization.* Figures 1.1 and 1.2 show some simple examples of realizations and optimal realizations.

$$\begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{matrix} a & b & c & d \end{matrix} \\ \begin{bmatrix} 0 & 3 & 4 & 3 \\ 3 & 0 & 3 & 4 \\ 4 & 3 & 0 & 3 \\ 3 & 4 & 3 & 0 \end{bmatrix} \end{array}$$

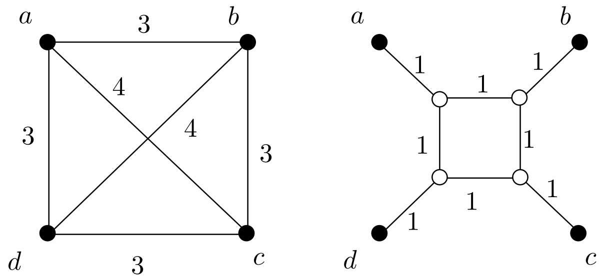


Figure 1.2: A metric space with four elements; its realization as a complete graph on four vertices (with total edge weight 20); and a realization of weight 8. It is not difficult to prove that this last realization is optimal, by considering to what extent paths joining pairs of elements can intersect while still preserving the distances of the metric exactly.

It seems reasonable to assume that every finite metric has some realization that is optimal since there always exists at least one realization, and if there are several possible realizations we can just calculate their total edge weights and choose the smallest. Though of course a necessary condition for this is that no metric has an infinite series of realizations with successively smaller total weight. This is indeed known to be true, and a formal proof is given in Chapter 2.

1.1 Optimal realizations

The problem of finding an optimal realization of a given metric was first investigated by Hakimi and Yau in 1964 [HY64]. In this paper, the authors showed that a symmetric matrix has a realization as a weighted graph if and only if it is a metric space, and observed that in general there are several possible realizations, motivating the definition of an optimal realization.

A natural next question to ask is if the optimal realization of a given metric is *unique*, or if there might be several graphs, necessarily with the same total edge weight, which are each optimal realizations. One remark is in order here: we disregard any auxiliary vertices of degree two (that is with two adjacent edges), since otherwise we could add a vertex anywhere on any edge of an optimal realization and obtain another optimal realization that is not essentially different.

Hakimi and Yau proved the following theorem: "If a metric space (X, d) has a realization T which is a *tree*, then T is optimal, and moreover T is the only tree which realizes (X, d)." A *tree* in the graph-theoretic sense is a connected graph with no circuits, or equivalently a graph such that for every pair of elements there is a unique path connecting them. Metrics that have tree realizations have widespread applications, which we will return to in Section 1.2.

Hakimi and Yau also gave procedures for reducing a realization to one with smaller total edge weight, though they noted that this does not give a complete procedure for finding an optimal realization, since in general one does not know when to stop searching for smaller realizations.

Having shown uniqueness for the special case of metrics realizable by a tree, they stated that they "believe that in general the optimal realization of any metric is unique", but had not been able to find a proof. The interested reader may wish to pause here and ponder how to prove or disprove this statement, while the impatient reader may flip ahead to Figure 1.6.

Around the same time, Imrich and Stoitskii [IS72] investigated optimal realizations and reached many of the same conclusions, specifically that tree realizations are unique and optimal. They also note some other cases where the optimal realization of a metric is unique, and that in general the question of uniqueness remained open.

1.2 An application: phylogenetic analysis

We now consider some applications of optimal realizations. Hakimi and Yau noted that optimal realizations can be used in the synthesis of electrical networks. More recent applications from the field of computer science include attempts to find an efficient algorithm for the *k-server problem*, minimizing the distance traveled by servers to handle requests by customers, c.f. [CL94]; compression software [LCL$^+$04] and discovering the structure of large networks such as the Internet, c.f. [CGGS01].

An application that has aroused considerable interest over the last few decades is in the field of modeling evolutionary relationships between organisms, or *phylogenetic analysis*. The idea of searching for "the tree of life" dates back to the publication of Charles Darwin's "On the Origin of Species" in 1859; although Darwin never explicitly attempts to construct a phylogenetic tree, it is clear that his theory of evolution implies the existence of one. The first explicit depiction of the tree of life was made seven years later, by Darwin's supporter Ernst Haeckel, and is reproduced in Figure 1.3.

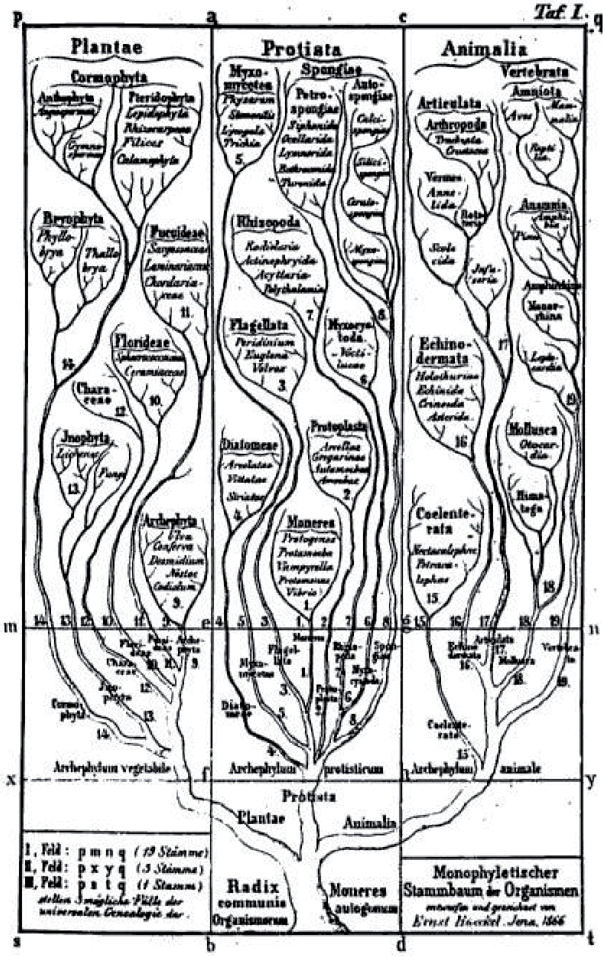


Figure 1.3: The first published depiction of the tree of life.

To construct this tree Haeckel consulted the naturalist systems that had been in place since the time of Linnaeus (1707-1778), where organisms are grouped together according to their similarities, and reinterpreted their static divisions of the natural world as a process of evolution, using the new assumption that similar species had evolved from a common ancestor. For roughly a hundred years this time-consuming method of classifying organisms by their *morphological* characteristics, such as the placement of leaves on a plant or the limb structure of an animal, remained the only method of constructing phylogenetic trees.

However, in the late 1960's advances in molecular biology made a completely new approach feasible: one could determine DNA or amino acid sequences of for example the same gene from a family of related organisms, obtain a mathematically well-defined distance between any pair of these organisms in terms of the number of differences in the sequences, and then reconstruct a phylogenetic tree from this distance. Figure 1.4 shows a sketch of the idea behind this method.

man	GATCACAGGT CTATCACCCT
ape	AATCACAGGA TTATCACCCT
dog	AACTTATGGA CAGTCACCCT
cat	AACATTTGGA CAGTCACCCT
bug	ATCATTAGGA CTATGGGGGG

$$
\begin{array}{c}
man\\ ape\\ dog\\ cat\\ bug
\end{array}
\begin{bmatrix}
0 & 3 & 10 & 10 & 14\\
3 & 0 & 9 & 9 & 13\\
10 & 9 & 0 & 2 & 10\\
10 & 9 & 2 & 0 & 10\\
14 & 13 & 10 & 10 & 0
\end{bmatrix}
$$

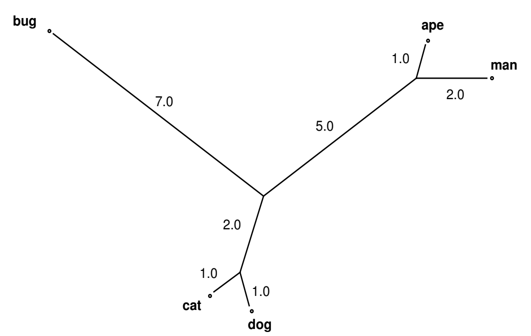


Figure 1.4: A toy example of constructing a phylogenetic tree from DNA sequences. The sequences are compared one pair at a time, and the number of positions in which they differ is calculated. Then one seeks to construct a tree for which these distances are exactly preserved. The tree in this figure was produced using the program `SplitsTree` [HB06] which produces phylogenetic trees and networks from sequence data or distance matrices.

There are many simplifications here; the underlying assumption is always that all of the currently existing species' sequences are derived by successive mutations from some original ancestral species. One simplification is that we are assuming that any discrepancy is caused by the smallest possible number of mutations; if we observe an "A" in one sequence and a "T" in the same position in another we are assuming that the ancestral species's sequence contained either an "A" or a "T" at that position which then mutated in one of the descendants. Another important point is that in the toy example above we are implicitly assuming that any mutation is equally likely, since any substitution is given the same weight (namely 1) when the distance matrix is calculated. However, we know from biology that some mutations are more probable than others, and thus it is often desirable to assign different weights to different substitutions, i.e. a substitution between nucleotides “A” and “C”, which have a similar chemical structure, might be given weight 1 while a substitution between “A” and “T” is given weight 2. It is also frequently the case that some sequence of nucleotides is deleted or inserted by mutation, and such a “gap” might be better modeled by assigning it a lower weight than a corresponding sequence of mismatches. Selecting an appropriate model to use is of course an entire research topic in itself; for our purposes here we observe that the only important consideration for results on optimal realizations to be applicable is that the distances resulting from the model actually constitute a metric according to our definition, i.e. they are symmetric and satisfy the triangle inequality.

The first paper where phylogenetic trees are constructed from amino acid sequences was published by Fitch and Margoliash in 1967 [FM67]. In this paper, the authors consider 20 species, from human to *Candida*, a type of yeast, and calculate the number of differing positions in their respective amino acid sequences of a protein called cytochrome *c*. From this distance, several phylogenetic trees are then constructed using previously known methods from *cluster analysis*.

An important distinction here is that Fitch and Margoliash's method will not in general provide a *realization* in our sense of the word; since their trees are constructed using average distances, it is very unlikely that all of the original distances will be preserved exactly. There are many other methods of this type, which given a metric space produce either one tree which is the best according to some criterion or a collection of different trees, but that are not exact realizations of the metric. A comprehensive list of available computer programs and web servers for phylogeny can be found at Joseph Felsenstein's web site [Fel]

As noted by Hakimi and Yau, if a metric space has a realization which is a tree, then this realization is the unique optimal realization; so a logical next question to ask is whether there is some way to determine if a given metric has a tree realization.

Such a test indeed exists and has come to be known as the *four-point condition*: for any four elements, say x, y, u and v in our set X, consider the three possible sums of pairs of these distances; $d(x, y) + d(u, v)$, $d(x, u)+d(v, y)$ and $d(x, v)+d(u, y)$. We say that the four-point condition is satisfied if the two largest of these three sums are equal to each other, as sketched in Figure 1.5.

It is relatively easy to see that any tree satisfies the four-point condition, and with some extra work it is possible to prove that the condition is also sufficient, so a metric space has a realization as a weighted tree if and only if the four-point condition is satisfied for all quartets of elements.

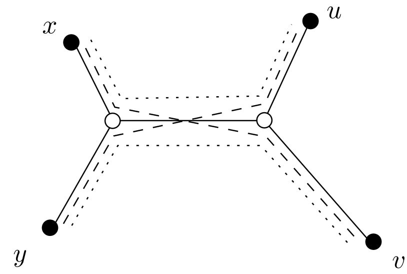


Figure 1.5: The four-point condition: the dotted lines represent the sum of the two distances $d(x, u)$ and $d(y, v)$, while the dashed lines represent $d(x, v) + d(y, u)$. These two sums are clearly equal and not less than the third possible sum of pairs, $d(x, y) + d(u, v)$.

This result was independently discovered by quite a few researchers in the late 1960's and early 1970's: Zaretskii [Zar65] established the result for metrics with integer distances; while the general result was proved by Simoes-Pereira [SP69], Patrinos and Hakimi [PH72], Dobson [Dob74] and Buneman, [Bun71], [Bun74].

If the four-point condition is satisfied it is straightforward to determine the unique tree realization, for example by using the "tree-popping" algorithm due to Meacham [Mea83], see also Semple and Steel [SS03]. Most of the approximation methods mentioned above will also output the exact tree realization if it exits.

Tree-like metrics are well understood; but in general we cannot expect a metric derived from real biological data to satisfy the four-point condition exactly. This is where optimal realizations, or more generally *phylogenetic networks*, may come in useful in various ways. If we are working under the hypothesis that our data have evolved from a branching process corresponding to a tree, then an optimal realization could be used to visualize how "tree-like" the data is. On the other hand, for some organisms a tree may not be the best model of evolution: processes such as hybridization in plants or horizontal gene transfer in bacteria, where species that have differentiated may recombine at a later date, might be better modeled by a network.

1.3 Some problems

As with most problems of this sort, the next reasonable question to ask is whether it is possible to find optimal realizations in a moderate amount of time.

A basic concept in the study of algorithms is the notion of *complexity classes* for problems, that is, whether there exists a reasonably efficient algorithm to solve a given problem. The classic reference on this subject is the book [GJ79] by Garey and Johnson. We will here only outline some of the most basic concepts.

A decision problem, i.e. a problem with a "yes" or "no" answer, is said to belong to complexity class P if there exists an algorithm that can find an answer in polynomial time, that is, if the running time of the algorithm is bounded by a polynomial in n where n is the size of the problem. The complexity class NP is the class of all decision problems where a suggested answer can be verified in polynomial time, and hence NP includes P as a subset. Among the problems in NP there is an important class of problems which we call *NP-complete*. Any of these problems can be transformed in polynomial time into any of the others, and moreover any other problem in NP can be transformed into an NP-complete problem in polynomial time. Hence the NP-complete problems are the hardest problems in the class NP, and, if a polynomial-time algorithm were to be found for any NP-complete problem, then it would be possible to solve all such problems in polynomial time, and hence the classes P and NP would be the same.

However it is widely believed, though not yet proven, that P is a proper subset of NP and hence that it is not possible to find polynomial-time algorithms to solve NP-complete problems.

We will say that a problem is *NP-hard* if it is at least as hard as any NP-complete problem, but is not necessarily in the class NP itself. This is typically the case when we are interested in a problem that is not itself a decision problem, but for example an optimization problem involving maximizing or minimizing some parameter. Such a problem can however be transformed into a decision problem for each specific value of the parameter, and hence must be at least as hard as this corresponding decision problem.

In 1988, Althöfer [Alt88] showed that the realization problem is NP-hard by reducing it from the problem of finding minimal *transversals* in certain graphs. The minimal transversal in any given graph is the complement of its maximum independent set, (the maximum subset of vertices for which the corresponding induced subgraph contains no edges) and finding maximum independent sets in graphs is known to be NP-complete [GJ79].

In the same year Winkler [Win88] independently showed the NP-hardness of the realization problem, in this case by reducing the problem of partitioning a graph into cliques (specifically, for a given graph G and a positive integer, does there exist a partition of the vertices of G into disjoint sets such that the induced subgraph on each set is a complete graph?), which is also known to be an NP-complete problem [Kar72]. Both authors also observe that, for the special case of metrics where all distances are integers, computing optimal realizations belongs to the class NP; since the number of vertices in an optimal realization is bounded above by n^4, and the total edge weight of a realization is possible to calculate in polynomial time, then it is possible to verify in polynomial time if a realization has a certain total edge weight.

We now return to Hakimi and Yau's question: is the optimal realization of any metric a unique graph, or might there be several possible graphs? The renewed interest in optimal realizations in the 1980's provided an answer to this question as well; in the form of two independently published, quite simple counterexamples. Dress' 1984 article [Dre84], which we will return to below, contained an example of a metric with two distinct optimal realizations, which was later reproduced in a slightly simplified form by Althöfer [Alt88] and is shown in Figure 1.6.

$$\begin{array}{c} \\ a \\ b \\ c \\ d \\ e \end{array} \begin{array}{c} \begin{array}{ccccc} a & b & c & d & e \end{array} \\ \begin{bmatrix} 0 & 2 & 4 & 2 & 2 \\ 2 & 0 & 2 & 4 & 2 \\ 4 & 2 & 0 & 2 & 2 \\ 2 & 4 & 2 & 0 & 2 \\ 2 & 2 & 2 & 2 & 0 \end{bmatrix} \end{array}$$

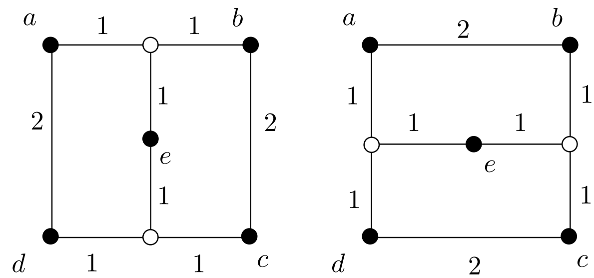


Figure 1.6: A metric with two distinct optimal realizations.

