UNIVERSITÀ DI PISA

DIPARTIMENTO DI MATEMATICA
CORSO DI LAUREA TRIENNALE IN MATEMATICA

TESI DI LAUREA TRIENNALE

Split Decomposition

Properties and Algorithms of the Split Decomposition Method in Phylogenetics

Relatore:
Roberto Grossi

Candidato:
Alessandro Moretti

Correlatore:
Veronica Guerrini

ANNO ACCADEMICO 2023/2024

Contents

Introduction

[...] *a fundamental problem that has been of interest since Charles Darwin first proposed the theory of evolution. Namely, how can one use the present-day characteristics of a group of species to infer, in their evolution from a common ancestor, the historical relationships between these species? Typically, these historical relationships are represented by an evolutionary (phylogenetic) tree and such representations were already suggested by Darwin in the nineteenth century. Determining this tree for different groups of species, or groups of populations, is fundamental to many questions in evolutionary biology as well as for related areas such as conservation genetics and epidemiology.* [...]

Over time these questions have been studied from many perspectives, particularly as the types of data available have increased. Initially, comparisons of the physical characteristics of species, such as their morphology or physiology, gave clues to their evolutionary relationships. However, there are processes that can mislead simplistic inferences. For example, the same characteristic can evolve independently in unrelated species (convergent evolution), or a characteristic can evolve and later disappear (reverse transitions). [...]

From the biologist's perspective, the field was revolutionized by the arrival of molecular data. This began with protein sequences in the late 1960s, genetic (DNA and RNA) sequences in the late 1970s, and most recently whole genome data. The abundance of these data has led to the resolution of many outstanding problems in biology, along with extensive revision in what had previously been believed. [...] – [SS03, Preface]

Various methods have been proposed to tackle the problem of phylogenetic reconstruction.

One approach is that of sequence-based methods, whose main representatives are maximum parsimony, maximum likelihood and Bayesian inference.
These methods deal directly with the sequences and are typically of probabilistic nature.

Another approach is constituted by distance-based methods, which instead operate on a distance matrix obtained by pairwise dissimilarities between the sequences. Examples are UPGMA and Neighbor-joining.

All the mentioned methods have one thing in common: they all produce a tree.

> *By definition, phylogenetic trees are well suited to represent evolutionary histories in which the main events are speciations (at the internal nodes of the tree) and descent with modification (along the edges of the tree). But such trees are less suited to model mechanisms of reticulate evolution, such as horizontal gene transfer, hybridization, recombination or reassortment. Moreover, mechanisms such as incomplete lineage sorting, or complicated patterns of gene duplication and loss, can lead to incompatibilities that cannot be represented on a tree. Although the analysis of individual genes or short stretches of genomic sequence often gives strong support to a phylogenetic tree, different genes or sequence segments usually support different trees.* – [HRS11, Preface]

There are some biological phenomena that lead to think that in several cases evolution is not best described by a tree.
Moreover, even in cases where a tree representation is suitable, there may be noise or incompatibilities within the data that makes it necessary to approximate the real tree.
Or we may be able to construct a lot of different trees with different methods, each giving some information about the evolutionary history; these trees may be incompatible with each other, thus it arises the problem of resolving these conflicts (usually this is done by a consensus tree).

In 1992, Bandelt and Dress proposed a method that try to solve these issues. In particular it is a non-approximative method that allows to display the inconsistencies in the data (through so called splits) and to what extent they are supported (by assigning an index or coefficient). Most notably, the output can be represented as a network (that is not necessarily a tree).

> *Phylogenetic analysis of molecular sequence data often is carried out by first calculating pairwise similarity coefficients, converting these into evolutionary distances, and finally applying some distance-matrix method in order to estimate an unrooted phylogenetic tree. Goodness-of-fit would be judged by comparing the evolutionary distances with the additive distances read off the estimated tree. So, data are fit to a best (or at least, near-optimal) tree, whether or not they bear any resemblance with additive tree data. In practice, one tries to avoid methodological artifacts by applying different tree approximation methods (some operating on sequence data, others using derived distances) and then putting up with a strict consensus tree. Still, one may fall into the trap of systematic error when the methods are subject to the same bias and all disguise true phylogenetic relationships.*
>
> *We therefore propose to accompany any phylogenetic analysis by a nonapproximative method as well that allows for conflicting alternative grouping (to some extent) and hence is able to detect some of those distinctive minor features in distance data which are dominated by others and not supported by estimated trees. This goal can be achieved by split decomposition, developed by Bandelt and Dress (1992), which may be regarded as a kind of factor analysis for distance matrices. It decomposes any dissimilarity matrix d into a number of "binary factors," described as "splits" weighted by "isolation indices," plus a residual indecomposable term (here interpreted as noise). For phylogenetic analysis split decomposition serves two purposes: (a) to exhibit tentative phylogenetic relationships even when they are overridden by parallel events, and (b) to detect groupings brought about by pronounced convergence or systematic error.* – [BD92b]

Introduction

This thesis focuses on the joint article, [BD92a], that builds the mathematical foundations for the split decomposition method.

I will explain the theory of split decomposition, as delineated in the original article, with some additions. In particular, I elaborated more on the omitted details, expanded the introductory chapter and added some results which were not present in the article.

In Chapter 1 we introduce one of the most important type of distance function and present some of its geometric properties.
In Chapter 2 we state the principal definitions of the theory and the intermediate results leading to the canonical decomposition theorem.
In Chapter 3 we see some properties of the set of splits (the main object of the theory) obtained by the split decomposition method.
In Chapter 4 we investigate the properties of a special class of functions, called totally decomposable, whose decomposition is particularly nice.
In Chapter 5 we show the application of the theory to the problem of phylogenetic reconstruction.
In Chapter 6 we analyze in more detail the algorithm of the split decomposition method.

As a final note, we want to mention that there are good reasons to be interested in the theory behind these methods: the wealth of data available today forces us to rely on automated processing, thus it is desirable to have guarantees about their soundness and robustness; phylogenetic techniques (in particular distance-based methods) are very flexible and find application even outside phylogenetics, thus it is necessary to have a theory that prescinds by the specific application; thinking about the challenges of the problems proposed by phylogenetics (and more in general, by biology) gives inspirations to explore new areas of mathematics [Coh04; Stu05].

> *Today, the field of phylogenetics—the reconstruction and analysis of phylogenetic trees and networks—is a flourishing area of interaction between mathematics, statistics, computer science, and biology.*
>
> [...] *Applications of these techniques extend well beyond 'reconstructing the past', and the methods are applied in areas such as epidemiology to investigate the origins, relationships, and future development of viruses such as influenza and HIV. Other areas where phylogenetic methods have found applications include ecology (for classifying new species), medicine, and some quite different areas of classification such as linguistics and cognitive psychology.*
>
> *Our interest in this book is the mathematical foundations of phylogenetics. These foundations date back at least to the pioneering work by Peter Buneman, David Sankoff, and others in the early 1970s. Curiously, Buneman's early paper (1971) dealt not with biology but rather with reconstructing the copying history of manuscripts. The data-driven expansion of phylogenetics during the 1980s and 1990s has led to the need for further mathematical development.* [...] – [SS03, Preface]

Part I

Theory

Chapter 1

Preliminaries

Let X be a set.

Definition (**pseudo-metric**)

A function $d : X \times X \to \mathbb{R}$ is a **pseudo-metric** on X if

- $d(x, x) = 0\,,$ $\qquad\qquad \forall\, x \in X$
- $d(x, y) \leq d(x, z) + d(y, z)\,,$ $\qquad \forall\, x, y, z \in X$

that is, it vanishes on the diagonal $\Delta_X = \{\, (x, y) \in X \times X \mid x = y \,\}$ and it satisfies the triangle inequality.

Definition (**metric**)

A function $d : X \times X \to \mathbb{R}$ is a **metric** on X if

- $d(x, y) = 0 \iff x = y\,,$ $\qquad \forall\, x, y \in X$
- $d(x, y) \leq d(x, z) + d(y, z)\,,$ $\qquad \forall\, x, y, z \in X$

In particular, a metric is a pseudo-metric that vanishes only on the diagonal.

Proposition 1.1

If $d : X \times X \to \mathbb{R}$ is a pseudo-metric, then

- $d(x, y) = d(y, x)\,, \qquad \forall\, x, y \in X$
- $d(x, y) \geq 0\,, \qquad \forall\, x, y \in X$

that is, it is a symmetric and non-negative function.

Proof

From triangle inequality on x, y, x and on y, x, y

$$d(x, y) \leq \cancel{d(x, x)} + d(y, x)$$
$$d(y, x) \leq \cancel{d(y, y)} + d(x, y)$$

that implies $d(x, y) = d(y, x)\,.$

From triangle inequality on x, x, y

$$0 = d(x, x) \leq d(x, y) + d(x, y) = 2\, d(x, y)$$

that implies $d(x, y) \geq 0\,.$

□

Let V be a vector space over $\mathbb{R}$.

Definition (convex set)

A subset $C \subseteq V$ is **convex** if $\forall\, v, w \in C$

$$(1 - \lambda)v + \lambda w \in C\,, \quad \forall\, \lambda \in [0, 1]$$

Definition ((linear) cone)

A subset $C \subseteq V$ is a **(linear) cone** if

$$v \in C \implies \lambda v \in C, \quad \forall\, \lambda \geq 0$$

Notation $\mathbb{R}^{X \times X} := \{\, f : X \times X \to \mathbb{R} \mid f \text{ function} \,\}$

$M(X) := \{\, d : X \times X \to \mathbb{R} \mid d \text{ pseudo-metric} \,\} \subseteq \mathbb{R}^{X \times X}$

Fact $\mathbb{R}^{X\times X}$ is a (real) vector space.

Proposition 1.2

$M(X)$ is a convex cone.

Proof

Let $d \in M(X)$ pseudo-metric on X and $\lambda \geq 0$. Then

- $d(x,x) = 0 \implies \lambda d(x,x) = 0\,, \quad \forall\, x \in X$
- $d(x,y) \leq d(x,z) + d(y,z) \implies$
  $\lambda d(x,y) \leq \big(d(x,z) + d(y,z)\big) = \lambda d(x,z) + \lambda d(y,z)\,,$
  $\forall\, x,y,z \in X$

So λd is a pseudo-metric, $\lambda d \in M(X)$, and $M(X)$ is a cone.

To show that $M(X)$ is convex,
we need to show that $\forall\, d_1, d_2 \in M(X)$

$$(1-\lambda)d_1 + \lambda d_2 \in M(X)\,, \quad \forall\, \lambda \in [0,1]$$

But since $M(X)$ is a cone,

$$(1-\lambda)d_1 \in M(X) \quad \text{and} \quad \lambda d_2 \in M(X)$$

so it suffice to show that $M(X)$ is closed under addition.

Let $d_1, d_2 \in M(X)$. Then

- $(d_1+d_2)(x,x) = \underbrace{d_1(x,x)}_{=0} + \underbrace{d_2(x,x)}_{=0} = 0\,, \quad \forall\, x \in X$
- $(d_1+d_2)(x,y) = d_1(x,y) + d_2(x,y) \leq$
  $\leq d_1(x,z) + d_1(y,z) + d_2(x,z) + d_2(y,z) =$
  $= d_1(x,z) + d_2(x,z) + d_1(y,z) + d_2(y,z) =$
  $= (d_1+d_2)(x,z) + (d_1+d_2)(y,z)\,, \quad \forall\, x,y,z \in X$

So $d_1 + d_2$ is a pseudo-metric in X.

□

Remark The set of all metrics is also a non-pointed convex cone
(same definition of cone but with $\lambda > 0$):
in fact, the zero function is not a metric.

Let us consider $\mathbb{R}^{X\times X}$ with the topology of pointwise convergence τ_p .

Proposition 1.3

If X is countable[1], then $M(X)$ is closed in $\left(\mathbb{R}^{X\times X}, \tau_p\right)$.

Proof

Let us show that $M(X)$ is sequentially closed.

Let us consider a convergent sequence of pseudo-metrics

$$d_n \to \overline{d}\ , \qquad d_n \in M(X),\ \forall\, n \in \mathbb{N}$$

and let us show that the limit is also a pseudo-metric,

$$\overline{d} \in M(X)$$

For the characterization of the pointwise convergence,

$$d_n \to \overline{d} \quad \Longleftrightarrow \quad d_n(\boldsymbol{x}) \to \overline{d}(\boldsymbol{x}), \quad \forall\, \boldsymbol{x} \in X \times X$$

Now for every $\boldsymbol{x} \in X \times X$,
$\{d_n(\boldsymbol{x})\}_{n\in\mathbb{N}}$ is a real-valued sequence, so $\forall\, x, y, z \in X$

- $d_n(x,x) = 0,\ \forall\, n \in \mathbb{N} \quad \Longrightarrow \quad \lim_{n\to+\infty} d_n(x,x) = \overline{d}(x,x) = 0$
- $d_n(x,y) \le d_n(x,z) + d_n(y,z),\ \forall\, n \in \mathbb{N}$, so in the limit
  $\overline{d}\,(x,y) \le\ \overline{d}\,(x,z) +\ \overline{d}\,(y,z)$

This shows that $\overline{d}$ is a pseudo-metric and $M(X)$ is sequentially closed.

Since X is countable (and so is $X \times X$), the space $\mathbb{R}^{X\times X} = \prod_{\boldsymbol{x}\in X\times X} \mathbb{R}$
is a countable product of first-countable spaces (namely $\mathbb{R}$),
so it is first-countable itself.

Now $M(X)$ is sequentially closed in a first-countable space,
so it is closed.

□

[1] Here countable means finite or countably infinite (*al più numerabile*).

Remark If X is uncountable,
then we cannot conclude that $\mathbb{R}^{X\times X}$ is first-countable.

Remark In the same hypothesis,
the set of metrics is not even sequentially closed in $\left(\mathbb{R}^{X\times X}, \tau_p\right)$.

In fact, let us consider a convergent sequence of metrics such that

$$d_n \to \overline{d} \quad \text{and} \quad d_n(\boldsymbol{x}) = \frac{1}{n},\ n \in \mathbb{N}$$

for some $\boldsymbol{x} \in (X \times X) \setminus \Delta_X$. Then

$$\overline{d}(\boldsymbol{x}) = \lim_{n \to +\infty} d_n(\boldsymbol{x}) = 0$$

but $\boldsymbol{x} \notin \Delta_X$, so $\overline{d}$ is not a metric.

From now on we will assume X finite of cardinality n.
We may sometimes identify X with $\{1, \ldots, n\}$, since they are in bijection.

Remark Sometimes it is useful to think of functions $d : X \times X \to \mathbb{R}$,
that is $d \in \mathbb{R}^{X\times X}$, as real-valued square matrices $D \in M(n, \mathbb{R})$.

In particular, a pseudo-metric corresponds to a symmetric matrix,
with 0 on the diagonal, non-negative entries elsewhere
and such that the elements satisfy the triangle inequality.

Let us consider a real vector space V.

Definition (**conic/conical combination**)

A point $v \in V$ is a **conical combination** of $v_1, \ldots, v_k \in V$ if

$$\exists\, \lambda_1, \ldots, \lambda_k \geq 0 \quad \text{such that} \quad v = \lambda_1 v_1 + \cdots + \lambda_k v_k$$

Definition (**extreme/extremal ray**)

An **extreme ray** of a cone $C \subseteq V$ is a subset $S \subseteq C$ of the form

$$S = \{\, \lambda r \mid \lambda \geq 0 \,\}$$

for some $r \in C \setminus \{0\}$, such that the elements of S cannot be expressed as finite conical combinations of elements of $C \setminus S$;

or equivalently, for every $v_1, \ldots, v_k \in C$ and $\lambda_1, \ldots, \lambda_k \geq 0$

$$\lambda_1 v_1 + \cdots + \lambda_k v_k \in S \quad \Longrightarrow \quad \exists\, i \in \{1, \ldots, k\} \,:\, v_i \in S$$

or equivalently, for every $v, w \in C$

$$v + w \,\in\, S \quad \Longrightarrow \quad v \in S \quad \text{or} \quad w \in S$$

We can identify an extreme ray with the associated vector r.

Definition (**simplicial cone**)

A cone $C \subseteq V$ is a simplicial cone if every complete[2] set of representatives of the extreme rays is linearly independent;

or equivalently, if

$$\#\,\{\text{extreme rays}\} = \dim_{\mathbb{R}} \langle C \rangle$$

where $\langle C \rangle$ is the vector subspace spanned by C.

[2] This is to be understood as every vector corresponds to an extreme ray and no two vectors correspond to the same extreme ray.

Definition (**split metric**)

Given a partition (or split) of X into two disjoint non-empty sets A and B, we call **split metric**[3] of $\{A, B\}$ the function $\delta_{A,B} : X \times X \to \mathbb{R}$ defined by

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

In other words $\delta_{A,B}$ equals 1 on elements that are separated by the split/cut given by $\{A, B\}$, and equals 0 on the elements that belong to the same "side".

Notation If $a \in X$, we denote the **trivial split metric**

$$\delta_a := \delta_{\{a\},\, X\setminus\{a\}}$$

Proposition 1.4

The split metrics are pseudo-metrics.

Proof

Let us fix a split $\{A, B\}$ of X and let us show that $\delta_{A,B} \in M(X)$. The vanishing on the diagonal is obvious.

Let us show the triangle inequality.

- $x, y, z \in A$
$$\underbrace{\delta_{A,B}(x,y)}_{0} \leq \underbrace{\delta_{A,B}(x,z)}_{0} + \underbrace{\delta_{A,B}(y,z)}_{0}$$
- $x, y \in A,\ z \in B$
$$\underbrace{\delta_{A,B}(x,y)}_{0} \leq \underbrace{\delta_{A,B}(x,z)}_{1} + \underbrace{\delta_{A,B}(y,z)}_{1}$$
- $x, z \in A,\ y \in B$
$$\underbrace{\delta_{A,B}(x,y)}_{1} \leq \underbrace{\delta_{A,B}(x,z)}_{0} + \underbrace{\delta_{A,B}(y,z)}_{1}$$
- $x \in A,\ y, z \in B$
$$\underbrace{\delta_{A,B}(x,y)}_{1} \leq \underbrace{\delta_{A,B}(x,z)}_{1} + \underbrace{\delta_{A,B}(y,z)}_{0}$$

Analogous cases switching A and B. □

[3] Other authors call them cut metrics, binary metrics or binary dissimilarities.

Lemma 1.5

If d is a pseudo-metric, then for every $x, y, z \in X$

$$d(x,y) = 0 \quad \Longrightarrow \quad d(x,z) = d(y,z)$$

Proof

We have the following triangle inequalities

$$\begin{cases} d(x,y) \leq d(x,z) + d(y,z) \\ d(x,z) \leq \cancel{d(x,y)} + d(y,z) \\ d(y,z) \leq \cancel{d(x,y)} + d(x,z) \end{cases}$$

From the last two we get the double inequality, hence the thesis.

□

Notation For every split $\{A, B\}$ of X, we define

$$\Gamma_{A,B} := \left\{ \gamma \in M(X) \;\middle|\; \gamma(x,y) = 0 \;\text{ if } x,y \in A \text{ or } x,y \in B \right\}$$

These are the pseudo-metrics that vanish
where the split metric $\delta_{A,B}$ vanishes (the other entries can be anything).

Lemma 1.6

These pseudo-metrics are multiples of the relative split metric

$$\gamma \in \Gamma_{A,B} \quad \Longrightarrow \quad \exists\, \lambda \geq 0 \;:\; \gamma = \lambda \delta_{A,B}$$

or in other words

$$\Gamma_{A,B} = \{\, \lambda \delta_{A,B} \mid \lambda \geq 0 \,\} = \operatorname{cone}(\delta_{A,B})$$

Proof

Let us fix $a_0 \in A$.
Since $\forall\, b, b' \in B,\ \gamma(b,b') = 0$, from Lemma 1.5 we have

$$\gamma(a_0, b) = \gamma(a_0, b')\,, \quad \forall\, b, b' \in B$$

Symmetrically, for $b_0 \in B$ we have

$$\gamma(b_0, a) = \gamma(b_0, a')\,, \quad \forall\, a, a' \in A$$

Thus $\gamma(a,b) = \gamma(a',b')\,, \quad \forall\, a, a' \in A,\ \forall\, b, b' \in B$. Call this value λ.
Since γ vanishes on all the other couples (because $\delta_{A,B}$ does), then

$$\gamma = \lambda \delta_{A,B}$$

□

Proposition 1.7

The split metrics are extreme rays of $M(X)$.[4]

Proof

Let us consider a split $\{A, B\}$ and the associated split metric $\delta_{A,B}$.

From the previous Lemma 1.6 we have

$$\Gamma_{A,B} = \{\, \lambda\delta_{A,B} \mid \lambda \geq 0 \,\}$$

so we need to show

$$\gamma_1 + \gamma_2 \in \Gamma_{A,B} \quad \Longrightarrow \quad \gamma_1 \in \Gamma_{A,B} \text{ or } \gamma_2 \in \Gamma_{A,B}$$

Suppose $\gamma_1 + \gamma_2 = \lambda\delta_{A,B}$. Then for every $a, a' \in A$

$$\gamma_1(a, a') + \gamma_2(a, a') = \lambda\delta_{A,B}(a, a') = 0$$

but since γ_1, γ_2 are non-negative, we have

$$\gamma_1(a, a') = \gamma_2(a, a') = 0$$

Idem for every $b, b' \in B$.

Thus γ_1 and γ_2 vanish where $\delta_{A,B}$ vanishes, that is

$$\gamma_1 \in \Gamma_{A,B} \text{ and } \gamma_2 \in \Gamma_{A,B}$$

□

Remark The number of split metrics is $2^{n-1} - 1$.

In fact, the set of split metrics is in bijection with the set of splits of X. In creating a split, for every element of X we have two choices: put it in A or put it in B.

We have 2^n possible arrangements. We need to subtract the cases corresponding to $(\emptyset, X)$ and $(X, \emptyset)$, and divide by two, since the split $\{A, B\}$ is the same as $\{B, A\}$.

In the end we get

$$\frac{2^n - 2}{2} = 2^{n-1} - 1$$

[4] For this reason, they may be also called extremal metrics.

