A Fast and Simple Algorithm for Computing Approximate Euclidean Minimum Spanning Trees

Sunil Arya
Hong Kong University of Science and Technology

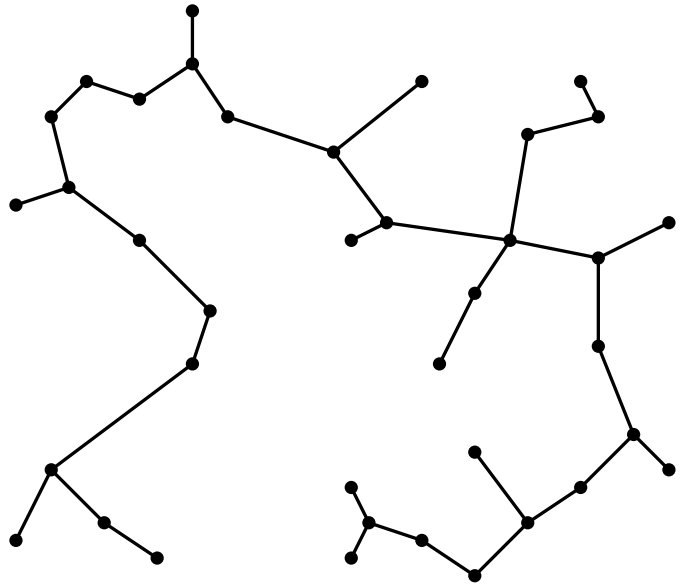
and

David Mount
University of Maryland

Euclidean Minimum Spanning Tree

Euclidean MST

Given a set P of n points in $\mathbb{R}^d$, the EMST of P is the minimum spanning tree of the complete graph on P, where each edge is weighted by the Euclidean distance between these points.



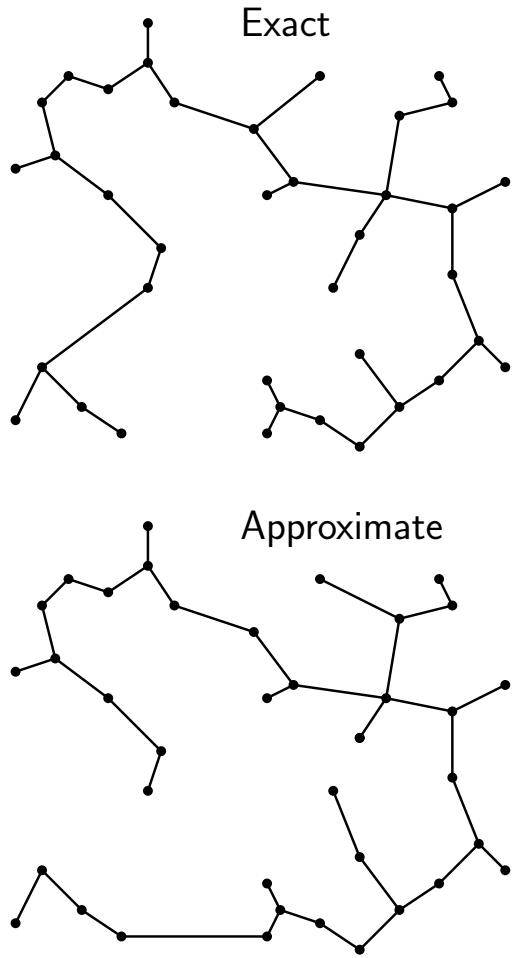
- d is constant
- Results generalize to common norms (L_1, L_∞)

Approximate EMST

As d grows, exact algorithms for the EMST run in nearly quadratic time.

ε-Approximate EMST

Return a spanning tree whose weight is at most $(1+\varepsilon) \cdot wt(\mathrm{MST}(P))$.