Dress conjectured that any metric has a finite number of optimal realizations. Again, the interested reader may want to attempt a proof or counterexample to this conjecture, or flip ahead to Figure 1.14. Also in 1984, Imrich, Simoes-Pereira and Zamfirescu's paper [ISPZ84] demonstrated that optimal realizations need not be unique, and moreover that the same metric can have different optimal realizations with different numbers of edges. Their example is reproduced in Figure 1.7. In Paper III and Paper IV we will return to this example and its implications for the calculation of optimal realizations.

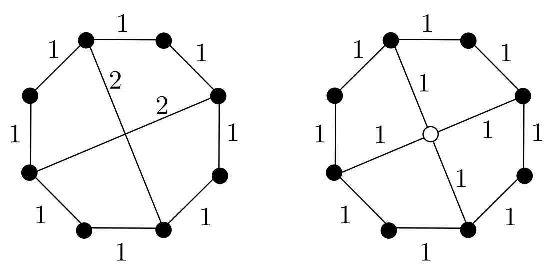


Figure 1.7: Two optimal realizations of the same metric on eight points.

1.4 A new mathematical framework

Motivated in part by phylogenetics, Dress [Dre84] laid many of the foundations for what has come to be known as *T-theory*: the study of trees, tight spans of metric spaces and related topics. After the publication of Dress' article [Dre84] in 1984, it was discovered that the tight-span construction had been previously studied by Isbell in 1965 [Isb65], where it was called the *injective hull*. Some years later, in 1994, it was rediscovered again by Chrobak and Larmore [CL94] in the context of theoretical computer science. Unless otherwise stated we will follow the terminology from [Dre84].

One of the most important ideas in [Dre84] is the concept of the *tight span* of a metric space. The tight span can be viewed as an abstract analog of the convex hull; it is the smallest *injective metric space*, a space into which a given metric space can be *isometrically* embedded, i.e. embedded so that all distances are exactly preserved. We will only give a brief description here, more details are in Section 2.3.

For a metric space (X, d) we consider the set of all maps f from the elements of X into the real numbers $\mathbb{R}$, and denote it by $\mathbb{R}^X$. Among these maps we consider a set of non-contracting maps which we will call $P(X, d)$, defined as

$$P(X, d) := \{f \in \mathbb{R}^X \mid f(x) + f(y) \geq d(x, y) \text{ for all } x, y \in X\}.$$

We then define the *tight span* of (X, d), which we denote by $T(X, d)$, as the bounded elements of $P(X, d)$. Figure 1.8 sketches this relationship for the simplest case of a metric with two elements.

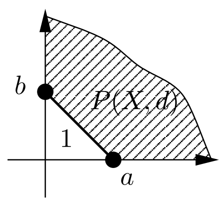


Figure 1.8: A sketch of the sets $P(X, d)$ and $T(X, d)$ for a metric space (X, d) with two elements a and b, with $d(a, b) = 1$. $P(X, d)$ is the unbounded region containing all maps f such that the sum of its distance to a and b is at least 1, while the tight span $T(X, d)$ is the line segment of length 1 joining a and b.

As the definition of the tight span can be rather hard to grasp, we will describe its structure for some very small examples. The tight span of any metric on two elements a and b is a line segment of length $d(a, b)$. For any metric on three elements, say a, b and c, we let $\delta_a = \frac{1}{2}(d(a, b) + d(a, c) - d(b, c))$, $\delta_b = \frac{1}{2}(d(a, b) + d(b, c) - d(a, c))$ and $\delta_c = \frac{1}{2}(d(b, c) + d(a, c) - d(a, b))$, any one of which may be equal to 0. Then the tight span consists of three line segments of these lengths joining each of the three elements to a central vertex, as in Figure 1.9. For a "generic"

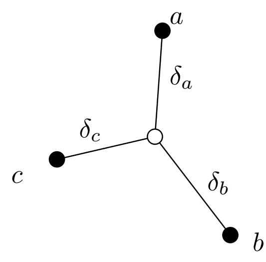


Figure 1.9: A sketch of the tight span of any metric d on three elements a, b and c.

metric on four elements the tight span consists of a rectangle and four line segments as sketched in Figure 1.10. The distances between points in this central rectangle are given by the ℓ_1 or *Manhattan* metric as the sum of their coordinates.

Some of the lengths in the figure may be zero for specific metrics; for example if the sums of distances $d(a, b) + d(c, d)$ and $d(a, c) + d(b, d)$ are equal then the rectangle collapses to a line segment, c.f. Figure 1.5. In fact, if a metric satisfies the four-point condition and hence has a tree as its optimal realization, then the tight span of the metric is precisely this same tree.

The similarity between these simplest tight spans and the simplest examples of optimal realizations, such tree and the examples in Figures 1.1 and 1.2 is not merely coincidental, as we will see in the next section.

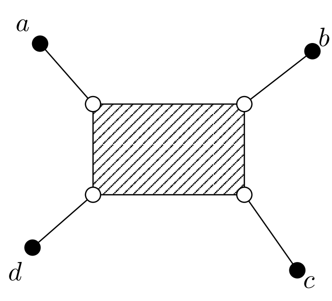


Figure 1.10: The tight span of a generic metric on four elements.

For more than four elements things become more complex; for metrics on five elements there are three possible "generic" tight spans. Figure 1.11 shows the "1-skeletons" of each of these three types, i.e. the cells of dimension 0 and 1. In Paper II and its appendix we calculate the optimal realizations of metrics of each of these types, as well as the "non-generic" metrics on five points. For metrics on six points there are 339 generic tight spans. [SY04] [Kåh05].

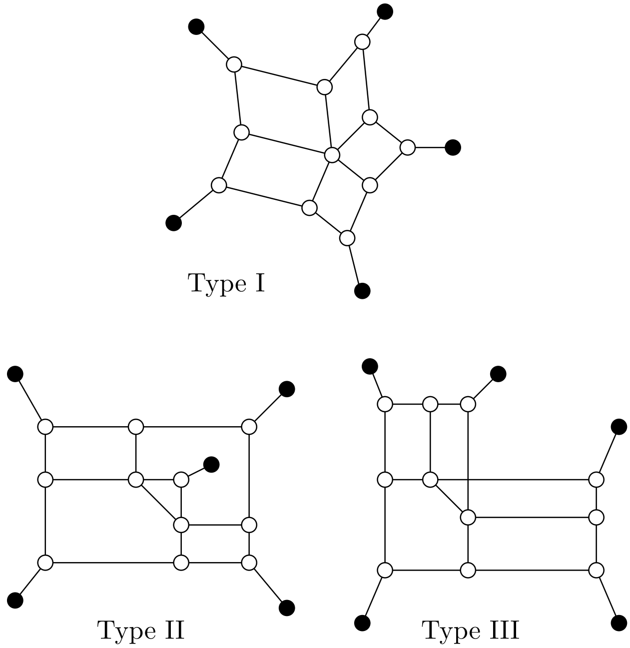


Figure 1.11: The three generic tight spans for metrics on five elements.

As a first step towards understanding the relation between optimal realizations and tight spans, Dress [Dre84] defines *hereditarily optimal realizations*. These realizations are not optimal (except in very special cases) but they avoid several of the difficulties with optimal realizations; they are unique and can be explicitly calculated (for all metrics in principle, and, for large classes of metrics, easily in practice).

For a metric space (X, d) we define a *hereditarily optimal* (or *h-optimal*) realization inductively with respect to $|X|$, the number of elements of X, in the following way: If $|X| = 2$ then any optimal realization of d is defined to be h-optimal; so the h-optimal realization of a metric space with two elements a and b is simply the graph with vertices a and b joined by an edge of weight $d(a, b)$. For larger sets we require that the property of h-optimality is preserved on all subsets: if $|X| = k$ with $k \geq 3$ and if h-optimal realizations have been defined previously for all metric spaces Y with $|Y| < k$, then a realization $G = (V, E, w)$ of (X, d) is defined to be h-optimal if for any proper subset Y of X there is some subgraph $G' = (V', E', w\,\big|_{E'})$ of G such that G' is a hereditarily optimal realization of $(Y, d|_Y)$ and $\sum_{e \in E} w(e)$ is minimal among all such graphs.

But the real power of hereditarily optimal realizations becomes evident only when we relate them to the tight span; one of the main results in [Dre84] is that the h-optimal realization of a metric is always a subgraph of the 1-skeleton of its tight span. We will discuss this in more detail in Section 2.3.

One of the main results in Paper I is the categorization of precisely when these two graphs are equal, that is when the 1-skeleton of the tight span of a metric is precisely its hereditarily optimal realization.

Returning momentarily to phylogenetic analysis, we observe that hereditarily optimal realizations can be useful tools for visualizing relationships present in biological data. As an example, Figure 1.12 shows the h-optimal realization of a metric derived from an alignment of DNA sequences from various species of buttercups, calculated using the package `Spectronet` [HLP$^+$02] on data from [HMLD01]. See also [BFSR95] for an application of h-optimal realizations (in the guise of *median networks*) to data from human mitochondria.

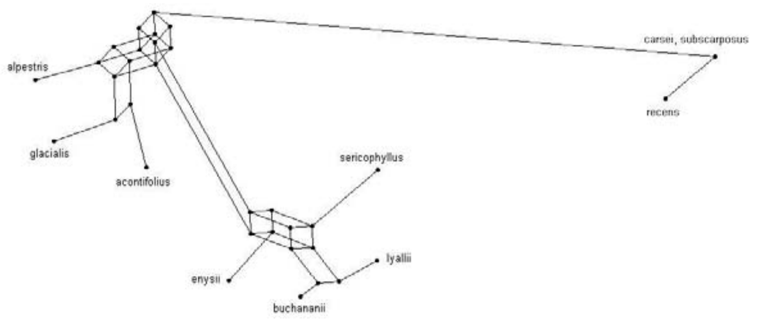


Figure 1.12: An h-optimal realization visualizing evolutionary relationships among ten species of buttercups.

1.5 The big question: how are hereditarily optimal realizations related to optimal realizations?

A logical next step is to investigate how hereditarily optimal realizations are related to optimal realizations of the same metric. As we saw in the previous section, for very small examples (all metrics with at most four elements), the optimal realization, the 1-skeleton of the tight span, and the h-optimal realization all coincide, as is the case for any tree-like metric. Might it be the case in general that information about optimal realizations can be deduced from the corresponding h-optimal realization?

As a concrete example, we begin by returning to the metric with two distinct optimal realizations that was studied in Figure 1.6. We can calculate the h-optimal realization of this metric directly from the definition in the previous section (though we shall see in Chapter 3.1 that there is also an easier method) to obtain the graph in Figure 1.13.

	a	b	c	d	e
a	0	2	4	2	2
b	2	0	2	4	2
c	4	2	0	2	2
d	2	4	2	0	2
e	2	2	2	2	0

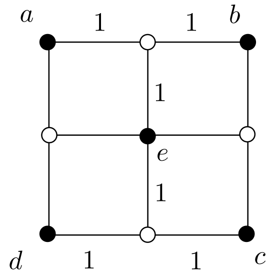


Figure 1.13: The hereditarily optimal realization of the metric with two distinct optimal realizations depicted in Figure 1.6.

Comparing this figure to Figure 1.6 is suggestive: is the (unique) h-optimal realization really just the union of all possible optimal realizations? Or conversely, can we obtain any optimal realization of a given metric from the corresponding h-optimal realization "merely" by deleting some of its edges? This is one of the major questions raised by Dress in [Dre84]. More precisely, his article contains the following three conjectures:

- Noting that the set of edge-weightings of a given graph that make it an optimal realization forms a convex polytope, it is conjectured that this polytope always consists precisely of one point, which in particular implies that any given metric has a finite number of optimal realizations.
- Any optimal realization of a given metric can be obtained from its hereditarily optimal realization by deleting some (possibly empty) subset of edges, and possibly thereafter ignoring any vertices of degree 2.
- In support of the above conjecture, Dress proves the existence of a distance-preserving map from the vertices of any optimal realization to the tight span of the corresponding metric, and conjectures that this map is always injective.

The first two of these conjectures were investigated by Althöfer in 1988 [Alt88], where both were disproved by an example of a metric with a continuum of optimal realizations, reproduced here as Figure 1.14. As Figures 1.15 and 1.16 demonstrate, only the two "extremal" optimal realizations of this metric can be found by deleting edges in the h-optimal realization. Dress's injectivity conjecture does hold for this example however, for any value of ϵ any distance-preserving map from the eight vertices of the optimal realization to the tight span is injective.

	a	b	c	d	e	f
a	0	1	2	3	3	2
b	1	0	1	3	4	3
c	2	1	0	2	3	3
d	3	3	2	0	1	2
e	3	4	3	1	0	1
f	2	3	3	2	1	0

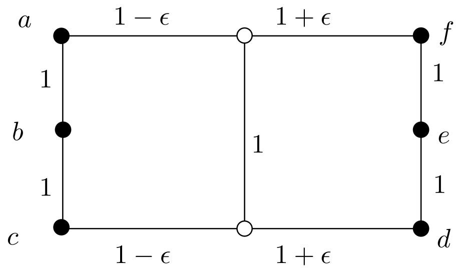


Figure 1.14: Althöfers counterexample to two of Dress' conjectures; for every ϵ between $-\frac{1}{2}$ and $\frac{1}{2}$ the graph shown is an optimal realization of this metric.

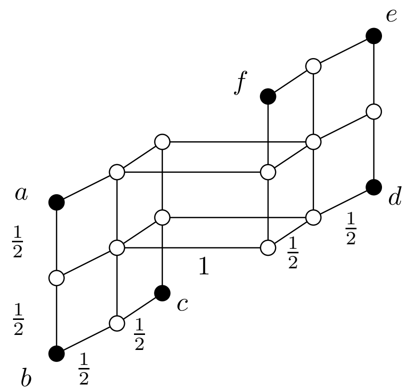


Figure 1.15: The hereditarily optimal realization of Althöfer's continuum example in Figure 1.14.

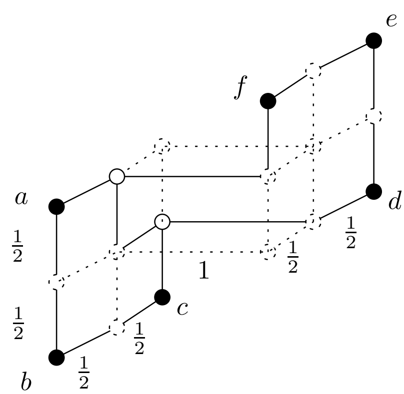


Figure 1.16: The optimal realization obtained by letting ϵ equal $\frac{1}{2}$ is indeed a subgraph of the hereditarily optimal realization of this example, as is the optimal realization with $\epsilon = -\frac{1}{2}$; however for any other value of ϵ the optimal realization cannot be found by deleting edges from the h-optimal realization.

So Dress' original conjecture concerning the relationship between optimal and h-optimal realizations does not hold, but, as noted by Althöfer, the counterexample immediately suggests a modified question:

Given a graph G and a set W of edge-weightings that make G an optimal realization of some metric, can the extremals of W be derived from the h-optimal realization of the metric?

Most of the remainder of this thesis is devoted to providing partial answers to this question for various classes of metrics.

Paper I sets the stage by providing a more detailed characterization of h-optimal realizations for certain "well-behaved" types of metrics.

In Paper II we characterize optimal realizations of all metrics with exactly five elements, and show that all possible optimal realizations of these can be found in the corresponding h-optimal realizations, so that the answer to our question is yes for any metric with at most five elements.

In Paper III we actually answer Althöfer's question; No, there exist metrics such that it is possible to find an optimal realization that is an extremal weighting but is still not a subgraph of the h-optimal realization. These examples also disprove Dress's injectivity conjecture. However, these counterexamples are in a sense pathological; each of the metrics considered also has another optimal realization that can be found in the h-optimal realization (and thus also satisfies Dress' injectivity conjecture).