Lemma 1.8

The functions $\varepsilon_{ij} : X \times X \to \mathbb{R}\,,\ i \neq j \in X$ defined by

$$\varepsilon_{ij}(\boldsymbol{x}) = \begin{cases} 1\,, & \text{if } \boldsymbol{x} = (i,j) \text{ or } \boldsymbol{x} = (j,i) \\ 0\,, & \text{otherwise} \end{cases}$$

can be expressed as linear combinations of the split metrics.

Proof

Consider the function given by $\delta_i + \delta_j - \delta_{ij}$, where

$$\delta_i := \delta_{\{i\},X\setminus\{i\}}\ , \quad \delta_j := \delta_{\{j\},X\setminus\{j\}}\ , \quad \delta_{ij} := \delta_{\{i,j\},X\setminus\{i,j\}}$$

Let us divide in cases (up to symmetry):

- (i,j)
$$\underbrace{\delta_i(i,j)}_{1} + \underbrace{\delta_j(i,j)}_{1} - \underbrace{\delta_{ij}(i,j)}_{0} = 2$$
- (i,y) with $y \neq j$
$$\underbrace{\delta_i(i,y)}_{1} + \underbrace{\delta_j(i,y)}_{0} - \underbrace{\delta_{ij}(i,y)}_{1} = 0$$
- (x,j) with $x \neq i$
$$\underbrace{\delta_i(x,j)}_{0} + \underbrace{\delta_j(x,j)}_{1} - \underbrace{\delta_{ij}(x,j)}_{1} = 0$$
- (x,y) with $(x,y) \neq (i,j)$
$$\underbrace{\delta_i(x,y)}_{0} + \underbrace{\delta_j(x,y)}_{0} - \underbrace{\delta_{ij}(x,y)}_{0} = 0$$

Thus this function coincide with $2\,\varepsilon_{ij}$, from which

$$\varepsilon_{ij} = \frac{1}{2}\,\delta_i + \frac{1}{2}\,\delta_j - \frac{1}{2}\,\delta_{ij}$$

□

Proposition 1.9

$\langle M(X)\rangle$ is a vector subspace of $\mathbb{R}^{X\times X}$ with the following properties:

- $\langle M(X)\rangle = \left\{ f : X\times X \to \mathbb{R} \;\middle|\; \begin{array}{c} f(x,y) = f(y,x),\ \forall\, x,y \in X \\ \text{and } f(x,x) = 0,\ \forall\, x \in X \end{array} \right\}$
- $\dim_{\mathbb{R}} \langle M(X)\rangle = \binom{n}{2}$

Proof

Let us indicate the set of symmetric functions vanishing on the diagonal with $S_0(X)$.

It is clear that $M(X) \subseteq S_0(X)$ and $S_0(X)$ is closed under linear combinations. So $\langle M(X)\rangle \subseteq S_0(X)$.

Notice that $\{\varepsilon_{ij}\}_{i<j}$ is a basis for $S_0(X)$.
In fact, the functions in $S_0(X)$ can be represented as symmetric matrices with zeroes on the diagonal; while the function ε_{ij} can be represented as a symmetric matrix with 1 in positions (i,j) and (j,i), and zeroes elsewhere.

This also shows that

$$\dim_{\mathbb{R}} S_0(X) \;=\; \#\{\varepsilon_{ij}\}_{i<j} \;=\; \binom{n}{2}$$

From the previous Lemma 1.8 we can express the functions $\{\varepsilon_{ij}\}_{i<j}$ as linear combinations of split metrics, thus the same holds for every function in $S_0(X)$; that is $S_0(X) \subseteq \langle M(X)\rangle$.

□

Remark For $n = 2$, say $X = \{a, b\}$, all the pseudo-metrics are multiple of the only split metric $\delta_{\{a\},\{b\}}$;
so $M(X)$ is just a one-dimensional ray.

Remark For $n = 3$, say $X = \{a, b, c\}$, the split metrics

$$\delta_a := \delta_{\{a\},\{b,c\}}\,, \quad \delta_b := \delta_{\{b\},\{a,c\}}\,, \quad \delta_c := \delta_{\{c\},\{a,b\}}$$

generate the cone $M(X)$.

In fact, let us consider a pseudo-metric $d \in M(X)$;
then we want to show that $\exists\, \lambda, \mu, \nu \geq 0$ such that

$$d = \lambda\delta_a + \mu\delta_b + \nu\delta_c$$

In particular, by evaluating

$$\begin{aligned} d_{a,b} &:= d(a,b) = \\ &= \lambda\delta_{\{a\},\{b,c\}}(a,b) + \mu\delta_{\{b\},\{a,c\}}(a,b) + \nu\delta_{\{c\},\{a,b\}}(a,b) = \\ &= \lambda \cdot 1 + \mu \cdot 1 + \nu \cdot 0 = \\ &= \lambda + \mu \end{aligned}$$

and analogously for the other couples,
we get the following system of equations

$$\begin{cases} d_{ab} = \lambda + \mu \\ d_{ac} = \lambda + \nu \\ d_{bc} = \mu + \nu \end{cases}$$

Solving for λ, μ, ν we get

$$\lambda = \frac{d_{ab} + d_{ac} - d_{bc}}{2}$$

$$\mu = \frac{d_{ab} - d_{ac} + d_{bc}}{2}$$

$$\nu = \frac{-d_{ab} + d_{ac} + d_{bc}}{2}$$

Moreover, from triangle inequality on d, we have $\lambda, \mu, \nu \geq 0$.
This shows that these split metrics are the only extreme rays
(every other pseudo-metric is a conical combination of them).

The same calculation also shows that the split metrics are
linearly independent in $\langle M(X) \rangle$. In fact, if

$$\lambda\delta_a + \mu\delta_b + \nu\delta_c = \mathbf{0}$$

where $\mathbf{0}$ is the identically zero pseudo-metric, then

$$\lambda = 0, \quad \mu = 0, \quad \nu = 0$$

In particular, for $n = 3$ the decomposition in split metrics is
unique.

Remark For $n \geq 4$, the decomposition in split metrics is not necessarily unique.

Consider $X = \{a, b, c, d\}$ and its split metrics

$$\delta_{\{a\},\{b,c,d\}} \qquad \delta_{\{b\},\{a,c,d\}} \qquad \delta_{\{c\},\{a,b,d\}} \qquad \delta_{\{d\},\{a,b,c\}}$$
$$\delta_{\{a,b\},\{c,d\}} \qquad \delta_{\{a,c\},\{b,d\}} \qquad \delta_{\{a,d\},\{b,c\}}$$

Then the pseudo-metric d defined by

$$d(x,y) := \begin{cases} 2\,, & \text{if } x \neq y \\ 0\,, & \text{if } x = y \end{cases}$$

can be expressed as

$$d = \delta_{\{a\},\{b,c,d\}} + \delta_{\{b\},\{a,c,d\}} + \delta_{\{c\},\{a,b,d\}} + \delta_{\{d\},\{a,b,c\}}$$

as well as

$$d = \delta_{\{a,b\},\{c,d\}} + \delta_{\{a,c\},\{b,d\}} + \delta_{\{a,d\},\{b,c\}}$$

Remark $M(X)$ is a simplicial cone if and only if $n = 2, 3$. In fact,

$$\#\{\text{extreme rays}\} \;\geq\; \#\{\text{split metrics}\} \;=\; 2^{n-1} - 1$$

and for $n \geq 4$ this last number is greater than

$$\dim_{\mathbb{R}} \langle M(X) \rangle = \binom{n}{2} = \frac{n(n-1)}{2}$$

But for $n = 2, 3$ the number of extreme rays coincide with the number of split metrics and also

$$n = 2, \qquad 2^{2-1} - 1 = 1 = \binom{2}{2}$$

$$n = 3, \qquad 2^{3-1} - 1 = 3 = \binom{3}{2}$$

Chapter 2

Decomposition via d-splits

Let $d : X \times X \to \mathbb{R}$ be a symmetric function.

Notation $uv := d(u, v)\,, \quad \forall\, u, v \in X$

In the following we will implicitly refer to d.

Definition **(isolation index)**

Let A, B be non-empty subset of X.
Then for every $a, a' \in A$ and $b, b' \in B$ we define

$$\beta_{\{a,a'\},\{b,b'\}} := \frac{1}{2}\Big(\max\{ab + a'b',\, a'b + ab',\, aa' + bb'\} - aa' - bb'\Big)$$

$$\alpha_{A,B} := \min_{\substack{a,a' \in A \\ b,b' \in B}} \beta_{\{a,a'\},\{b,b'\}}$$

We call $\alpha_{A,B} = \alpha^d_{A,B}$ the **isolation index** with respect to d.

Remark We have $\beta_{\{a,a'\},\{b,b'\}} \geq 0$ for every $a, a' \in A$, $b, b' \in B$, since we are subtracting a term that is inside the maximum:

$$\max\{ab + a'b',\, a'b + ab',\, aa' + bb'\} \;\geq\; aa' + bb'$$

It follows that also $\alpha_{A,B} \geq 0\,.$

Moreover if $\beta_{\{a,a'\},\{b,b'\}} = 0$ <u>for some</u> $a, a' \in A$, $b, b' \in B$, then also $\alpha_{A,B} = 0$.

Remark If A and B intersect, then $\alpha_{A,B} = 0$.

In fact, let $x \in A \cap B$. Then

$$\begin{aligned}\beta_{\{x\},\{x\}} &= \frac{1}{2}\Big(\max\{xx + xx,\ xx + xx,\ xx + xx\} - xx - xx\Big)\\ &= \frac{1}{2}(xx + xx - xx - xx) = 0\end{aligned}$$

and so $\alpha_{A,B} = 0$.

Proposition 2.1

If d is a pseudo-metric, then for every $t, u, v, w \in X$

$$\alpha_{\{t,u\},\{v,w\}} = \beta_{\{t,u\},\{v,w\}}$$

Proof

By (reverse) triangle inequality and the fact that d vanishes on the diagonal, we observe that

$$\begin{aligned}\beta_{\{t\},\{v,w\}} &= \frac{1}{2}\Big(\max\{tv + tw,\ tv + tw,\ \cancel{tt} + vw\} - \cancel{tt} - vw\Big)\\ &= \frac{1}{2}(tv + tw - vw)\\ &\le \frac{1}{2}(tv + tv)\\ &= \frac{1}{2}\Big(\max\{tv + tv,\ tv + tv,\ \cancel{tt} + \cancel{vv}\} - \cancel{tt} - \cancel{vv}\Big)\\ &= \beta_{\{t\},\{v\}}\end{aligned}$$

where we used $tv + tw \ge vw$ in the first line
and $tw - vw \le tv$ in the second one.

With analogous calculations we get

$$\begin{aligned}\beta_{\{t\},\{v,w\}} &\le \beta_{\{t\},\{v\}}\,, & \beta_{\{t\},\{v,w\}} &\le \beta_{\{t\},\{w\}}\\ \beta_{\{u\},\{v,w\}} &\le \beta_{\{u\},\{v\}}\,, & \beta_{\{u\},\{v,w\}} &\le \beta_{\{u\},\{w\}}\\ \beta_{\{t,u\},\{v\}} &\le \beta_{\{t\},\{v\}}\,, & \beta_{\{t,u\},\{v\}} &\le \beta_{\{u\},\{v\}}\\ \beta_{\{t,u\},\{w\}} &\le \beta_{\{t\},\{w\}}\,, & \beta_{\{t,u\},\{w\}} &\le \beta_{\{u\},\{w\}}\end{aligned}$$

Again by (reverse) triangle inequality

$$\begin{aligned}
tv + uw - tu - vw &\leq tv + tw - vw = 2\beta_{\{t\},\{v,w\}} \\
&\leq uv + uw - vw = 2\beta_{\{u\},\{v,w\}} \\
&\leq tv + uv - tu = 2\beta_{\{t,u\},\{v\}} \\
&\leq tw + uw - tu = 2\beta_{\{t,u\},\{w\}} \\
uv + tw - tu - vw &\leq tv + tw - vw = 2\beta_{\{t\},\{v,w\}} \\
&\leq uv + uw - vw = 2\beta_{\{u\},\{v,w\}} \\
&\leq tv + uv - tu = 2\beta_{\{t,u\},\{v\}} \\
&\leq tw + uw - tu = 2\beta_{\{t,u\},\{w\}}
\end{aligned}$$

Since

$$\beta_{\{t,u\},\{v,w\}} = \frac{1}{2}\Big(\max\{tv + uw,\ uv + tw,\ tu + vw\} - tu - vw\Big)$$

we have proven that $\beta_{\{t,u\},\{v,w\}}$ is smaller than
all other possible β indices for the quartet $\big\{\{t,u\},\{v,w\}\big\}$.

This implies by definition of the isolation index

$$\alpha_{\{t,u\},\{v,w\}} = \beta_{\{t,u\},\{v,w\}}$$

□

Corollary 2.2

If d is a pseudo-metric and $\min\{|A|, |B|\} \geq 2$
(that is both A and B have at least 2 elements), then

$$\alpha_{A,B} = \alpha_{\{a,a'\},\{b,b'\}}$$

for some $a \neq a' \in A$ and $b \neq b' \in B$.

Proof

Let $\big\{\{a,a'\},\{b,b'\}\big\}$ be the quartet that minimizes the β index.

Then, for the previous inequalities about β indices
and the hypothesis on the number of elements of A and B,
we conclude that $a \neq a'$ and $b \neq b'$.

Using the previous Proposition 2.1

$$\alpha_{A,B} = \beta_{\{a,a'\},\{b,b'\}} = \alpha_{\{a,a'\},\{b,b'\}}$$

□

Definition (**splits and d-splits**)

A **partial split** (of X) is an unordered pair $\{A, B\}$ with $A, B \subseteq X$.
A **(total) split** (of X) is partial split $\{A, B\}$ such that $A \cup B = X$.

A partial/total ***d*-split** is a partial/total split $\{A, B\}$ such that $\alpha^d_{A,B} > 0$.
Notice that in this case A and B must be disjoint.

The sets A and B are called **parts** of the split.
We call a split **trivial** if one of the parts contains only one element.
A partial split of the form $\big\{\{a, a'\}, \{b, b'\}\big\}$ is called **quartet**.

We denote the set of all splits of X with $\mathcal{S}(X)$
and the set of all d-splits of X with $\mathcal{S}_d(X)$.

Definition (d-**convex set**)

A set $S \subseteq X$ is ***d*-convex** if

$$\forall\, x, y \in S,\ \forall\, z \in X, \qquad xz + zy = xy \implies z \in S$$

Proposition 2.3

If $\{A, B\}$ is a d-split, then A and B are d-convex.

Proof

Let $x, y \in A$ and $z \in X$ such that $xz + zy = xy$.

Since $X = A \cup B$, we have $z \in A$ or $z \in B$.

If by absurd $z \in B$, then

$$\begin{aligned}\beta_{\{x,y\},\{z\}} &= \frac{1}{2}\Big(\max\{xz + yz,\ xz + yz,\ xy + \cancel{zz}\} - xy - \cancel{zz}\Big) \\ &= \frac{1}{2}(xz + yz - xy) = \frac{1}{2}(xy - xy) = 0\end{aligned}$$

that implies $\alpha_{A,B} = 0$. ↯

Then it must be $z \in A$ and so A is d-convex.

□

Definition (**extension of d-splits**)

We say that a partial d-split $\{A, B\}$ **extends** another partial d-split $\{A', B'\}$ if $A \supseteq A'$ and $B \supseteq B'$ (or $A \supseteq B'$ and $B \supseteq A'$).
We denote it $\{A, B\} \succcurlyeq \{A', B'\}$.

Notice that $\alpha_{A,B} \leq \alpha_{A',B'}$ (it's a minimum on a larger set).

Lemma 2.4

Let $\{A, B\}$ be a partial d-split.

Then for every $a_1, a_2 \in A,\ b_1, b_2 \in B,\ x \in X \setminus (A \cup B)$

$$\alpha_{\{a_1,a_2,x\},\{b_1,b_2\}} + \alpha_{\{a_1,a_2\},\{b_1,b_2,x\}} \leq \beta_{\{a_1,a_2\},\{b_1,b_2\}}$$

Proof

Suppose by absurd that exist $a_1, a_2,\ b_1, b_2,\ x$ such that

$$\alpha_{\{a_1,a_2,x\},\{b_1,b_2\}} + \alpha_{\{a_1,a_2\},\{b_1,b_2,x\}} > \beta_{\{a_1,a_2\},\{b_1,b_2\}}$$

Observe that $\beta_{\{a_1,a_2\},\{b_1,b_2\}} > 0$ since $\{A, B\}$ is a partial d-split, so also

$$\alpha_{\{a_1,a_2,x\},\{b_1,b_2\}} > 0\,, \qquad \alpha_{\{a_1,a_2\},\{b_1,b_2,x\}} > 0$$

because isolation indices are not negative.

From this, and by definition of isolation index, we have

$$\begin{aligned} \beta_{\{a_1,x\},\{b_1,b_2\}} &\geq \alpha_{\{a_1,a_2,x\},\{b_1,b_2\}} > 0 \\ \beta_{\{a_1,a_2\},\{x,b_i\}} &\geq \alpha_{\{a_1,a_2\},\{b_1,b_2,x\}} > 0\,, \quad i = 1, 2 \end{aligned}$$

Let $\{i, j\} = \{1, 2\}$ so that

$$\beta_{\{a_1,x\},\{b_1,b_2\}} = \frac{1}{2}\left(a_1 b_j + x b_i - a_1 x - b_1 b_2\right)$$

in fact, since $\beta_{\{a_1,x\},\{b_1,b_2\}} > 0$, the maximum

$$\max\{a_1 b_1 + x b_2,\ a_1 b_2 + x b_1,\ a_1 x + b_1 b_2\}$$

cannot be $a_1 x + b_1 b_2$.

For the same reason we have

$$\beta_{\{a_1,a_2\},\{x,b_i\}} = \frac{1}{2}\Big(\max\{a_1x + a_2b_i,\, a_1b_i + a_2x\} - a_1a_2 - xb_i\Big)$$
$$\beta_{\{a_1,a_2\},\{b_1,b_2\}} = \frac{1}{2}\Big(\max\{a_1b_1 + a_2b_2,\, a_1b_2 + a_2b_1\} - a_1a_2 - b_1b_2\Big)$$

From the previous inequalities we have

$$\beta_{\{a_1,x\},\{b_1,b_2\}} + \beta_{\{a_1,a_2\},\{x,b_i\}} \geq$$
$$\geq \alpha_{\{a_1,a_2,x\},\{b_1,b_2\}} + \alpha_{\{a_1,a_2\},\{b_1,b_2,x\}} > \beta_{\{a_1,a_2\},\{b_1,b_2\}}$$

and by substituting the expressions for the β indices

$$\frac{1}{2}\Big(a_1b_j + \cancel{xb_i} - a_1x - \cancel{b_1b_2}$$
$$+ \max\{a_1x + a_2b_i,\, a_1b_i + a_2x\} - a_1a_2 - \cancel{xb_i}\Big) >$$
$$> \frac{1}{2}\Big(\max\{a_1b_1 + a_2b_2,\, a_1b_2 + a_2b_1\} - a_1a_2 - \cancel{b_1b_2}\Big)$$

simplifying we obtain

$$a_1b_j - a_1x + \max\{a_1x + a_2b_i,\, a_1b_i + a_2x\}$$
$$> \max\{a_1b_1 + a_2b_2,\, a_1b_2 + a_2b_1\}$$

If by absurd $\max\{a_1x + a_2b_i,\, a_1b_i + a_2x\} = a_1x + a_2b_i$,
the previous inequality would become

$$a_1b_j - \cancel{a_1x} + \cancel{a_1x} + a_2b_i > \max\{a_1b_1 + a_2b_2,\, a_1b_2 + a_2b_1\}$$

but since $a_1b_j + a_2b_i$ is a term of the maximum on the right,
this is a contradiction. ↯

So we must have $a_1x + a_2b_i < a_1b_i + a_2x$ and thus

$$a_1b_1 + a_1b_2 - a_1x + a_2x = a_1b_j - a_1x + a_1b_i + a_2x$$
$$> \max\{a_1b_1 + a_2b_2,\, a_1b_2 + a_2b_1\}$$

For this to be true we need

$$a_1b_k + a_2x > a_1x + a_2b_k\,, \quad k = 1, 2$$

By symmetry, interchanging a_1 and a_2 (notice that we can do this because the absurd hypothesis is symmetrical in a_1, a_2),
we obtain the other strict inequality. ↯ □

In the following theorem we use the notation

$$\sum\{\, a \mid \dots \} \;:=\; \sum_{\dots} a$$

Theorem 2.5 **([BD92a, Theorem 1])**

Let $\{A_0, B_0\}$ be a partial d-split. Then

$$\sum\Big\{\, \alpha_{A,B} \;\Big|\; \{A,B\} \in \mathcal{S}(X)\,,\ \{A,B\} \succcurlyeq \{A_0,B_0\} \Big\} \;\le\; \alpha_{A_0,B_0}$$

Proof

Let $a_1, a_2 \in A_0$ and $b_1, b_2 \in B_0$ such that

$$\alpha_{A_0,B_0} = \beta_{\{a_1,a_2\},\{b_1,b_2\}}$$

From the previous Lemma 2.4, we have for every $x \in X \setminus (A_0 \cup B_0)$

$$\begin{aligned} \alpha_{A_0\cup\{x\},\,B_0} + \alpha_{A_0,\,B_0\cup\{x\}} &\le \alpha_{\{a_1,a_2,x\},\{b_1,b_2\}} + \alpha_{\{a_1,a_2\},\{b_1,b_2,x\}} \\ &\le \beta_{\{a_1,a_2\},\{b_1,b_2\}} \\ &= \alpha_{A_0,B_0} \end{aligned}$$

This also proves the theorem in the case $|X \setminus (A_0 \cup B_0)| = 1$.

We prove the general case by induction on $|X \setminus (A_0 \cup B_0)|$.

We have already seen the base case.

Let $|X \setminus (A_0 \cup B_0)| = n > 1$ and $x \in X \setminus (A_0 \cup B_0)$.