Prior Results - Computing the Whole Tree

- Exact in $\mathbb{R}^d$:
  - Yao (1982): $\widetilde{O}\left(n^{2-\frac{1}{2^d+1}}\right)$
  - Agarwal *et al.* (1991): $\widetilde{O}(n^{2-\frac{4}{d}})$
- ε-Approximate in $\mathbb{R}^d$: (Asymptotic d)
  - Har-Peled, *et al.* (2012): Roughly $\widetilde{O}\left(d \cdot n^{1+\frac{1}{1+\varepsilon}}\right) \rightarrow$ using LSH
- ε-Approximate in $\mathbb{R}^d$: (Constant d)
  - Vaidya (1991): $\widetilde{O}(\varepsilon^{-d} \cdot n)$
  - Callahan and Kosaraju (1995): $\widetilde{O}(\varepsilon^{-\frac{d}{2}} \cdot n) \rightarrow$ using WSPDs
  - Arya and Chan (2014): $\widetilde{O}(\varepsilon^{-(\frac{d}{3}+O(1))} \cdot n) \rightarrow$ using DVDs

Ideal: Nearly linear in n and low dependency on ε: $\frac{1}{\varepsilon^d} \rightarrow \frac{1}{\varepsilon^{d/2}} \rightarrow \frac{1}{\varepsilon^{d/3}} \rightarrow \cdots$

Prior Results - Estimating the Weight

ε-Approximating the weight of the MST (w.h.p.) in sublinear time:

- Chazelle *et al.* (2005) $O(DW\varepsilon^{-3})$
  (for graphs of degree D and edge-weight spread W)
- Czumaj *et al.* (2005) Roughly $O(\sqrt{n} \cdot \varepsilon^{-\frac{d}{2}})$ for EMST
  (assuming appropriate geometric oracles)
- Czumaj and Sohler (2009) $O(n \cdot \mathrm{poly}(1/\varepsilon))$ for MST in metric space
  (presented as an $n \times n$ distance matrix)

Main Result

Existing approximation algorithms either:

- Have ε-factors that grow exponentially with dimension (ε^{-d} or $\varepsilon^{-\frac{d}{2}}$)

  — or —

- Just estimate the weight of the EMST

ε dependencies are a major issue. Is it possible to reduce them?

Main result:

- ε-approximate EMSTs in $\mathbb{R}^d$ in $\widetilde{O}(\varepsilon^{-2}n)$ time (constants depend on d)
- Simple, deterministic algorithm (quadtrees, well-separated pairs)
- Exploits a simple EMST lower bound (Czumaj *et al.* 2005) and an amortization trick

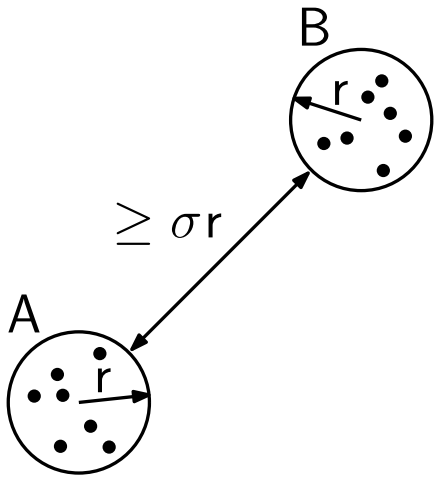
Roadmap

- Preliminaries - Well-separated pairs, quadtrees, EMST lower bound
- Algorithm - Getting closest pairs on the cheap
- Analysis - On the importance of being sloppy

Well-Separated Pairs

Well-Separated Pair:

Given a separation parameter $\sigma \geq 1$, two point sets A and B are σ-well separated if they can be enclosed within balls of radius r such that the closest distance between these balls is at least σr



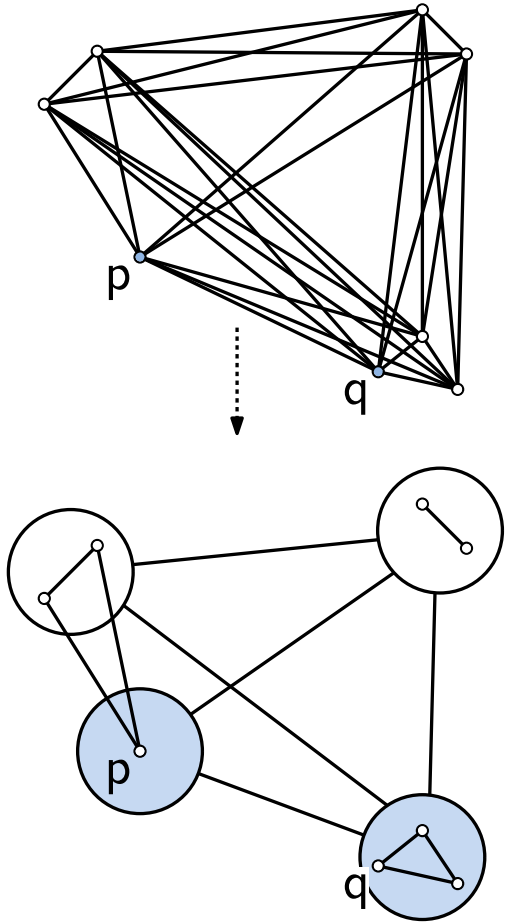
Well-Separated Pair Decomposition

Well-Separated Pair Decomposition (WSPD):