Finally, in Paper IV we look at the next natural modification of Althöfer's question: does any metric have *some* optimal realization that can be found in its h-optimal realization? In particular, we look at optimal realizations with as many shortest paths as possible, and show that such optimal realizations can always be mapped injectively into the tight span (and hence into the h-optimal realization), and that — at least for the class of metrics where the tight span is at most 2-dimensional — any such "extremal" optimal realization can indeed be found in the h-optimal realization.

Chapter 4 of this thesis provides a collection of examples that demonstrate various properties of optimal realizations, and that may be useful for future research in this area.

2. Some basic results

In this chapter we summarize some previous results concerning optimal realizations, the tight span and the split decomposition of a metric.

2.1 Terminology

For completenes, we begin by recalling some basic definitions concerning *polytopes* and graphs.

2.1.1 Polyhedra and polytopes

Following Ziegler [Zie95], we define a *polyhedron* in $\mathbb{R}^n$, for some $n \in \mathbb{N}$, to be the intersection of a finite collection of halfspaces in $\mathbb{R}^n$. A bounded polyhedron is called a *polytope.* For an n-dimensional polytope P we define a *face* of P to be the intersection of P with one of its supporting hyperplanes, thus P itself and the empty set are also faces. A face of dimension 0 is called a *vertex* of P.

A *polyhedral complex* is a finite collection C of polyhedra such that each face of any polytope $P \in C$ is also in C, and such that the intersection $P \cap Q$ of two polyhedra in C is a face of both P and Q. The elements of a polyhedral complex will be called its *cells.* If every cell in C is a polytope we call C a *polytopal complex.*

2.1.2 Graphs

The terminology used in graph theory is notoriously nonstandard; in what follows we will use the definitions below.

- A *graph* $G = (V, E)$ is an ordered pair consisting of a set V of *vertices*, and a set E of *edges* which are subsets of V with size 1 or 2. An edge of size 1 is called a *loop.*
- Two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ are said to be *isomorphic* if there exists a bijection $\Psi : V_1 \to V_2$ such that $\{u, v\} \in E_1$ if and only if $\{\Psi(u), \Psi(v)\} \in E_2$.
- The *complete graph* on n vertices, denoted by K_n, is the graph with n vertices and all two-element subsets of V as edges.
- A graph G is called *bipartite* if its vertex set V can be partitioned into two disjoint sets X and Y such that every edge in G is of the form

$\{x, y\}$ where $x \in X$, $y \in Y$. The *complete bipartite graph* $K_{m,n}$ is the graph with $m+n$ vertices ($|X| = m$ and $|Y| = n$) and all possible edges having one element in X and the other in Y.
- A *weighted* graph is an ordered triple $G = (V, E, w)$ where $G = (V, E)$ is a graph and $w : E \to \mathbb{R}_{\geq 0}$ is a function, which we will call the *weight function* of G.
- A *path* $v_1, v_2, \ldots, v_k$ in a graph G is a sequence of distinct vertices $v_1, v_2, \ldots, v_k \in V$ such that $\{v_i, v_{i+1}\} \in E$ for all $i \in \{1, \ldots, k-1\}$. We define the length of a path to be $\sum_i w(\{v_i, v_{i+1}\})$; if G is unweighted, edges are assumed to have weight 1.
- The *(graph-theoretic) distance* $d_G(u, v)$ between two vertices u and v in a graph G is defined to be the length of a shortest possible path $u, \ldots, v$ in G. Any such path will be called a *geodesic*.
- A *cycle* is a graph which is a "closed path", i.e. a graph consisting of a path $v_1, \ldots, v_k$ and an additional edge $\{v_k, v_1\} \in E$. A cycle on n vertices (or equivalently, n edges), will be called an n-cycle.

2.2 More on optimal realizations

As promised in the introduction, in this section we prove formally that optimal realizations always exist. We also recall two basic lemmas, which will be used repeatedly in our investigations of optimal realizations, and which have elementary proofs.

Theorem 2.1 ([ISPZ84] Theorem 2.2, [Dre84] p. 392). *Every finite metric space (X, d) has an optimal realization.*

Proof. We begin by showing that any optimal realization has a bounded number of vertices: Let G be a realization of a metric space (X, d) with $|X| = n$. We can assume that G is the union of the $\binom{n}{2}$ shortest paths between each pair of elements in X, and that these paths are in lexicographic order. Now, for any such shortest path, we can for each other pair of elements in X choose some shortest path such that these two paths at most intersect in an interval, since intersection in several places would imply that we had parallel sets of edges of the same weight, implying that we could remove one such set of edges from the path and obtain a smaller realization. Such an intersection of two paths along an interval gives rise to at most two vertices, and hence each such path has at most n^2 vertices, (arising from the possible intersections of this path with shortest paths between any other two elements) so the total number of vertices in an optimal realization is bounded above by n^4.

Next, we observe that the set W of edge weights in any graph G that realizes d is compact in $\mathbb{R}$: The weights of the individual edges are bounded below by 0, and bounded above by the maximum of the distances $d(x, y)$ in (X, d). To show that W is a closed polyhedron, we observe that it is given by inequalities of the type $w(\{x, v_1\}) + w(\{v_1, v_2\}) + \cdots + w(\{v_k, y\}) \geq d(x, y)$, where x, y is a pair of elements in X and $x, v_1, v_2, \ldots, v_k, y$ is a path in G. Moreover, for any pair x, y there must exist some path where equality holds.

The total edge weight for any finite weighted graph is a continuous function, $f : \mathbb{R}^E \to \mathbb{R}$, of the weights of its edges, and since a continuous image of a compact space is compact we must have that the set of total lengths is compact, and so f must attain its infimum value. Thus there must exist some realization which has this minimal total edge weight. □

Lemma 2.2. *[ISPZ84, Lemma 3.1.i] Suppose that (X, d) is a finite metric space, and that $x, y, z \in X$ are distinct with*

$$d(x, y) + d(y, z) = d(x, z).$$

Suppose also that $\widetilde{xy}$ and $\widetilde{yz}$ are geodesics between x, y and y, z, respectively, in some realization of d. Then the only vertex that $\widetilde{xy}$ and $\widetilde{yz}$ have in common is y.

Lemma 2.3. *[ISPZ84, Lemma 3.1.ii] Suppose that (X, d) is a finite metric space, and that $x, y, u, v \in X$ are distinct with*

$$d(x, y) + d(u, v) < \max\{d(x, u) + d(y, v), d(x, v) + d(y, u)\}.$$

Then in any realization of d every geodesic between x and y is disjoint from every geodesic between u and v.

2.3 More on the tight span

This section collects some useful facts concerning the tight span $T(X, d)$ of a metric space (X, d). Although $T(X, d)$ can be defined for arbitrary (i.e. not necessarily finite) metric spaces, we will continue to restrict our attention to finite metric spaces for clarity.

Recall that we defined the unbounded polyhedron $P(X, d)$ by

$$P(X, d) := \{f \in \mathbb{R}^X \mid f(x) + f(y) \geq d(x, y) \text{ for all } x, y \in X\}$$

(where $\mathbb{R}^X$ is defined to be the set of all maps f from X into $\mathbb{R}$). Note in particular that $f(x) \geq 0$ for all f in $P(X, d)$ and all x in X, since $f(x) + f(x) \geq d(x, x) = 0$.

We now define the *tight span* $T(X, d)$ more formally as the set of functions in $P(X, d)$ that are minimal with respect to the pointwise partial ordering of $\mathbb{R}^X$ given by $f \leq g$ if and only if $f(x) \leq g(x)$ for all $x \in X$.

To any $f \in P(X,d)$, we can associate its *tight-equality graph*, which we will denote by $K(f)$, which has vertex set X and edge set

$$\left\{ \{x,y\} \in \binom{X}{2} \mid f(x) + f(y) = xy. \right\}$$

Hence $K(f)$ represents the set of pairs of elements in X for which the inequality in the definition of $P(X,d)$ is an equality. Note that there may be loops in this graph. Tight-equality graphs are a useful tool for investigating properties of the tight span; among their many properties we will especially make note of the following.

We have $f \in T(X,d)$ if and only if no vertex in $K(f)$ has degree 0.

Let $[f]$ denote the smallest face of the polyhedron $P(X,d)$ which contains the function $f \in P(X,d)$. We observe that, since at the boundary of the face $[f]$ new equality relations can be introduced but not destroyed, we have

$$[f] = \{g \in P(X,d) \,:\, K(f) \subseteq K(g)\}.$$

This can be used to show another important fact:

Lemma 2.4. *[Dre89] The dimension* $\dim[f]$ *of the face* $[f]$ *coincides with the number of bipartite connected components of* $K(f)$.

We now examine the relation between the tight span and hereditarily optimal realizations in more detail. Define a weighted graph $\Gamma_d := (V_d, E_d, w_d)$ with vertex set V_d defined by

$$V_d = \{f \in P(X,d) : K(f) \text{ is connected and not bipartite}\},$$

edge set E_d defined to be

$$E_d := \{\{f,g\} \in \binom{V_d}{2} : K((f+g)/2) \text{ is connected and bipartite}\},$$

and weight function

$$w_d : E_d \to \mathbb{R}_{>0}$$

given by

$$w_d(\{f,g\}) := \max_{x \in X} |f(x) - g(x)|.$$

As one of the main theorems in [Dre84] Dress shows that this graph Γ_d is a hereditarily optimal realization of (X,d), and moreover, any other hereditarily optimal realization of (X,d) is essentially isomorphic to Γ_d in the sense that it becomes isomorphic after removal of all vertices of degree two.

Note that the set V_d above is necessarily a subset of all elements f in $P(X,d)$ such that $K(f)$ has no bipartite connected components, and hence V_d is a subset of the vertices (or 0-cells) of $P(X,d)$, or equivalently of the vertices of $T(X,d)$. Similarly, E_d is necessarily a subset of all elements $\frac{f+g}{2}$ in $P(X,d)$ such that $K(\frac{f+g}{2})$ has exactly one bipartite connected component.

So, if we consider the *1-skeleton* of the tight span, i.e. the weighted graph $T(X,d)^{(1)} = (F_0, F_1, w_d)$ that has vertex set F_0 consisting of the vertices of $P(X,d)$, edge set F_1 consisting of those $\{f,g\} \in \binom{F_0}{2}$ for which f and g are the vertices of a 1-dimensional face in $P(X,d)$, and weighting w_d as defined for Γ_d above, then we have that the graph Γ_d is a subgraph of the graph $T(X,d)^{(1)}$ for any metric space (X,d). Thus it is possible, at least in principle, to calculate the h-optimal realization of any metric by using software for computing polyhedra and polytopes, for instance `polymake` [GJ05].

In Paper I we characterize precisely when these two graphs are equal. For this characterization we will need some previous results concerning *splits* and the *split-decomposition*.

2.4 Splits

Let us begin by returning to the toy example of a phylogenetic tree from Section 1.2. We observe that removing any one edge from this tree would split the elements of X into two nonempty sets, as sketched in Figure 2.1. Note also that the distance between any pair of elements in our set of animals corresponds precisely to the sum of the lengths of the edges which split them.

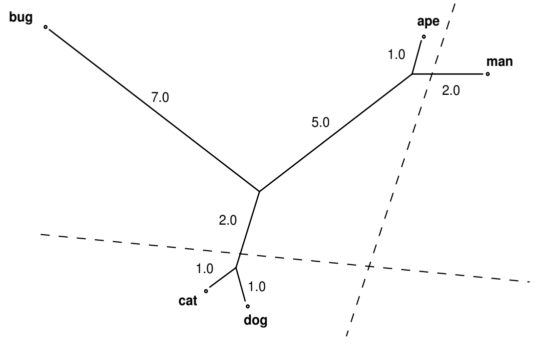


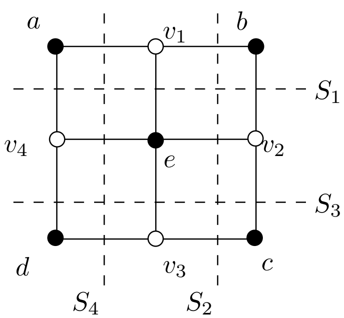
Figure 2.1: Two compatible splits: {cat,dog}|{bug,ape,man} and {cat,dog,bug,ape}|{man} on a phylogenetic tree; the intersection {cat,dog} ∩ {man} is empty while the other three possible intersections are nonempty.

Formally, we define a *split* $S = A|B$ of a set X to be a bipartition of X into two nonempty subsets A and B, which we call *blocks*. Note that any two splits $A_1|B_1$ and $A_2|B_2$ corresponding to edges of this tree (or in fact any phylogenetic tree) satisfy the condition that exactly one of the four possible intersections $A_1 \cap B_1$, $A_1 \cap B_2$, $A_2 \cap B_1$ or $A_2 \cap B_2$ is empty. Two such splits are said to be *compatible*. There is a one-to-one correspondence between collections of splits such that every pair is compatible and phylogenetic trees (formally, trees such that every vertex of degree less than 2 is labeled by an element of X); this is known as the *Splits-equivalence theorem* and is due to Buneman [Bun71]. If we consider a collection of splits where every split is assigned some nonzero weight, then this corresponds in the natural way to a phylogenetic tree with each edge given the weight of the corresponding split.

2.4.1 The Buneman graph and the Buneman complex

Buneman's construction of a tree from a set of compatible splits was generalized by Barthélemy [Bar89], who associated a general graph, later dubbed the *Buneman graph*, to an arbitrary set of splits of a set X.

Let $\mathcal{S}$ be a set of splits of X. Then the Buneman graph $B^{(1)} = (B_0, B_1)$ associated to $\mathcal{S}$ is defined as follows: Each vertex in B_0 is constructed by choosing, for each split in $\mathcal{S}$, exactly one block, in such a way that each pair of chosen blocks from any pair of splits has a nonempty intersection. The edge set B_1 consists of those unordered pairs from the set B_0 that differ by precisely one of the blocks that they contain. Each edge then corresponds to a split $A|B$. A simple example is shown in Figure 2.2.



$a\ :\{\{a,b\},\{a,d,e\},\{a,b,e\},\{a,d\}\}$
$b\ :\{\{a,b\},\{b,c\},\{a,b,e\},\{b,c,e\}\}$
$c\ :\{\{c,d,e\},\{b,c\},\{c,d\},\{b,c,e\}\}$
$d\ :\{\{c,d,e\},\{a,d,e\},\{c,d\},\{a,d\}\}$
$e\ :\{\{c,d,e\},\{a,d,e\},\{a,b,e\},\{b,c,e\}\}$
$v_1:\{\{a,b\},\{a,d,e\},\{a,b,e\},\{b,c,e\}\}$
$v_2:\{\{c,d,e\},\{b,c\},\{a,b,e\},\{b,c,e\}\}$
$v_3:\{\{c,d,e\},\{a,d,e\},\{c,d\},\{b,c,e\}\}$
$v_4:\{\{c,d,e\},\{a,d,e\},\{a,b,e\},\{a,d\}\}$

Figure 2.2: An example of the Buneman graph, where $X = \{a,b,c,d,e\}$ and the splits in $\mathcal{S}$ are $S_1 = \{a,b\}|\{c,d,e\}$, $S_2 = \{b,c\}|\{a,d,e\}$, $S_3 = \{c,d\}|\{a,b,e\}$ and $S_4 = \{a,d\}|\{b,c,e\}$.

As was the case for tree-like metrics described by splits, we can easily extend this definition so that each split is given a weighting, which each edge in the Buneman graph corresponding to a given split is weighted by.

For a set of splits of size $|\mathcal{S}| = m$ the Buneman graph can also be seen as a subgraph of a m-dimensional hypercube $H(X,d)$: For some ordering $\mathcal{S}(X) = \{S_1, \ldots, S_m\}$ of the splits, let $H(X,d)$ be the unit hypercube

$[0,1]^m \subset \mathbb{R}^m$ equipped with the ℓ_1-type metric

$$d_H(x,y) = \alpha_{S_1}|x_1 - y_1| + \alpha_{S_2}|x_2 - y_2| + \cdots + \alpha_m|x_m - y_m|.$$

Loosely speaking, for split number i, $S_i = A_i|B_i$, we associate each coordinate x_i with a normalized distance to the split block A_i. This representation thus implicitly assumes that we have oriented each split "from A_i to B_i". We now define the *Buneman complex* $B(X,d)$ as a subcomplex of the hypercube complex $H(X,d)$ as follows: $(x_1, x_2, \ldots, x_m)$ is an element of $B(X,d)$ if and only if

$$\begin{aligned} x_i &\geq \lceil x_j \rceil \text{ whenever } A_i \subset A_j \\ x_i &\geq \lceil 1 - x_j \rceil \text{ whenever } A_i \subset B_j. \end{aligned}$$

This can be interpreted geometrically as stating that if we have two splits such that a block of one split is completely contained in a block of another, then we must cross out of the smaller split block completely before starting to cross the larger split. Then the 1-skeleton of $B(X,d)$ is isomorphic to the *Buneman graph*, c.f. [DHHM97].