Observe that the d-splits extending $\{A_0, B_0\}$ are exactly the collection of those extending $\big\{A_0 \cup \{x\}, B_0\big\}$ and those extending $\big\{A_0, B_0 \cup \{x\}\big\}$. In fact, since d-splits are partitions, if $\{A, B\}$ extends $\{A_0, B_0\}$ – say WLOG $A \supseteq A_0$ and $B \supseteq B_0$ – it must be $x \in A$ or $x \in B$; that is $A \supseteq A_0 \cup \{x\}$ or $B \supseteq B_0 \cup \{x\}$.

Moreover,

$$\begin{aligned} \big|\, X \setminus \big((A_0 \cup \{x\}) \cup B_0\big) \,\big| &= n-1 \\ \big|\, X \setminus \big(A_0 \cup (B_0 \cup \{x\})\big) \,\big| &= n-1 \end{aligned}$$

So we can apply the inductive step:

$$\sum\left\{\alpha_{A,B} \;\middle|\; \{A,B\}\in\mathcal{S}(X)\,,\ \{A,B\}\succcurlyeq\{A_0,B_0\}\right\}=$$
$$=\sum\left\{\alpha_{A,B} \;\middle|\; \{A,B\}\in\mathcal{S}(X)\,,\ \{A,B\}\succcurlyeq\left\{A_0\cup\{x\},B_0\right\}\right\}$$
$$\sum\left\{\alpha_{A,B} \;\middle|\; \{A,B\}\in\mathcal{S}(X)\,,\ \{A,B\}\succcurlyeq\left\{A_0,B_0\cup\{x\}\right\}\right\}\leq$$
$$\leq\ \alpha_{A_0\cup\{x\},\,B_0}+\alpha_{A_0,\,B_0\cup\{x\}}\ \leq\ \alpha_{A_0,B_0}$$

where the last inequality is the one proved at the beginning.

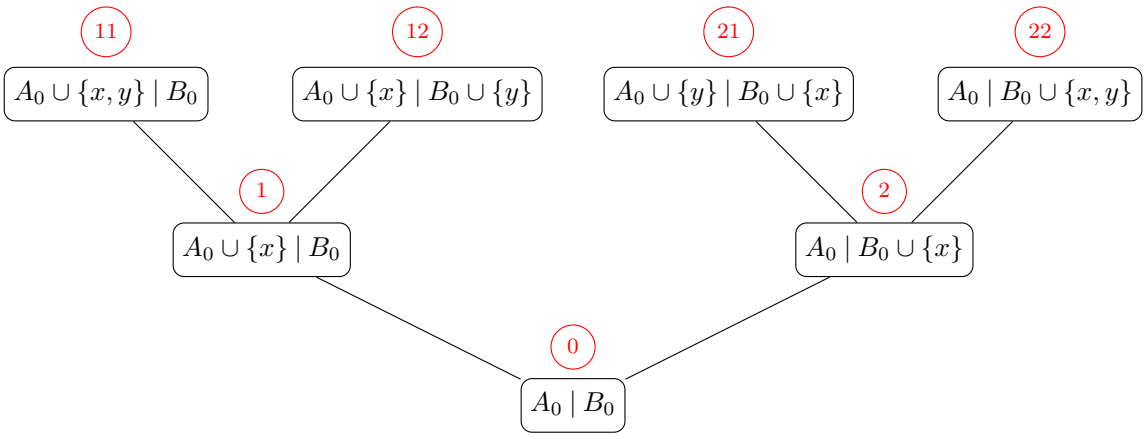
□

Remark We can substitute splits with d-splits in the sum (splits that are not d-splits do not contribute).

Remark The proof of Theorem 2.5 tells us that we can build a tree of partial splits inductively by fixing a partial split as the root and adding one element at a time to each part of the partial split, obtaining two children.

In this tree the sum of the isolation indices of the children is less than or equal to the isolation index of the parent.
In particular the sum of the isolation indices of all the nodes that share a common ancestor is less than or equal to the isolation index of the common ancestor.
As a consequence the sum of the isolation indices of all the nodes at the same level is less than or equal to the isolation index of the root (Theorem 2.5 is just a special case where the level is that of the leaves).

Observe that such tree is not unique, but depends on the order in which the elements are added.
In particular if $\{A,B\}$ is an extension of $\{A_0,B_0\}$, than it exists an extension tree rooted in $\{A_0,B_0\}$ that contains $\{A,B\}$.



$$\left.\begin{array}{l} \alpha_{11} + \alpha_{12} \leq \alpha_1 \\ \alpha_{21} + \alpha_{22} \leq \alpha_2 \end{array} \qquad \alpha_1 + \alpha_2 \leq \alpha_0 \right\} \quad \sum_{\substack{i=1,2 \\ j=1,2}} \alpha_{ij} \leq \alpha_0$$

Corollary 2.6 ([BD92a, Corollary 1])

Let $\{A, B\}$ a d-split and let $a_1, a_2 \in A$ and $b_1, b_2 \in B$ such that

$$\alpha_{A,B} = \beta_{\{a_1,a_2\},\{b_1,b_2\}}$$

Then $\{A, B\}$ is the unique d-split that extends $\big\{\{a_1, a_2\}, \{b_1, b_2\}\big\}$.

Proof

Observe that $\big\{\{a_1, a_2\}, \{b_1, b_2\}\big\}$ is a partial d-split.

In fact, if by absurd

$$\alpha_{\{a_1,a_2\},\{b_1,b_2\}} = \beta_{\{a_1,a_2\},\{b_1,b_2\}} = 0$$

then $\alpha_{A,B} = 0$, but $\{A, B\}$ is a d-split. ↯

By applying Theorem 2.5

$$\begin{aligned} \alpha_{A,B} &\leq \sum \left\{ \alpha_{A',B'} \;\middle|\; \begin{array}{c} \{A', B'\} \in \mathcal{S}_d(X)\,, \\ \{A', B'\} \succcurlyeq \big\{\{a_1, a_2\}, \{b_1, b_2\}\big\} \end{array} \right\} \\ &\leq \alpha_{\{a_1,a_2\},\{b_1,b_2\}} \\ &\leq \beta_{\{a_1,a_2\},\{b_1,b_2\}} = \alpha_{A,B} \end{aligned}$$

Since the sum is equal to $\alpha_{A,B}$ and $\{A, B\}$ is a member of the range of the sum, this means that $\{A, B\}$ is the only d-split that extends $\big\{\{a_1, a_2\}, \{b_1, b_2\}\big\}$.

□

Definition (split metric)

Given A, B non-empty disjoint subsets of X,
the **split metric** $\delta_{A,B}$ on $A \cup B$ is defined as

$$\delta_{A,B}(u, v) := \begin{cases} 0\,, & \text{if } u, v \in A \text{ or } u, v \in B \\ 1\,, & \text{otherwise} \end{cases}$$

This definition is slightly different from that given previously,
because we do not require $A \cup B = X$ (we now admit *partial* split metrics).

Proposition 2.7

The split $\{A, B\}$ is the only $\delta_{A,B}$-split and its isolation index is 1.

Proof

In fact, let us consider $\{A, B\}$ and another split $\{A', B'\}$,
with the following intersections:

$$A \cap A'\,, \quad A \cap B'\,, \quad B \cap A'\,, \quad B \cap B'$$

Since $A \cup B = A' \cup B' = X$ and $A, B, A', B' \neq \emptyset$,
at least two of these intersection must be non-empty.

We can divide in cases.

If all the intersections are non-empty, we can pick

$$a_1 \in A \cap A'\,, \quad a_2 \in A \cap B'\,, \quad b_1 \in B \cap A'\,, \quad b_2 \in B \cap B'$$

then, with respect to $\delta_{A,B}$, we have

$$\begin{aligned} a_1 b_1 + a_2 b_2 &= 1 + 1 = 2 \\ a_1 b_2 + a_2 b_1 &= 1 + 1 = 2 \\ a_1 a_2 + b_1 b_2 &= 0 + 0 = 0 \end{aligned}$$

$$\beta_{\{a_1,b_1\},\{a_2,b_2\}} = \frac{1}{2}\left(\max\begin{Bmatrix} a_1b_1 + a_2b_2, \\ a_1b_2 + a_2b_1, \\ \cancel{a_1a_2 + b_1b_2} \end{Bmatrix} - a_1b_1 - a_2b_2\right) = 0$$

so $\alpha_{A',B'} = 0$.

If only three of the intersections are non-empty
– say WLOG $A \cap A'$, $B \cap A'$, $B \cap B'$ – then we can pick

$$a_1 \in A \cap A', \quad b_1 \in B \cap A', \quad b_2 \in B \cap B'$$

and then, with respect to $\delta_{A,B}$, we have

$$a_1b_1 + b_2b_2 = 1 + 0 = 1$$
$$a_1b_2 + b_1b_2 = 1 + 0 = 1$$

$$\beta_{\{a_1,b_1\},\{b_2\}} = \frac{1}{2}\left(\max\begin{Bmatrix} a_1b_2 + b_1b_2, \\ a_1b_1 + b_2b_2 \end{Bmatrix} - a_1b_1 - b_2b_2\right) = 0$$

so $\alpha_{A',B'} = 0$.

If only two intersections are non-empty, then $\{A, B\} = \{A', B'\}$.

In fact, say WLOG $A \cap B'$, $B \cap A' = \emptyset$.
Since $A \cup B = X$, then we have $A' \subseteq A$ and $B' \subseteq B$.
But since $A' \cup B' = X$, we also have $A \subseteq A'$ and $B \subseteq B'$,
so it must be $A' = A$ and $B' = B$.

Now <u>for every</u> $a_1, a_2 \in A$ and $b_1, b_2 \in B$, with respect to $\delta_{A,B}$,

$$a_1b_1 + a_2b_2 = 1 + 1 = 2$$
$$a_1b_2 + a_2b_1 = 1 + 1 = 2$$
$$a_1a_2 + b_1b_2 = 0 + 0 = 0$$

$$\beta_{\{a_1,a_2\},\{b_1,b_2\}} = \frac{1}{2}\left(\max\begin{Bmatrix} a_1b_1 + a_2b_2, \\ a_1b_2 + a_2b_1, \\ \cancel{a_1a_2 + b_1b_2} \end{Bmatrix} - \cancel{a_1a_2 - b_1b_2}\right) = 1$$

so $\alpha_{A,B} = 1$.

□

Definition (**split-prime function**)

A function $d : X \times X \to \mathbb{R}$ is **split-prime** if it does not admit any d-split.

$$\mathcal{S}_d(X) = \emptyset$$

Definition (**dissimilarity function**)

A function $d : X \times X \to \mathbb{R}$ is a **dissimilarity function** if

- $d(x, y) = d(y, x)\,, \qquad \forall\, x, y \in X$
- $d(x, x) = 0\,, \qquad \forall\, x \in X$
- $d(x, y) \geq 0\,, \qquad \forall\, x, y \in X$

In practice, it is a pseudo-metric without triangle inequality.

Theorem 2.8 ([BD92a, Theorem 2])

Let $d : X \times X \to \mathbb{R}$ be a symmetric function.
Let $\lambda_S \in \mathbb{R}$ such that $\lambda_S \leq \alpha_S^d$ if S is a d-split
and $\lambda_S = 0$ otherwise.

Then

$$d' \;:=\; d \;-\; \sum_{S \in \mathcal{S}(X)} \lambda_S \cdot \delta_S$$

is a symmetric function such that

$$\alpha_S^{d'} = \alpha_S^d - \lambda_S\,, \quad \forall\, S \in \mathcal{S}(X)$$

In addition, if d is a dissimilarity function (or a pseudo-metric), then d' is also a dissimilarity function (or a pseudo-metric).

Proof

It suffices to prove the assertions for

$$d' = \widetilde{d} := d - \lambda \cdot \delta_{A_0, B_0}$$

where $\{A_0, B_0\}$ is a d-split and $\lambda \leq \alpha_{A_0,B_0}^d$.

Then the general case follows by subtracting one split metric at a time (formally induction on the number of non-zero λ's).

We use the following notation: $uv = d(u,v), \quad \forall\, u,v \in X$

$$\alpha = \alpha^d\,, \quad \beta = \beta^d, \qquad \widetilde{\alpha} = \alpha^{\widetilde{d}}\,, \quad \widetilde{\beta} = \beta^{\widetilde{d}}$$

Clearly $\widetilde{d}$ is a symmetric function (since d and δ are such).

For every $u, v \in X$ we have

$$\widetilde{d}(u,v) = \begin{cases} uv\,, & \text{if } \{u,v\} \subseteq A_0 \text{ or } \{u,v\} \subseteq B_0 \\ uv - \lambda\,, & \text{otherwise} \end{cases}$$

Suppose that d is a dissimilarity function.

Then $\widetilde{d}(u,u) = d(u,u) = 0$. For $u \in A_0$ and $v \in B_0$ we have

$$\lambda \le \alpha_{A_0,B_0} \le \beta_{\{u\},\{v\}} = uv$$

thus $\widetilde{d}(u,v) = uv - \lambda \ge 0$, so $\widetilde{d}$ is a dissimilarity function.

Now suppose that d is a pseudo-metric.
We have to verify the triangle inequality for $\widetilde{d}$.

Let $u, v, w \in X$. If they all belong to A_0 or all to B_0,
then d and $\widetilde{d}$ agrees on $\{u,v,w\}$ and we are done.
Otherwise, say WLOG $u, v \in A_0$ and $w \in B_0$, then we get

$$\begin{aligned} \lambda \le \alpha_{A_0,B_0} &\le \beta_{\{u,v\},\{w\}} \\ &= \frac{1}{2}\Big(\max\{uw + vw,\, uv + \cancel{uvu}\} - uv - \cancel{uvu}\Big) \\ &= \frac{1}{2}\,(uw + vw - uv) \end{aligned}$$

where we used the triangle inequality for d, and by rearranging

$$\underbrace{uv}_{\widetilde{d}(u,v)} \;\le\; \underbrace{(uw - \lambda)}_{\widetilde{d}(u,w)} + \underbrace{(vw - \lambda)}_{\widetilde{d}(v,w)}$$

For the remainder of the proof d is just a symmetric function.

Let $\{t,u\},\{v,w\} \subseteq X$.
We claim that

$$\widetilde{\beta}_{\{t,u\},\{v,w\}} = \begin{cases} \beta_{\{t,u\},\{v,w\}} - \lambda\,, & \text{if } \{A_0,B_0\} \succcurlyeq \big\{\{t,u\},\{v,w\}\big\} \\ \beta_{\{t,u\},\{v,w\}}\,, & \text{otherwise} \end{cases}$$

Suppose $\{A_0,B_0\} \succcurlyeq \big\{\{t,u\},\{v,w\}\big\}$.

Since $\{A_0,B_0\}$ is a d-split, the max get simplified

$$\beta_{\{t,u\},\{v,w\}} = \frac{1}{2}\Big(\max\{tv+uw,\, tw+uv\} - tu - vw\Big) > 0$$

and we get the following inequalities

$$\begin{aligned} \lambda \;&\leq\; \alpha_{A_0,B_0} \;\leq\; \beta_{\{t,u\},\{v,w\}} \\ 2\lambda \;&\leq\; \max\begin{Bmatrix} tv+uw \\ tw+uv \end{Bmatrix} - tu - vw \\ tu+vw \;&\leq\; \max\begin{Bmatrix} (tv-\lambda)+(uw-\lambda) \\ (tw-\lambda)+(uv-\lambda) \end{Bmatrix} \end{aligned}$$

from which

$$\begin{aligned} \widetilde{\beta}_{\{t,u\},\{v,w\}} &= \frac{1}{2}\left(\max\begin{Bmatrix} \tilde{d}(t,v)+\tilde{d}(u,w) \\ \tilde{d}(t,w)+\tilde{d}(u,v) \\ \tilde{d}(t,u)+\tilde{d}(v,w) \end{Bmatrix} - \tilde{d}(t,u) - \tilde{d}(v,w)\right) \\ &= \frac{1}{2}\left(\max\begin{Bmatrix} (tv-\lambda)+(uw-\lambda) \\ (tw-\lambda)+(uv-\lambda) \\ tu+vw \end{Bmatrix} - tu - vw\right) \\ &= \frac{1}{2}\left(\max\begin{Bmatrix} tv+uw \\ tw+uv \end{Bmatrix} - 2\lambda - tu - vw\right) \\ &= \beta_{\{t,u\},\{v,w\}} - \lambda \end{aligned}$$

Suppose instead that WLOG $t, v \in A_0$ and $u, w \in B_0$.

Since $\{A_0, B_0\}$ is a d-split, the max get simplified

$$\begin{aligned}\beta_{\{t,v\},\{u,w\}} &= \frac{1}{2}\Big(\max\{tu + vw,\, tw + vu\} - tv - uw\Big) \\ &\geq \alpha_{A_0,B_0} \geq \lambda\end{aligned}$$

from which, similarly as before, we get the inequality

$$tv + uw \leq \max\{tu + vw - 2\lambda,\, tw + uv - 2\lambda\}$$

from which

$$\begin{aligned}\widetilde{\beta}_{\{t,u\},\{v,w\}} &= \frac{1}{2}\left(\max\begin{Bmatrix} tv + uw \\ (tw - \lambda) + (uv - \lambda) \\ (tu - \lambda) + (vw - \lambda)\end{Bmatrix} - (tu - \lambda) - (vw - \lambda)\right) \\ &= \frac{1}{2}\left(\max\begin{Bmatrix} tw + uv \\ tu + vw\end{Bmatrix} - \cancel{2\lambda} - tu - vw + \cancel{2\lambda}\right) \\ &= \beta_{\{t,u\},\{v,w\}}\end{aligned}$$

If instead A_0 or B_0 contains three of t, u, v, w
– say WLOG $t, u, v \in A_0$ and $w \in B_0$ – then

$$\begin{aligned}\widetilde{\beta}_{\{t,u\},\{v,w\}} &= \frac{1}{2}\left(\max\begin{Bmatrix} tv + (uw - \lambda) \\ (tw - \lambda) + uv \\ tu + (vw - \lambda)\end{Bmatrix} - tu - (vw - \lambda)\right) \\ &= \frac{1}{2}\left(\max\begin{Bmatrix} tv + uw \\ tw + uv \\ tu + vw\end{Bmatrix} \cancel{-\lambda} - tu - (vw \cancel{-\lambda})\right) \\ &= \beta_{\{t,u\},\{v,w\}}\end{aligned}$$

If instead all of t, u, v, w are contained in A_0 or all in B_0,
then $\widetilde{d}$ acts as d on these elements, so

$$\widetilde{\beta}_{\{t,u\},\{v,w\}} = \beta_{\{t,u\},\{v,w\}}$$

Finally, we claim that for every split $\{A, B\}$

$$\widetilde{\alpha}_{A,B} = \begin{cases} \alpha_{A,B} - \lambda\,, & \text{if } \{A,B\} = \{A_0, B_0\} \\ \alpha_{A,B}\,, & \text{otherwise} \end{cases}$$

Since, as has been proven before, for every $t, u \in A_0\,,\ v, w \in B_0$

$$\widetilde{\beta}_{\{t,u\},\{v,w\}} = \beta_{\{t,u\},\{v,w\}} - \lambda$$

then, by taking the minimum, $\widetilde{\alpha}_{A_0,B_0} = \alpha_{A_0,B_0} - \lambda\,.$

Now suppose $\{A, B\} \neq \{A_0, B_0\}\,.$

We are going to prove the double inequality between $\alpha_{A,B}$ and $\widetilde{\alpha}_{A,B}$.

Choose $a, a' \in A$ and $b, b' \in B$ such that

$$\alpha_{A,B} = \beta_{\{a,a'\},\{b,b'\}}$$

We claim that $\{A_0, B_0\}$ cannot extend $\big\{\{a, a'\}, \{b, b'\}\big\}$.

In fact, if by absurd it is an extension, then there are two cases.

If $\alpha_{A,B} = 0\,,$ then $\beta_{\{a,a'\},\{b,b'\}} = \alpha_{A,B} = 0\,.$
Since $\{A_0, B_0\} \succcurlyeq \big\{\{a, a'\}, \{b, b'\}\big\}\,,$ we have $\alpha_{A_0,B_0} = 0\,.$
But $\{A_0, B_0\}$ is a d-split. ↯

If $\alpha_{A,B} > 0\,,$ then $\{A, B\}$ is a d-split extending $\big\{\{a, a'\}, \{b, b'\}\big\}$.
But by Corollary 2.6, such a d-split extension is unique,
so $\{A, B\} = \{A_0, B_0\}$. ↯

Since we proved that $\{A_0, B_0\} \not\succcurlyeq \big\{\{a, a'\}, \{b, b'\}\big\}\,,$ we get

$$\alpha_{A,B} = \beta_{\{a,a'\},\{b,b'\}} = \widetilde{\beta}_{\{a,a'\},\{b,b'\}} \geq \widetilde{\alpha}_{A,B}$$

Now let $t, u \in A$ and $v, w \in B\,.$

If $\{A_0, B_0\}$ does not extend $\big\{\{t, u\}, \{v, w\}\big\}$, then

$$\alpha_{A,B} \leq \beta_{\{t,u\},\{v,w\}} = \widetilde{\beta}_{\{t,u\},\{v,w\}}$$

Otherwise, if $\{A_0, B_0\}$ does extend $\big\{\{t,u\},\{v,w\}\big\}$, then

$$\begin{aligned}\lambda &\le \alpha_{A_0,B_0}\\ \alpha_{A,B} + \lambda &\le \alpha_{A_0,B_0} + \alpha_{A,B}\end{aligned}$$

$$\begin{aligned}\alpha_{A,B} &\le (\alpha_{A,B} + \alpha_{A_0,B_0}) - \lambda\\ &\le \alpha_{\{t,u\},\{v,w\}} - \lambda\\ &\le \beta_{\{t,u\},\{v,w\}} - \lambda\\ &= \widetilde{\beta}_{\{t,u\},\{v,w\}}\end{aligned}$$

where we used Theorem 2.5, since $\{A, B\}$ and $\{A_0, B_0\}$ are two extensions of $\big\{\{t,u\},\{v,w\}\big\}$, which is a partial d-split:
in fact, if by absurd $\alpha_{\{t,u\},\{v,w\}} = 0$, then also $\alpha_{A_0,B_0} = 0$. ↯

Since we proved that for every $t, u \in A,\ v, w \in B$

$$\alpha_{A,B} \le \widetilde{\beta}_{\{t,u\},\{v,w\}}$$

then

$$\alpha_{A,B} \le \min \widetilde{\beta}_{\{t,u\},\{v,w\}} = \widetilde{\alpha}_{A,B}$$

□

Corollary 2.9 (**split decomposition / canonical decomposition**)

For any symmetric function $d : X \times X \to \mathbb{R}$ we can write

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

where d_0 is a split-prime (symmetric) function.

