A $(1+\varepsilon)$-approximation for Ultrametric Embedding in $\Theta(n^2)$ time

joint work with Gabriel Bathie

ULTRAMETRIC

(X, d) metric space

$$d(x,y) \leq d(x,z) + d(z,y)$$

Triangle inequality

ULTRAMETRIC

(X, d) metric space

$$d(x,y) \leq d(x,z) + d(z,y)$$

Triangle inequality

$$\Downarrow$$

$$\Delta(x,y) \leq \max(\Delta(x,z), \Delta(z,y))$$

Ultrametric inequality

(X, Δ) ultrametric space

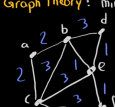
A few examples

Topology : discrete metric

Number Theory : p-adic numbers

$$d(x,y) = p^{-v_p(x-y)}$$

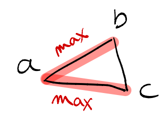
Graph Theory : minmax paths



$$\Delta(x,y) = \min_{p: x \to y} \max_{e \in p} w(e)$$

$$\Delta(b,c) = 2$$

$\forall a, b, c \in X$



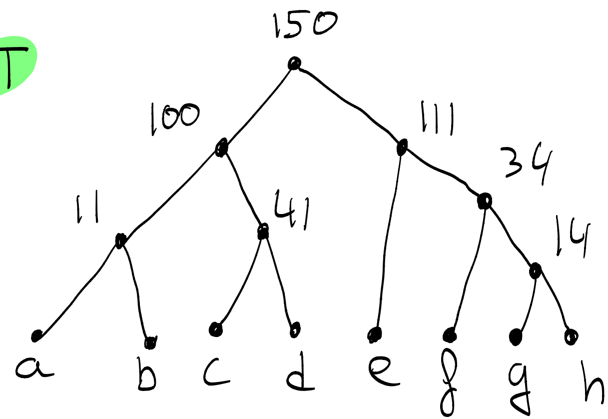
is isoceles "on the max"

+

extension to any cycles

(X, Δ) ultrametric space $\qquad \Delta(x,y) \leq \max(\Delta(x,z), \Delta(z,y))$

as a tree T



$\omega : \text{nodes} \longrightarrow \mathbb{R}^+$

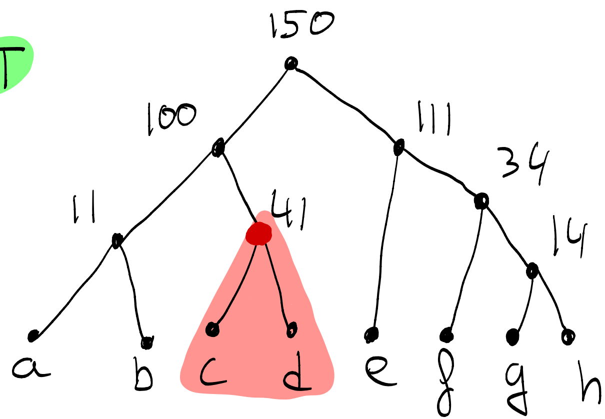
non-increasing from root to leaves

$$\Delta_{T,\omega}(x,y) = \omega(LCA(x,y))$$

(X, Δ) ultrametric space $\qquad \Delta(x,y) \leq \max(\Delta(x,z), \Delta(z,y))$

as a tree T

a node = one cluster



ω : nodes $\longrightarrow \mathbb{R}^+$

non-increasing from root to leaves

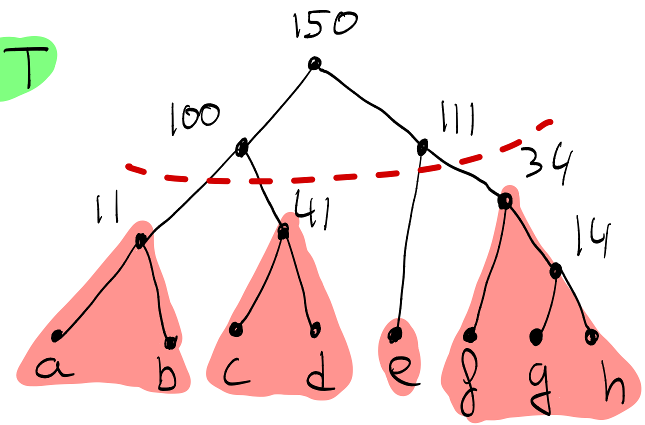
$$\Delta_{T,\omega}(x,y) = \omega(LCA(x,y))$$

(X, Δ) ultrametric space $\qquad \Delta(x,y) \leq \max(\Delta(x,z), \Delta(z,y))$

as a tree T

a node = one cluster

hierarchy of clusters !