2.4.2 Split-decomposition theory

In this subsection we consider a question that in a sense is the converse of the above: given an arbitrary metric, not necessarily corresponding to a collection of splits, compatible or otherwise, can we describe the metric with the aid of splits?

This is the essence of *split-decomposition theory*, which originates in the work of Bandelt and Dress in 1992 [BD92].

To any split S of X we associate the *split (pseudo-)metric* δ_S, defined by

$$\delta_S(x,y) = \begin{cases} 0 & \text{if } x,y \in A \text{ or } x,y \in B \\ 1 & \text{otherwise,} \end{cases}$$

i.e. $\delta_S(x,y) = 1$ if S separates x and y and 0 otherwise.

For any (not necessarily distinct) $a, a', b, b' \in X$ we define

$$\begin{aligned} \alpha(aa'|bb') := &\max\{d(a,b) + d(a',b'), d(a',b) + d(a,b'), d(a,a)' + d(b,b')\} \\ &- d(a,a') - d(b,b'). \end{aligned}$$

Next, for a split $A|B$ of X, we define the *isolation index* $\alpha_{A,B}$ to be the quantity

$$\alpha_{A,B} := \frac{1}{2} \min_{a,a' \in A, b,b' \in B} \alpha(aa'|bb').$$

This isolation index measures the "minimum distance across the split" as sketched in Figure 2.3.

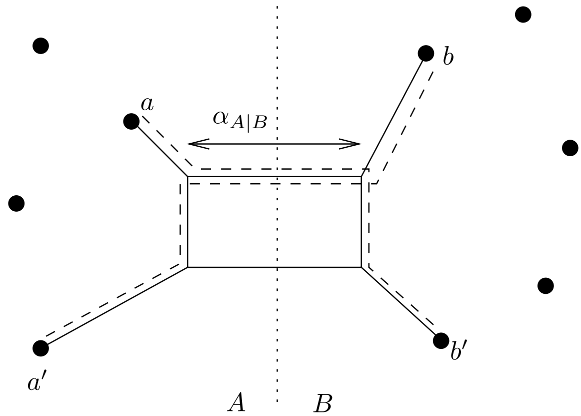


Figure 2.3: A sketch visualizing the isolation index of a split $A|B$. The nine black points are the elements of the set X; the two dotted lines trace the sum $d(a,b)' + d(a',b)$ which is the maximum of the three possible pairs. Subtracting $d(a,a')$ and $d(b,b')$, then taking the minimum over all pairs $a, a' \in A$, $b, b' \in B$ and dividing by 2 gives the isolation index $\alpha_{A|B}$ of the split $A|B$.

Note that $\alpha_{A,B}$ is always nonnegative. If in addition $\alpha_{A,B} \neq 0$, we call $A|B$ a *d-split*.

There are metrics for which no d-splits exist; we will call such metrics *split-prime*. An example of a split-prime metric is the metric induced by the complete bipartite graph $K_{2,3}$, shown in Figure 2.4. It can be shown [[BD92], Lemma 1] that this is (up to a scalar) the only split-prime metric on five points.

	a	b	c	d	e
a	0	1	1	1	2
b	1	0	2	2	1
c	1	2	0	2	1
d	1	2	2	0	1
e	2	1	1	1	0

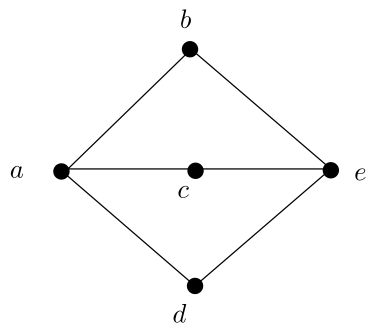


Figure 2.4: The metric induced by $K_{2,3}$ (with all edges of weight 1) is split-prime. In Chapter 3.2 we will show that $K_{2,3}$ is the unique optimal realization of this metric.

We denote the set of all splits of X by $\mathcal{S}(X)$, and the set of all d-splits by $\mathcal{S}_d$. The main result of split-decomposition theory is the following theorem, which states that we can express any metric in a canonical way as a sum of split metrics and a *split-prime residue* which we call d_0.

Theorem 2.5. *[BD92] Every metric d on a finite set X can be written as the sum*

$$d = d_0 + \sum_{S \in \mathcal{S}(X)} \alpha_S \delta_S,$$

where d_0 is split-prime.

If $d_0 = 0$ we say that the metric d is *totally split-decomposable.* We note that for a metric to be totally split-decomposable, for any four points $\{t, u, v, w\} \in X$ all three indices $\alpha(tu|vw)$, $\alpha(tv|uw)$ and $\alpha(tw|uv)$ cannot simultaneously be positive. The collection of d-splits corresponding to a split-decomposable metric d is *weakly compatible*, i.e. there do not exist four points $\{t, u, v, w\} \in X$ and three splits S_1, S_2, S_3 separating them into the three possible pairs of two, as depicted in Figure 2.5.

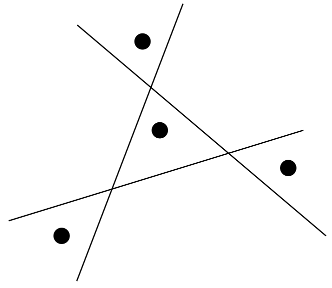


Figure 2.5: The forbidden configuration for a set of splits to be weakly compatible: three splits that separate four points.

Totally split-decomposable metrics can be characterized as follows:

Theorem 2.6. *[BD92] A metric d on a set X is totally split-decomposable if and only if, for any five points $\{t, u, v, w, x\} \in X$,*

$$\alpha(tu|vw) \leq \alpha(tx|vw) + \alpha(tu|vx).$$

In particular, this implies that a metric d on a set X is totally split-decomposable if and only if, for every five-element subset Y of X, the restriction of d to Y (denoted by $d|_Y$) is totally split-decomposable.

3. Summary of papers

3.1 When is the 1-skeleton of the tight span a hereditarily optimal realization? — Paper I [DHLM06]

This paper investigates hereditarily optimal realizations in more detail; in particular, as we saw in Section 2.3, the hereditarily optimal realization of a metric is always a subgraph of the 1-skeleton of its tight span, and in this paper we show that these two graphs coincide precisely when the metric is totally split-decomposable.

We first show that the five-point condition for total split-decomposability can be reformulated in terms of the tight-equality graph $K(f)$.

Lemma 3.1. *For a metric d on a set X, the following statements are equivalent:*

(i) The metric d is not totally split-decomposable.

(ii) There exists some $Y \subseteq X$ with $Y = \{x, y, u, v, t\}$ and some $f \in T(d|_Y)$ with $K(f) = \{\{x, y\}, \{v, u\}, \{u, t\}, \{t, v\}\}$.

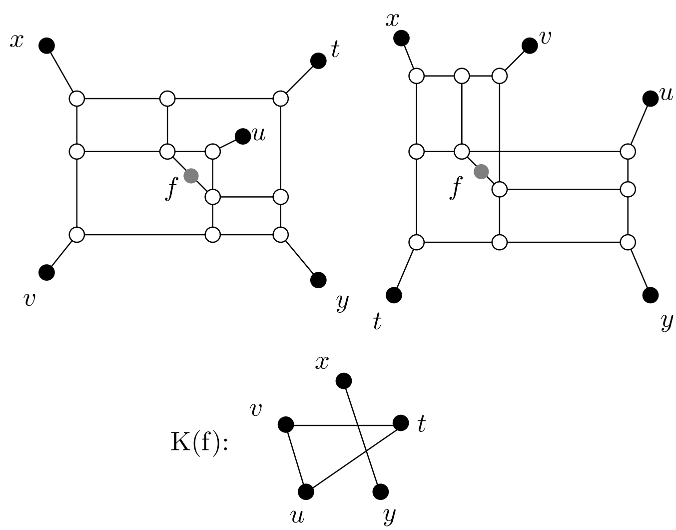


Figure 3.1: For each of the two types of metrics on five points which are not totally split-decomposable, there exists an element f in the tight span with a $K(f)$-graph which is the disjoint union of K_3 and K_2.

Figure 3.1 illustrates this condition for the two types of non-split-decomposable metrics on five points. This lemma enables us to characterize totally split-decomposable metrics as those metrics such that every cell of dimension 1 in $T(X,d)$ has a connected $K(f)$ graph.

Proposition 3.2. *For a metric d defined on a finite set X, the following statements are equivalent:*
(i) The metric d is totally split-decomposable.
(ii) For every $f \in T(d)$ with $\dim[f] = 1$, the graph $K(f)$ is connected.

This proposition in turn leads to the main result of this paper, that total split-decomposability of a metric d on a set X is equivalent to the h-optimal realization of d consisting of precisely all 0-cells and 1-cells in $T(X,d)$. (Note that the notation in this paper differs from that used in the rest of this thesis: the h-optimal realization Γ_d is denoted by H_d and the 1-skeleton $T(X,d)^{(1)}$ of the tight span by H^d.)

Theorem 3.3. *For any metric d on X, the following statements are equivalent:*
(i) d is totally split-decomposable,
(ii) $\Gamma_d = T(X,d)^{(1)}$, that is, $V_d = F_0$ and $E_d = F_1$.

Finally, we consider the implications of the above theorem for the subclass of totally split-decomposable metrics known as *consistent* metrics, which in addition satisfy the following *six-point condition*: For every subset Y of X with $|Y| = 6$ there exists some pair of distinct elements $u, v \in Y$ such that

$$0 \leq \max\{d(x,u) + d(y,v), d(x,v) + d(y,u)\} - d(x,y) - d(u,v)$$

for all $x, y \in Y - \{u, v\}$ (and hence for all $x, y \in Y$). This condition can be shown to be equivalent to requiring that (X,d) does not contain an *octahedral split system* as depicted in Figure 3.2. For consistent metrics, it is shown in [DHM02] that the tight span is isomorphic as a polytopal complex to the Buneman complex, and hence that the Buneman graph is isomorphic to $T(X,d)^{(1)}$. Combining this result with our previous theorem immediately gives our final theorem.

Theorem 3.4. *If d is a consistent metric on X, then Γ_d is isomorphic (as a weighted graph, containing X in its vertex set) to the weighted Buneman graph $B_d^{(1)}$. In particular, $B_d^{(1)}$ is an h-optimal realization of d.*

Since the Buneman graph can be easily constructed, this theorem implies that h-optimal realizations can be easily found for the class of consistent metrics. Besides facilitating further studies of the relation

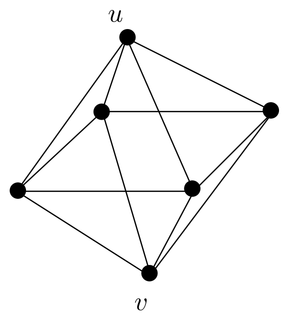


Figure 3.2: The metric d_G induced by this graph corresponds to an octahedral collection of splits, formed by considering the four possible splits of the six vertices of the octahedron into two blocks of size three. If we consider any pair of opposing vertices u and v in this octahedron, we note that $\max\{d_G(x,u) + d_G(y,v), d_G(x,v) + d_G(y,u)\} = 2$ for all vertices x, y in $V(G) \setminus \{u,v\}$, while $d_G(x,y) + d_G(u,v) \geq 3$, and thus the six-point condition is violated.

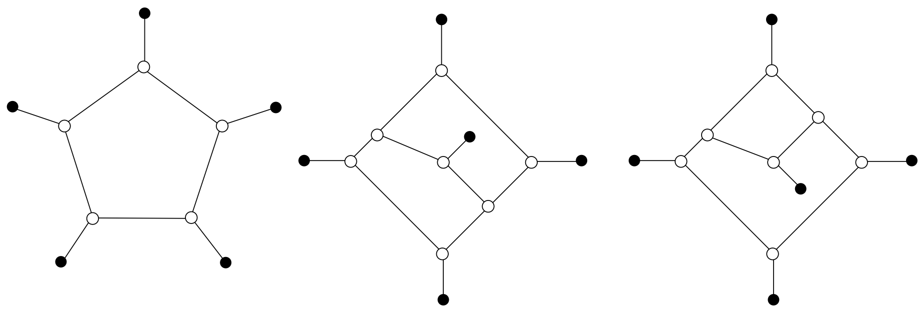
between h-optimal and optimal realizations, this can be useful in phylogenetic analysis since, for metrics constructed from biological data, the non-consistent component has been found in practice to be relatively small, thus subtracting this component can yield a reasonably close approximation of the metric.

3.2 Althöfer's conjecture holds for all metrics on five elements. . . — Paper II [KLM07b]

It is not hard to show that for all metrics on four or fewer elements (which are necessarily consistent) the optimal realization is unique, and equal to the h-optimal realization and to the 1-skeleton of the tight span. In this paper we take the natural next step of investigating optimal realizations of metrics on five elements.

It has been observed previously (see e.g. [SY04]) that the classification of the three types of tight spans of metrics on five points as in Figure 1.11 leads to a subdivision, denoted MF_5, of the cone MET_5 of five-point metrics. We show that MF_5 can be further subdivided into subcones of optimal realizations. We shall call a metric d *generic* if it lies in the interior of a maximum cone in MF_5. We show that there are precisely three nonisomorphic unlabled graphs that can be optimal realizations of generic 5-point metrics, in particular we prove the following:

Theorem 3.5. *Suppose that G is an optimal realization of some generic metric $d \in MET_5$. Then G must be in one of the three isomorphism classes (a)–(c) pictured in Figure 3.3. Moreover, if d is in the interior of a Type I or Type II cone, then G must be in class (a) or class (b),* *respectively, whereas if d is in the interior of a Type III cone then G can be either in class (b) or in class (c).*



(a) (b) (c)

Figure 3.3: Three classes of optimal realizations on $|X| = 5$ points. For each graph the five black vertices correspond to the elements of X.

From the proof of this theorem it also follows that there exist metrics on five points having more than one optimal realization, but that the maximum possible number of optimal realizations for a fixed metric on five points is three (where one is in class (c) above, and the other two are different labelings in class (b)).

It is also straightforward to verify that, for any generic metric on five points, any optimal realization is a subgraph of the corresponding h-optimal realization, and thus that the answer to Althöfer's question is "yes" for these generic 5-point metrics. To fully verify the quite natural assumption that for any metric on at most 5 points Althöfer's question has a positive answer, it thus remains only to consider the non-generic 5-point metrics. This is accomplished by straightforward case-checking, and included in this thesis as an appendix, Paper IIa, to Paper II.

3.3 . . . but not in general — Paper III [KLM07a]

In this paper we show that the answer to Althöfer's question in the general case is "no": there exist metrics with an optimal realization which is an extremal weighting, but which cannot be found by deleting edges in the hereditarily optimal realization.

We begin by proving a theorem that makes it simpler to construct examples of metrics with known optimal realizations, via the *underlying graph*, $UG(d)$, of a given metric d on a set X. This graph $UG(d)$ is defined as the weighted graph with vertex set X, edge set consisting of the pairs $\{x, y\} \in X$ for which $d(x, z) + d(z, y) > d(x, y)$ for all $z \in X$

distinct from x, y, and edge-weighting defined by giving each edge $\{x, y\}$ weight $d(x, y)$. This graph is always a realization of d, and with the aid of Lemmas 2.2 and 2.3 we characterize the metrics d for which $UG(d)$ is an optimal realization, in particular extending a theorem in [ISPZ84]:

Theorem 3.6. *Suppose that (X, d) is a finite metric space. Then $UG(d)$ is an optimal realization of d if and only if the following two conditions hold:*

(i) for every pair of edges $\{x, y\}, \{y, z\}$ in $UG(d)$ with x, z distinct, $d(x, z) = d(x, y) + d(y, z)$.

(ii) for every disjoint pair of edges $\{x, y\}, \{u, v\}$ in $UG(d)$, either $d(x, y) + d(u, v) < \max\{d(x, u) + d(y, v), d(x, v) + d(y, u)\}$ or $d(x, y) + d(u, v) = d(x, u) + d(y, v) = d(x, v) + d(y, u)$.

With the aid of this theorem we construct an infinite family of metrics d_k, where $k \geq 1$, on $|X_k| = 2^{k+2}$ elements, such that $UG(d_k)$ is an optimal realization of (X_k, d_k). For each of these metrics d_k we show that there are several optimal realizations possible, at least one of which cannot be found in the h-optimal realization. Figure 3.4 shows the two optimal realizations of the metric d_1, namely $UG(d_1)$ and the h-optimal realization of d_1. Clearly, the graph $UG(d_1)$ is not a subgraph of the h-optimal realization.