We call d_0 the **split-prime residue** of d.

Proof

Apply Theorem 2.8 with $\lambda_S = \alpha_S^d,\ \forall\, S \in \mathcal{S}_d(X)$. Let

$$d_0 := d - \sum_{S \in \mathcal{S}(X)} \alpha_S^d \cdot \delta_S$$

Then, if S is a d-split $\alpha_S^{d_0} = \alpha_S^d - \lambda_S = \alpha_S^d - \alpha_S^d = 0$
otherwise $\alpha_S^{d_0} = \alpha_S^d - \lambda_S = 0 - 0 = 0$.

□

Remark We can sum over only the d-splits, since the others do not contribute (they have zero coefficient).

Remark The residue of a split-prime function coincides with the function itself (there are no splits on which to decompose).

Corollary 2.10 **([BD92a, Corollary 2])**

Suppose $d : X \times X \to \mathbb{R}$ is a pseudo-metric.
Let $\lambda \geq 0$ and $x \in X$. Consider

$$d' := d + \lambda \cdot \delta_x$$

Then d' and d have the same split-prime residue and

$$\alpha_S^{d'} = \begin{cases} \alpha_S^d + \lambda\,, & \text{if } S = \big\{\{x\},\, X \setminus \{x\}\big\} \\ \alpha_S^d\,, & \text{otherwise} \end{cases}$$

Proof

If $\lambda = 0$, then $d' = d$, and there is nothing to prove.
So we can assume WLOG that $\lambda > 0$.

Observe that for every $u, v \in X \setminus \{x\}$ we have

$$\begin{aligned} \beta^{d'}_{\{x\},\{u,v\}} &= \frac{1}{2}\left(\max\left\{\begin{matrix}(xu+\lambda)+(xv+\lambda)\\ \cancel{xx}+uv\end{matrix}\right\} - \cancel{xx} - uv\right) \\ &= \frac{1}{2}\left(xu + xv + 2\lambda - uv\right) \\ &= \lambda + \frac{1}{2}\left(xu + xv - uv\right) \geq \lambda \end{aligned}$$

where we used the triangle inequality for d.

Therefore $\alpha^{d'}_{\{x\},X\setminus\{x\}} \geq \lambda > 0$ and thus $\big\{\{x\},\, X \setminus \{x\}\big\}$ is a d'-split.
Since $d = d' - \lambda \cdot \delta_x$, we get the thesis by applying Theorem 2.8 to d'.

Moreover, since $\alpha^{d'}_{\{x\},X\setminus\{x\}} = \alpha^{d}_{\{x\},X\setminus\{x\}} + \lambda$

$$\begin{aligned} (d')_0 &= d' - \sum \alpha_S^{d'} \cdot \delta_S - \alpha^{d'}_{\{x\},X\setminus\{x\}} \cdot \delta_x \\ &= (d + \cancel{\lambda \cdot \delta_x}) - \sum \alpha_S^{d} \cdot \delta_S - \big(\alpha^{d}_{\{x\},X\setminus\{x\}} \cancel{+\lambda}\big) \cdot \delta_x = d_0 \end{aligned}$$

where the sums range over the splits different from $\big\{\{x\},\, X \setminus \{x\}\big\}$. □

Proposition 2.11

In the same hypothesis of Theorem 2.8,
for every partial split T we have

$$\alpha_T^{d'} = \alpha_T^d - \sum \lambda_S$$

where the sum ranges over all the (total) splits that extend T.

Proof

We can recycle the proof of Theorem 2.8 with some minor variations.
For brevity the steps that are identical are omitted.

Consider a d-split $\{A_0, B_0\}$, $\lambda \leq \alpha^d_{A_0,B_0}$ and

$$\widetilde{d} := d - \lambda \cdot \delta_{A_0,B_0}$$

We claim that for every partial split $\{A, B\}$ we have

$$\widetilde{\alpha}_{A,B} = \begin{cases} \alpha_{A,B} - \lambda\,, & \text{if } \{A_0, B_0\} \succcurlyeq \{A, B\} \\ \alpha_{A,B}\,, & \text{otherwise} \end{cases}$$

Suppose $\{A_0, B_0\} \succcurlyeq \{A, B\}$.

We know from the first part of the proof that

$$\widetilde{\beta}_{\{t,u\},\{v,w\}} = \begin{cases} \beta_{\{t,u\},\{v,w\}} - \lambda\,, & \text{if } \{A_0, B_0\} \succcurlyeq \big\{\{t, u\}, \{v, w\}\big\} \\ \beta_{\{t,u\},\{v,w\}}\,, & \text{otherwise} \end{cases}$$

For every $t, u \in A$, $v, w \in B$,
we have that $\{A_0, B_0\} \succcurlyeq \big\{\{t, u\}, \{v, w\}\big\}$, so

$$\widetilde{\beta}_{\{t,u\},\{v,w\}} = \beta_{\{t,u\},\{v,w\}} - \lambda$$

Then, by taking the minimum on A, B, we have $\widetilde{\alpha}_{A,B} = \alpha_{A,B} - \lambda$.

Now suppose $\{A_0, B_0\} \not\succcurlyeq \{A, B\}$.

We claim that $\{A_0, B_0\}$ cannot extend the quartet that realizes $\alpha^d_{A,B}$.

Suppose by absurd that $\{A_0, B_0\}$ is indeed such an extension.

In order to apply the Corollary 2.6 in the case $\alpha_{A,B} > 0$,
we need to restrict $\{A_0, B_0\}$ to $A \cup B$.

Since both $\alpha_{A_0,B_0} > 0$ and $\alpha_{A,B} > 0$,
we have that A_0, B_0 and A, B respectively are disjoint;
so we can suppose WLOG that $A_0, A \ni a, a'$ and $B_0, B \ni b, b'$.

Then $\{A_0 \cap A, B_0 \cap B\}$ extends $\big\{\{a, a'\}, \{b, b'\}\big\}$ and it is a d-split (on $A \cup B$) because restriction of the d-split $\{A_0, B_0\}$.
Since also $\{A, B\}$ is a d-split on $A \cup B$ extending $\big\{\{a, a'\}, \{b, b'\}\big\}$, from Corollary 2.6 we have

$$\{A_0, B_0\} \succcurlyeq \{A_0 \cap A, B_0 \cap B\} = \{A, B\} \quad \lightning$$

In order to apply Theorem 2.5, consider the extension tree rooted in $\big\{\{t, u\}, \{v, w\}\big\}$ that contains $\{A, B\}$ (see the remark of the theorem).

We backtrack the total split $\{A_0, B_0\}$ to an ancestor partial split $\{A', B'\}$ at the same level of $\{A, B\}$; observe that $\{A', B'\} \neq \{A, B\}$ because $\{A_0, B_0\}$ is not an extension of $\{A, B\}$.

Given S, T partial splits in the tree, we denote with $[\alpha_S]^T$ the sum of the isolation indices of all the partial splits on the same level of S that have T as common ancestor (we may omit T if it is the root). Then

$$\alpha_{A_0,B_0} \leq [\alpha_{A_0,B_0}]^{A',B'} \leq \alpha_{A',B'}$$
$$\alpha_{A,B} + \alpha_{A',B'} \leq [\alpha_{A,B}] = [\alpha_{A',B'}] \leq \alpha_{\{t,u\},\{v,w\}}$$

from which

$$\alpha_{A,B} + \alpha_{A_0,B_0} \leq \alpha_{A,B} + \alpha_{A',B'} \leq \alpha_{\{t,u\},\{v,w\}}$$

As before, we get the thesis by induction on the number of split metrics.

□

Corollary 2.12

For every partial split T of X we have

$$\alpha_T^{d_0} = \alpha_T^d - \sum \alpha_S^d$$

where the sum ranges over all the d-splits that extend T.

Proof

Apply Proposition 2.11 with $\lambda_S = \alpha_S^d,\ \forall\, S \in \mathcal{S}_d(X)$.

□

Corollary 2.13 **([BD92a, Corollary 3])**

Every partial d_0-split is also a partial d-split;

every partial d-split which does not extend to a (total) d-split, is also a partial d_0-split.

Proof

Let T be a partial d_0-split. Then by Corollary 2.12

$$\alpha_T^d = \underbrace{\alpha_T^{d_0}}_{>0} + \sum \underbrace{\alpha_S^d}_{\geq 0} > 0$$

Let T be a partial d-split which does not extend to a total d-split. Then by Corollary 2.12

$$\alpha_T^{d_0} = \underbrace{\alpha_T^d}_{>0} - \underbrace{\sum \alpha_S^d}_{=0} > 0$$

□

Chapter 3

Weak compatibility

Let $d : X \times X \to \mathbb{R}$ be a symmetric function.

Remark Let $t, u, v, w \in X$.
Then at least one of the following indices must be zero

$$\alpha_{\{t,u\},\{v,w\}} \quad \alpha_{\{t,v\},\{u,w\}} \quad \alpha_{\{t,w\},\{u,v\}}$$

In fact, suppose

$$\max\{tu + vw, tv + uw, tw + uv\} = tu + vw$$

then $\beta_{\{t,u\},\{v,w\}} = 0$ and so $\alpha_{\{t,u\},\{v,w\}} = 0$.

The other cases are analogous.

Definition (**weak compatibility**)

Three splits S_1, S_2, S_3 of X are **weakly compatible**
if there are no four points $t, u, v, w \in X$ such that

$$S_1 \succcurlyeq \big\{\{t, u\}, \{v, w\}\big\}, \quad S_2 \succcurlyeq \big\{\{t, v\}, \{u, w\}\big\}, \quad S_3 \succcurlyeq \big\{\{t, w\}, \{u, v\}\big\}$$

A set of splits $\mathcal{S}$ of X is **weakly compatible**
if its splits are (triplewise) weakly compatible.

Remark Subsets of weakly compatible sets are weakly compatible.

Proposition 3.1

The set of all d-splits $\mathcal{S}_d(X)$ is weakly compatible.

Proof

Let $t, u, v, w \in X$.
From the previous remark we can suppose WLOG $\alpha_{\{t,u\},\{v,w\}} = 0$.

For every split S that extends $\big\{\{t,u\},\{v,w\}\big\}$ we have

$$0 \leq \alpha_S \leq \alpha_{\{t,u\},\{v,w\}} = 0$$

that implies $\alpha_S = 0$, that is S is not a d-split.
In other words, there are no d-splits that extend $\big\{\{t,u\},\{v,w\}\big\}$.

□

Theorem 3.2 **([BD92a, Theorem 3])**

Let $\mathcal{S}$ be a set of weakly compatible splits of X.
For each $S \in \mathcal{S}$, let $\lambda_S > 0$ and consider

$$d := \sum_{S \in \mathcal{S}} \lambda_S \cdot \delta_S$$

Then $\mathcal{S} = \mathcal{S}_d(X)$ and $\alpha_S = \lambda_S$, $\forall\, S \in \mathcal{S}$.

Proof

Notice that d is a conical combination of split metrics
(which are pseudo-metrics), so it is a pseudo-metric.

Let $\{A, B\} \in \mathcal{S}$.
Thanks to Corollary 2.2, pick $t, u \in A$ and $v, w \in B$ such that

$$\alpha_{\{t,u\},\{v,w\}} = \alpha_{A,B}$$

Consider the sets

$$\mathcal{S}_0 = \Big\{ S \in \mathcal{S} \;\Big|\; S \succcurlyeq \big\{\{t,u\},\{v,w\}\big\} \Big\}$$

$$\mathcal{S}_1 = \Big\{ S \in \mathcal{S} \;\Big|\; S \succcurlyeq \big\{\{t,v\},\{u,w\}\big\} \Big\}$$

$$\mathcal{S}_2 = \Big\{ S \in \mathcal{S} \;\Big|\; S \succcurlyeq \big\{\{t,w\},\{u,v\}\big\} \Big\}$$

If all three sets are non-empty, then there exist three splits that violates the weak compatibility assumption.
So at least one of them is empty, say WLOG $\mathcal{S}_2$.

All splits in $\mathcal{S} \setminus (\mathcal{S}_0 \cup \mathcal{S}_1 \cup \mathcal{S}_2)$ equally contribute to each of the three distances $tu + vw$, $tv + uw$, $tw + uv$;
so we can ignore them in the calculation of the isolation index:

$$\begin{aligned}
\alpha_{A,B} = \alpha_{\{t,u\},\{v,w\}} = \beta_{\{t,u\},\{v,w\}} &= \\
&= \frac{1}{2}\Big(\max\{tu+vw,\, tv+uw,\, tw+uv\} - tu - vw\Big) \\
&= \max\left\{ \sum_{S\in\mathcal{S}_1\cup\underbrace{\mathcal{S}_2}_{=\emptyset}} \lambda_S\,,\ \sum_{S\in\mathcal{S}_0\cup\underbrace{\mathcal{S}_2}_{=\emptyset}} \lambda_S\,,\ \sum_{S\in\mathcal{S}_0\cup\mathcal{S}_1} \lambda_S \right\} - \sum_{S\in\mathcal{S}_1\cup\underbrace{\mathcal{S}_2}_{=\emptyset}} \lambda_S \\
&= \sum_{S\in\mathcal{S}_0\cup\mathcal{S}_1} \lambda_S \ - \sum_{S\in\mathcal{S}_1} \lambda_S \\
&= \sum_{S\in\mathcal{S}_0} \lambda_S \\
&\geq \lambda_{A,B} > 0
\end{aligned}$$

So for every $S \in \mathcal{S}$ we have $\alpha_S \geq \lambda_S > 0$ and S is a d-split, therefore $\mathcal{S} \subseteq \mathcal{S}_d(X)$.

Writing the canonical decompostion of d, we get

$$\begin{aligned}
d &= d_0 + \sum_{S\in\mathcal{S}_d(X)} \alpha_S \cdot \delta_S \\
&\geq \sum_{S\in\mathcal{S}} \alpha_S \cdot \delta_S \\
&\geq \sum_{S\in\mathcal{S}} \lambda_S \cdot \delta_S = d
\end{aligned}$$

where we used that $d_0 \geq 0$, since it is a pseudo-metric.

So the inequalities hold as equal. Thus

$$d_0 = 0\,, \qquad \begin{cases} \alpha_S = \lambda_S\,, & \forall\, S \in \mathcal{S} \\ \alpha_S = 0\,, & \text{otherwise} \end{cases}$$

that is if $S \notin \mathcal{S}$, then S is not a d-split.

We conclude that $\mathcal{S} = \mathcal{S}_d(X)$. □

Corollary 3.3 ([BD92a, Corollary 4])

Let $\mathcal{S}$ be a set of weakly compatible splits of X.

Then the split metrics $\{\delta_S\}_{S \in \mathcal{S}}$ are linearly independent.
Also $|\mathcal{S}| \leq \binom{n}{2}$, where $n = |X|$.

Proof

In order to prove the linear independence, suppose

$$\sum_{S \in \mathcal{S}} \lambda_S \cdot \delta_S = 0$$

for some $\lambda_S \in \mathbb{R}$ for each $S \in \mathcal{S}$.

Let

$$\mathcal{S}^+ := \{ S \in \mathcal{S} \mid \lambda_S > 0 \}$$
$$\mathcal{S}^- := \{ S \in \mathcal{S} \mid \lambda_S < 0 \}$$

We can decompose the previous expression in

$$\sum_{S \in \mathcal{S}^+} \lambda_S \cdot \delta_S \; + \sum_{S \in \mathcal{S}^-} \lambda_S \cdot \delta_S = 0$$

Consider the pseudo-metric

$$d := \sum_{S \in \mathcal{S}^+} \lambda_S \cdot \delta_S = \sum_{S \in \mathcal{S}^-} (-\lambda_S) \cdot \delta_S$$

Observe that both $\mathcal{S}^+$ and $\mathcal{S}^-$ are weakly compatible,
because subsets of $\mathcal{S}$ which is weakly compatible.
Applying Theorem 3.2 to the first expression of d we get

$$\mathcal{S}^+ = \mathcal{S}_d(X)$$

and doing the same with the second expression we get

$$\mathcal{S}^- = \mathcal{S}_d(X)$$

So $\mathcal{S}^+ = \mathcal{S}^-$. But $\mathcal{S}^+$ and $\mathcal{S}^-$ are disjoint due to how they are defined.
So they are both empty: $\mathcal{S}^+ = \mathcal{S}^- = \emptyset$.

We conclude that $\lambda_S = 0, \; \forall\, S \in \mathcal{S}$.
Therefore the split metrics $\{\delta_S\}_{S \in \mathcal{S}}$ are linearly independent.

Since $\delta_S \in M(X)$, $\forall\, S \in \mathcal{S}$ and $\dim_{\mathbb{R}} \langle M(X) \rangle = \binom{n}{2}$ then

$$|\mathcal{S}| = \#\{\delta_S\}_{S \in \mathcal{S}} \leq \binom{n}{2}$$

□

Remark Since $\mathcal{S}_d(X)$ is weakly compatible (Proposition 3.1), from Corollary 3.3 we have that the number of d-splits is at most $\binom{n}{2}$.

A brute force approach to find all the d-splits of X would be to compute the isolation indices of all the splits of X and discard those that are zero. But, since $|\mathcal{S}(X)| = 2^{n-1} - 1$, this is an exponential algorithm.

We can instead use a more "inductive" approach:
suppose that the $d|_{Y \times Y}$-splits of a proper subset $Y \subset X$ of size k have already been determined. Then pick any $x \in X \setminus Y$ and check

- if $\{Y, \{x\}\}$ is a partial d-split
- if $\{A, B \cup \{x\}\}$ and $\{A \cup \{x\}, B\}$ are partial d-split for any $\{A, B\}$ d-split of Y

In this way we obtain all the d-splits of $Y \cup \{x\}$.

Crucially, we just have to check at most $2 \cdot \binom{k}{2} + 1$ splits at each step, thanks to the previous remark.
So we can compute the d-splits of X (and their decomposition) in polynomial time.
In particular, this algorithm has complexity $O(n^6)$ (see Chapter 6 for the details of the calculation).

Corollary 3.4 **([BD92a, Corollary 5])**

Let $d : X \times X \to \mathbb{R}$ be a symmetric function.

Then the residue d_0 is linearly independent from $\{\delta_S\}_{S \in \mathcal{S}_d(X)}$.
In particular, if there are $\binom{n}{2}$ d-splits, then $d_0 = 0$.

If d is a pseudo-metric,
then d_0 is linearly independent from $\{\delta_S\}_{S \in \mathcal{S}_d(X)} \cup \{\delta_x\}_{x \in X}$.
If there are $\binom{n}{2} - n$ non-trivial d-splits, then $d_0 = 0$.

Proof

Suppose that

$$d_0 = \sum_{S \in \mathcal{S}_d(X)} \lambda_S \cdot \delta_S$$

so that

$$\begin{aligned} d &= d_0 + \sum_{S \in \mathcal{S}(X)} \alpha_S^d \cdot \delta_S \\ &= \sum_{S \in \mathcal{S}_d(X)} \lambda_S \cdot \delta_S + \sum_{S \in \mathcal{S}_d(X)} \alpha_S^d \cdot \delta_S \\ &= \sum_{S \in \mathcal{S}_d(X)} \left(\alpha_S^d + \lambda_S \right) \cdot \delta_S \end{aligned}$$

Let

$$\begin{aligned} \mathcal{S}^+ &:= \{ S \in \mathcal{S}_d(X) \mid \lambda_S \geq 0 \} \\ \mathcal{S}^- &:= \{ S \in \mathcal{S}_d(X) \mid \lambda_S < 0 \} \end{aligned}$$

Observe that $\mathcal{S}_d(X) = \mathcal{S}^+ \sqcup \mathcal{S}^-$.

Consider the pseudo-metric

$$d' := \sum_{S \in \mathcal{S}^+} \left(\alpha_S^d + \lambda_S \right) \cdot \delta_S + \sum_{S \in \mathcal{S}^-} \alpha_S^d \cdot \delta_S$$

Applying Theorem 3.2 we get

$$\alpha_S^{d'} = \begin{cases} \alpha_S^d + \lambda_S \,, & S \in \mathcal{S}^+ \\ \alpha_S^d \,, & S \in \mathcal{S}^- \end{cases}$$

We can write d' as

$$\begin{aligned} d' &= \sum_{S \in \mathcal{S}^+} \left(\alpha_S^d + \lambda_S \right) \cdot \delta_S + \sum_{S \in \mathcal{S}^-} \left(\alpha_S^d + \lambda_S - \lambda_S \right) \cdot \delta_S \\ &= \sum_{S \in \mathcal{S}^+} \left(\alpha_S^d + \lambda_S \right) \cdot \delta_S + \sum_{S \in \mathcal{S}^-} \left(\alpha_S^d + \lambda_S \right) \cdot \delta_S - \sum_{S \in \mathcal{S}^-} \lambda_S \cdot \delta_S \\ &= \sum_{S \in \mathcal{S}_d(X)} \left(\alpha_S^d + \lambda_S \right) \cdot \delta_S - \sum_{S \in \mathcal{S}^-} \lambda_S \cdot \delta_S \\ &= d - \sum_{S \in \mathcal{S}^-} \lambda_S \cdot \delta_S \end{aligned}$$

Applying Theorem 2.8 we get

$$\alpha_S^{d'} = \begin{cases} \alpha_S^d \,, & S \in \mathcal{S}^+ \\ \alpha_S^d - \lambda_S \,, & S \in \mathcal{S}^- \end{cases}$$

Thus $\lambda_S = 0$ for every split S, proving the linear independence.

The second assertion follows from Corollary 3.3 applied to $\mathcal{S}_d(X)$ and the fact that $\dim_{\mathbb{R}} \langle M(X) \rangle = \binom{n}{2}$.