Given a point set P, a σ-WSPD is a set of pairs $\{\{A_i, B_i\}\}_{i=1}^{k}$ of subsets of P such that:

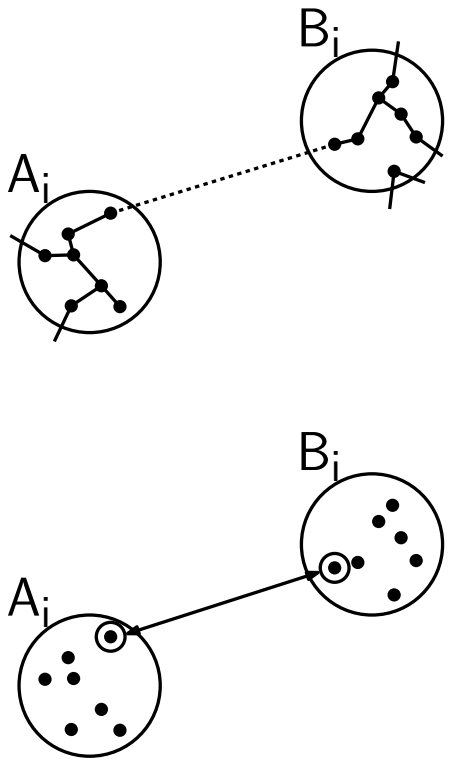
(1) for $1 \leq i \leq k$, A_i and B_i are σ-well separated

(2) for any $p, q \in P$, there is exactly one pair $\{A_i, B_i\}$ such that $p \in A_i$ and $q \in B_i$



Useful Observations (Callahan and Kosaraju (1995))

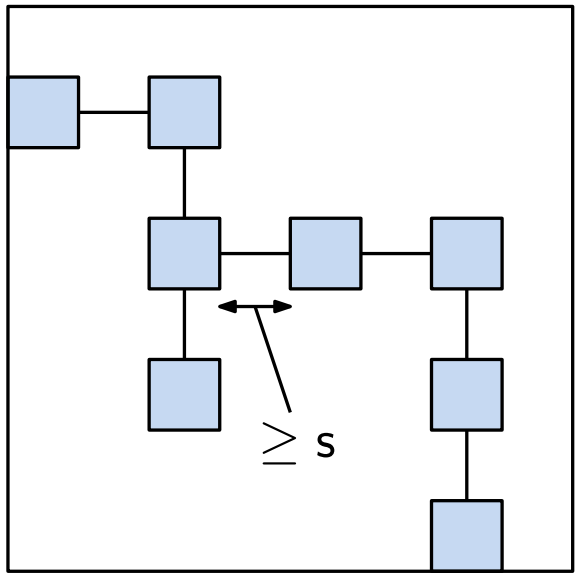
- A 2-WSPD of size $O(n)$ can be constructed in time $O(n \log n)$
- Each pair of a 2-WSPD contributes at most one edge to the EMST
- Given a 2-WSPD for P, form a graph G from the closest pair from each (A_i, B_i)
  - $|G| = O\left(n\left(2\sqrt{d}\right)^d\right) = O(n)$
  - $\mathrm{MST}(G) = \mathrm{EMST}(P)$
- ε-approximate closest pairs yield a graph G_ε
  - $wt(\mathrm{MST}(G_\varepsilon)) \leq (1+\varepsilon) \cdot wt(\mathrm{EMST}(P))$



Fast EMST Lower Bound

Lemma [Czumaj *et al.* 2005]

Consider a grid of side length s in $\mathbb{R}^d$. Let m be the number of grid boxes containing at least one point of P. Then there is a constant c (depending on d) such that $wt(\mathrm{MST}(P)) \geq s\,m/c$.