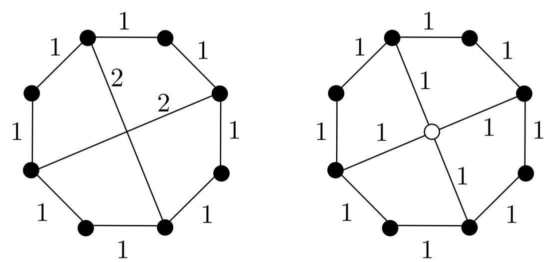


Figure 3.4: Two optimal realizations of the metric d_1; the graph in (a) is the UG-graph of d_1, which is an optimal realization of d_1 by Theorem 3.6, and the graph in (b) is the h-optimal realization Γ_{d_1} of d_1.

Figure 3.5 depicts one of the possible optimal realizations of the metric d_2. Again, this graph is not a subgraph of the h-optimal realization, and this example also disproves a conjecture of Dress [Dre84]: Having proved the existence, for any graph $G = (V, E, w)$ that is an optimal realization of some metric space (X, d), of a map $\psi : V \to T(X, d)$ which satisfies $\psi(x) = d(x, y)(y \in X)$ for all $x \in X$ and $\|\psi(u), \psi(v)\| = w(\{u, v\})$ for all $\{u, v\} \in E$, it was conjectured that this map is always injective, since this would be a necessary condition for the optimal realization to be a subgraph of the tight span.

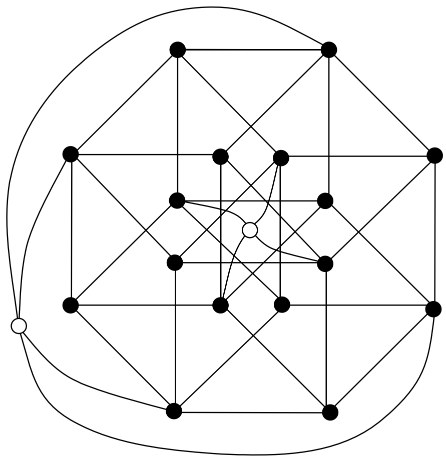


Figure 3.5: An optimal realization of the metric d_2 (all edges have weight 1), which provides a counterexample to the conjecture that any map from the vertices of an optimal realization to the corresponding tight span that preserves the distance d on X and the weights of edges is always injective; the two white auxiliary vertices are mapped to the same element of the tight span $T(X_2, d_2)$.

3.4 Where do we go from here? — Paper IV [Les07]

As the counterexamples in Paper III show, we cannot in general find even an extremal weighting of any optimal realization in the h-optimal realization. But these counterexamples are in a sense pathological; for each of the metrics considered there always exists *some* optimal realization that is a subgraph of the h-optimal realization. So a natural modification of our original question is to ask whether this holds in general. In Paper IV we consider the following:

Conjecture 3.7. *Every finite metric space (X, d) has some optimal realization G obtainable as a subgraph of $T^{(1)}(X, d)$.*

A necessary condition for this conjecture to hold is that the embedding ψ of the vertices of G into $T(X, d)$ as considered in Paper III is indeed injective, so we must introduce some extra requirement on optimal realizations so that injectivity holds.

We define an optimal realization G of a metric space (X, d) to be *extremal* if the number of shortest paths between elements of X in G is maximal, so for instance the graph on the right in Figure 3.4 is an extremal optimal realization.

Defining a *(general) realization* of (X, d) as an isometric embedding $\phi : X \to Y$ into a metric space (Y, d'), and an X*-contraction* of a realization Y ($\phi : X \to Y$) into another realization Y' ($\phi' : X \to Y'$) to be a contraction such that $\psi|_X$ is the identity, we can prove:

Theorem 3.8. *If $G = (V, E, w)$ is an extremal optimal realization of a given metric (X, d), then any X-contraction $\psi : G \to Y$ is necessarily injective.*

In general, injectivity of X-contractions is not sufficient to establish Conjecture 3.7, as shown by i.e. Althöfer's example in Figure 1.14; we must also ensure that the vertices of an optimal realization are mapped to vertices in the tight span. Considering the hypercube complex $H(X, d)$ as in Section 2.4.1, we can show:

Theorem 3.9. *Assume that G is an extremal optimal realization, with minimum number of vertices, which admits an X-contraction ξ of G into $H(X, d)$. Then all vertices in G must map to vertices in $H(X, d)$, i.e. $\xi(V(G)) \subseteq H^{(0)}(X, d)$.*

By [DHM01], for a *2-compatible metric space* (X, d), the tight span $T(X, d)$ is at most 2-dimensional, and *isometric* to the Buneman complex $B(X, d) \subset H(X, d)$. Thus we immediately have:

Corollary 3.10. *If (X, d) is a 2-compatible metric space then Conjecture 3.7 holds, i.e. we can contract the realization G into the realization $T^{(1)}(X)$.*

In the more general case, we do not yet have a good condition for the existence of an X-contraction of G into $H(X, d)$ as in Theorem 3.9, though it would seem reasonable that any totally split-decomposable metric would admit such a contraction. This would be interesting to investigate further.

We also note that the proof of Theorem 3.9 would seem to suggest that if a metric d is split-prime, then d has only a finite number of optimal realizations.

4. Examples of optimal realizations

In this chapter we examine in more detail how optimal realizations of some totally split-decomposable metrics can be calculated. In particular, we will provide examples which demonstate that

- there exist optimal realizations such that the corresponding polytope of edge-weightings is a 3-dimensional cube,
- the difference between the maximal and minimal number of edges in different optimal realizations of the same metric can be arbitrarily large,
- there exists a metric which is not consistent, but where the optimal realizations can still be derived by deleting edges from the h-optimal realization.

It is hoped that these examples will be useful for continued research in this area. To simplify notation, we will often denote a distance $d(x, y)$ simply by xy, when the metric d is clear from context, and any geodesic joining x and y by $\widetilde{xy}$.

4.1 A useful lemma

Lemma 4.1. *If d is a totally split-decomposable metric on a set X, then for every split $A|B$ of X with isolation index α, no geodesic $\widetilde{aa'}$ between any two vertices $a, a' \in A$ can at any point come within distance α of any geodesic $\widetilde{bb'}$ between any two vertices $b, b' \in B$ in any realization of (X, d).*

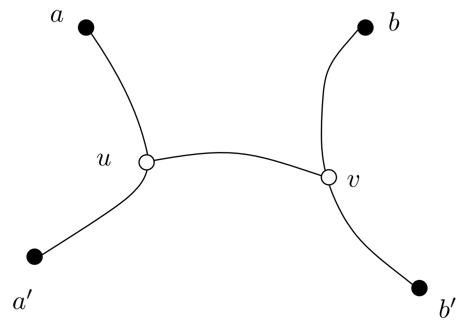
Proof. Let G be a realization of (X, d). Let u be a point on some geodesic $\widetilde{aa'}$ between a pair of vertices in A, and v be on some geodesic $\widetilde{bb'}$ between vertices in B. Since G is connected there must be some path between u and v, having length $d_G(u, v)$.

Hence (writing xy for $d_G(x, y)$) we have

$$ab + a'b' \leq au + uv + vb + a'u + uv + vb' = aa' + bb' + 2uv$$
$$ab' + a'b \leq au + uv + vb' + a'u + uv + vb = aa' + bb' + 2uv$$

and thus

$$\max\{ab + a'b', a'b + ab', aa' + bb'\} \leq aa' + bb' + 2uv.$$



So we have

$$\beta := \max\{ab + a'b', a'b + ab', aa' + bb'\} - aa' - bb' \leq 2uv$$

and

$$\alpha_{A,B} = \min_{a,a' \in A, b,b' \in B} \{\frac{1}{2}\beta\} \leq \min_{a,a' \in A, b,b' \in B} \{uv\}$$

and hence no path between u and v can have length less than α. Since this holds for all u and v the lemma is proved. □

4.2 Optimal realizations of circular metrics

In this section we investigate optimal realizations of the class of metrics known as *circular* metrics, which in turn is a subclass of the class of consistent metrics.

A split $A|B$ is said to be of size k if $\min(\#A, \#B) = k$; if $k = 1$ we call $A|B$ a *trivial* split. A collection $\mathcal{S}$ of splits on a set X is called a *split system.* A split system $\mathcal{S}$ on X is called *circular* if it can be represented by a circular configuration of the elements of X in the plane where all splits are represented by straight lines through the circle. More precisely, $\mathcal{S}$ is circular if the elements of X can be labeled $x_0, x_1, \ldots, x_n = x_0$ so that every split in $\mathcal{S}$ is of the form $\{x_{i+1}, x_{i+2}, \ldots, x_j\}|\{x_{j+1}, \ldots, x_{i-1}, x_i\}$ where we without loss of generality assume $i < j$.

Lemmas 2.2 and 2.3 can be used to prove the following lemma.

Lemma 4.2. *Any realization of a circular split system with no trivial splits must contain a subgraph homeomorphic to a cycle, labeled in accordance with the circular ordering and weights given by the corresponding metric.*

Proof. Given a circular split system with no trivial splits on a set X, for any three consecutive $p, q, r \in X$ there can be no split that separates q from both p and r (since this would be a trivial split). The total weight of all splits separating p and r must then equal the sum of the total weights of the splits that separate p and q and those that separate q and r. Hence Lemma 2.2 implies that q is the only common point of any geodesics $\widetilde{pq}$ and $\widetilde{qr}$ in any realization of d, and so any realization of d must contain a subgraph homeomorphic to the path $p - q - r$. Moreover, Lemma 2.3 implies that for distinct $p, q, r \in X$, $p', q', r' \in X$ these paths do not intersect, so all realizations must contain our desired cycle. □

4.2.1 A metric with three variable parameters

Consider the circular split system on 6 points given by taking all possible nontrivial circular splits, i.e. all six 2-splits and all three 3-splits. We will prove that the realization of the corresponding circular metric $d = \sum_{S \in \mathcal{S}} \delta_s$ shown in Figure 4.1 is optimal, and thus that this metric can be thought of as a generalization of Althöfer's continuum example (Figure 1.14), with three parameters that can vary independently. The polytope corresponding to all optimal weightings of this graph is a 3-cube, and its extremals can be derived by edge-deletion from the corresponding h-optimal realization (which is isomorphic to the 1-skeleton of the tight span since this metric is consistent) as depicted in Figure 4.2.

	a	b	c	d	e	f
a	0	3	6	7	6	3
b	3	0	3	6	7	6
c	6	3	0	3	6	7
d	7	6	3	0	3	6
e	6	7	6	3	0	3
f	3	6	7	6	3	0

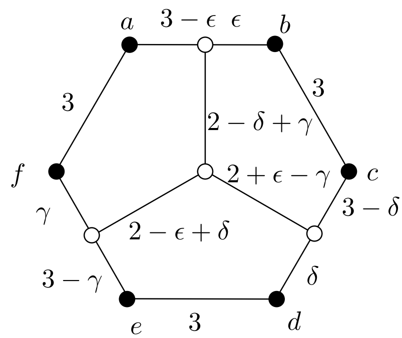


Figure 4.1: The metric corresponding to the six possible circular splits of size 2 and the three possible circular splits of size 3 on {a,b,c,d,e,f}. The graph on the right is an optimal realization for all $1 \leq \gamma, \delta, \epsilon \leq 2$.

Proof. By Lemma 4.2, any realization of d must contain a cycle $a - b - c - d - e - f$ where all edges have weight 3, call it C. To obtain an optimal realization G of d from C we must add geodesics $\widetilde{ad}$, $\widetilde{be}$, and $\widetilde{cf}$, each of length 7. We claim that the total weight added to C by these three geodesics must be at least 6 in any realization of d. Let us first consider a geodesic $\widetilde{ad}$. Lemma 4.1 implies that such a geodesic cannot come within distance 1 of any geodesic $\widetilde{bc}$ or $\widetilde{ef}$, specifically the edges $\{b, c\}$ and $\{e, f\}$ of C. There are two possible cases to consider up to symmetry; either $\widetilde{ad}$ intersects the edges $\{a, b\}$ and $\{d, e\}$, call this Case

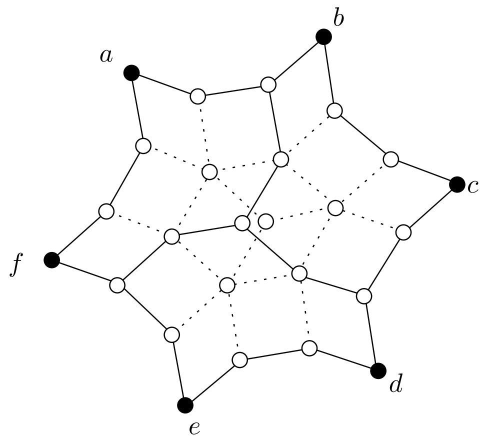


Figure 4.2: The hereditarily optimal realization of the metric in Figure 4.1 with the optimal realization corresponding to $\gamma, \delta, \epsilon = 1$ as a subgraph.

1, or $\widetilde{ad}$ intersects $\{a, b\}$ and $\{c, d\}$, call this Case 2. These two cases are sketched in Figure 4.3.

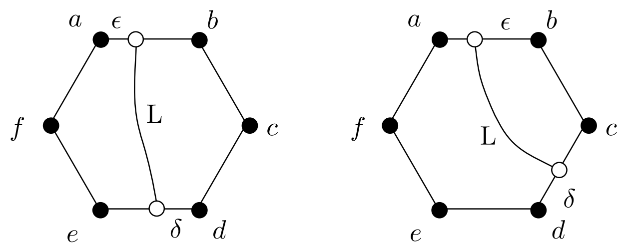


Figure 4.3: The two possible ways (up to symmetry) of adding a geodesic $\widetilde{ad}$. Since $d(a, d) = 7$, we have $L = 7 - \epsilon - \delta$.

For Case 1, we observe that $d(b, e) = 7$ implies that $3 - \epsilon + 7 - \epsilon - \delta + 3 - \delta \geq 7$ and thus $\epsilon + \delta \leq 3$, so the weight of the path L is at least 4 in any such realization of (X, d). For Case 2, L must have weight at least 3 since $3 - \epsilon \leq 2$ and $\delta \leq 2$.

Next we consider adding a path from c to f of length 7 to either of these two graphs. We claim that adding such a path to a graph as in Case 1 results in a graph with weight strictly greater than $\|C\| + 6$, and hence no realization of this type can be optimal.

Up to symmetry, we again have two possible cases, sketched in Figure 4.4. For the case where $\widetilde{cf}$ intersects $\{b, c\}$, we have $M + O + Q \geq 4$ by the reasoning above, and by symmetry also $N + O + P \geq 4$. Next observe that a path from a to c via M and N must have length at least 6, and hence $\epsilon + M + N \geq 4$. (Recall that $\widetilde{cf}$, including the edge of weight N, cannot come within distance 1 of the edge $\{a, b\}$.) Similarly,

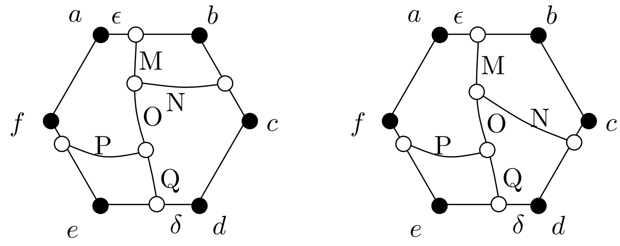


Figure 4.4: The two possible ways of adding a geodesic $\widetilde{cf}$ in Case 1.

$\delta+Q+P \geq 4$. Adding these four inequalities gives $M+N+O+P+Q \geq \frac{1}{2}(16-(\epsilon+\delta)) > 6$. The other case is similar; note that $N+O+P \geq 3$ and $\epsilon+M+N \geq 5$.

Returning to Case 2 in Figure 4.3, we note that adding a path $\widetilde{cf}$ joining the edge pairs $\{b,c\}$ and $\{e,f\}$, $\{c,d\}$ and $\{a,f\}$ or $\{b,c\}$ and $\{a,f\}$ would lead to a situation symmetric to that in Case 1, and hence no such realization can be optimal.

So it remains only to consider adding a path $\widetilde{cf}$ joining the edge pairs $\{c,d\}$ and $\{e,f\}$, as sketched in Figure 4.5. We can assume that this path follows the path $\widetilde{ad}$ for some distance, since otherwise we would have a higher total weight.