Consider

$$d^* := d + \sum_{x \in X} \delta_x$$

By Corollary 2.10, $(d^*)_0 = d_0$. Also notice that

$$\mathcal{S}_{d^*}(X) = \mathcal{S}_d(X) \cup \bigcup_{x \in X} \{\{x\}, X \setminus \{x\}\}$$

We get the thesis by applying the first part to d^*.

□

Definition (**trace**)

Let $\mathcal{S}$ be a collection of splits of X and $Y \subseteq X$ a subset of X.

Then the **trace** of $\mathcal{S}$ on Y is the set

$$\mathcal{S}|_Y := \left\{ \{A \cap Y,\, B \cap Y\} \;\middle|\; \begin{array}{c} \{A, B\} \in \mathcal{S}, \\ A \not\supseteq Y \text{ and } B \not\supseteq Y \end{array} \right\}$$

In practice, it is the collection of the restrictions of the splits of $\mathcal{S}$ to Y
such that they are still splits; in fact, the parts of a split cannot be empty.
Since splits are partitions of X,
this is equivalent to ask that Y is not contained in neither part.

Remark Clearly $\mathcal{S}(Y) = \mathcal{S}(X)|_Y$.

Given a partial split $\{A, B\}$ of Y, we have $\alpha^d_{A,B} = \alpha^{d|_{Y\times Y}}_{A,B}$:
in fact d and $d|_{Y\times Y}$ coincide on $A \cup B$, since $A, B \subseteq Y$.

For this reason, we may refer to the $d|_{Y\times Y}$-splits as d-splits of Y
and indicate them with $\mathcal{S}_d(Y)$.

Moreover $\mathcal{S}_d(Y) \supseteq \mathcal{S}_d(X)|_Y$, because by restricting
the isolation index cannot get lower.

Corollary 3.5 (**[BD92a, Corollary 6]**)

Let $\mathcal{S}$ and $\mathcal{T}$ be collections of weakly compatible splits of X.

Then $\mathcal{S} = \mathcal{T}$ if and only if
their traces are identical on every 4-subset of X.

Proof

The ($\Rightarrow$) implication is obvious.

Consider the pseudo-metrics
(since they are conical combinations of split metrics)

$$d_1 := \sum_{S \in \mathcal{S}} \delta_S \,, \qquad d_2 := \sum_{T \in \mathcal{T}} \delta_T$$

Observe that, given $Y = \{t, u, v, w\}$ a 4-subset of X, we have

$$d_1|_Y = d_2|_Y$$

because they depend only on the restriction of the split metrics on Y, and $\mathcal{S}, \mathcal{T}$ have the same trace on 4-subsets.

If we consider a split of Y, for example $\{\{t, u\}, \{v, w\}\}$, thanks to Proposition 2.1, we have

$$\alpha^{d_1}_{\{t,u\},\{v,w\}} = \beta^{d_1}_{\{t,u\},\{v,w\}} = \beta^{d_2}_{\{t,u\},\{v,w\}} = \alpha^{d_2}_{\{t,u\},\{v,w\}}$$

In particular, the isolation index on quartets is positive with respect to d_1 if and only if it is positive with respect to d_2.

This is true also for generic splits because the isolation index on a generic split is the minimum of the isolation indices on appropriate quartets.
In particular, the d_1-splits coincide with the d_2-splits.

By Theorem 3.2 we have

$$\mathcal{S} = \mathcal{S}_{d_1}(X) = \mathcal{S}_{d_2}(X) = \mathcal{T}$$

□

Proposition 3.6

Let $\mathcal{S}$ be a collection of weakly compatible splits of X.

Then for every $S \in \mathcal{S}$ there exists a partial split $\{\{t, u\}, \{v, w\}\}$ such that S is its unique split extension in $\mathcal{S}$.

Proof

Let $S \in \mathcal{S}$ and $\{\{t, u\}, \{v, w\}\}$ such that

$$\alpha^d_S = \beta^d_{\{t,u\},\{v,w\}}$$

with respect to the pseudo-metric $d = \sum_{S \in \mathcal{S}} \delta_S$.

Since by Theorem 3.2 it holds $\mathcal{S} = \mathcal{S}_d(X)$, then S is a d-split.
Applying Corollary 2.6 we conclude that S is the unique d-split (that is equivalent to saying element of $\mathcal{S}$) extending $\{\{t, u\}, \{v, w\}\}$.

□

Chapter 4

Total decomposability

Definition (**total decomposability**)

A symmetric function $d : X \times X \to \mathbb{R}$ is **totally decomposable** if its split-prime residue is zero $d_0 = 0$;

or equivalently, if it can be written as

$$d = \sum_{S \in \mathcal{S}_d(X)} \alpha_S^d \cdot \delta_S$$

Remark A totally decomposable function is a pseudo-metric.

In fact, it is a conical combination of split metrics (which are pseudo-metrics).

Remark We have already seen in Chapter 1 that in the cases $n = 2, 3$ all the pseudo-metrics are totally decomposable (observe that *a posteriori* the coefficients found are exactly the isolation indices of the corresponding split metric).

We now prove that this holds also for $n = 4$, but not for $n \geq 5$.

Lemma 4.1

Suppose d is a pseudo-metric on $X = \{t, u, v, w\}$ and

$$\alpha_{\{t,u\},\{v,w\}} = \alpha_{\{t,v\},\{u,w\}} = \alpha_{\{t,w\},\{u,v\}} = 0$$

Then, for every $x, y \in X$

$$d(x, y) = \alpha_{\{x\},X\setminus\{x\}} + \alpha_{\{y\},X\setminus\{y\}}$$

Proof

We can assume WLOG that $x = t$ and $y = u$.

From Proposition 2.1 we have

$$\alpha_{\{t,u\},\{v,w\}} = \frac{1}{2}\Big(\max\{tv + uw,\ tw + uv,\ tu + vw\} - tu - vw\Big)$$
$$\alpha_{\{t,v\},\{u,w\}} = \frac{1}{2}\Big(\max\{tu + vw,\ tw + uv,\ tv + uw\} - tv - uw\Big)$$
$$\alpha_{\{t,w\},\{u,v\}} = \frac{1}{2}\Big(\max\{tu + vw,\ tv + uw,\ tw + uv\} - tw - uv\Big)$$

and the hypothesis that they are all zero implies

$$\max\{tu + vw,\ tv + uw,\ tw + uv\} = tu + vw = tv + uw = tw + uv$$

By rearranging we get also

$$tu = tv + uw - vw = tw + uv - vw$$

Observe that by triangle inequality

$$\beta_{\{t\},\{u,v\}} = \frac{1}{2}\Big(\max\{tu + tv,\ tv + tu,\ \cancel{tt} + uv\} - \cancel{tt} - uv\Big)$$
$$= \frac{1}{2}(tu + tv - uv)$$

and similarly

$$\beta_{\{t\},\{u,w\}} = \frac{1}{2}(tu + tw - uw)$$
$$\beta_{\{t\},\{v,w\}} = \frac{1}{2}(tv + tw - vw)$$

So we can rewrite the thesis as

$$\begin{aligned} tu &= \alpha_{\{t\},\{u,v,w\}} + \alpha_{\{u\},\{t,v,w\}} \\ &= \frac{1}{2}\min\left\{\begin{array}{l} tu+tv-uv, \\ tu+tw-uw, \\ tv+tw-vw \end{array}\right\} + \frac{1}{2}\min\left\{\begin{array}{l} tu+uv-tv, \\ tu+uw-tw, \\ uv+uw-vw \end{array}\right\} \end{aligned}$$

We apply the equalities proven at the beginning in the various cases.

$$(tu + \cancel{tv} - \cancel{uv}) + (tu + \cancel{uv} - \cancel{tv}) = 2\,tu$$

$$(tu + \cancel{tw} - \cancel{uw}) + (tu + \cancel{uw} - \cancel{tw}) = 2\,tu$$

$$\begin{aligned} &(tu + tv - uv) + (tu + uw - tw) = \\ &\quad = 2\,tu + (tv + uw) - (tw + uv) = 2\,tu \end{aligned}$$

$$\begin{aligned} &(tu + tw - uw) + (tu + uv - tv) = \\ &\quad = 2\,tu + (tw + uv) - (tv + uw) = 2\,tu \end{aligned}$$

$$\begin{aligned} &(tv + tw - vw) + (uv + uw - vw) = \\ &\quad = (tv + uw - vw) + (tw + uv - vw) = 2\,tu \end{aligned}$$

$$\begin{aligned} &(\cancel{tv} + tw - vw) + (tu + uv - \cancel{tv}) = \\ &\quad = tu + (tw + uv - vw) = 2\,tu \end{aligned}$$

$$\begin{aligned} &(tv + \cancel{tw} - vw) + (tu + uw - \cancel{tw}) = \\ &\quad = tu + (tv + uw - vw) = 2\,tu \end{aligned}$$

$$\begin{aligned} &(tu + tv - \cancel{uv}) + (\cancel{uv} + uw - vw) = \\ &\quad = tu + (tv + uw - vw) = 2\,tu \end{aligned}$$

$$\begin{aligned} &(tu + tw - \cancel{uw}) + (uv + \cancel{uw} - vw) = \\ &\quad = tu + (tw + uv - vw) = 2\,tu \end{aligned}$$

□

Proposition 4.2

The following conditions are equivalent:

(i) every pseudo-metric on X is totally decomposable

(ii) there are no non-zero split-prime pseudo-metrics on X

(iii) the cone generated by the split metrics coincide with $M(X)$

Proof

(i $\Leftrightarrow$ ii) If all the split-prime functions are the zero function,
then the residue is also the zero function, since it is split-prime.

Otherwise, if the residue is always the zero function,
then all the split-prime functions are the zero function
because they coincide with their residue.

(i $\Leftrightarrow$ iii) A pseudo-metric is totally decomposable if and only if
it can be written as a conical combination of split metrics.

□

Proposition 4.3

Every pseudo-metric on 4 elements is totally decomposable.

Proof

This is equivalent to say that if $|X| = 4$,
then there are no non-zero split-prime pseudo-metrics on X.

Suppose by absurd that d is a split-prime pseudo-metrics on X.
Then all its isolation indices (relative to total splits) are equal to 0.
But for Lemma 4.1 this implies that d is the zero function.

□

Remark If we discard the split metric $\delta_{\{i,j\},\{h,k\}}$,
where $\{i,j,h,k\} = \{t,u,v,w\} =: X$, such that

$$ij + hk = \max\{tu + vw,\ tv + uw,\ tw + uv\}$$

(that is $\alpha_{\{i,j\},\{h,k\}} = 0$),
then we have a unique decomposition in split metrics for $n = 4$

$$d = \sum_{S \in \mathcal{S}_d(X)} \alpha_S^d \cdot \delta_S$$

because the remaining split metrics are a vector basis of $M(X)$
(they are linearly independent and in the right number).

Proposition 4.4

The metric on 5 elements $\hat{d}$ induced by the graph $K_{2,3}$ is split-prime.

Proof

It suffice to prove that there are no splits into two disjoint $\hat{d}$-convex subsets.

□

Remark In general, given a proper subset $Y \subset X$,
it is not true that the restriction of the residue
coincides with the residue of the restriction.

$$d_0|_{Y\times Y} \neq \left(d|_{Y\times Y}\right)_0$$

In fact, the restriction of the metric $\hat{d}$
(that coincide with its residue since it is split-prime)
is not the zero function because it is never 0 outside the diagonal.
But its restrictions are totally decomposable,
so their residues are the zero function.

Theorem 4.5 ([BD92a, Theorem 6])

Let $d : X \times X \to \mathbb{R}$ be a symmetric function with zero diagonal. Then the following conditions are equivalent:

(i) d is totally decomposable

(ii) for every partial split T

$$\alpha_T = \sum \left\{ \alpha_S \mid S \in \mathcal{S}_d(X),\ S \succcurlyeq T \right\}$$

(iii) for all $t, u, v, w, x \in X$

$$\alpha_{\{t,u\},\{v,w\}} = \alpha_{\{t,u,x\},\{v,w\}} + \alpha_{\{t,u\},\{v,w,x\}}$$

(iv) for all $t, u, v, w, x \in X$

$$\alpha_{\{t,u\},\{v,w\}} \leq \alpha_{\{t,x\},\{v,w\}} + \alpha_{\{t,u\},\{v,x\}}$$

Proof

(i $\Rightarrow$ ii) By definition of total decomposability, we have

$$d = \sum_{S \in \mathcal{S}_d(X)} \alpha_S \cdot \delta_S$$

For any proper subset $Y \subset X$

$$d|_{Y \times Y} = \sum_{S \in \mathcal{S}_d(X)} \alpha_S \cdot \delta_S|_{Y \times Y} = \sum \left\{ \lambda_T \cdot \delta_T \;\middle|\; T \in \mathcal{S}(Y) \right\}$$

where $\lambda_T = \sum \left\{ \alpha_S \;\middle|\; S \in \mathcal{S}_d(X),\ S \succcurlyeq T \right\}$.

From Theorem 2.5, we have $\lambda_T \leq \alpha_T$ for every T split of Y. Thus

$$\left\{ T \in \mathcal{S}(Y) \;\middle|\; \lambda_T > 0 \right\} \subseteq \mathcal{S}_d(Y)$$

so it is weakly compatible.

Applying Theorem 3.2 to this set and $d|_{Y \times Y}$ we get $\alpha_T = \lambda_T$.

(ii ⇒ iii) Let $t, u, v, w, x \in X$. Then

$$\begin{aligned}\alpha_{\{t,u\},\{v,w\}} &= \sum \left\{ \alpha_S \;\middle|\; S \in \mathcal{S}_d(X),\ S \succcurlyeq \{\{t,u\},\{v,w\}\} \right\} \\ &= \sum \left\{ \alpha_S \;\middle|\; S \in \mathcal{S}_d(X),\ S \succcurlyeq \{\{t,u,x\},\{v,w\}\} \right\} \\ &\quad + \sum \left\{ \alpha_S \;\middle|\; S \in \mathcal{S}_d(X),\ S \succcurlyeq \{\{t,u\},\{v,w,x\}\} \right\} \\ &= \alpha_{\{t,u,x\},\{v,w\}} + \alpha_{\{t,u\},\{v,w,x\}}\end{aligned}$$

(iii ⇒ iv) Let $t, u, v, w, x \in X$. Then

$$\begin{aligned}\alpha_{\{t,u\},\{v,w\}} &= \alpha_{\{t,u,x\},\{v,w\}} + \alpha_{\{t,u\},\{v,w,x\}} \\ &\leq \alpha_{\{t,x\},\{v,w\}} + \alpha_{\{t,u\},\{v,x\}}\end{aligned}$$

(iv ⇒ iii) Let $t, u, v, w, x \in X$. By Theorem 2.5 we have

$$\alpha_{\{t,u,x\},\{v,w\}} + \alpha_{\{t,u\},\{v,w,x\}} \leq \alpha_{\{t,u\},\{v,w\}}$$

On the other hand, thanks to Proposition 2.1,

$$\begin{aligned}\alpha_{\{t,u,x\},\{v,w\}} &= \min\{\alpha_{\{t,u\},\{v,w\}}, \alpha_{\{t,x\},\{v,w\}}, \alpha_{\{u,x\},\{v,w\}}\} \\ \alpha_{\{t,u\},\{v,w,x\}} &= \min\{\alpha_{\{t,u\},\{v,w\}}, \alpha_{\{t,u\},\{v,x\}}, \alpha_{\{t,u\},\{w,x\}}\}\end{aligned}$$

Applying condition (iv) with respect to x and either

$$t, u; v, w\,, \quad t, u; w, u\,, \quad u, t; v, w\,, \quad u, t; w, v$$

we get

$$\begin{aligned}\alpha_{\{t,u\},\{v,w\}} &\leq \alpha_{\{t,x\},\{v,w\}} + \alpha_{\{t,u\},\{v,x\}} \\ \alpha_{\{t,u\},\{w,v\}} &\leq \alpha_{\{t,x\},\{w,v\}} + \alpha_{\{t,u\},\{w,x\}} \\ \alpha_{\{u,t\},\{v,w\}} &\leq \alpha_{\{u,x\},\{v,w\}} + \alpha_{\{u,t\},\{v,x\}} \\ \alpha_{\{u,t\},\{w,v\}} &\leq \alpha_{\{u,x\},\{w,v\}} + \alpha_{\{u,t\},\{w,x\}}\end{aligned}$$

Therefore

$$\begin{aligned}
\alpha_{\{t,u,x\},\{v,w\}} + \alpha_{\{t,u\},\{v,w,x\}} &= \min \left\{ \begin{array}{c} \alpha_{\{t,u\},\{v,w\}} + \alpha_{\{t,u\},\{v,w\}}, \\ \alpha_{\{t,x\},\{v,w\}} + \alpha_{\{t,u\},\{v,w\}}, \\ \alpha_{\{u,x\},\{v,w\}} + \alpha_{\{t,u\},\{v,w\}}, \\ \alpha_{\{t,u\},\{v,w\}} + \alpha_{\{t,u\},\{v,x\}}, \\ \alpha_{\{t,x\},\{v,w\}} + \alpha_{\{t,u\},\{v,x\}}, \\ \alpha_{\{u,x\},\{v,w\}} + \alpha_{\{t,u\},\{v,x\}}, \\ \alpha_{\{t,u\},\{v,w\}} + \alpha_{\{t,u\},\{w,x\}}, \\ \alpha_{\{t,x\},\{v,w\}} + \alpha_{\{t,u\},\{w,x\}}, \\ \alpha_{\{u,x\},\{v,w\}} + \alpha_{\{t,u\},\{w,x\}} \end{array} \right\} \\
&\geq \min \left\{ \begin{array}{c} \alpha_{\{t,u\},\{v,w\}}, \\ \alpha_{\{t,x\},\{v,w\}} + \alpha_{\{t,u\},\{v,x\}}, \\ \alpha_{\{t,x\},\{v,w\}} + \alpha_{\{t,u\},\{w,x\}}, \\ \alpha_{\{u,x\},\{v,w\}} + \alpha_{\{t,u\},\{v,x\}}, \\ \alpha_{\{u,x\},\{v,w\}} + \alpha_{\{t,u\},\{w,x\}} \end{array} \right\} \\
&\geq \alpha_{\{t,u\},\{v,w\}}
\end{aligned}$$

where we lower bounded the expressions in the first line with $\alpha_{\{t,u\},\{v,w\}}$, and used the previous inequalities in the second line.

(iii $\Rightarrow$ i) **Omitted**.

□

Corollary 4.6

Let d a totally decomposable pseudo-metric, $\{A, B\}$ a d-split and $a_1, a_2 \in A$, $b_1, b_2 \in B$.

Then $\{A, B\}$ is the only d-split extension of $\{\{a_1, a_2\}, \{b_1, b_2\}\}$ if and only if $\alpha_{\{a_1,a_2\},\{b_1,b_2\}} = \alpha_{A,B}$.

Proof

($\Rightarrow$) By Theorem 4.5

$$\begin{aligned}
\alpha_{\{a_1,a_2\},\{b_1,b_2\}} &= \sum \left\{ \alpha_S \;\middle|\; S \in \mathcal{S}_d(X),\ S \succcurlyeq \{\{a_1, a_2\}, \{b_1, b_2\}\} \right\} \\
&= \alpha_{A,B}
\end{aligned}$$

since we are supposing $\{A, B\}$ is the only d-split extension.

($\Leftarrow$) By Theorem 4.5

$$\begin{aligned}\alpha_{A,B} &= \alpha_{\{a_1,a_2\},\{b_1,b_2\}} \\ &= \sum \left\{ \alpha_S \;\middle|\; S \in \mathcal{S}_d(X),\ S \succcurlyeq \{\{a_1,a_2\},\{b_1,b_2\}\} \right\}\end{aligned}$$

and since $\alpha_{A,B}$ is a term of the sum, it must be the only one.

□

Consider $a_1, a_2, a_3, a_4 \in X$ such that $\alpha_{\{a_1,a_2\},\{a_3,a_4\}} > 0$ and the sets

$$A := \{\, x \in X \mid \alpha_{\{a_1,a_2\},\{a_3,a_4,x\}} = 0 \,\}$$
$$B := \{\, x \in X \mid \alpha_{\{a_1,a_2,x\},\{a_3,a_4\}} = 0 \,\}$$

Suppose that the following identity holds for all $x \in X,\ x \neq a_1, a_2, a_3, a_4$

$$\alpha_{\{a_1,a_2\},\{a_3,a_4\}} = \alpha_{\{a_1,a_2,x\},\{a_3,a_4\}} + \alpha_{\{a_1,a_2\},\{a_3,a_4,x\}}$$

Then we have $a_1, a_2 \in A$ and $a_3, a_4 \in B$.

Notice that all the extensions of $\{\{a_1,a_2\},\{a_3,a_4\}\}$ with at least one element of A in the second part have isolation index equal to 0.
In fact $\alpha_{\{a_1,a_2\},\{a_3,a_4,x\}} = 0$ and by extending the isolation index cannot increase.
Idem for extensions of $\{\{a_1,a_2\},\{a_3,a_4\}\}$ with at least one element of B in the first part.

So between the split extensions of $\{\{a_1,a_2\},\{a_3,a_4\}\}$,
the only possibly non-zero isolation index is that of the split $\{A,B\}$
(that is all elements of A in the first part and all elements of B in the second part).

If d is totally decomposable, then by Theorem 4.5

$$\alpha_{\{a_1,a_2\},\{a_3,a_4\}} = \sum \left\{ \alpha_S \;\middle|\; S \in \mathcal{S}_d(X),\ S \succcurlyeq \{\{a_1,a_2\},\{a_3,a_4\}\} \right\} = \alpha_{A,B}$$

This suggest a polynomial algorithm to check whether a symmetric function d is totally decomposable and to compute the d-splits and their isolation indices:

- check the identity

$$\alpha_{\{a_1,a_2\},\{a_3,a_4\}} = \alpha_{\{a_1,a_2,x\},\{a_3,a_4\}} + \alpha_{\{a_1,a_2\},\{a_3,a_4,x\}}$$

  for all quartets $\big\{\{a_1,a_2\},\{a_3,a_4\}\big\}$ which have non-zero isolation index
- if the identity holds, then check if $A \cup B = X$
- if this is true, then $\{A, B\}$ is a d-split and its isolation index is

$$\alpha_{\{a_1,a_2\},\{a_3,a_4\}}$$

Due to the 5-point condition, this algorithm has complexity $\mathcal{O}(n^5)$.