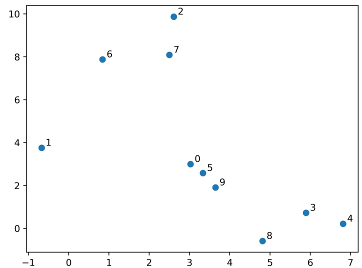
ω : nodes $\longrightarrow \mathbb{R}^+$

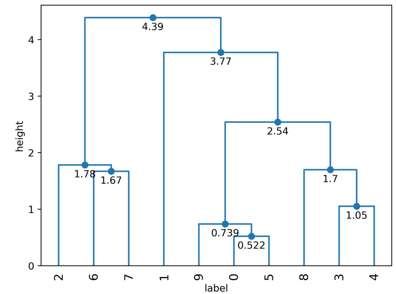
non-increasing from root to leaves

$$\Delta_{T,\omega}(x,y) = \omega(LCA(x,y))$$

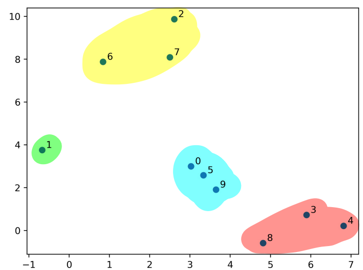
Euclidean space → Ultrametric space

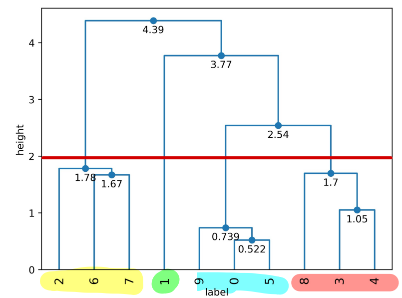
(X, ℓ_2) → (X, Δ)





hierchical clustering !





Goal Find an embedding that preserves relation between points

$$\Delta(x,y) \approx \ell_2(x,y)$$

BUF_∞ Best Ultrametric fit

Input (X,d) a metric space

Output Δ ultrametric such that

$$d(x,y) \leq \Delta(x,y) \leq \rho_{OPT} \cdot d(x,y)$$

minimal
called distortion

$$BUF_\infty \approx \min_{\Delta} \left\| \frac{\Delta}{d} \right\|_\infty$$

with $\Delta \geq d$

Goal Find an embedding that preserves relation between points

$$\Delta(x,y) \approx \ell_2(x,y)$$

BUF_∞ Best Ultrametric fit

Input (X,d) a metric space

Output Δ ultrametric such that

$$d(x,y) \leq \Delta(x,y) \leq \rho_{OPT} \cdot d(x,y)$$

minimal

called distortion

$$BUF_\infty \approx \min_{\Delta} \left\| \frac{\Delta}{d} \right\|_\infty \longrightarrow BUF_p \quad \min_{\Delta} \left\| \frac{\Delta}{d} \right\|_p$$

with $\Delta \geq d$ with $\Delta \geq p$

most popular methods: agglomerative algorithms

- average linkage
- single linkage
- Ward's method
- ...

Drawbacks

- $\Omega(n^2)$ running time in the best case
- most often $\Omega(n^2)$ memory
- not clear what loss functions these algos optimize

Theorem (Farach-Kannan-Warnow)

1. Optimal embedding in $O(n^2)$
2. Lower bound of $\Omega(n^2)$

Algorithmica (1995) 13: 155–179

Algorithmica

A Robust Model for Finding Optimal Evolutionary Trees

M. Farach,[1] S. Kannan,[2] and T. Warnow[3]

Abstract. Constructing evolutionary trees for species sets is a fundamental problem in computational biology. One of the standard models assumes the ability to compute distances between every pair of species, and seeks to find an edge-weighted tree T in which the distance d_{ij}^T in the tree between the leaves of T corresponding to the species i and j exactly equals the observed distance, d_{ij}. When such a tree exists, this is expressed in the biological literature by saying that the distance function or matrix is *additive*, and trees can be constructed from additive distance matrices in $O(n^2)$ time. Real distance data is hardly ever additive, and we therefore need ways of modeling the problem of finding the best-fit tree as an optimization problem.

