COMPSCI 590.1: Graph Algorithms Febuary 25, 2016

Lecture 13 Well Separated Pair Decomposition

Lecturer: Pankaj K. Agarwal *Scribe: Rex Ying*

1 Overview

In the previous lecture, we covered an approximation algorithm that answers a $(1+\epsilon)$-approximate nearest neighbor query in logarithmic time, using only linear space. This algorithm makes use of the compressed quadtree data structure, which clusters the set of distances between query point and the points in the given set, $\{\|q-p_i\| \,|1 \le i \le n\}$, into $O(\log n + \frac{1}{\epsilon^d})$ clusters. In this lecture, we introduce the notion of well-separated pair decomposition (WSPD) that clusters the set of all pairwise distances in a given point set and discuss a few applications of WSPD.

2 Well-separated Pair Decomposition

The objective is to cluster the set of all $\binom{n}{2}$ distances between points in a given point set into a few clusters, such that all distances in each cluster are roughly the same.

A naive way to cluster the pairwise distances is to use the algorithm in the previous lecture to cluster the set of distances $\{\|p_j - p_i\| \,|1 \le i \le n\}$ for each $1 \le j \le n$. This results in $O(n\log n + \frac{n}{\epsilon^d})$ clusters. In this lecture, we modify the algorithm to achieve linear number of clusters.

2.1 Algorithm

In this subsection, we outline the procedure for computing a WSPD, give the time complexity, and prove that the number of clusters is linear. We start by defining the WSPD.

Definition 1. *Given two point sets A and B and a constant $0 < \delta < 1$, we say (A, B) is δ-separated if*

$$\max\{\text{diam}(A), \text{diam}(B)\} \le \delta d(A, B),$$

where $d(A, B) = \min_{(a,b)\in A\times B} \|a-b\|$.

Intuitively this is shown in figure 1, where the distance between the enclosing disks is at least $\frac{D}{\delta}$. By triangle inequality, for any representative we choose in A and B, we have

$$\forall (a,b) \in A \times B, \quad \|a-b\| \le (1+2\delta)\,\|\text{rep}(A) - \text{rep}(B)\|\,. \tag{1}$$

Definition 2. *An ϵ-separated pair decomposition (ϵ-WSPD) of a given point set S is a family of pairs $\mathcal{F} = \{(A_1, B_1), \ldots, (A_s, B_s)\}$ such that*

- *A_i, B_i are ϵ-separated for any $1 \le i \le s$.*
- *For any pair $p, q \in S$, there is a unique i such that $p \in A_i$, $q \in B_i$ (or vice-versa).*

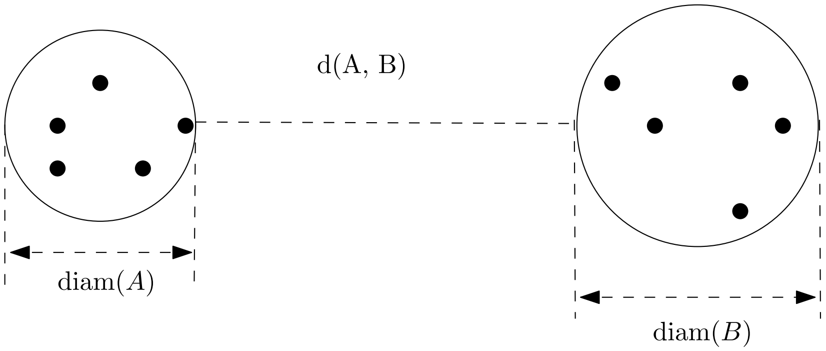

Figure 1: δ-separated condition

We say the size of the ϵ-WSPD is the size of the family $|\mathcal{F}|$, which obeys the following theorem.

Theorem 1. *Given $S \subset \mathbb{R}^d$ and $\epsilon > 0$, an ϵ-WSPD of size $O(\frac{n}{\epsilon^d})$ can be computed in time $O(n \log n + \frac{n}{\epsilon^d})$.*

Fact 2. *Although $s = |\mathcal{F}|$ is linear to the number of points, the quantity $\sum_{i=1}^{s} |A_i| + |B_i|$ can be arbitrarily large, unless the spread of S is $O(n^{O(1)})$, in which case $\sum_{i=1}^{s} |A_i| + |B_i| = O(\frac{n}{\epsilon^d} \log n)$.*

The algorithm for constructing the WSPD first constructs a compressed quadtree. Define Δ_u to be the diameter of the quadtree node u. We use the following routine:

```
WSPD(u, v) {
  if $\Delta_u < \Delta_v$
    swap(u, v)
  end if
  if $\Delta_u \leq \epsilon d(\square_u, \square_v)$
    F = F $\cup \{(u,v)\}$
  else
    foreach children $w$ of $u$
      WSPD($w$, $v$)
    end for
  end if
}
```

Intially, the algorithm sets $S = \emptyset$, and calls WSPD(root, root).