Question: Does it exist an $\mathcal{O}(n^5)$ algorithm for the general case?

Definition (**compatibility**)

Given two splits $\{A, B\}$ and $\{A', B'\}$, we say that they are **compatible** if one of the following four intersections is empty

$$A \cap A', \quad A \cap B', \quad B \cap A', \quad B \cap B'$$

We say that a set of splits is **compatible** if its splits are (pairwise) compatible.

Remark Subsets of compatible sets are compatible.

Remark A compatible set of splits is weakly compatible.

In fact, consider for every quartet $\big\{\{t, u\}, \{v, w\}\big\}$ the sets

$$\mathcal{S}_0 = \Big\{ S \in \mathcal{S} \;\Big|\; S \succcurlyeq \big\{\{t, u\}, \{v, w\}\big\} \Big\}$$

$$\mathcal{S}_1 = \Big\{ S \in \mathcal{S} \;\Big|\; S \succcurlyeq \big\{\{t, v\}, \{u, w\}\big\} \Big\}$$

$$\mathcal{S}_2 = \Big\{ S \in \mathcal{S} \;\Big|\; S \succcurlyeq \big\{\{t, w\}, \{u, v\}\big\} \Big\}$$

Weak compatibility is equivalent to ask that at least *one* of these sets is empty.

Notice that two splits from two different sets are not compatible: in fact, if WLOG

$$\{A, B\} \in \mathcal{S}_0\,, \qquad A \ni t, u\,,\ B \ni v, w$$
$$\{A', B'\} \in \mathcal{S}_1\,, \qquad A' \ni t, v\,,\ B' \ni u, w$$

then

$$A \cap A' = \{t\}\,, \quad A \cap B' = \{u\}\,, \quad B \cap A' = \{v\}\,, \quad B \cap B' = \{w\}$$

Thus compatibility is equivalent to ask that at most one of the sets is non-empty (that is at least *two* must be empty).

Definition (four-point condition)

We say that d satisfies the **four-point condition**
if for any four points $t, u, v, w \in X$ it holds

$$tu + vw \leq \max\{\, tv + uw,\, tw + uv \,\}$$

Proposition 4.7

A symmetric function d satisfies the four-point condition
if and only if, for any four points $t, u, v, w \in X$,
it exists a permutation of t, u, v, w in i, j, h, k respectively such that

$$ij + hk \leq ih + jk = ik + jh$$

Proof

($\Leftarrow$) Consider the permutation that gives

$$ij + hk \leq ih + jk = ik + jh$$

Then we have

$$\begin{aligned} ij + hk &\leq \max\{\, ih + jk,\, ik + jh \,\} \\ ih + jk &\leq \max\{\, ij + hk,\, ik + jh \,\} \\ ik + jh &\leq \max\{\, ij + hk,\, ih + jk \,\} \end{aligned}$$

($\Rightarrow$) Applying the four-point condition to $t, u; v, w$, $t, v; u, w$ and $t, w; u, v$ we get

$$\begin{aligned} tu + vw &\leq \max\{\, tv + uw,\, tw + uv \,\} \\ tv + uw &\leq \max\{\, tu + vw,\, tw + uv \,\} \\ tw + uv &\leq \max\{\, tv + uw,\, tu + vw \,\} \end{aligned}$$

Let L, M, N be such that $\{L, M, N\} = \{tu + vw,\, tv + uw,\, tw + uv\}$
and $L \leq M \leq N$ (that is we order the three expressions).

If by absurd $N > M$ (and thus $N > L$) then

$$N \not\leq \max\{L, M\}$$

that violates the four-point condition. ↯

□

Remark If d is a pseudo-metric satisfying the four-point condition and $\alpha_{\{t,u\},\{v,w\}} > 0$, then

$$tu + vw < tv + uw = tw + uv$$

In fact, $tu + vw$ cannot be the maximum among the three (otherwise the isolation index would be zero for Proposition 2.1).

Proposition 4.8

If d is a totally decomposable pseudo-metric, then

$$\mathcal{S}_d(Y) = \mathcal{S}_d(X)|_Y\,, \quad \forall\, Y \subseteq X$$

Proof

Let $Y \subseteq X$. Clearly a d-split of X is also a d-split of Y, because by restricting the isolation index cannot decrease. So $\mathcal{S}_d(Y) \supseteq \mathcal{S}_d(X)|_Y$.

If $T \in \mathcal{S}_d(Y)$, from Theorem 4.5

$$0 < \alpha_T = \sum \left\{ \alpha_S \mid S \in \mathcal{S}_d(X),\ S \succcurlyeq T \right\}$$

that means there are d-splits on X which are extensions of T. Moreover, the restriction of these extensions to Y gives exactly T. Thus $\mathcal{S}_d(Y) \subseteq \mathcal{S}_d(X)|_Y$. □

Corollary 4.9 **([BD92a, Corollary 7])**

Let d be a pseudo-metric on X.

Then d is totally decomposable and any two d-splits are compatible if and only if d satisfies the four-point condition.

Proof

($\Rightarrow$) Let $t, u, v, w \in X$. We are assuming that d is totally decomposable and $\mathcal{S}_d(X)$ is compatible.

Since compatibility is preserved by restriction, from the previous Proposition 4.8 we get that $\mathcal{S}_d(Y)$ is compatible for all subsets Y of X. In particular $\mathcal{S}_d(\{t, u, v, w\})$ is compatible.

By compatibility, at least two quartets cannot be d-splits – suppose WLOG $\{\{t,v\},\{u,w\}\}$ and $\{\{t,w\},\{u,v\}\}$. This means that their isolation index is 0 and, by Proposition 2.1, this implies that

$$\max\{tu+vw,\ tv+uw,\ tw+uv\} = tv+uw = tw+uv$$

If $\alpha_{\{t,u\},\{v,w\}} = 0$, then for the same reason

$$tu+vw = tv+uw = tw+uv$$

If $\alpha_{\{t,u\},\{v,w\}} > 0$, since $tu+vw$ cannot be the maximum

$$tu+vw < tv+uw = tw+uv$$

In both cases the four-point condition is satisfied.

($\Leftarrow$) Let $t,u,v,w \in X$. If $\alpha_{\{t,u\},\{v,w\}} = 0$, then for every $x \in X$

$$0 = \alpha_{\{t,u\},\{v,w\}} \leq \alpha_{\{t,x\},\{v,w\}} + \alpha_{\{t,u\},\{v,x\}}$$

since isolation indices are not negative.

Now suppose $\alpha_{\{t,u\},\{v,w\}} > 0$. We are assuming that d satisfies the four-point condition, thus for the previous remark

$$tu+vw < tv+uw = tw+uv$$

Then for every $x \in X$

$$\begin{aligned}
\alpha_{\{t,u\},\{v,w\}} &= \frac{1}{2}\Big(\max\{tv+uw,\ tw+uv,\ tu+vw\} - tu - vw\Big) \\
&= \frac{1}{2}\big(tw+uv-tu-vw\big) \\
&= \frac{1}{2}\big(tw+xv-tx-vw\big) + \frac{1}{2}\big(tx+uv-tu-vx\big) \\
&\leq \frac{1}{2}\Big(\max\{tv+xw,\ tw+xv,\ tx+vw\} - tx - vw\Big) \\
&\quad + \frac{1}{2}\Big(\max\{tv+ux,\ tx+uv,\ tu+vx\} - tu - vx\Big) \\
&= \alpha_{\{t,x\},\{v,w\}} + \alpha_{\{t,u\},\{v,x\}}
\end{aligned}$$

By Theorem 4.5 this proves that d is totally decomposable.

Suppose by absurd to have two incompatible d-splits $\{A, B\}$ and $\{A', B'\}$.
Then we can suppose that exist $t, u, v, w \in X$ such that

$$t, u \in A\,, \quad v, w \in B\,, \quad t, v \in A'\,, \quad u, w \in B'$$

Since d is totally decomposable, its restriction to $\{t, u, v, w\}$ is a conical combination of split metrics that splits 2-2 (that is they extend one of the quartets) and 3-1 (the trivial split metrics).

But the latter contribute equally to the three distances

$$tu + vw\,, \qquad tv + uw\,, \qquad tw + uv$$

In fact, if $x \in \{t, u, v, w\}$, then

$$\delta_x(t, u) + \delta_x(v, w) = \delta_x(t, v) + \delta_x(u, w) = \delta_x(t, w) + \delta_x(u, v) = 1$$

otherwise, if $x \notin \{t, u, v, w\}$, then

$$\delta_x(t, u) + \delta_x(v, w) = \delta_x(t, v) + \delta_x(u, w) = \delta_x(t, w) + \delta_x(u, v) = 0$$

Since d-splits are weakly compatible (Proposition 3.1) and

$$\{A, B\} \succcurlyeq \big\{\{t, u\}, \{v, w\}\big\}\,, \qquad \{A', B'\} \succcurlyeq \big\{\{t, v\}, \{u, w\}\big\}$$

then $\big\{\{t, w\}, \{u, v\}\big\}$ does not have any d-split extension.

So the only different contributions come from split metrics that split like $\delta_{A,B}$ and $\delta_{A',B'}$. But

$$\begin{aligned}
\delta_{A,B}(t, u) + \delta_{A,B}(v, w) &= 0 \qquad & \delta_{A',B'}(t, u) + \delta_{A',B'}(v, w) &= 2\\
\delta_{A,B}(t, v) + \delta_{A,B}(u, w) &= 2 \qquad & \delta_{A',B'}(t, v) + \delta_{A',B'}(u, w) &= 0\\
\delta_{A,B}(t, w) + \delta_{A,B}(u, v) &= 2 \qquad & \delta_{A',B'}(t, w) + \delta_{A',B'}(u, v) &= 2
\end{aligned}$$

As a consequence we have

$$tw + uv \;>\; \begin{matrix} tv + uw \\ tu + vw \end{matrix}$$

violating the four-point condition. ↯

□

Part II

Method

Chapter 5

Graphical representation

So far we have seen essentially an abstract theory about (pseudo-)metric spaces. This work would not be complete without addressing the relevance of this theory with respect to the problem of phylogenetic reconstruction, as outlined in the Introduction.

So we are going to make an informal discussion of the connections, stating the principal results without proof.

The content of this chapter is mainly based on [HRS11].
We will assume familiarity with some basics notions and terminology from graph theory.
In the following we will use distance function basically as a synonym for dissimilarity function.

Basic definitions

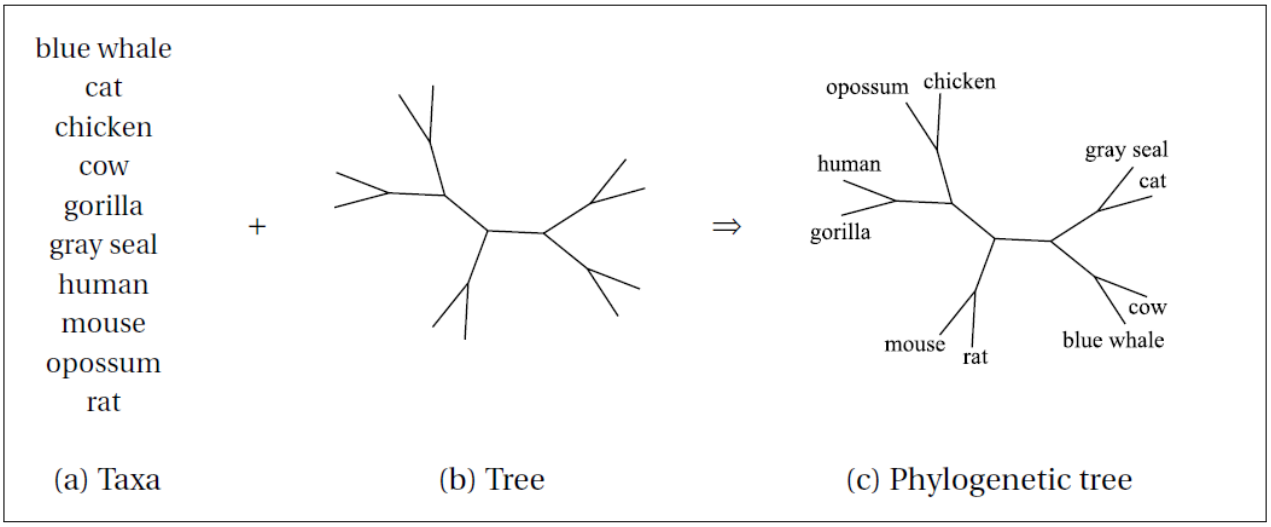
Phylogenetics is the study of the evolutionary history and relationships among or within groups of organisms.
The goal of phylogenetic inference is to produce a diagram that represents an *hypothesis* of relationships that reflects the evolutionary history of a group of organisms.

Biologists call **taxon** (plural **taxa**) the taxonomic unit that represents some species/group/individual organism whose evolutionary history is of interest.

So in this context X is a (necessarily finite) set of taxa.

Phylogenetic trees

An (unrooted)[1] **phylogenetic tree** is a (graph-theoretical) tree, in which all nodes have degree $\neq 2$, together with a labeling that assigns exactly one taxon to every leaf, and none to any internal node.



Source: [HRS11, Figure 3.2]

In order to quantify the effect of evolution, we assign to each edge a **weight** or **length**.[2]

A weighted phylogenetic tree T induces a **tree distance** D_T on each pair of taxa given by the length of the unique path between the leaves labeled by those taxa. Notice that this distance is actually a metric.

We say that a distance function D is an **additive distance** (or **tree-like**) if there exists a phylogenetic tree T such that $D = D_T$.

A characterization of tree metrics is given in [Bun74]:

Fact A distance function is additive if and only if it satisfies the four-point condition.

[1] From the point of view of phylogenetics it would be more desirable to have *rooted* trees, since they give an explicit direction of time (namely from the root to the leaves); but from a theoretical and algorithmic point of view unrooted trees are easier to work with.

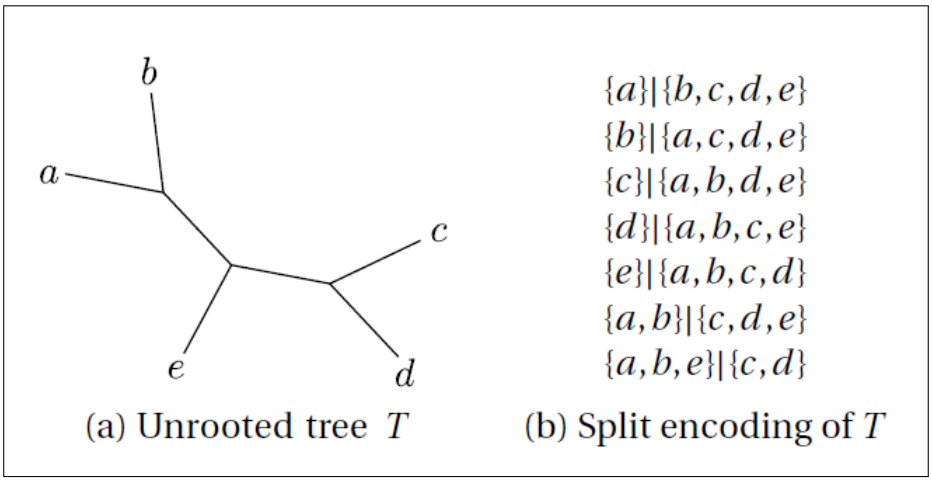
[2] This number often is proportional to the estimated number of mutations occurred along the edge or is correlated to evolutionary time in some other way.

Trees and splits

Any edge of a phylogenetic tree defines a split of the underlying taxon set X: deleting the edge produces two subtrees, and their taxon labels constitute the two parts of the split.
In fact, since every leaf in a phylogenetic tree is labeled by some taxon, the two parts are non-empty (the two subtrees must contain some leaves); and, since each taxon occurs precisely once, it follows that the two parts are disjoint and cover all the taxa.

If the edges of the tree have lengths/weights, then these can be assigned to the corresponding split.

The **split encoding** of an (unrooted) phylogenetic tree is the set of all the splits represented by its edges.



Source: [HRS11, Figure 5.2]

Fact A tree can be uniquely reconstructed from its split encoding.

Thus we can say that a tree T **represents** (or **realizes**) a set of splits $\mathcal{S}$ if and only if $\mathcal{S}$ coincide with the split encoding of T.

A classical result by Buneman [Bun72] gives us a characterization of the set of splits induced by a tree.

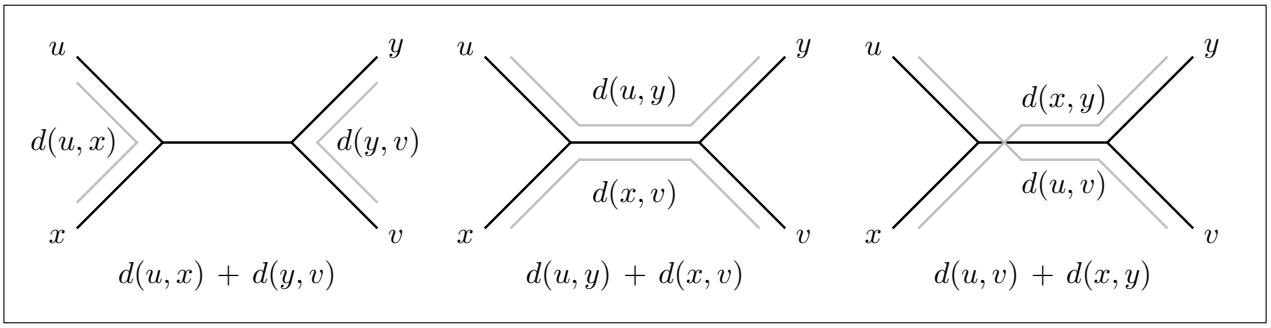
Fact Let $\mathcal{S}$ be a set of splits on X that contains all trivial splits on X.

Then there exists a unique unrooted phylogenetic tree T that realizes $\mathcal{S}$ if and only if $\mathcal{S}$ is compatible.

Towards split decomposition

We have defined splits for distances and splits for trees.
Now the non-trivial fact to prove is that the split encoding of a tree coincides with the splits of the induced tree metric, and the edge weights coincide with the isolation indices.

To give an idea of why this would be true, consider a tree metric on 4 elements realized by the following tree



It is clear that

$$d(u,x) + d(y,v) \;<\; d(u,y) + d(x,v) \;=\; d(u,v) + d(x,y)$$

Moreover, the internal edge, corresponding to the split $\{\{u,x\},\{y,v\}\}$, has length given by

$$\frac{1}{2}\big(d(u,y) + d(x,v) - d(u,x) - d(y,v)\big) = \beta_{\{u,x\},\{y,v\}} = \alpha_{\{u,x\},\{y,v\}}$$

where we used Proposition 2.1.

Fact Let d be a tree metric.

Then a split $\{A,B\}$ is a d-split if and only if
for every $a,a' \in A$ and $b,b' \in B$

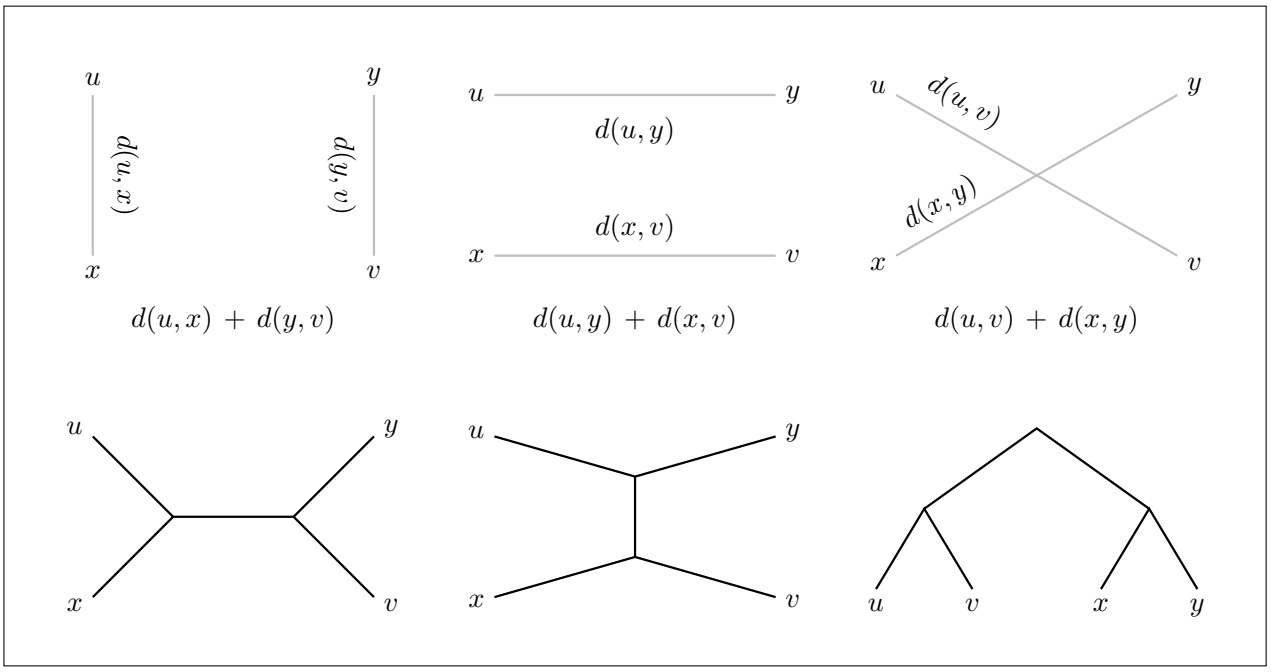
$$aa' + bb' < ab + a'b' = ab' + a'b$$

In this case the split corresponds to a unique edge of length $\alpha_{A,B}$ in the tree realizing d.

We want to understand how to reconstruct the tree starting from the pairwise distances between the taxa.

Consider 4 taxa $u, v, x, y \in X$, and their pairwise distances.
There are exactly three tree topologies that could realize the quadruple, which correspond to the three quartets

$$\{\{u,x\},\{y,v\}\}\,, \quad \{\{u,y\},\{x,v\}\}\,, \quad \{\{u,v\},\{x,y\}\}$$



If d is a tree metric, then we just need a way to choose among these topologies.
In particular, we need to select the splits that will constitute the tree.
It can be done by picking the splits that respect the characterizing condition.