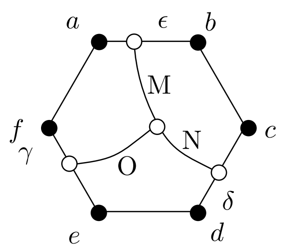


Figure 4.5: Adding a geodesic $\widetilde{cf}$ in Case 2.

Since $d(a,d) = d(b,e) = d(c,f) = 7$ the following inequalities must hold:

$$M+N+3-\epsilon+\delta \geq 7$$
$$M+O+\epsilon+3-\gamma \geq 7$$
$$O+N+\gamma+3-\delta \geq 7.$$

Adding these three inequalities gives $M+N+O \geq 6$ and thus the smallest possible weight added to the cycle C is 6, as claimed. Recalling that $\delta \leq 2$ and $\epsilon \geq 1$, we note that by symmetry we also have $\gamma \leq 2$, $\epsilon \leq 2$ and $\delta, \gamma \geq 1$. Thus the graph in Figure 4.1 is an optimal realization of d as claimed. Recall however that we assumed that $\widetilde{ad}$ intersected $\widetilde{ab}$.

The symmetric possibility, i.e. $\widetilde{ad}$ intersecting $\widetilde{af}$, gives the only other optimal realization of (X, d); the three internal edges are rotated by $\frac{\pi}{3}$. □

4.2.2 Metrics with all s-splits on $n = 2sk$ vertices

Let $|X| = n = 2sk$, where $s, k \geq 2$, and let $d_{s,k}$ be the metric on X corresponding to all circular splits (of equal weight, without loss of generality weight 1) of size s, namely

$$d_{s,k}(x_i, x_j) = \begin{cases} 2|i-j| & \text{for } |i-j| \leq s \\ 2s & \text{else (indices modulo } n). \end{cases}$$

By Lemma 4.2, any optimal realization of $d_{s,k}$ must contain a subgraph homeomorphic to the cycle C_n with all edge weights 2. Clearly, C_n contains geodesics $\widetilde{x_i, x_j}$, for all pairs $x_i, x_j \in X$ with $|i - j| \leq s$ and no others, so a realization of the entire metric must add geodesics $\widetilde{x_i, x_j}$ for all pairs with $|i - j| > s$. In particular, we must add the sk *opposite* geodesics $\widetilde{x_i, x_{i+sk}}$, $i \in \{1, \ldots, sk\}$ (note that $|i - (i + sk)| = sk > s$ since $k \geq 2$). Let us now consider the possible intersection between some opposite geodesic $\widetilde{x_i, x_{i+sk}}$ and the cycle C_n.

Lemma 4.3. *Adding a single opposite geodesic to the cycle C_n contributes an edge weight of at least $2s - 2 = 2(s - 1)$.*

Proof. Since the split system corresponding to $d_{s,k}$ contains the s-split $\{x_{i+1}, \ldots, x_{i+s}\}|X \setminus \{x_{i+1}, \ldots, x_{i+s}\}$ and the path $x_{i+1}, \ldots, x_{i+s}$ along C_n is of length $2s$ and hence a geodesic $\widetilde{x_{i+1}, x_{i+s}}$, Lemma 4.1 implies that $\widetilde{x_i, x_{i+sk}}$ cannot come within distance 1 of x_{i+1}, and hence can intersect C_n by at most a distance of 1. This same reasoning of course applies at the vertex x_{i+sk}. □

So let us consider the possible intersections between two distinct opposite geodesics $\widetilde{x_i, x_{i^*}}$ and $\widetilde{x_j, x_{j^*}}$ where $i^* = i + sk$.

Lemma 4.4. *For two opposite geodesics $\widetilde{x_i, x_{i^*}}$ and $\widetilde{x_j, x_{j^*}}$, if*

$$s \leq |i - j| \leq s(k - 1),$$

then $\widetilde{x_i, x_{i^*}}$ and $\widetilde{x_j, x_{j^*}}$ can intersect in at most one point in any realization of $d_{s,k}$.

Proof. First observe that

$$|i^* - j| = |i - i^*| - |i^* - j| = sk - |i - j|$$

and hence $|i - j| \leq sk - s = s(k - 1)$ is equivalent to $|i^* - j| \geq s$.

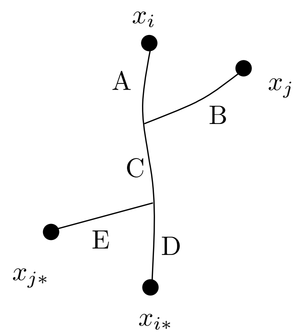


Figure 4.6: The intersection of two opposite geodesics.

So let us consider some pair $\widetilde{x_i, x_{i^*}}$, $\widetilde{x_j, x_{j^*}}$ such that $s \leq |i-j| \leq s(k-1)$. If $\widetilde{x_i, x_{i^*}}$ and $\widetilde{x_j, x_{j^*}}$ intersect, as sketched in Figure 4.6. Hence

$$\begin{aligned}
A + C + D &= 2s && \text{(geodesic } \widetilde{x_i, x_{i^*}}) \\
B + C + E &= 2s && \text{(geodesic } \widetilde{x_j, x_{j^*}}) \\
A + B &\geq 2s && \text{(since } |i-j| \geq s \text{ so } d(x_i, x_j) = 2s) \\
D + E &\geq 2s && \\
B + C + D &\geq 2s && \text{(since } |i^* - j| \geq s) \\
A + C + E &\geq 2s. &&
\end{aligned}$$

Combining these inequalities gives

$$\begin{aligned}
A &= B \\
D &= E \\
A, B, D, E &\geq s \\
C &= 0
\end{aligned}$$

Specifically, any intersection can be in at most one vertex. □

Lemma 4.5. *For two opposite geodesics $\widetilde{x_i, x_{i^*}}$ and $\widetilde{x_j, x_{j^*}}$, if*

$$|i - j| = s - l \text{ where } 1 \leq l < s$$

then $\widetilde{x_i, x_{i^*}}$ and $\widetilde{x_j, x_{j^*}}$ can intersect for a distance of at most $2l$ in any realization of $d_{s,k}$.

Proof. First observe that $|i-j| = s-l$ implies $|i^* - j^*| = s - l$ and $|i^* - j| = |i - j^*| = |i - i^*| - |i - j| = sk - (s - l) = s(k-1) + l \geq s$.

So let us consider some pair $\widetilde{x_i, x_{i^*}}$, $\widetilde{x_j, x_{j^*}}$ such that $|i-j| = s - l$.

Referring again to Figure 4.6, we have

$$
\begin{aligned}
A + C + D &= 2s && \text{(geodesic } \widetilde{x_i, x_{i^*}}\text{)}\\
B + C + E &= 2s && \text{(geodesic } \widetilde{x_j, x_{j^*}}\text{)}\\
A + B &\geq 2(s-l) && \text{(since } |i-j| = s-l \text{ so } d(x_i, x_j) = 2(s-l)\text{)}\\
D + E &\geq 2(s-l) &&\\
B + C + D &\geq 2s && \text{(since } |i^* - j| \geq s\text{)}\\
A + C + E &\geq 2s. &&
\end{aligned}
$$

Combining these inequalities gives

$$
\begin{aligned}
A &= B\\
D &= E\\
A, B, D, E &\geq (s-l)\\
C &\leq 2s - 2(s-l) = 2l
\end{aligned}
$$

□

For small values of s we can now easily construct optimal realizations of $d_{s,k}$. First consider the case $s = 2$, i.e. the metric d corresponding to all splits of size 2 on $n = 4k$ elements. Using the above lemmas it is straightforward to show that optimal realizations of $d_{2,k}$ can be obtained from n-cycles with all edges having weight 2, by joining the midpoints of alternate edges on the cycle by edges of weight 1 to a central auxiliary vertex, as shown in Figure 4.7 for $n = 8$. Hence, for a fixed n, the metric $d_{2,k}$ has precisely two optimal realizations, each of which is indeed a subgraph of the h-optimal realization, since the h-optimal realization is an n-cycle where the midpoint of every edge is joined to the central vertex. Note that these metrics are in addition 2-compatible, and hence this result also follows from Corollary 3.10 in Paper IV.

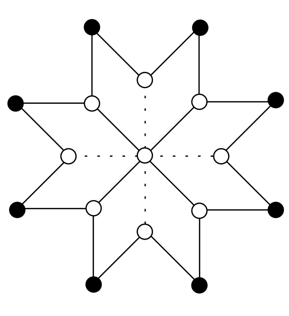


Figure 4.7: An optimal realization of the metric corresponding to all splits of size 2 on $n = 8$ elements as a subgraph of its h-optimal realization.

Perhaps more interesting is the case $s = 3$, i.e. $n = 6k$ for some $k \geq 2$, and $d_{3,k}$ is the metric corresponding to all circular 3-splits on X, namely:

$$d_{3,k}(x_i, x_j) = \begin{cases} 0 & \text{if } i = j \\ 2 & \text{if } |i-j| = 1 \pmod{n} \\ 4 & \text{if } |i-j| = 2 \pmod{n} \\ 6 & \text{else.} \end{cases}$$

Figure 4.8 depicts two optimal realizations of each of the first two metrics in this family, one with $2n$ edges and one with $2n + 2n/3$ edges, hence demonstrating that there can be an arbitrarily large gap between the maximal and minimal number of edges in different optimal realizations of the same metric. Note that these two optimal realizations are not the only possibilities, by combining the two types we observe that any metric in this family has at least $3 \cdot 2^{n/3}$ distinct optimal realizations, all of which can be found as subgraphs of the h-optimal realization.

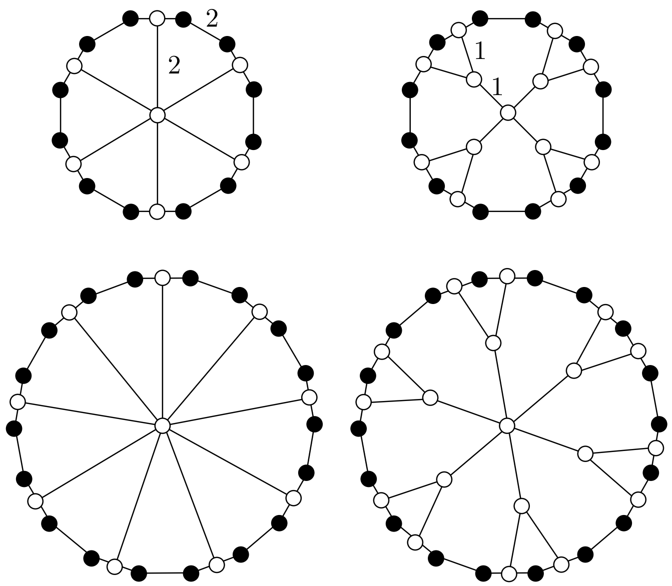


Figure 4.8: Two optimal realizations of $d_{3,k}$, for $k = 2$ (so $n = 6k = 12$) and $k = 3$. Note that the realizations on the left have $n + n/2$ edges on the outer cycle and $n/2$ internal edges (each of weight 2) while the realizations on the right have $n + 2n/3$ edges on the outer cycle and $n/3 + 2n/3 = n$ internal edges.

To prove the optimality of the realizations in Figure 4.8, let $G = (V, E, w)$ be any realization of $d_{3,k}$. By Lemma 4.2, G must contain a subgraph homeomorphic to the cycle $\{x_1, x_2, \ldots, x_n\}$ with all edge weights equal to 2, denote this cycle by C and note that $\|C\| = 2n$.

Next we consider adding the $\frac{n}{2}$ opposite geodesics $\widetilde{x_i x_{i^*}}$, where we without loss of generality assume $i \leq \frac{n}{2}$. We claim that the total weight added to C must be at least n in any realization. Any individual such geodesic must add an additional weight of $2(s-1) = 4$ to C by Lemma 4.3. So it remains to consider possible intersections between opposite geodesics.

By Lemma 4.4, no two opposite geodesics $\widetilde{x_i x_{i^*}}$ and $\widetilde{x_j x_{j^*}}$ with $|i-j| \geq 3$ intersect in more than one point. This in turn implies that no set of four or more opposite geodesics intersects at more than one point, and that any set of three opposite geodesics intersecting at more than one point must consist of three geodesics such that the endpoints x_i, x_j, x_k form a consecutive sequence in X. Moreover, Lemma 4.5 implies that any two consecutive opposite geodesics have a maximum shared weight of 4, while two opposite geodesics with $|i-j| = 2$ have a maximum shared weight of 2. Note that if we assume that no set of three consecutive opposite geodesics shares a nonzero weight, then the total weight added to C must be at least $4 \times \frac{1}{2} \times \frac{n}{2} = n$ and our claim is proved, so assume that there is some such set of three consecutive geodesics.

Observe that since $\widetilde{x_i x_{i^*}}$ can intersect any edge on the outer cycle for a distance of at most 1, there must be internal edges in G along the path from x_i to the shared part of the geodesics with weight at least 1, and similarly for x_k, x_{i^*} and x_{k^*}. So the added edges include a subgraph homeomorphic to a binary tree, as sketched in Figure 4.9, with a minimum weight of $4 \times 1 + 2 = 6$. Hence, any subset of three

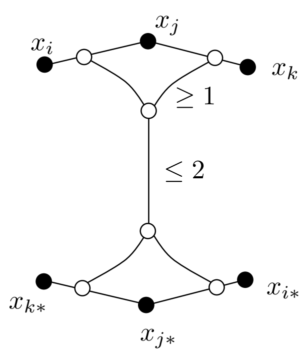


Figure 4.9: The intersection of three opposite geodesics $\widetilde{x_i x_{i^*}}$, $\widetilde{x_j x_{j^*}}$ and $\widetilde{x_k x_{k^*}}$.

intersecting opposite geodesics contributes an internal edge weight of at least 6. Summing over all $(n/2)/3$ such subsets, we have that the total internal edge weight must be at least $6n/6 = n$ as required, and no combination of 2-element subsets and 3-element subsets of opposite geodesics can lead to an added weight of less than n.

For a final example of this type, we refer the reader to the cover of this thesis: the graph depicted there in red is the optimal realization of the metric corresponding to all circular splits of size 4 on $n = 16$ elements, shown as a subgraph of its h-optimal realization.

4.3 The octahedron metric

In this section we consider the optimal realization of the metric depicted in Figure 4.10, which has the octahedron as its underlying graph (as defined in Section 3.3). This is a metric on six points which does not correspond to a circular split system. This metric has been studied previously, see e.g. [HKM05], [SY04], and its tight span is known to be a rhombic dodecahedron (with all edges having weight 1/2).

$$\begin{array}{c} \\ a \\ b \\ c \\ d \\ e \\ f \end{array} \begin{array}{c} \begin{array}{cccccc} a & b & c & d & e & f \end{array} \\ \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 2 \\ 1 & 0 & 1 & 2 & 1 & 1 \\ 1 & 1 & 0 & 1 & 2 & 1 \\ 1 & 2 & 1 & 0 & 1 & 1 \\ 1 & 1 & 2 & 1 & 0 & 1 \\ 2 & 1 & 1 & 1 & 1 & 0 \end{bmatrix} \end{array}$$

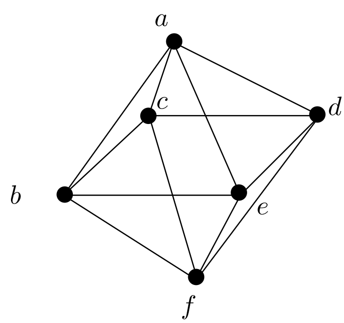


Figure 4.10: The octahedron metric and its UG graph.

Note that this metric is totally split-decomposable, since it can be thought of as the sum of four splits each having weight 1/2:

$$d_{oct} = \frac{1}{2}(\delta_{abc|def} + \delta_{abe|cdf} + \delta_{ade|bcf} + \delta_{acd|bef}).$$

However, by definition d_{oct} is not consistent (as defined in Section 3.1) since it contains an octahedron. Since the metric is totally split-decomposable, Theorem 1 of Paper I implies that its hereditarily optimal realization is isomorphic to the 1-skeleton of the tight span of this metric, i.e. the rhombic dodecahedron.

The octahedron metric has two isomorphic optimal realizations, each of which is a subgraph of this hereditarily optimal realization, as sketched in Figure 4.11.

To prove optimality of these realizations, we begin by noting that since the split system corresponding to d_{oct} contains a split $\{a, b, c\}|\{d, e, f\}$, Lemma 4.1 implies that shortest paths within one block of this split are disjoint from shortest paths in the other block, which in turn implies that any realization of d_{oct} must contain a subgraph homeomorphic to the graph in Figure 4.12.