Proof:

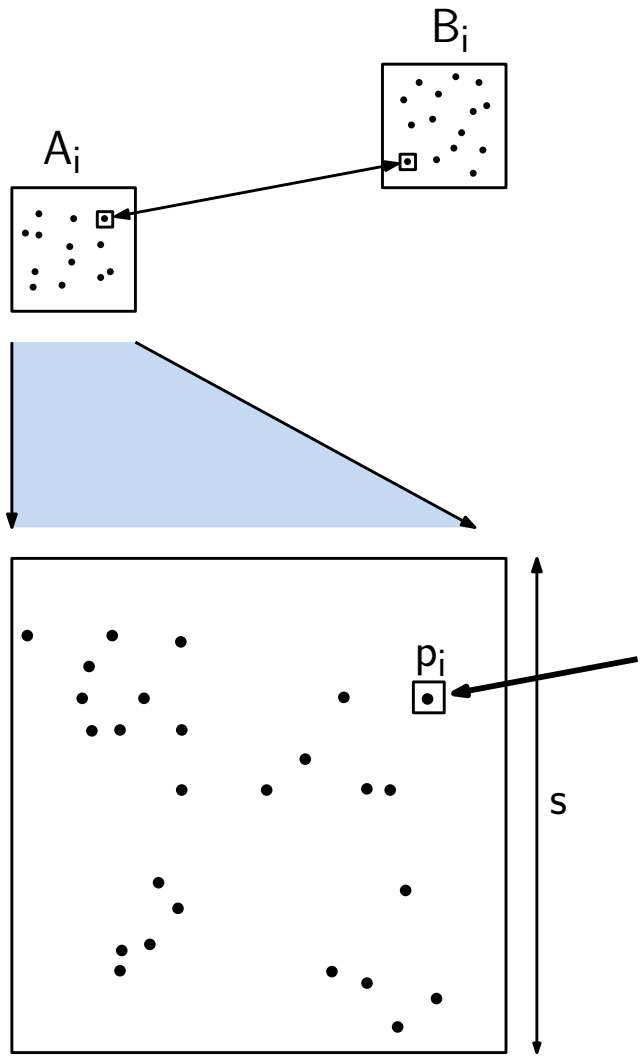
- Color the grid with 2^d colors. Boxes of the same color are separated by distance $\geq s$
- Some color class has at least $m/2^d$ boxes
- The cost of connecting these boxes is $\Omega(s\,m)$

Roadmap

- Preliminaries - Well-separated pairs, quadtrees, EMST lower bound
- Algorithm - Getting closest pairs on the cheap
- Analysis - On the importance of being sloppy

Simple (Slow) Algorithm

- Compute a 2-WSPD for P
- Each box stores a representative point
- For each WSP (A_i, B_i):
  - Let s be the box size. Subdivide A_i and B_i until the box diameter $\leq \varepsilon s$
  - $(p_i, q_i) \leftarrow$ closest pair of box representatives
- $G \leftarrow$ closest pairs. Return $\mathrm{MST}(G)$



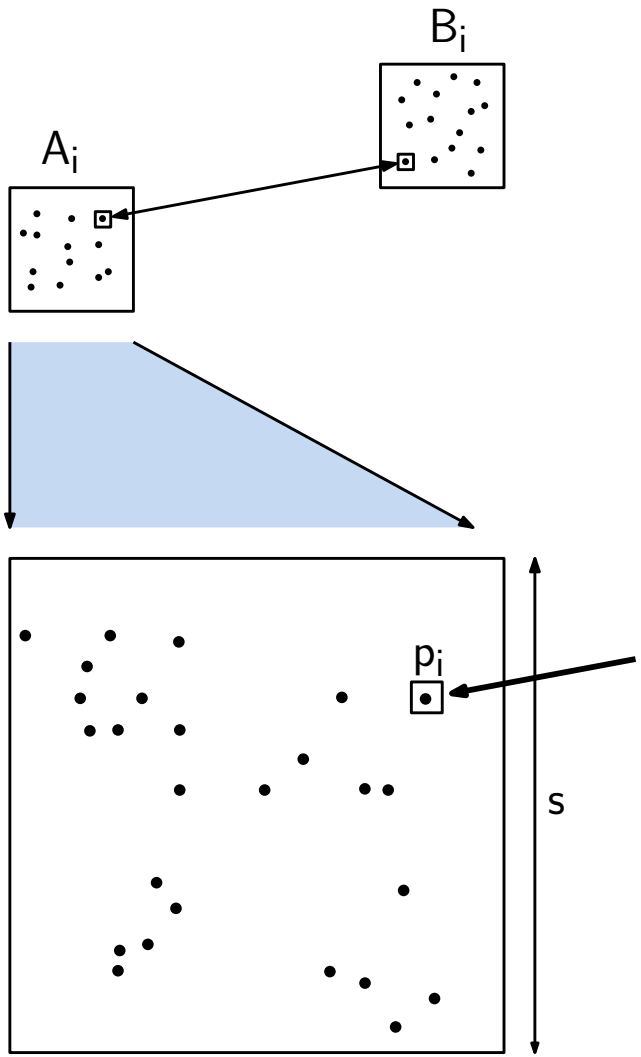
Slow! $O(n/(\varepsilon^d)^2) = O(n \cdot \varepsilon^{-2d})$.