The correctness of 1 follows from this construction algorithm. The complete proof can be found in [HP11]. The main idea is to use a packing argument.

Lemma 3. *If $(S_u, S_v) \in \mathcal{F}$,*

$$\max\{\Delta_u, \Delta_v\} \leq \min\{\Delta_{p(u)}, \Delta_{p(v)}\},$$

where $S_u = S \cap \square_u$ and $p(u)$ denote the parent node of u.

Proof. Without loss of generality, suppose $\Delta_u > \Delta_v$.

If $\min\{\Delta_{p(u)}, \Delta_{p(v)}\} = \Delta_{p(u)}$, since $\Delta_u \leq \Delta_{p(u)}$ by property of quadtree, the lemma is true.

If $\min\{\Delta_{p(u)}, \Delta_{p(v)}\} = \Delta_{p(v)}$, suppose the pair (u, v) is produced by the WSPD algorithm after a sequence of recursive calls of pairs:

$$(u_0, v_0), (u_1, v_1), \ldots, (u_t, v_t) = (u, v). \tag{2}$$

By property of quadtree, the diameters of u_i and v_i decreases as i increases. Suppose j is the first index such that $v_j = v$, i.e. $v_{j-1} = p(v)$. Then

$$\Delta_{p(v)} = \Delta_{v_{j-1}} \geq \Delta_{u_{j-1}} \geq \Delta_u. \tag{3}$$

In either case, the lemma holds. □

Corollary 4. *If $(S_u, S_v) \in \mathcal{F}$, the corresponding quadtree nodes u and v are either at the same level or one level apart in the quadtree.*

Lemma 5. *The number of squares of size Δ that are at most l away from a given point is $O(\frac{l^d}{\Delta^d})$.*

Proof of Theorem 1. Suppose WSPD(u, v) was called recursively by WSPD(u, v'), i.e. $v' = p(v)$ and $\Delta_{v'} \geq \Delta_u$, then charge the pair (u, v) to the node v'.

Since u and v' are not ϵ-separated,

$$\Delta_{v'} \geq \epsilon d(u, v) \Rightarrow d(u, v) < \frac{1}{\epsilon}\Delta_{v'}. \tag{4}$$

Let $N_{v'} = \{u | (u, v) \in \mathcal{F}, v' = p(v)\}$. All nodes in $N_{v'}$ are disjoint by property of quadtree. By corollary 4 and lemma 5,

$$|N_{v'}| = O\left(\frac{1}{\epsilon^d}\right). \tag{5}$$

Summing over all n nodes in the quadtree $\mathcal{T}$,

$$\sum_{v' \in \mathcal{T}} |N_{v'}| = O\left(\frac{n}{\epsilon^d}\right). \tag{6}$$

The time taken to compute WSPD(root, root) is linear in the size of the $\mathcal{F}$, together with the time to construct the compressed quadtree, the total running time is thus $O(n \log n + \frac{n}{\epsilon^d})$.

□

2.2 Applications

In this subsection, we give some of the applications of WSPD.

Example 1. A graph $G = (S, E)$ is a *ϵ-spanner* if

$$\forall p, q \in S, \quad d_G(p, q) \leq (1 + \epsilon) \|p - q\|. \tag{7}$$

That is, we want to construct a small-size graph such that the shortest path distance in G for any two points is roughly their Euclidean distance.

Spanners are widely used for designing communication networks, which require to use few edges to connect the network but the cost of reaching one node from another is still not too much greater than that in a complete graph of S.

Remark 1. *The Delaunay triangulation of S is a 2-spanner. But we can compute a better spanner using WSPD.*

Theorem 6. *$\forall \epsilon > 0$, an ϵ-spanner on S of size $O(\frac{n}{\epsilon^d})$ can be constructed in $O(n \log n + \frac{n}{\epsilon^d})$ time.*

Proof. Set $\delta = \frac{\epsilon}{16}$. Construct a δ-WSPD $\mathcal{F}$ of S. For each pair $(A_i, B_i) \in \mathcal{F}$, choose $(p_i, q_i) \in A_i \times B_i$. Let $E = \{(p_i, q_i) | 1 \leq i \leq |\mathcal{F}|\}$. We use induction on the $\binom{n}{2}$ distances to show that $G = (S, E)$ is an ϵ-spanner.