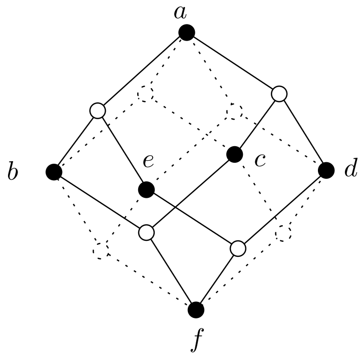


Figure 4.11: Each of the two optimal realizations of the octahedron metric, shown as solid and dotted lines respectively, is a subgraph of the rhombic dodecahedron.

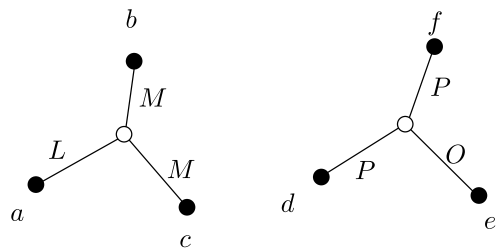


Figure 4.12: The first step in determining the optimal realization of the octahedron metric. When adding a shortest path $\widetilde{ac}$, we observe that we can assume that this path intersects $\widetilde{ab}$ (possibly only at the vertex a) since otherwise we would have a realization of greater weight. Also note that $ab = ac = 1$ implies $L + M = 1$, while $bc = 1$ implies $M \geq \frac{1}{2}$, and similar considerations give $O + P = 1$ and $P \geq \frac{1}{2}$

Now consider adding a shortest path $\widetilde{bf}$. We claim that such a path cannot intersect any edges of the graph in Figure 4.12 at any points other than the vertices b and f. To show this, first note that since $ab+bf = af$, $\widetilde{bf}$ cannot intersect $\widetilde{ab}$ at any point other than the vertex b by Lemma 2.2. Moreover, since $\{a, c, d\}|\{b, e, f\}$ and $\{a, d, e\}|\{b, c, f\}$ are d_{oct}-splits of X, this path is disjoint from any shortest path $\widetilde{ac}$ or $\widetilde{de}$, respectively. It now remains only to show that $\widetilde{bf}$ cannot intersect the edge adjacent to the vertex d in Figure 4.12. Assume the existence of a point x on this edge which is on a shortest path $\widetilde{bf}$, and let $d_{oct}(x, d) = P'$. But then $d_{oct}(b, f) = 1 = Q + P - P' + P$, while $d_{oct}(b, d) = 2$ implies that $Q + P' \geq 2$, which is a contradiction, and hence our claim is proved.

Now let us try to add a path $\widetilde{cf}$. By symmetry of b and c this path cannot intersect any edges in the subgraph in Figure 4.12 at any points other than the vertices c and f. It is possible however that it intersects the new path $\{b, f\}$, and hence any realization must contain a subgraph homeomorphic to the one depicted in Figure 4.13.

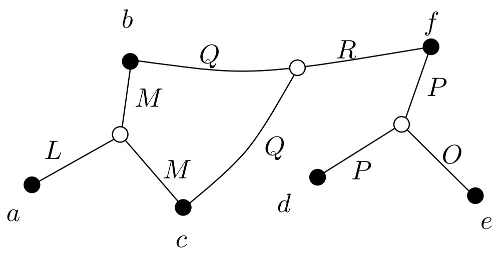


Figure 4.13: Adding paths $\widetilde{bf}$ and $\widetilde{cf}$. Since $\{b, f\}$ and $\{c, f\}$ were assumed to be shortest paths we have $Q + R = 1$, and since $d_{oct}(b, c) = 1$ we have $Q \geq \frac{1}{2}$.

Now consider adding a shortest path $\widetilde{ae}$. Lemma 2.2 implies that such a path is disjoint from any path $\widetilde{ac}$ or $\widetilde{ef}$, and moreover $\{a, d, e\}|\{b, c, f\}$ being a d_{oct}-split implies that it is also disjoint from any paths $\widetilde{bc}$, $\widetilde{bf}$ or $\widetilde{cf}$. We are thus left with two possibilities, shown in Figure 4.14. We will consider Case (a) in detail.

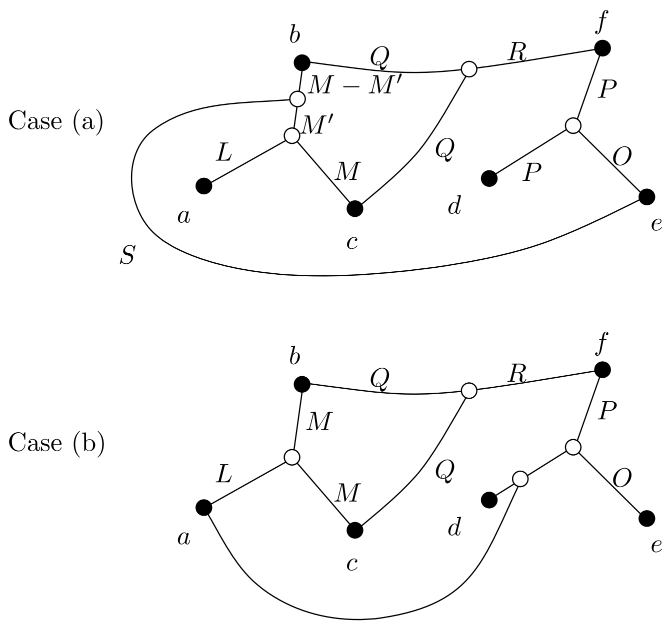


Figure 4.14: Adding a shortest path $\widetilde{ae}$ leads to two cases.

With edge weights as in Case (a) of Figure 4.14, we note that, since $\widetilde{ac}$ and $\widetilde{ae}$ were assumed to be shortest paths and since $d_{oct}(c, e) = 2$, we have the system of relations

$$\begin{aligned} L + M' + S &= 1 \\ L + M &= 1 \\ S + M' + M &\geq 2 \end{aligned}$$

to which the only possible solution is that $L = 0$, $M = 1$ and $S = M - M'$. Redrawing according to these specifications gives us Figure 4.15.

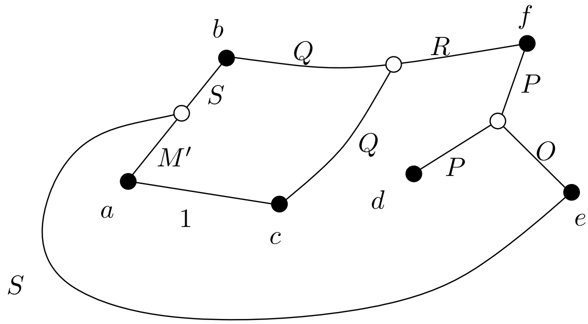


Figure 4.15: Requiring $\widetilde{ae}$ to be a shortest path in this case implies that $\widetilde{ab}$ and $\widetilde{ac}$ do not intersect other than at a.

Now add a path $\widetilde{cd}$. Lemma 2.2 implies that this path does not intersect any paths $\widetilde{de}$ or $\widetilde{bc}$ at any points other than c and d, while various d_{oct}-splits imply that $\widetilde{cd}$ is disjoint from any shortest paths $\widetilde{ab}$, $\widetilde{ae}$, $\widetilde{ef}$ and $\widetilde{bf}$. Hence such a path must be added as shown in Figure 4.16.

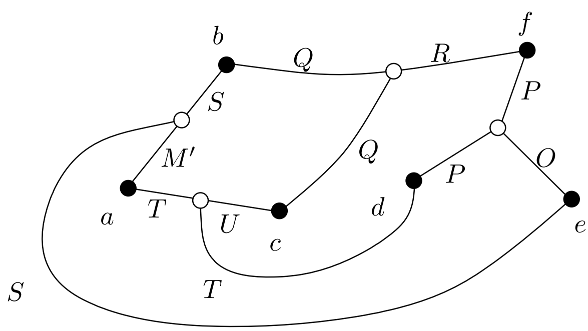


Figure 4.16: Adding a path $\widetilde{cd}$. Similar considerations as in previous cases give bounds $T + U = 1$ and $T \geq \frac{1}{2}$.

The graph in Figure 4.16 has total edge weight

$$M' + O + 2P + 2Q + R + 2S + 2T + U = \frac{9}{2} + P + Q + S \geq 6.$$

Letting every edge have weight $\frac{1}{2}$ gives a realization of d_{oct} with total weight 6, which must therefore be an optimal realization. We have thus arrived at the graph corresponding to the dotted lines in Figure 4.11.

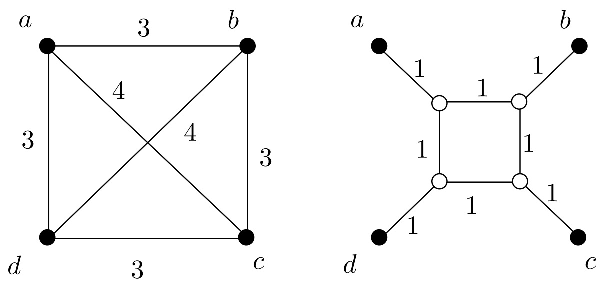
Considering the symmetric possibility of an optimal realization by continuing from Case (b) in Figure 4.14 we obtain the other possible optimal realization, shown as the solid lines in Figure 4.11.

5. Summary in Swedish

Optimala och ärftligt optimala realiseringar av metriska rum

Denna avhandling behandlar problemet att finna en *optimal realisering* av ett givet *metriskt rum*. Ett metriskt rum är en mängd X (i vårt fall oftast ändlig) tilsammans med ett avståndsmått d definierat på varje par x, y av element i X, som är icke-negativt, symmetriskt, samt uppfyller triangelolikheten. En realisering av det metriska rummet (X, d) är en viktad graf $G = (V, E, w)$ som uppfyller att $X \subseteq V$ och där det för varje par av element i X gäller att $d_G(x, y)$, det kortaste avståndet mellan x och y i G, är precis $d(x, y)$. Om summan $\sum_{e \in E} w(e)$ av alla kantvikter i G är minimal bland alla möjliga realiseringar säger vi att G är *optimal*.

$$\begin{array}{c|cccc} & a & b & c & d \\ \hline a & 0 & 3 & 4 & 3 \\ b & 3 & 0 & 3 & 4 \\ c & 4 & 3 & 0 & 3 \\ d & 3 & 4 & 3 & 0 \end{array}$$



Figur 5.1: En metrik d på en mängd X med fyra element, en realisering av (X, d) och den optimala realiseringen av (X, d).

Problemet att finna optimala realiseringar har studerats sedan åtminstone 1960-talet, och tillämpningar finns inom bland annat elektronik och datalogi. Ett viktigt tillämpningsområde som uppstått under de senaste decennierna är *fylogenetisk analys*, eller läran om evolutionärt släktskap. I typfallet tänker vi oss mängden X som ett antal arter och avståndsmåttet d som något biologiskt relevant avstånd mellan dem, exempelvis kan vi jämföra deras DNA-sekvenser för någon viss gen och låta avståndet vara det minsta antalet mutationer som krävs för att omvandla en sekvens till en annan. En optimal realisering motsvarar då ett *fylogenetiskt träd* eller ett *fylogenetiskt nätverk*, där ett fylogenetiskt träd är en graf som representerar släktskapet mellan arterna samt är ett träd i grafteoretisk mening, d.v.s. saknar cykler.

Ofta är det träd som är intressanta inom biologin eftersom vi föreställer oss evolutionen som en process där nya arter sucessivt uppstår genom delning av någon tidigare art, och två arter som en gång delats aldrig återkorsas. Men för vissa organismer, exempelvis växter eller virus, är korsningar mellan arter möjliga och då kan ett nätverk, där vi alltså tillåter cykler, vara en bättre modell. Även i fall där en trädstruktur vore en rimlig hypotes kan ett allmännare nätverk ge insikter om exempelvis hur "träd-likt" indata är.

Sökandet efter optimala realiseringar är dock behäftat med vissa problem. Det har bevisats att det är *NP-svårt* att finna en optimal realisering av en given metrik, det vill säga att det är mycket osannolikt att det går att finna en effektiv algoritm för att lösa problemet. En ytterligare komplikation är att optimala realiseringar inte nödvändigtvis är unika; en viss metrik kan ha flera grafer (med samma sammanlagda kantvikt) som är optimala realiseringar.

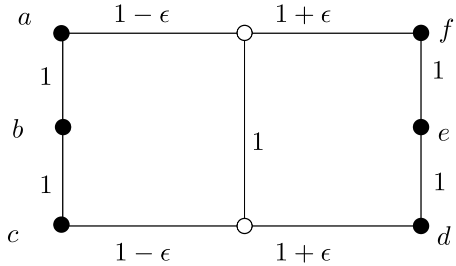
Ett angreppsätt på detta problem bygger på arbeten av Andreas Dress. Han definierar *ärftligt* optimala realiseringar med hjälp av induktion, enligt följande: För en metrik på en mängd X med högst två element så är varje optimal realisering också ärftligt optimal. Om antalet element n i X är större än eller lika med 3, och vi antar att ärfligt optimala realiseringar har definierats tidigare på alla metriker med högst $k < n$ element, så är en realisering $G = (V, E, w)$ ärftligt optimal om det för varje äkta delmängd Y av X med k element finns någon delgraf $G' = (V', E', w \big|_{E'})$ till G sådan att G' är en ärftligt optimal realisering av $(Y, d|_Y)$ och dessutom $\sum_{e \in E} w(e)$ är minimal bland alla sådana grafer. Dessa ärfligt optimala realiseringar är i allmänhet inte optimala utan har större sammanlagd kantvikt, men de har fördelen att de är unika. De kan också beräknas explicit, för alla metriker i princip och för vissa klasser av metriker enkelt i praktiken, genom att de är nära besläktade med det så kallade *strama höljet* av metriken ifråga. Detta strama hölje definieras enligt följande: betrakta först mängden

$$P(X, d) := \{f \in \mathbb{R}^X \mid f(x) + f(y) \geq d(x, y) \text{ for all } x, y \in X\},$$

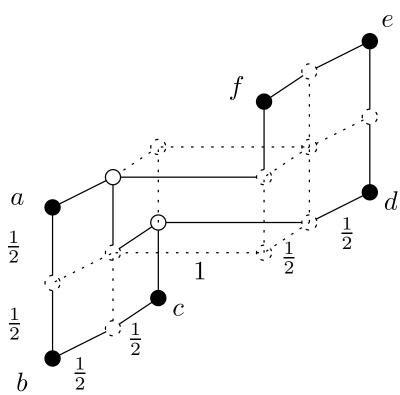
och definiera det strama höljet $T(X, d)$ som de funktioner i $P(X, d)$ som är minimala i den punktvisa partialordningen av $\mathbb{R}^X$ som ges av, för alla $f, g \in \mathbb{R}^X$, $f \leq g \iff f(x) \leq g(x)$ för varje $x \in X$.

Dress förmodade att varje optimal realisering av en given metrik kunde hittas som en delgraf till den ärftligt optimala, något som dock motbevisades av Ingo Althöfer som gav exempel på en metrik med ett kontinuerligt spektrum av optimala realiseringar (figur 5.2), olika viktningar av samma oviktade graf, varav endast ändpunkterna är delgrafer till den ärftligt optimala realiseringen (figur 5.3). Men detta gav förstås genast upphov till en ny fråga: givet en graf G som kan viktas så att den är en optimal realisering av en given metrik, är de extremala viktingarna av G delgrafer till den ärftligt optimala realiseringen?

	a	b	c	d	e	f
a	0	1	2	3	3	2
b	1	0	1	3	4	3
c	2	1	0	2	3	3
d	3	3	2	0	1	2
e	3	4	3	1	0	1
f	2	3	3	2	1	0



Figur 5.2: Althöfers exempel på en metrik med ett kontinuerligt spektrum av optimala realiseringar.



Figur 5.3: De optimala realiseringarna av Althöfers exempel där $\epsilon = \pm\frac{1}{2}$ är delgrafer till den ärfligt optimala realiseringen (streckad i figuren), men inte de övriga optimala realiseringarna.

Denna avhandling innehåller en inledande del som beskriver problemet och dess historiska utveckling, fyra artiklar som behandlar olika aspekter av Althöfer's fråga, samt dessutom ett antal exempel som belyser olika aspekter av optimala realiseringar.