Worst case arises when pairs have many boxes.

By Lower-Bound Lemma, MST cost is high.

A Smarter/Sloppier Algorithm

- Compute a 2-WSPD for P
- Each box stores a representative point
- For each (A_i, B_i) approximate the closest pair:
  - Let s be the box size. Subdivide A_i and B_i until either:
    - Box diameter $\leq \varepsilon s$ — or —
    - The number of nonempty boxes $\geq c/\varepsilon$ (for some constant c)
  - $(p_i, q_i) \leftarrow$ closest pair of box representatives
- $G \leftarrow$ closest pairs. Return $\mathrm{MST}(G)$



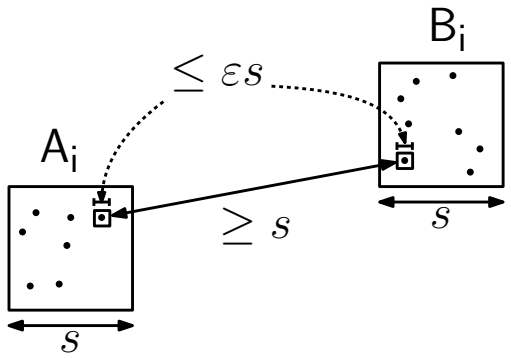
Roadmap

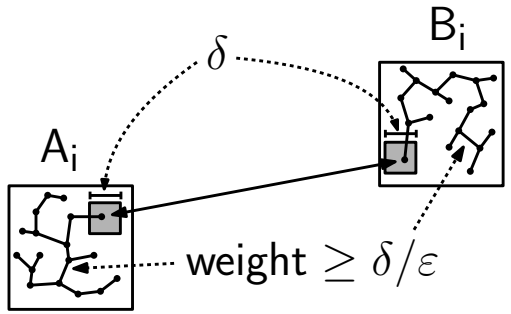
- Preliminaries - Well-separated pairs, quadtrees, EMST lower bound
- Algorithm - Getting closest pairs on the cheap
- Analysis - On the importance of being sloppy

Approximation Analysis (First Attempt)

- Case 1: Box diameters $\leq \varepsilon s$:
  - Absolute error $\leq \varepsilon s \leq \varepsilon \cdot \mathrm{dist}(A_i, B_i)$
  - Relative error $\leq \varepsilon$
- Case 2: Number of nonempty boxes $\geq c/\varepsilon$:
  - Let δ be the diameters of the boxes
  - Absolute error $\lessapprox \delta$
  - By Lemma, weight of MST within box $\geq \delta(c/\varepsilon)/c = \delta/\varepsilon$
  - Relative error is $\lessapprox \varepsilon$ (amortized over the box)

. . . hey, aren't you multiply charging?