In this paper we present several natural and realistic ways of modeling the inaccuracies in the distance data. In one model we assume that we have upper and lower bounds for the distances between pairs of species and try to find an additive distance matrix between these bounds. In a second model we are given a partial matrix and asked to find if we can fill in the unspecified entries in order to make the entire matrix additive. For both of these models we also consider a more restrictive problem of finding a matrix that fits a tree which is not only additive but also *ultrametric*. Ultrametric matrices correspond to trees which can be rooted so that the distance from the root to any leaf is the same. Ultrametric matrices are desirable in biology since the edge weights then indicate evolutionary time. We give polynomial-time algorithms for some of the problems while showing others to be NP-complete. We also consider various ways of "fitting" a given distance matrix (or a pair of upper- and lower-bound matrices) to a tree in order to minimize various criteria of error in the fit. For most criteria this optimization problem turns out to be NP-hard, while we do get polynomial-time algorithms for some.

BUF_p $\quad \min_{\Delta} \left\| \frac{\Delta}{d} \right\|_p$

with $\Delta \geq p$

p	complexity	approx
1	NP-hard APX-hard	$\log n$
2	NP-hard	$\sqrt{\log n \log\log n}$
p	???	$(\log n \log\log n)^{1/p}$
∞	$\Theta(n^2)$	we are here!

open questions here

Theorem (Bathie, L.)

For any $c \geq 1$, there is an algorithm that computes a c-approximation of BUF_∞ in time $\tilde{O}(n^{1+1/c})$ and memory $\tilde{O}(n^{1+\frac{1}{c}})$.

$$c = 1 + \varepsilon \longrightarrow \tilde{O}\left(n^{2-\varepsilon+O(\varepsilon^2)}\right)$$

subquadratic!

Before

$\sqrt{2} \cdot c$ - approx in time $\tilde{O}(n^{1+12/c^2})$

$\rightarrow$ at best $\sqrt{2} \cdot \sqrt{12} \approx 4{,}90$ - approx in subquadratic time.

Theorem (Bathie, L.)

For any $c \geq 1$, there is an algorithm that computes a c-approximation of BUF_∞ in time $\tilde{O}(n^{1+1/c})$ and memory $\tilde{O}(n^{1+\frac{1}{c}})$.

$c = 1+\varepsilon \longrightarrow \tilde{O}\left(n^{2-\varepsilon+O(\varepsilon^2)}\right)$

subquadratic!

+ performs well on experiments!

Before

$\sqrt{2} \cdot c$ – approx in time $\tilde{O}\left(n^{1+12/c^2}\right)$

$\rightarrow$ at best $\sqrt{2} \cdot \sqrt{12} \approx 4{,}90$ – approx in subquadratic time.

Farach-Kannan-Warnow Algorithm

Input (X, d)

Output optimal ultrametric for BUF_∞

1. Compute a minimum spanning tree T — $\tilde{O}(|X|^2)$

2. Compute the cut weights of edges in T — $\tilde{O}(|X|^2)$

3. Output a cartesian tree $= \Delta$ — $\tilde{O}(|X| \cdot \log |X|)$

Approximation Algorithm

Input ~~(X, d)~~ (X, ℓ_2)

Output ~~optimal~~ approx ultrametric for BUF_∞

γ - Kruskal tree

1. Compute a ~~minimum spanning~~ tree T — $\forall \gamma \geq 1,\ \tilde{O}\left(|X|^{1+\frac{1}{\gamma^2}}\right)$

2. Compute β- cut weights of edges in T — $\forall \beta \geq 1,\ \tilde{O}\left(|X|^{1+\frac{1}{\beta^2}}\right)$

3. Output a cartesian tree $= \Delta$ — $\tilde{O}(|X| \cdot \log |X|)$

Claim this outputs a $\gamma \cdot \beta$-approx

The End

Thanks!