Artikel I — "Ärftligt optimala realiseringar av konsistenta metriker"

Dress, A., Huber, K. T., Lesser, A., Moulton, V. (2006) Hereditarily optimal realizations of consistent metrics

I denna artikel undersöker vi ärftligt optimala realiseringar närmare, och kategoriserar för vilka metriker den ärftligt optimala realiseringen är precis 1-skelettet av det strama höljet. Det visar sig att dessa två grafer är lika om och endast om metriken d är *fullständigt splitt-uppdelningsbar*; vilket i sin tur innebär att den kan skrivas som

$$d = \sum_{A|B \in \mathcal{S}(X)} \alpha_{A|B} \delta_{A|B},$$

där varje $A|B$ är en *splitt*, en partition av mängden X i två icke-tomma delmängder A och B, $\mathcal{S}(X)$ är en mängd av splittar, var och en med någon vikt $\alpha_{A|B} > 0$, och *splitt-metriken* $\delta_{A|B}$ ges av

$$\delta_{A|B}(x,y) = \begin{cases} 0 & \text{om } x,y \in A \text{ eller } x,y \in B \\ 1 & \text{annars,} \end{cases}$$

d.v.s. $\delta_{A|B}(x,y) = 1$ om $A|B$ delar på x och y, och 0 annars.

En underkategori till de fullständigt splitt-uppdelningsbara metrikerna är de *konsistenta* metrikerna, som dessutom uppfyller följande *sexpunkts villkor*: För varje delmängd Y av X med 6 element finns det något par $u, v \in Y$ av distinkta element sådant att

$$0 \leq \max\{d(x,u) + d(y,v), d(x,v) + d(y,u)\} - d(x,y) - d(u,v)$$

för alla $x, y \in Y - \{u, v\}$ (och därmed för alla $x, y \in Y$).

Det är känt att för konsistenta metriker så är det strama höljet isomorft som komplex med det så kallade *Buneman-komplexet*, vars 1-skelett *Buneman-grafen* används inom fylogenetisk analys. Våra resultat innebär även att för konsistenta metriker är Buneman-grafen isomorf med den ärfligt optimala realiseringen.

Artikel II — "Optimala realiseringar av generiska metriker på 5 punkter"

Koolen, J., Lesser, A., Moulton, V. (2007) Optimal realizations of generic 5-point metrics

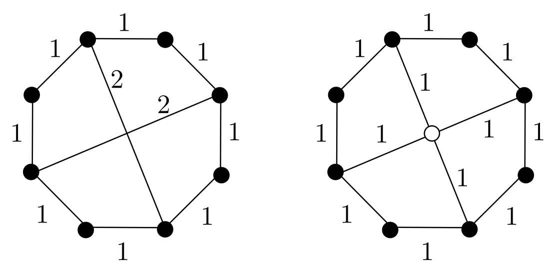
I denna artikel visas att Althöfer's förmodan är sann för alla metriker på fem punkter (och därmed är förmodan visad för alla metriker med högst fem punkter). Det är tidigare känt att konen av alla metriker på fem punkter, som betecknas MET_5, kan delas upp i totalt 102 delkoner baserat på strukturen hos det strama höljet, och att det bland dessa 102 delkoner finns tre symmetriklasser, det vill säga tre *generiska* typer av metriker. Vi visar att denna uppdelning i delkoner kan förfinas ytterligare, genom att beräkna de möjliga optimala realiseringarna för var och en av de tre generiska metriktyperna.

Av denna klassificering framgår bland annat att en metrik på fem punkter kan ha högst tre möjliga optimala realiseringar, som var och en är en delgraf till motsvarande ärftligt optimala realisering. Till artikeln hör också ett appendix, enbart publicerat i denna avhandling, där optimala realiseringar av samtliga övriga, icke-generiska, metriker på fem punkter beräknas, och det verifieras att dessa också är delgrafer av motsvarande ärftligt optimala realiseringar.

Artikel III — "Angående sambandet mellan ärfligt optimala realiseringar av metriska rum och det strama höljet"
Koolen, J., Lesser, A., Moulton, V. (2007) Concerning the Relationship Between Realizations and Tight Spans of Finite Metrics

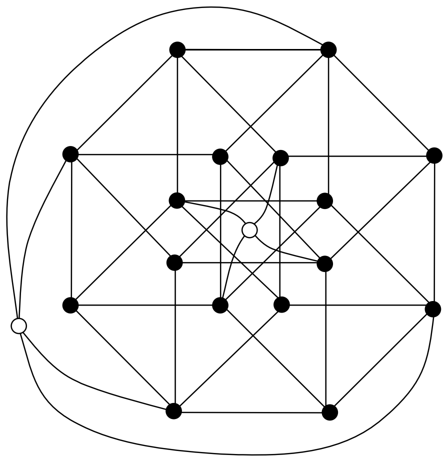
I denna artikel visas att svaret på Althöfer's fråga i det allmänna fallet är "nej". Först visar vi en sats som gör det möjligt att konstruera exempel på optimala realiseringar: För en metrik d på en mängd X definieras den *underliggande grafen* (UG) som den graf med hörnmängd X, kanter mellan par $\{x, y\}$ av hörn sådana att $d(x, y) < d(x, z)+d(z, y)$ för alla $z \neq x, y$ i X och där varje kant $\{x, y\}$ ges vikten $d(x, y)$. Vi ger nödvändiga och tillräckliga villkor för när grafen UG är en optimal realisering.

Med hjälp av ovanstående kan vi konstruera en familj av metriker som var och en har flera optimala realiseringar, och där vissa av dessa inte går att finna som delgrafer i motsvarande ärftligt optimala. Figur 5.4 och figur 5.5 visar de första två metrikerna i denna familj.



Figur 5.4: Två optimala realiseringar av samma metrik. Grafen (a) är UG-grafen, och (b) är den ärftligt optimala realiseringen. Eftersom (a) inte är en delgraf av (b) har vi alltså visat att svaret på Althöfer's fråga är "nej".

Dessa exempel visar även att inbäddningen av hörnen av en optimal realisering in i den ärftligt optimala realiseringen inte nödvändigtvis är injektiv, något som motsäger en tidigare förmodan av Dress och visar att sambandet mellan optimala och ärfligt optimala realiseringar kan vara ytterligare mer komplicerat än man tidigare trott.



Figur 5.5: En optimal realisering av nästa metrik i serien med motexempel till Althöfer's fråga. De två vita hörnen avbildas på samma element i det strama höljet, och därmed på samma element i den ärfligt optimala realiseringen, så inbäddningen är alltså inte injektiv.

Artikel IV — "Extremala optimala realiseringar"

Lesser, A. (2007) Extremal optimal realizations

I denna artikel undersöks en svagare version av den ursprungliga frågan: finns det alltid *någon* optimal realisering som går att finna som en delgraf av den ärftligt optimala?

Vi definierar först en *extremal* optimal realisering av ett metriskt rum (X, d) som en optimal realisering som innehåller maximalt antal kortaste stigar mellan elementen i X, så i Figur 5.4 är den högra grafen extremal. Vi visar att för sådana realiseringar är inbäddningen av hörnen in i det strama höljet alltid injektiv. Dessutom visas att om det finns en kontraktion från en extremal optimal realisering G in i Buneman-komplexet $B(X, d)$ så avbildas samtliga hörn i G på hörn i $B(X, d)$.

Det är tidigare känt att $B(X, d)$ och $T(X, d)$ är isometriska om och endast om d är fullständigt splitt-uppdelningsbar och dessutom $T(X, d)$ har dimension 2, det vill säga d är 2-*kompatibel*. I kombination med resultaten ovan får vi omedelbart att för 2-kompatibla metriker är varje extremal optimal realisering en delgraf av den ärftligt optimala.

Förhoppningsvis kan dessa resultat vara till hjälp även för senare studier av det mer allmänna fallet med fullständigt splitt-uppdelningsbara metriker.

6. Acknowledgments

Many thanks are due to…

…my network of advisors. In chronological order; Vincent Moulton, for taking me on as a PhD student in the fall of 2002 and handing me a photocopied article with "here is the question" scribbled in the margin, and for good work together both before and after his move to East Anglia in the summer of 2004; Pierre Flener, my original co-advisor, for support, good advice and a great project idea which worked out well without me; Svante Janson, who took over as my formal main advisor after Vincent's move, for mathematical advice and general support; Anders Johansson, who has been my informal main advisor since 2005, for many stimulating mathematical discussions, particularly concerning Paper IV; and Lars-Erik Persson, who has been my mentor and co-advisor since 2006, for generously sharing his extensive experience on everything related to doing a mathematics PhD.

…my co-authors Katharina Huber, Andreas Dress and Jack Koolen.

…everyone at the Linnaeus Centre for Bioinformatics, where this research was undertaken; our (regrettably short-lived) mathematics research group: Vincent, Katharina, Stefan Grünewald, Eva Freyhult and Johan Kåhrström; and the rest of the LCB's interdisciplinary and international researchers for a friendly and stimulating work environment where there are always opportunities for interesting conversation (unless the topic is football).

…the department of Mathematics, where I have been enrolled as a PhD student, for always making me feel welcome there as well.

…the BMC computing department, especially Emil Lundberg for many hours spent installing uncooperative math software packages.

…the teachers who have inspired and encouraged me in the past, especially Jörgen Backelin and Dagny Lorendahl.

…all my friends; those who also can put "Dr." in front of their names, or soon will be able to, for knowing what it's like, and everyone else for showing me how many other things there are that are worth doing.

…and of course, to my parents Frances and Martin Lesser for their love, support, encouragement and for letting me get my doctoral degree in any subject I wanted; to my grandparents Francine and Murray Stone for always believing that I could do *anything* I wanted; and to my husband Jonas Lesser, for all of the above and so much more.

References

[Alt88] Ingo Althöfer, *On optimal realizations of finite metric spaces by graphs*, Discrete and Computational Geometry **3** (1988), 103–122.

[Bar89] J. P. Barthelemy, *From copair hypergraphs to median graphs with latent vertices*, Discrete Mathematics **76** (1989), 9–28.

[BD92] Hans-Jürgen Bandelt and Andreas Dress, *A canonical decomposition theory for metrics on a finite set*, Advances in Mathematics **92** (1992), 47–105.

[BFSR95] Hans-Jürgen Bandelt, Peter Forster, Bryan C. Sykes, and Martin B. Richards, *Mitochondrial portraits of human population using median networks*, Genetics **141** (1995), 743–753.

[Bun71] Peter Buneman, *The recovery of trees from measures of dissimilarity*, Mathematics in the Archeological and Historical Sciences (D.G. Kendall and P. Tautu, eds.), Edinburgh University Press, 1971, pp. 387–395.

[Bun74] Peter Buneman, *A note on the metric properties of trees*, Journal of Combinatorial Theory. Series B **17** (1974), 48–50.

[CGGS01] Fan Chung, Mark Garrett, Ronald Graham, and David Shallcross, *Distance realization problems with applications to internet tomography*, Journal of Computer and System Sciences **63** (2001), 432–448.

[CL94] Marek Chrobak and Lawrence L. Larmore, *Generosity helps or an 11-competitive algorithm for three servers*, Journal of Algorithms **16** (1994), no. 2, 234–263.

[DHHM97] Andreas Dress, Michael Hendy, Katharina T. Huber, and Vincent Moulton, *On the number of vertices and edges in the buneman graph*, Annals of Combinatorics **1** (1997), 329–337.

[DHLM06] Andreas Dress, Katharina T. Huber, Alice Lesser, and Vincent Moulton, *Hereditarily optimal realizations of consistent metrics*, Annals of Combinatorics **10** (2006), no. 1, 63–76.

[DHM01] Andreas Dress, Katharina T. Huber, and Vincent Moulton, *Totally decomposable metrics of combinatorial dimension two*, Annals of Combinatorics **5** (2001), 99–112.

[DHM02] Andreas Dress, Katharina T. Huber, and Vincent Moulton, *An explicit computation of the injective hull of certain finite metric spaces in terms of their associated buneman complex*, Advances in Mathematics **168** (2002), 1–28.

[Dob74] Annette J. Dobson, *Unrooted trees for numerical taxonomy*, Journal of Applied Probability **11** (1974), 32–42.

[Dre84] Andreas Dress, *Trees, tight extensions of metric spaces, and the cohomological dimension of certain groups: A note on combinatorial properties of metric spaces*, Advances in Mathematics **53** (1984), 321–402.

[Dre89] Andreas Dress, *Towards a classification of transitive group actions on finite metric spaces*, Advances in Mathematics **74** (1989), 163–189.

[Fel] Joseph Felsenstein, *Phylogeny programs*, `http://evolution.genetics.washington.edu/phylip/software.html`.

[FM67] Walter M. Fitch and Emanuel Margoliash, *Construction of phylogenetic trees*, Science **155** (1967), 279–284.

[GJ79] Michael R. Garey and David Stifler Johnson, *Computers and intractability: A guide to the theory of np-completeness*, Freeman, New York, 1979.

[GJ05] Ewgenij Gawrilow and Michael Joswig, *Geometric reasoning with polymake*, `arXiv:math.CO/0507273`, 2005.

[HB06] Daniel H. Huson and David Bryant, *Application of phylogenetic networks in evolutionary studies*, Molecular Biology and Evolution **23** (2006), no. 2, 254–267.

[HKM05] Katharina T. Huber, Jacobus H. Koolen, and Vincent Moulton, *The tight span of an antipodal metric space — part i: Combinatorial properties*, Discrete Mathematics **303** (2005), 65–79.

[HLP+02] Katharina T. Huber, Michael Langton, David Penny, Vincent Moulton, and Michael Hendy, *Spectronet: A package for computing spectra and median networks*, Applied Bioinformatics **1** (2002), no. 3, 159–161.

[HMLD01] Katharina T. Huber, Vincent Moulton, Pete Lockhart, and Andreas Dress, *Pruned median networks: a technique for reducing the complexity of median networks*, Molecular Phylogenetics and Evolution **19** (2001), 302–310.

[HY64] Seifollah Louis Hakimi and Stephen Sik Sang Yau, *Distance matrix of a graph and its realizability*, Quarterly of Applied Mathematics **22** (1964), 305–317.

[IS72] Wilfried Imrikh and E. D. Stotskii, *Optimal imbeddings of metrics in graphs*, Siberian Mathematical Journal **13** (1972), no. 3, 382–387, Translated from Sibirskii Matematicheskii Zhurnal, Vol. 13, No. 3, pp. 558–565, May–June, 1972.

[Isb65] John R. Isbell, *Six theorems about injective metric spaces*, Commentarii Mathematici Helvetici **39** (1965), 65–76.

[ISPZ84] Wilfried Imrich, J. M. S. Simoes-Pereira, and Christina M. Zamfirescu, *On optimal embeddings of metrics in graphs*, Journal of Combinatorial Theory. Series B **36** (1984), 1–15.

[Kar72] Richard M. Karp, *Reducibility among combinatorial problems*, Complexity of Computer Computations (Raymond E. Miller and James W. Thatcher, eds.), Plenum Press, 1972, pp. 85–103.

[Kåh05] Johan Kåhrström, *Two results in t-theory*, Licentiate Thesis, Uppsala University, 2005.

[KLM07a] Jack Koolen, Alice Lesser, and Vincent Moulton, *Concerning the relationship between realizations and tight spans of finite metrics*, Journal of Discrete and Computational Geometry, to appear (2007).

[KLM07b] Jack Koolen, Alice Lesser, and Vincent Moulton, *Optimal realizations of generic 5-point metrics*, Isaac Newton Institute preprint NI07070, submitted (2007).

[LCL$^+$04] Ming Li, Xin Chen, Xin Li, Bin Ma, and Paul M. B Vitanyi, *The similarity metric*, IEEE Transactions on Information Theory **50** (2004), 3250–3264.

[Les07] Alice Lesser, *Extremal optimal realizations*, Uppsala University Department of Mathematics preprint 2007:53, 2007.

[Mea83] C. A. Meacham, *Theoretical and computational considerations of the compatibility of qualitative taxonomic characters*, Numerical Taxonomy (J. Felsenstein, ed.), Springer, Berlin, 1983, pp. 304–314.

[PH72] A. N. Patrinos and S. L. Hakimi, *The distance matrix of a graph and its tree realization*, Quarterly of Applied Mathematics **30** (1972), 255–269.

[SP69] J. M. S. Simoes-Pereira, *A note on the tree realizability of a distance matrix*, Journal of Combinatorial Theory **6** (1969), 303–310.

[SS03] Charles Semple and Mike Steel, *Phylogenetics*, Oxford University Press, 2003.

[SY04] Bernd Sturmfels and Josephine Yu, *Classification of six-point metrics*, Electronic Journal of Combinatorics **11** (2004), #R44.

[Win88] Peter Winkler, *The complexity of metric realization*, SIAM Journal on Discrete Mathematics **1** (1988), 552–559.

[Zar65] K. A. Zaretskii, *Constructing trees from the set of distances between pendant vertices*, Uspehi Matematiceskih Nauk **20** (1965), 90–92, (in Russian).

[Zie95] Günter M. Ziegler, *Lectures on polytopes*, Springer, New York, 1995.