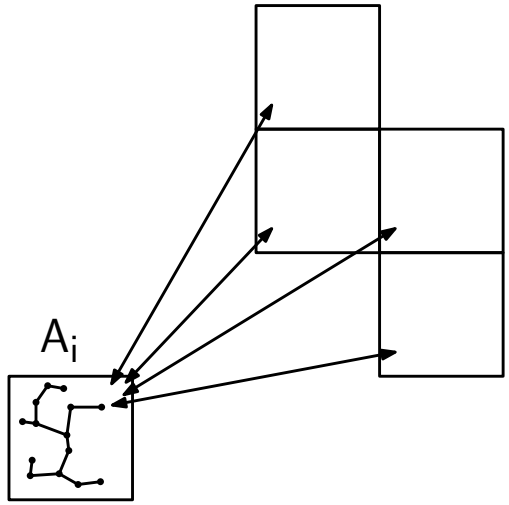


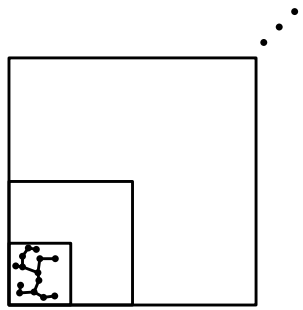


Approximation Analysis (Finer Points)

We charge the same MST edge multiple times:

- Multiple WSPs share the same quadtree box
  — each box is in $O\left(\sqrt{d}\right)^d = O(1)$ WSPs
  $\rightarrow$ increase c by this constant
- Multiple tree levels charge the same edge
  $\rightarrow$ further increase c by tree height
  — $\times\ O\left(\log \frac{n}{\varepsilon}\right)$ [Arora (1998)]
- Reducing the log factor
  — A more refined analysis reduces the log factor to $O\left(\log \frac{1}{\varepsilon}\right)$





Execution Time

- Build the quadtree and WSPD: $O(n \log n)$
- Find the approximate closest pair for each WSP:
  - $O(n)$ well-separated pairs
  - $O(\varepsilon^{-1} \log \frac{1}{\varepsilon})$ boxes per pair
  - $O(\varepsilon^{-2} \log^2 \frac{1}{\varepsilon})$ representative pairs per WSP
- Compute the MST of G: $O(n \log n)$
- Total time: $O\big(n \log n + (\varepsilon^{-2} \log^2 \frac{1}{\varepsilon}) n\big) = \widetilde{O}(\varepsilon^{-2} n)$

Concluding Remarks

- Summary:
  - ε-approximate EMSTs in $\mathbb{R}^d$ in $\widetilde{O}(\varepsilon^{-2}n)$ time
  - Simple, deterministic algorithm (quadtrees, well-separated pairs)
- Caveats:
  - EMST minimizes the bottleneck (max) edge cost — but ours does not
  - Big-Oh hides constant factors that grow exponentially with dimension
- Further Work:
  - Implementation (we're working on it)
  - Approximate minimum bottleneck spanning tree in similar time?
  - Reduce ε^{-2} (while maintaining simplicity)?
  - Is the $\log \frac{1}{\varepsilon}$ factor needed (or artifact of analysis)?

Bibliography

- P. K. Agarwal, H. Edelsbrunner, O. Schwartzkopf, and E. Welzl, Euclidean minimum spanning trees and bichromatic closest pairs, Discr. and Comp. Geom., 6, 1991, 407–422
- S. Arora, Polynomial Time Approximation Schemes for Euclidean Traveling Salesman and Other Geometric Problems, J. of the ACM, 45, 1998, 753–782
- S. Arya and T. M. Chan. Better ε-dependencies for offline approximate nearest neighbor search, Euclidean minimum spanning trees, and ε-kernels. Proc. 30th SoCG, 2014, 416–425.
- P. B. Callahan and S. R. Kosaraju, A Decomposition of Multidimensional Point Sets with Applications to k-Nearest-Neighbors and n-Body Potential Fields, J. of the ACM, 42, 1995, 67–90
- A. Czumaj, F. Ergün, L. Fortnow, A. Magen, I. Newman, R. Rubinfeld, and C. Sohler, Approximating the Weight of the Euclidean Minimum Spanning Tree in Sublinear Time, SIAM J. Comput., 35, 2005, 91–109
- A. Czumaj and C. Sohler, Estimating the Weight of Metric Minimum Spanning Trees in Sublinear Time, SIAM J. Comput., 39, 2009, 904–922
- S. Har-Peled, P. Indyk, and R. Motwani. Approximate nearest neighbor: Towards removing the curse of dimensionality. Theory of Computing, 8, 2012, 321–350.
- A. C. Yao, On Constructing Minimum Spanning Trees in k-Dimensional Spaces and Related Problems, SIAM J. Comput., 11, 1982, 721–736