Fix an i, suppose p is a representative of A_i, and q is a representative of B_i, as illustrated in figure 2.

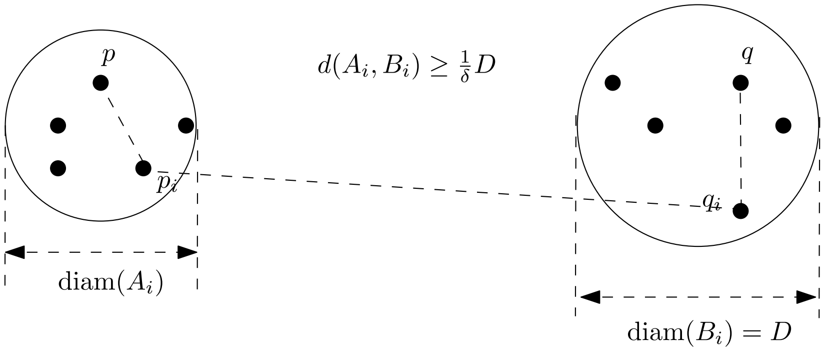

Figure 2: The inductive step

$$d_G(p, q) = d_G(p, p_i) + \|p_i, q_i\| + d_G(q, q_i). \tag{8}$$

By the property of δ-WSPD,

$$\|p - p_i\| \leq \delta \|p_i - q_i\|. \tag{9}$$

By the inductive hypothesis,

$$d_G(p, p_i) \leq (1 + \epsilon) \|p - p_i\| \leq (1 + \epsilon)\delta \|p_i - q_i\|. \tag{10}$$

Similarly, $(1 + \epsilon) \|q - q_i\| \leq (1 + \epsilon)\delta \|p_i - q_i\|$. By triangle inequality,

$$\|p_i - q_i\| \leq \|p - q\| + \|p - p_i\| + \|q - q_i\| \leq \|p - q\| + 2\delta \|p - q\| = (1 + 2\delta) \|p - q\|. \tag{11}$$

Since $\delta \leq \frac{\epsilon}{16}$, it can be argued that $\|p - p_i\| < \|p - q\|$. Therefore

$$\begin{aligned} d_G(p, q) &\leq d_G(p, p_i) + d_G(q, q_i) + \|p_i - q_i\| & (12) \\ &\leq 2(1 + \epsilon)\delta \|p_i - q_i\| + \|p_i - q_i\| & (13) \\ &\leq (1 + 2(1 + \epsilon)\delta)(1 + 2\delta) \|p - q\| & (14) \\ &\leq (1 + \epsilon) \|p - q\|. & (15) \end{aligned}$$

□

Example 2. Suppose p has weight w_p, compute

$$\sum_{p, q \in S, p \neq q} \frac{w_p w_q}{\|p - q\|}. \tag{16}$$

This is widely used in molecular biology and physics where one needs to track pairwise interactions between large number of entities.

The naive exact algorithm takes $O(n^2)$ time, which is often not practical when the size of S is large. Using WSPD, one can compute the interactions between each pair of sets (A_i, B_i) for $1 \leq i \leq |\mathcal{F}|$ to approximate the quantity, thus reducing the time complexity to $O(\frac{n}{\epsilon^d})$. One can read [CK95] further for details.

Other examples of applications of WSPD include approximating all-pair nearest neighbor, diameter of point sets [HP01] and sequence of alignment problems [AD12].

3 Summary

In this lecture, we covered the concept of the well-separated pair decomposition, which utilizes the data structure compressed quadtree to cluster the set of all pairwise distances of a point set. We showed that the algorithm clusters distances into linear number of clusters using a packing argument, and proved its run time. In applications, WSPD is essential in designing efficient approximation algorithms that require computation of distances within a point set.

References

[AD12] Carola Wenk Anne Driemel, Sariel Har-Peled. Approximating the frchet distance for realistic curves in near linear time. *Discrete and Computational Geometry*, 48(1):94–127, 2012.

[CK95] 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. Assoc. Comput. Mach.*, 42(1):67–90, 1995.

[HP01] S. Har-Peled. A practical approach for computing the diameter of a point set. In *Proceedings of the seventeenth annual symposium on Computational geometry, ACM*, pages 177–186, 2001.

[HP11] S. Har-Peled. *Geometric Approximation Algorithms*. American Mathematical Society, 2011.