However, in practical applications, distances are empirically derived estimates subject to noise; as a consequence, they do not satisfy exactly the four-point condition (sometimes not even the triangle inequality).
So what can we say in the general case?

Suppose

$$d(u,x) + d(y,v)\,,\ d(u,y) + d(x,v) \ <\ d(u,v) + d(x,y)$$

that is the third distance is the greatest.

We could use again the characterizing condition for tree metrics:
choose the splits $\{A, B\}$ such that, for all $a, a' \in A$ and $b, b' \in B$,

$$aa' + bb' < ab + a'b' = ab' + a'b$$

Such splits would be compatible, thus representable as a tree.

Unfortunately, the trees obtained in this way are far from being resolved[3].
In other words, different elements are not distinguishable because there are very few internal edges.

In our example, none of the three non-trivial splits would be chosen, and we would get a star-tree comprised only of trivial splits.

We can relax the condition and not require the two greatest distances to be equal: choose the splits $\{A, B\}$ such that, for all $a, a' \in A$ and $b, b' \in B$,

$$aa' + bb' < ab + a'b',\ ab' + a'b$$

or equivalently

$$aa' + bb' < \min\{ab + a'b',\ ab' + a'b\}$$

This corresponds to choosing only one topology, namely the one associated with the smallest distance (in our example, the one on the left).

The result is a set of compatible splits whose realization is the so called Buneman tree. Historically, this was the first distance-based method that possess all the following desirable properties:[4]

- the output is a tree, and the correct tree if given tree-like data
- it is continuous, that is slightly different inputs result in similar trees
- it is homogeneous, that is rescaling the input results in a rescaled output
- it is equivariant, that is consistent with permutations of the taxon labels (in other words the output does not depend on the order in which the taxa are processed)
- it can be computed in polynomial time (with respect to the number of taxa)

[3] A tree is resolved if all its internal nodes have degree equal to 3.

[4] The popular Neighbor-joining, for example, does not satisfy the fourth property of equivariance.

However, "*the price paid for continuity*", is that the resulting tree is often highly unresolved. There are some improvements that have been proposed to address this problem, like the refined Buneman tree [MS99; BM99].

The split decomposition method takes another approach: instead of choosing only one topology, it allows to choose up to *two* topologies.

This is done by further relaxing the condition to:
choose the splits $\{A, B\}$ such that, for all $a, a' \in A$ and $b, b' \in B$,

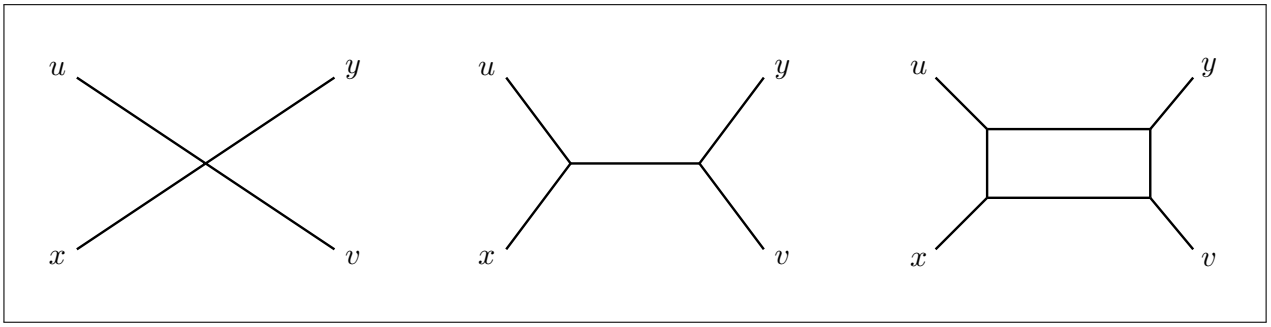
$$aa' + bb' < \max\{ab + a'b', ab' + a'b\}$$

Notice that this is equivalent to the defining condition for d-splits.
As such, the result is a collection of weakly compatible splits (not necessarily compatible).

In our example, the third distance is the largest and thus the corresponding topology is the "most unlikely" [5]. Consequently, this is the one we discard and we pick the first two.

We can display both topologies in one single network, where parallel edges corresponds to the same split.

Here's a picture that summarizes the kinds of diagrams obtainable by applying the three previous conditions.



From left to right: four-point condition, Buneman tree, split decomposition.

[5] If d were a tree metric, then this would be the wrong topology since we are not accounting for the length of the edge between the two internal nodes.
In the general case, we can regard this heuristic as a form of parsimony principle.

Proposition 5.1

If d is a pseudo-metric satisfying the four-point condition, then it can be written as

$$d = \sum_{S \in \mathcal{S}_d(X)} \alpha_S^d \cdot \delta_S$$

Moreover, a collection of splits $\mathcal{S}$ is of the form $\mathcal{S} = \mathcal{S}_d(X)$, where d is a pseudo-metric satisfying the four-point condition, if and only if $\mathcal{S}$ is compatible.

Proof

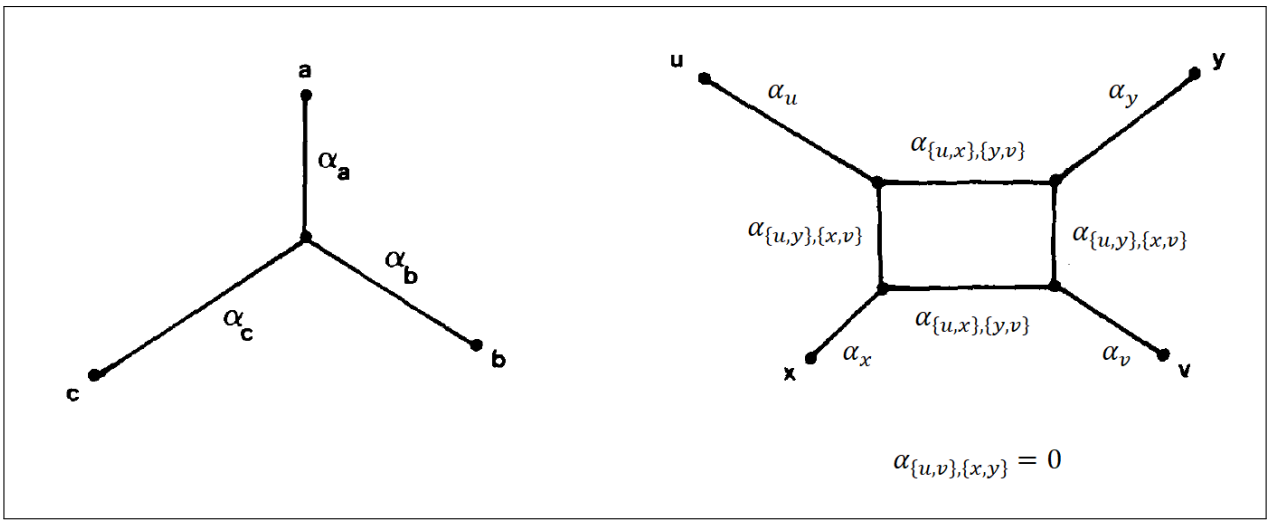
The first assertion is a consequence of Theorem 2.8 and Corollary 4.9.

($\Rightarrow$) It is a consequence of Corollary 4.9.

($\Leftarrow$) Since compatible implies weakly compatible, it is a consequence of Theorem 3.2 and Corollary 4.9. □

In other words, tree metrics are totally decomposable, and a collection of splits coincide with the d-splits of some tree metric if and only if they are compatible.

In Proposition 4.3 we have seen that pseudo-metrics on 4 elements are totally decomposable, but their d-splits are not necessarily compatible.
Thus they cannot be always represented as trees, unlike pseudo-metrics on 3 elements. However they can be represented as networks.



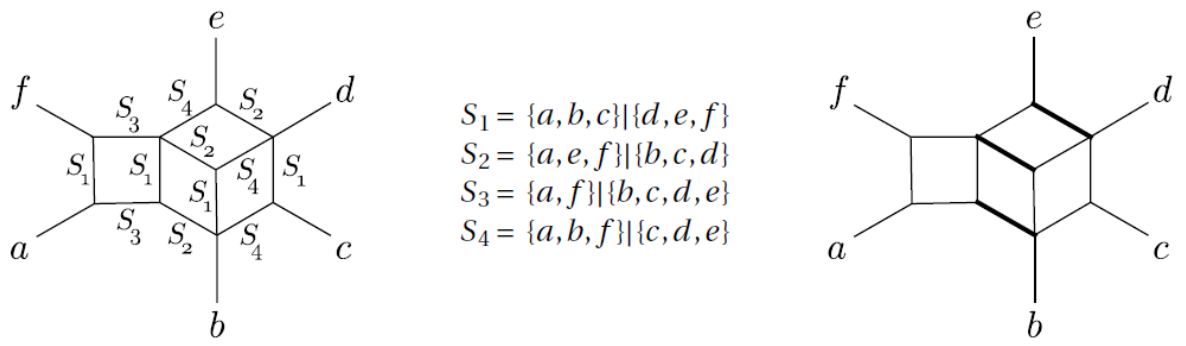
Source: adapted from [BD92a]

These observations motivate us to discard the residue also in the general case, and focus only on the totally decomposable part. Our hope is that we can build similar networks that approximate the original distance function.

Split networks

What kind of networks we are talking about exactly?
We want something akin to phylogenetic trees, that can be built from splits.

The precise definition is quite involved but the idea is that each split is represented by one or multiple edges (usually drawn parallel to each other) with the property that removing the edges corresponding to the same split produces exactly two connected components.
Such networks are called **split networks**.[6]



(a) Split network N (b) Non-trivial splits (c) Split $S_2 = \frac{\{a,e,f\}}{\{b,c,d\}}$

Source: [HRS11, Figure 5.6]

The remarkable result is that a generic collection of splits can always be represented by a split network [HRS11, §5.6].

Fact For any set of splits $\mathcal{S}$ there exists a unique split network that represents $\mathcal{S}$, called the **canonical split network** or **Buneman graph** associated with $\mathcal{S}$.

There are algorithms that take in input a set of splits and give in output a split network that represents those splits. In particular, the network computed by the convex hull algorithm is the Buneman graph.

[6] The attentive reader may ask why not use the name *phylogenetic networks*.
The reason is there are various different kinds of networks employed in phylogenetics, and also there is not much consensus on the terminology [HRS11, §4.2].

Methods for split networks

In this framework the role of split decomposition is to give a collection of splits with certain properties, namely weak compatibility.

This is a nice property because the split networks associated with weakly compatible splits are often quite close to being planar, as they usually have only a few edges crossing over each other and do not contain any "high-dimensional cubes", which may occur for completely unrestricted sets of splits.

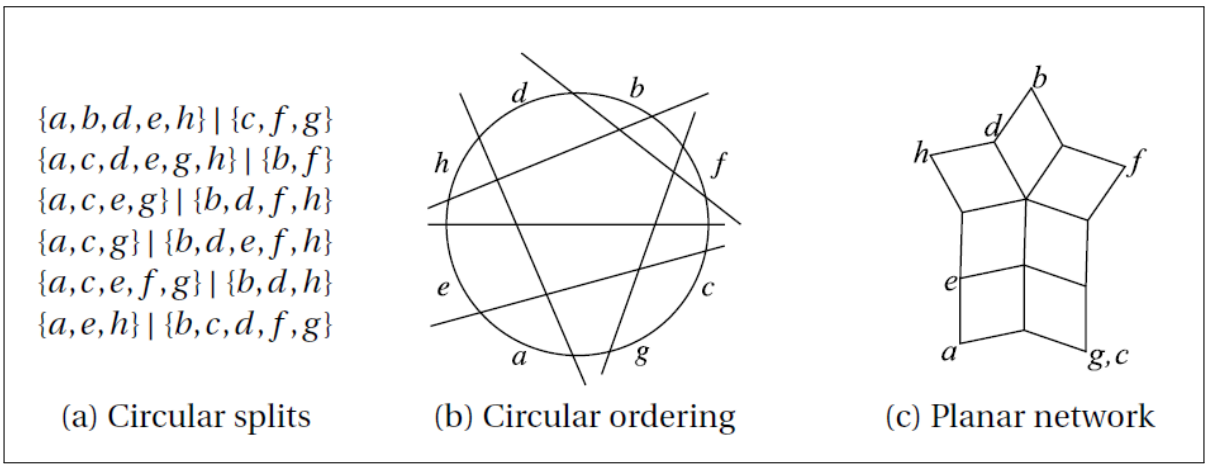
One of the strength and limitation of the split decomposition method is that it produces relatively few splits.
In fact, as the number of taxa increases, also the probability that isolation indices are 0 increases (it just takes one of the β indices to be 0); as a consequence the number of d-splits becomes very contained.

It makes sense to ask: what are the maximal sets of weakly compatible splits? The answer is circular splits.

A set of splits is called **circular** if the taxa can be placed around a circle in such a way that each split can be realized by a line through the circle; this line separates the plane into two half-planes corresponding to the two parts of the split.

Circular split have a very nice property:

Fact A set of splits is circular if and only if it can be represented as an outer-labeled planar split network.



(a) Circular splits (b) Circular ordering (c) Planar network

Source: [HRS11, Figure 5.9]

A successor of the split decomposition method, called NeighborNet [BM04; BH23], produces a collection of weighted circular splits.

> *Neighbor-net is an attractive method for computing split networks for the following reasons: First, the resulting networks are outer-labeled planar and thus easy to draw and to read. Secondly, the algorithm is quite fast. Thirdly, it produces resolved networks even for quite large numbers of taxa, unlike the split decomposition method, which rapidly loses resolution as the number of taxa increases.* – [HRS11, §10.4.5]

Both methods have the nice property that they produce a tree when given tree-like data.

Chapter 6

Split decomposition algorithm

We now pass to examine in more detail the algorithm to compute the split decomposition, based on what has been outlined in Chapter 3.

I also provide an implementation in Matlab;
the full code can be found in Appendix A.

Data structures

Given some ordering of X, we identify X with $\{1, \dots, n\}$, where $n = |X|$.

The obvious way to represent a symmetric function $d : X \times X \to \mathbb{R}$
is with a symmetric square matrix $D \in M(n, \mathbb{R})$ such that

$$D_{ij} = d(i, j)\,, \quad \forall\, i, j \in X$$

It is less obvious how to represent splits.

One way would be storing the indices of the elements contained in each part. This method has some issues: we need to store in memory data relative to *two* sets for each split; or we could store the elements of only one of the part, but it may not be easy to pass to the complement.

An alternative is to identify each split with one of its part (and get the other part as the complement), but for each element we annotate whether they belong or not to the part considered.

Thus we can represent a split with a binary array, where 1 in position i means that i belongs to the part and 0 means it belongs to the complement: for example [1 0 1 0 0] would represent the split $\{\{1,3\},\{2,4,5\}\}$ as encoded by its first part $\{1,3\}$; while the binary complement [0 1 0 1 1] represents the *same* split, but by encoding the other part $\{2,4,5\}$.

Notice that what we have done is a bit of a trick: we are still using two piece of information to represents the split, namely one of the part and the length of the array (that is the total number of elements).
In fact, with this method we cannot represent a generic *partial* split, since we would not know on which subset of X it is defined (or in other words what is the union of the parts).

However this is not a problem since, as we will see, there is no need to represent all partial splits, but only those obtained by iteratively adding elements to a base split.

We will use as ordering for adding the elements the same ordering used for X: that is $\{\{1\},\{2\}\}$ is the base split, and we will consider only subsets of the form $\{1,\dots,m\} \subseteq \{1,\dots,n\}$.
In particular, in this context an array of length $m < n$ will represent a split of the subset $\{1,\dots,m\}$.

In order to avoid unnecessary technical details for the moment, we will think of splits as sets of indices in the pseudo-code and use the binary array representation in the final implementation.

For ease of notation we use $A \mid B$ for the split $\{A, B\}$ and $A \cup x$ for $A \cup \{x\}$.

The β index

The formula for the β index relative to the quartet $t, u \mid v, w$ is

$$\beta_{t,u\,|\,v,w} = \frac{1}{2}\Big(\max\{tu + vw,\, tv + uw,\, tw + uv\} - tu - vw\Big)$$

This is straightforward enough, but we can make a small improvement to avoid unnecessary numerical cancellation: if

$$\max\{tu + vw,\, tv + uw,\, tw + uv\} = tu + vw$$

then we do not need to do the subtraction,
because we already know the result is 0.

Algorithm 1: β index

```
1.  function BETA (t, u, v, w)
2.      s1 ← D(t, u) + D(v, w)
3.      s2 ← D(t, v) + D(u, w)
4.      s3 ← D(t, w) + D(u, v)
5.      m ← MAX (s1, s2, s3)
6.      if m = s1 then
7.          b ← 0
8.      else
9.          b ← 1/2 (m − s1)
10.     return b
```

The α index

The formula for the α index relative to the split $A \mid B$ is

$$\alpha_{A\,|\,B} := \min_{\substack{t,\,u \in A \\ v,w \in B}} \beta_{t,u\,|\,v,w}$$

We can exploit the following symmetries of the β index

$$\beta_{t,u\,|\,v,w} = \beta_{u,t\,|\,v,w} = \beta_{t,u\,|\,w,v} = \beta_{u,t\,|\,w,v}\,, \quad \forall\, t,u,v,w \in X$$

to skip about $3/4$ of quartets.

To do so we need to compute the β index over all unordered pairs.

We use the notation $\binom{S}{2}$ for the set of unordered pairs of S.

Moreover, if just one of the β index is 0,
then we do not need to check further since also the α index would be 0.

Algorithm 2: α index

```
1.  function ALPHA (A | B)
2.      a ← ∞                                  ▷ initialize for the first min
3.      for all {t, u} ∈ (A 2) do              ▷ cycle on unordered pairs of A
4.          for all {v, w} ∈ (B 2) do          ▷ cycle on unordered pairs of B
5.              b ← BETA (t, u, v, w)
6.              if b = 0 then
7.                  a ← 0
8.                  return a
9.              else
10.                 a ← MIN (a, b)
11.     return a
```

The d-splits

The algorithm is of inductive nature

- start with $Y = \{y, y'\}$ where $y \neq y' \in X$
- until $Y = X$
  - pick any $x \in X \setminus Y$
  - for every $A \mid B$ d-split of Y,
    - check if $A \cup x \mid B$ is a partial d-split
    - check if $A \mid B \cup x$ is a partial d-split
  - check if $Y \mid x$ is a partial d-split
  - update Y to $Y \cup x$

We need to specify what to do after these checks and how to choose x.
Since we have an ordering, we can just cycle on all the elements
(we use k to emphasize this choice).

Algorithm 3: d-splits

```
1.  function DSPLIT
2.      Y ← {1, 2},   L_old ← 1 | 2,   L_new ← ∅
3.      for k ← 3 to n do
4.          for all A | B ∈ L_old do                ▷ current d-splits of Y
5.              if ALPHA(A ∪ k | B) > 0 then
6.                  INSERT(L_new, A ∪ k | B)
7.              if ALPHA(A | B ∪ k) > 0 then
8.                  INSERT(L_new, A | B ∪ k)
9.              DELETE(L_old, A | B)
10.         if ALPHA(Y | k) > 0 then
11.             INSERT(L_new, Y | k)
12.         Y ← Y ∪ k,   L_old ← L_new,   L_new ← ∅
13.     return L_old
```

This algorithm, as is written, has some inefficiencies:
when we compute the isolation index of $A \cup k \mid B$ or $A \mid B \cup k$,
we are calling the function ALPHA that recalculate over all quartets.
But if we store the value of $\alpha_{A \mid B}$ previously computed,
we only need to check the quartets that contain k.

So we can substitute the call to ALPHA in line 5 with a function ALPHA_SX defined as

Algorithm 4: α index

```
1.  function ALPHA_SX (A ∪ {k} | B, α_{A|B})
2.      a ← α_{A|B}                          ▷ initialize with the old value
3.      for all t ∈ A ∪ {k} do               ▷ cycle only on one element
4.          for all {v, w} ∈ C(B, 2) do
5.              b ← BETA (t, k, v, w)        ▷ quartet t, k | v, w
6.              if b = 0 then
7.                  a ← 0
8.                  return a
9.              else
10.                 a ← MIN (a, b)
11.     return a
```

and the call to ALPHA in line 8 with a similar function ALPHA_DX
with the loops swapped

```
for all {t, u} ∈ C(A, 2) do
    for all v ∈ B ∪ {k} do               ▷ cycle only on one element
        b ← BETA (t, u, v, k)            ▷ quartet t, u | v, k
```

Notice that for the partial splits of the form $Y \mid k$,
all the quartets contain k, so they are all "new".

The decomposition

If we store the isolation indices computed during the execution of DSPLIT, we can also compute the split-prime residue of D.

Algorithm 5: split-prime residue

```
1. function RESIDUE (S_d(X), {α_S}_{S ∈ S_d(X)})
2.     D_0 ← D
3.     for all S ∈ S_d(X) do
4.         D_0 ← D_0 − α_S · Δ_S
5.     return D_0
```

Here Δ_S is the matrix representing the split metric associated to the split S.

In the representation of splits as sets of indices, the calculation of Δ_S requires, for every couple $i, j \in X$, to cycle over one of the part and check whether i, j are both present/absent; if this is true we assign value 0, otherwise value 1. This, on average, has computational cost

$$\underbrace{\frac{n}{2}}_{\substack{\#\text{ of elements} \\ \text{in one part}}} \times \underbrace{\frac{n(n-1)}{2}}_{\substack{\#\text{ of unique} \\ \text{elements of } \Delta_S}} = O(n^3)$$

even accounting for the symmetry of the matrix.

But in the representation as binary arrays we just need an **xor** (we use the notation $[A]$ to emphasize that here A is an array).

Algorithm 6: split metric

```
1. function SPLIT_METRIC ([A])
2.     for all i ∈ [A] do
3.         for all j ∈ [A] do
4.             Δ_A(i, j) ← i xor j
```

This has cost of only $O(n^2)$.

The final algorithm looks like this

Algorithm 7: split decomposition

1. **function** SPLIT_DECOMP
2. $Y \leftarrow \{1,2\}$, $L_{\text{old}} \leftarrow \left[\, 1 \mid 2 \,,\, \text{ALPHA}\,(1 \mid 2)\,\right]$, $L_{\text{new}} \leftarrow \emptyset$
3. **for** $k \leftarrow 3$ to n **do**
4. **for all** $\left[\, A \mid B \,,\, \alpha_{A \mid B}\,\right] \in L_{\text{old}}$ **do**
5. $a \leftarrow \text{ALPHA\_SX}\left(A \cup k \mid B,\, \alpha_{A \mid B}\right)$
6. **if** $a > 0$ **then**
7. INSERT $\left(L_{\text{new}},\, \left[\, A \cup k \mid B \,,\, a \,\right]\right)$
8. $a \leftarrow \text{ALPHA\_DX}\left(A \mid B \cup k,\, \alpha_{A \mid B}\right)$
9. **if** $a > 0$ **then**
10. INSERT $\left(L_{\text{new}},\, \left[\, A \mid B \cup k \,,\, a \,\right]\right)$
11. DELETE $\left(L_{\text{old}},\, \left[\, A \mid B \,,\, \alpha_{A \mid B} \,\right]\right)$
12. $a \leftarrow \text{ALPHA}\,(\,Y \mid k)$
13. **if** $a > 0$ **then**
14. INSERT $\left(L_{\text{new}},\, \left[\, Y \mid k \,,\, a \,\right]\right)$
15. $Y \leftarrow Y \cup k$, $L_{\text{old}} \leftarrow L_{\text{new}}$, $L_{\text{new}} \leftarrow \emptyset$
16. $D_0 \leftarrow D$
17. **for all** $\left[\, A \mid B \,,\, \alpha_{A \mid B}\,\right] \in L_{\text{old}}$ **do**
18. $D_0 \leftarrow D_0 - \alpha_{A \mid B} \cdot \Delta_{A \mid B}$
19. **return** L_{old}, D_0

Correctness

We claim that at the end of each iteration of the outer loop, the list L_{old} contains all the d-splits of $Y \cup k$. We prove it by induction on k.

For the base case, before entering the outer loop,
we are adding the only possible d-split of $\{1, 2\}$, which is $1 \mid 2$.

For each $A \mid B$ d-split of Y,
it is clear that $A \cup k \mid B$, $A \mid B \cup k$ and $Y \mid k$ are splits of $Y \cup k$;
and we are adding to the list only those that have non-zero isolation index.

This proves that at the end those in the list are indeed d-splits of $Y \cup k$.

Now suppose S is a d-split of $Y \cup k$.

If $S = Y \mid k$, then we know that it will be added from what said above.

Otherwise S is an extension of a d-split of Y.
In fact, the restriction of S to Y is a split of Y, because neither of its parts can contain Y; otherwise S should be $Y \mid k$, which we excluded, or $Y \cup k \mid \emptyset$, which is not a split. Moreover, the isolation index of the restriction cannot be 0, because otherwise also its extension S should have index 0, which contradicts the hypothesis that S is a d-split.

This proves that every d-split of $Y \cup k$, different from $Y \mid k$, is an extension of a d-split of Y, which for inductive hypothesis is already in the old list; and is necessarily obtained by adding k to one of the part, so it is among those computed by the algorithm.

Computational complexity

Consider Y of cardinality m and let us count the quartets that contain k.

In $Y \mid k$, we have two choices for the elements in Y,
but we want only the unordered pairs, so there are

$$\frac{m \cdot m - m}{2} + m = \frac{m\,(m+1)}{2}$$

Now suppose to have a split $A \mid B$ such that

$$|A| = l \quad \text{and} \quad |B| = m - l$$

Then in $A \cup k \mid B$ we have one choice for the elements of A
and two choices for the elements of B (still only unordered pairs), so

$$l \cdot \frac{(m-l)(m-l+1)}{2}$$

Analogously in $A \mid B \cup k$, we have two choices for the elements of A
(still only unordered pairs) and one choice for the elements of B, so

$$\frac{l\,(l+1)}{2} \cdot (m-l)$$

In total from the split $A \mid B$ we have

$$\begin{aligned}
&\frac{1}{2}\,l\,(m-l)\big[(m-\not{l}+1)+\not{l}+1\big] \\
&\qquad = \frac{1}{2}\,l\,(m-l)(m+2) =: f_m(l)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial l} f_m(l) &= \frac{1}{2}\,(m-l)(m+2) - \frac{1}{2}\,l\,(m+2) \\
&= \frac{1}{2}\,\big[(m-l)-l\big](m+2) \\
&= \frac{1}{2}\,(m-2\,l)(m+2)
\end{aligned}$$

$$\frac{\partial}{\partial l} f_m(l) = 0 \quad \Longleftrightarrow \quad l = \frac{m}{2}$$

$$\begin{aligned}\max_l f_m(l) &= \frac{1}{2}\frac{m}{2}\left(m-\frac{m}{2}\right)(m+2)\\ &= \frac{1}{2}\frac{m}{2}\frac{m}{2}(m+2)\\ &= \frac{m^3}{8}+\frac{m^2}{4}\end{aligned}$$

So we found an upperbound for the number of quartets
that depends only on the cardinality of Y.

Since there can be at most $\binom{m}{2}$ d-splits on Y,
the number of "new" β indices we need to compute is at most

$$\frac{m\,(m+1)}{2}+\left(\frac{m^3}{8}+\frac{m^2}{4}\right)\cdot\binom{m}{2}$$

For each of these quartets, we need to do: 2 comparisons (for the maximum in the β index), 3 additions, 1 subtraction and 1 other comparison (against the current minimum).
Moreover, for each split we need to make 1 further comparison,
to check if the isolation index is non-zero.

In total the number of operations is at most

$$\begin{aligned}&(6+1)\left[\frac{m\,(m+1)}{2}+\left(\frac{m^3}{8}+\frac{m^2}{4}\right)\cdot\frac{m\,(m-1)}{2}\right]+\left(1+\frac{m\,(m-1)}{2}\right)\\ &=\frac{1}{16}\left(7\,m^5+7\,m^4-14\,m^3+64\,m^2+48\,m+16\right)\end{aligned}$$

Summing over all the Y's of the outer loop, we get

$$\sum_{m=2}^{n-1}\ldots=\frac{7}{96}\,n^6-\frac{21}{160}\,n^5-\frac{49}{192}\,n^4+\frac{23}{12}\,n^3-\frac{145}{192}\,n^2+\frac{73}{480}\,n-9$$

Notice how the upperbound on the total number of arithmetic operations
is a polynomial of grade 6 with somewhat small coefficients.

In particular, the total complexity of the algorithm is $O(n^6)$.

Conclusion

In the previous chapters we studied the theory of split decomposition which is the foundation for the split decomposition method.

We now want to mention possible directions that deserve further investigation:

- further explore the fundamentals about distance functions (especially pseudo-metrics), in particular
  - geometric properties of the cone $M(X)$, see also [Avi80; Avi81; AD91]
  - relationships with $(0,1)$-matrices, see also [Ans80]
  - relationships with graph theory (metrics induced by graphs, graph embeddability, tree realizability etc.), see also [IS72; Mul82; SZ82]
- further study the properties of split networks, in particular
  - characterize maximal sets of splits that possess desirable properties (like being representable as low-dimensional networks)
  - what happens if substitute splits with multi-splits,
    that is (non necessarily binary) graph cuts?
- further study possible application of numerical linear algebra (after all, our distance function is just a matrix)
  - one notable mention in this direction is a recent publication about NeighborNet [BH23]

- further explore the effect of perturbations of the input matrix on these methods, in particular
  - there are some results about trees in ∞-norm, for example in [SS03, §7.7; Dre+12, §10.2]; what about other norms?
  - to what extent we can consider the residue of the split decomposition as error/noise?
  - this area seems related to the topic of self-correcting codes, there may be some interesting connections

Appendix A

Matlab implementation

β index

```
function b = beta (D, t,u,v,w)
% function b = beta (D, t,u,v,w)
%
% returns the beta index of the quartet t,u|v,w
%
% D = distance matrix (symmetric square matrix, size nxn)
% t,u,v,w = taxa (integer numbers between 1 and n)

s1 = D(t,u) + D(v,w);
s2 = D(t,v) + D(u,w);
s3 = D(t,w) + D(u,v);
m = max ([s1, s2, s3]);

if m == s1      % avoid unnecessary numerical cancellation
    b = 0;
else
    b = (m - s1) / 2;
end

end
```

α index

```
function a = alfa (D, A)
% function a = alfa (D, A)
%
% returns the isolation index of the split A|B
%   where B is the complement of A
%
% D = distance matrix (symmetric square matrix, size nxn)
% A = one of the parts of the split (binary vector, size 1xm)
%   the other part is obtained by taking ...
%   ... the binary complement of the vector

% calculate the indices of the elements in A ...
iA = find  (A);
% ... and in B (complement of A)
iB = find (~A);

% initialize with the maximum distance
a = max (D, [], "all");

% check only unordered couples
TA = table2array (combinations (iA, iA));
TB = table2array (combinations (iB, iB));

for rA = TA'
    t = rA (1);
    u = rA (2);

    for rB = TB'
        v = rB (1);
        w = rB (2);

        b = beta (D, t,u,v,w);

        if b == 0                % no need to check further
            a = 0; return
        else
            a = min ([a, b]);
        end
    end
end

end
```

Split decomposition

```
function [D0, dS, adS] = split_decomp (D)
% function [D0, dS, adS] = split_decomp (D)
%
% compute the split decomposition of D
%
% D = distance matrix (symmetric square matrix, size nxn)
%
% D0 = split-prime residue (symmetric square matrix, size nxn)
% dS = d-splits (binary matrix, rows = splits, size Nxn)
% adS = isolation indices (vector, size Nx1)

n = size (D);

S{2} = [1, 0];          % start from 1|2
a{2} = alfa (D, S{2});

for k = 3 : n

    % build the new (partial) splits

    % duplicate the previous splits A|B
    S{k} = repmat (S{k-1}, 2, 1);
    sz = size (S{k});

    % for every split add the new element ...
    % ... to the left (A,k|B) or the right (A|B,k)
    c = [ones( sz(1)/2, 1); zeros( sz(1)/2, 1)];
    S{k} = [S{k}, c];

    % add the trivial split Y|k
    r = [ones( 1, sz(2)), 0];
    S{k} = [S{k}; r];

    % number of splits
    N = size (S{k}, 1);


    % compute the isolation index of every split

    % unoptimized
    % a{k} = zeros (N, 1);              % preallocation
    % for i = 1 : N
    %   a{k}(i) = alfa ( D, S{k}(i,:) );
    % end

```

```
46      % optimized
47      a{k} = repmat (a{k-1}, 2, 1);
48          % initialize with previous indices
49      half = (N-1) / 2;
50      for i = 1 : half
51          a{k}(i) = alfa_sx ( D, S{k}(i,:), a{k}(i) );  % A,k|B
52      end
53      for i = half+1 : N-1
54          a{k}(i) = alfa_dx ( D, S{k}(i,:), a{k}(i) );  % A|B,k
55      end
56      a{k}(N) = alfa ( D, S{k}(N,:) );                  % Y|k
57
58
59      % remove non d-splits
60
61      idx = (a{k} == 0);
62      S{k} (idx, :) = [];
63      a{k} (idx, :) = [];
64  end
65
66  dS  = S{n};      % (total) d-splits ...
67  adS = a{n};      % ... and their indices
68
69
70  D0 = D;
71  for i = 1 : size (dS, 1)
72      D0 = D0 - adS (i) * split_metric ( dS(i,:) );
73  end
74
75  end
```

```
function delta = split_metric (A)
% function delta = split_metric (A)
%
% returns the split metric of the split A|B
%   where B is the complement of A
%
% A = one of the parts of the split (binary vector, size 1xm)
%   the other part is obtained by taking ...
%   ... the binary complement of the vector

delta = xor (A, A');

end
```

```
function a = alfa_sx (D, Ak, init)

iAk = find  (Ak);
iB  = find (~Ak);

a = init;
TB = table2array (combinations (iB, iB));
k = size (Ak, 2);

for t = iAk
    for rB = TB'
        v = rB (1);
        w = rB (2);

        b = beta (D, t,k,v,w);

        if b == 0
            a = 0; return
        else
            a = min ([a, b]);
        end
    end
end

end
```

```
function a = alfa_dx (D, Ak, init)

iAk = find  (Ak);
iB  = find (~Ak);

a = init;
TAk = table2array (combinations (iAk, iAk));
k = size (Ak, 2);

for rAk = TAk'
    t = rAk (1);
    u = rAk (2);

    for v = iB
        b = beta (D, t,u,v,k);

        if b == 0
            a = 0; return
        else
            a = min ([a, b]);
        end
    end
end

end
```

Bibliography

[BD92a] Hans-Jürgen Bandelt and Andreas W.M. Dress. "A canonical decomposition theory for metrics on a finite set". In: *Advances in Mathematics* 92.1 (1992), pp. 47–105. ISSN: 0001-8708. DOI: https://doi.org/10.1016/0001-8708(92)90061-O. URL: https://www.sciencedirect.com/science/article/pii/000187089290061O.

[BD92b] Hans-Jürgen Bandelt and Andreas W.M. Dress. "Split decomposition: A new and useful approach to phylogenetic analysis of distance data". In: *Molecular Phylogenetics and Evolution* 1.3 (1992), pp. 242–252. ISSN: 1055-7903. DOI: https://doi.org/10.1016/1055-7903(92)90021-8. URL: https://www.sciencedirect.com/science/article/pii/1055790392900218.

Books

[Dre+12] Andreas Dress et al. *Basic Phylogenetic Combinatorics*. Cambridge University Pres, 2012. ISBN: 9781139019767. DOI: https://doi.org/10.1017/CBO9781139019767. URL: https://www.cambridge.org/core/books/basic-phylogenetic-combinatorics/D2477E14E5C24973BF11BA8049BF9828.

[Fel04] Joseph Felsenstein. *Inferring Phylogenies*. Sinauer Associates, 2004. ISBN: 9780878931774.

[HRS11] Daniel H. Huson, Regula Rupp, and Celine Scornavacca. *Phylogenetic Networks: Concepts, Algorithms and Applications*. Cambridge University Press, 2011. ISBN: 9780511974076. DOI: https://doi.org/10.1017/CBO9780511974076. URL: https://www.cambridge.org/core/books/phylogenetic-networks/45EB919453CD2F05AB3B58950C4A1415.

[SS03] Charles Semple and Mike Steel. *Phylogenetics*. Vol. 24. Oxford lecture series in mathematics and its applications. Oxford University Press, 2003. ISBN: 9780198509424. DOI: https://doi.org/10.1093/oso/9780198509424.001.0001. URL: https://academic.oup.com/book/53370.

Related Articles

[BM99] D. Bryant and V. Moulton. "A polynomial time algorithm for constructing the refined Buneman tree". In: *Applied Mathematics Letters* 12.2 (1999), pp. 51–56. ISSN: 0893-9659. DOI: https://doi.org/10.1016/S0893-9659(98)00148-7. URL: https://www.sciencedirect.com/science/article/pii/S0893965998001487. Available at https://www.maths.otago.ac.nz/~dbryant/Papers/99RefinedBuneman.pdf.

[BH23] David Bryant and Daniel H. Huson. "NeighborNet: improved algorithms and implementation". In: *Frontiers in Bioinformatics* 3 (2023). ISSN: 2673-7647. DOI: https://doi.org/10.3389/fbinf.2023.1178600. URL: https://www.frontiersin.org/journals/bioinformatics/articles/10.3389/fbinf.2023.1178600. Available at https://www.frontiersin.org/journals/bioinformatics/articles/10.3389/fbinf.2023.1178600/pdf, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10548196/pdf/fbinf-03-1178600.pdf.

[BM04] David Bryant and Vincent Moulton. "Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks". In: *Molecular Biology and Evolution* 21.2 (2004), pp. 255–265. ISSN: 0737-4038. DOI: https://doi.org/10.1093/molbev/msh018. eprint: https://academic.oup.com/mbe/article-pdf/21/2/255/4016168/msh018.pdf. URL: https://academic.oup.com/mbe/article/21/2/255/1187993. Available at https://academic.oup.com/mbe/article-pdf/21/2/255/4016168/msh018.pdf.

[Bun72] Peter Buneman. "The recovery of trees from measures of dissimilarity". In: *Mathematics in the Archaeological and Historical Sciences*. Ed. by F.R. Hodson, D.G. Kendall, and P. Tautu. Edinburgh: Edinburgh University Press, 1972, pp. 387–395. Available at https://homepages.inf.ed.ac.uk/opb/homepagefiles/phylogeny-scans/manuscripts.pdf.

[Bun74] Peter Buneman. "A note on the metric properties of trees". In: *Journal of Combinatorial Theory, Series B* 17.1 (1974), pp. 48–50. ISSN: 0095-8956. DOI: https://doi.org/10.1016/0095-8956(74)90047-1. URL: https://www.sciencedirect.com/science/article/pii/0095895674900471. Available at https://homepages.inf.ed.ac.uk/opb/homepagefiles/phylogeny-scans/metricproperties.pdf.

[MS99] Vincent Moulton and Mike Steel. "Retractions of finite distance functions onto tree metrics". In: *Discrete Applied Mathematics* 91.1 (1999), pp. 215–233. ISSN: 0166-218X. DOI: https://doi.org/10.1016/S0166-218X(98)00128-0. URL: https://www.sciencedirect.com/science/article/pii/S0166218X98001280. Available at https://core.ac.uk/download/pdf/82561918.pdf, https://www.math.canterbury.ac.nz/~m.steel/Non_UC/files/research/retractions.pdf.

Other Articles

[Ans80] R.P. Anstee. "Properties of (0, 1)-matrices with no triangles". In: *Journal of Combinatorial Theory, Series A* 29.2 (1980), pp. 186–198. ISSN: 0097-3165. DOI: https://doi.org/10.1016/0097-3165(80)90008-4. URL: https://www.sciencedirect.com/science/article/pii/0097316580900084.

[Avi80] David Avis. "On the Extreme Rays of the Metric Cone". In: *Canadian Journal of Mathematics* 32.1 (1980), pp. 126–144. DOI: https://doi.org/10.4153/CJM-1980-010-0. URL: https://www.cambridge.org/core/journals/canadian-journal-of-mathematics/article/on-the-extreme-rays-of-the-metric-cone/9F2921585542D703EE309BB751736BF8. Available at https://cgm.cs.mcgill.ca/~avis/doc/avis/Av80c.pdf.

[Avi81] David Avis. "Hypermetric Spaces and the Hamming Cone". In: *Canadian Journal of Mathematics* 33.4 (1981), pp. 795–802. DOI: https://doi.org/10.4153/CJM-1981-061-5. URL: https://www.cambridge.org/core/journals/canadian-journal-of-mathematics/article/hypermetric-spaces-and-the-hamming-cone/46146AC3338EF3A80BA1B5E23BE9A304. Available at https://cgm.cs.mcgill.ca/~avis/doc/avis/Av81a.pdf.

[AD91] David Avis and Michel Deza. "The cut cone, L^1 embeddability, complexity, and multicommodity flows". In: *Networks* 21.6 (1991), pp. 595–617. DOI: https://doi.org/10.1002/net.3230210602. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/net.3230210602. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/net.3230210602. Available at https://cgm.cs.mcgill.ca/~avis/doc/avis/AD91a.ps.

[Coh04] Joel E. Cohen. "Mathematics Is Biology's Next Microscope, Only Better; Biology Is Mathematics' Next Physics, Only Better". In: *PLOS Biology* 2.12 (2004). ISSN: 1545-7885. DOI: https://doi.org/10.1371/journal.pbio.0020439. URL: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0020439. Available at https://journals.plos.org/plosbiology/article/file?id=10.1371/journal.pbio.0020439&type=printable, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC535574/pdf/pbio.0020439.pdf.

[DHM01] A. Dress, K.T. Huber, and V. Moulton. "Metric spaces in pure and applied mathematics". In: *Documenta Mathematica* Extra Vol. (2001), pp. 121–139. ISSN: 1431-0643. URL: http://eudml.org/doc/121785. Available at https://www.emis.de/journals/DMJDMV/lsu/dress-huber-multon.pdf, https://www.kurims.kyoto-u.ac.jp/EMIS/journals/DMJDMV/lsu/dress-huber-multon.pdf.

[IS72] V. Imrikh and É. Stotskii. "Optimal imbeddings of metrics in graphs". In: *Siberian Mathematical Journal* 13.3 (1972), pp. 382–387. ISSN: 1573-9260. DOI: https://doi.org/10.1007/BF00968113. URL: https://link.springer.com/article/10.1007/BF00968113.

[Mul82] Henry Martyn Mulder. "Interval-regular graphs". In: *Discrete Mathematics* 41.3 (1982), pp. 253–269. ISSN: 0012-365X. DOI: https://doi.org/10.1016/0012-365X(82)90021-8. URL: https://www.sciencedirect.com/science/article/pii/0012365X82900218. Available at https://www.researchgate.net/profile/Henry-Martyn-Mulder/publication/220190794_Interval-regular_graphs/links/59ddd22c0f7e9b53c1ae1d98/Interval-regular-graphs.pdf.

[SZ82] J.M.S. Simões-Pereira and Christina M. Zamfirescu. "Submatrices of non-tree-realizable distance matrices". In: *Linear Algebra and its Applications* 44 (1982), pp. 1–17. ISSN: 0024-3795. DOI: https://doi.org/10.1016/0024-3795(82)90001-5. URL: https://www.sciencedirect.com/science/article/pii/0024379582900015. Available at https://library.navoiy-uni.uz/files/submatrices%20of%20non-tree-realizable%20distance%20matrices.pdf, https://core.ac.uk/download/pdf/81944889.pdf.

[Stu05] Bernd Sturmfels. "Can biology lead to new theorems?" In: *Annual report of the Clay Mathematics Institute* 2005 (2005), pp. 13–26. URL: https://www.claymath.org/library/annual_report/ar2005/05report_featurearticle.pdf. Available at https://www.claymath.org/library/annual_report/ar2005/05report_featurearticle.pdf, https://math.berkeley.edu/~bernd/ClayBiology.pdf.