

# Faster, Deterministic and Space Efficient Subtrajectory Clustering

Ivor van der Hoog  

Department of Applied Mathematics and Computer Science, Technical University of Denmark, Denmark

Thijs van der Horst  

Department of Information and Computing Sciences, Utrecht University, the Netherlands  
Department of Mathematics and Computer Science, TU Eindhoven, the Netherlands

Tim Ophelders  

Department of Information and Computing Sciences, Utrecht University, the Netherlands  
Department of Mathematics and Computer Science, TU Eindhoven, the Netherlands

---

## Abstract

Given a trajectory  $T$  and a distance  $\Delta$ , we wish to find a set  $C$  of curves of complexity at most  $\ell$ , such that we can cover  $T$  with subcurves that each are within Fréchet distance  $\Delta$  to at least one curve in  $C$ . We call  $C$  an  $(\ell, \Delta)$ -clustering and aim to find an  $(\ell, \Delta)$ -clustering of minimum cardinality. This problem variant was introduced by Akitaya *et al.* (2021) and shown to be NP-complete. The main focus has therefore been on bicriteria approximation algorithms, allowing for the clustering to be an  $(\ell, \Theta(\Delta))$ -clustering of roughly optimal size.

We present algorithms that construct  $(\ell, 4\Delta)$ -clusterings of  $\mathcal{O}(k \log n)$  size, where  $k$  is the size of the optimal  $(\ell, \Delta)$ -clustering. We use  $\mathcal{O}(n^3)$  space and  $\mathcal{O}(kn^3 \log^4 n)$  time. Our algorithms significantly improve upon the clustering quality (improving the approximation factor in  $\Delta$ ) and size (whenever  $\ell \in \Omega(\log n / \log k)$ ). We offer deterministic running times improving known expected bounds by a factor near-linear in  $\ell$ . Additionally, we match the space usage of prior work, and improve it substantially, by a factor super-linear in  $n\ell$ , when compared to deterministic results.

**2012 ACM Subject Classification** Theory of computation  $\rightarrow$  Computational Geometry

**Keywords and phrases** Fréchet distance, clustering, set cover

**Funding** *Ivor van der Hoog*: This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 899987, and from the Carlsberg Foundation via Eva Rotenberg’s Young Researcher Fellowship CF21-0302 “Graph Algorithms with Geometric Applications”.

*Tim Ophelders*: Partially supported by the Dutch Research Council (NWO) under the project number VI.Veni.212.260.

**Acknowledgements** We thank Jacobus Conradi for pointing out an error in a preprint of this paper.

## 1 Introduction

In subtrajectory clustering, the goal is to partition an input trajectory  $T$  with  $n$  vertices into subtrajectories and group them into *clusters* such that all subtrajectories within a cluster have low Fréchet distance to one another. Clustering under the Fréchet distance is a natural application of the Fréchet distance and a well-studied topic [9, 10, 11, 15, 16] with applications in, for example, map reconstruction [6, 7]. In recent years, several variants of this algorithmic problem have been proposed [1, 3, 5, 8]. Regardless of the variant, the subtrajectory clustering problem has been shown to be NP-complete [1, 3, 8].

We focus on the problem variant proposed by Akitaya, Brüning, Chambers, and Driemel [3]. Given a trajectory  $T$  and a distance  $\Delta$ , and some  $\ell$ , they compute what we call an  $(\ell, \Delta)$ -clustering  $C$  of  $T$ . Each cluster  $Z \in C$  is a set of subtrajectories together with a center curve

| # Clusters                         | $\Delta' =$ | Time  | Space                           | Source  |
|------------------------------------|-------------|---|---------------------------------|---------|
| $\mathcal{O}(k\ell^2 \log(k\ell))$ | $19\Delta$  | $\tilde{\mathcal{O}}(k\ell^4 \lambda^2 + n\lambda)$ | $\mathcal{O}(n + \lambda)$      | [3]     |
| $\mathcal{O}(k\ell \log k)$        | $11\Delta$  | $\tilde{\mathcal{O}}(kn^3 \ell)$                    | $\tilde{\mathcal{O}}(n^3)$      | [5]     |
| $\mathcal{O}(k \log n)$            | $11\Delta$  | $\tilde{\mathcal{O}}(kn^4 \ell + n^4 \ell^2)$       | $\tilde{\mathcal{O}}(n^4 \ell)$ | [13]    |
| $\mathcal{O}(k \log n)$            | $4\Delta$   | $\mathcal{O}(kn^3 \log^4 n)$                        | $\mathcal{O}(n^3)$              | Thm. 16 |

■ **Table 1** Prior work and our result. The first two (red) rows indicate randomized results.  $k$  denotes the smallest  $(\ell, \Delta)$ -clustering size of  $T$ .  $\lambda$  denotes the arc length of  $T$  relative to  $\Delta$ .

(the ‘reference curve’  $P_Z$ ) of complexity at most  $\ell$ . Each curve in a cluster must have Fréchet distance at most  $\Delta$  to the center and each point on  $T$  must be present in at least one cluster. The goal is to compute an  $(\ell, \Delta)$ -clustering of minimum cardinality. Note that the parameter  $\ell$  is necessary to not trivialize the problem. Indeed, if  $P_Z$  may have arbitrary complexity, then a trivial  $(n, 0)$ -clustering exists consisting of a single cluster  $Z$  where  $Z = \{T\}$ .

Akitaya *et al.* [3] propose a bicriteria approximation scheme: Given  $\ell$  and  $\Delta$ , let  $k$  be the minimum size of an  $(\ell, \Delta)$ -clustering of  $T$ . The goal is to compute an  $(\ell, \Theta(\Delta))$ -clustering of size  $\mathcal{O}(f(k))$ . This paradigm was studied in [3, 5, 13] and previous results are summarised in Table 1. [3] computes an  $(\ell, 19\Delta)$ -clustering of  $\mathcal{O}(k\ell^2 \log(k\ell))$  size. The running time and space bounds depend on the *arc length* of  $T$  relative to  $\Delta$ . Brüning, Conradi and Driemel [5] compute an  $(\ell, 11\Delta)$ -clustering of  $\mathcal{O}(k\ell \log k)$  size (where the hidden constant is exceptionally large). Their algorithm uses  $\tilde{\mathcal{O}}(n^3)$  space and has  $\tilde{\mathcal{O}}(kn^3)$  expected running time. Recently, Conradi and Driemel [13] improve both the size and the quality of the clustering. They compute an  $(\ell, 11\Delta)$ -clustering of  $\mathcal{O}(k \log n)$  size in  $\tilde{\mathcal{O}}(n^4 \ell)$  space and  $\tilde{\mathcal{O}}(kn^4 \ell + n^4 \ell^2)$  time.

**Results.** We present a bicriteria approximation algorithm that uses  $\mathcal{O}(kn^3 \log^4 n)$  time and  $\mathcal{O}(n^3)$  space, and computes an  $(\ell, 4\Delta)$ -clustering of size  $\mathcal{O}(k \log n)$ . When compared to previous works [3, 5, 13] our results:

- obtain deterministic results and improve the running time by a factor near-linear in  $\ell$ ,
- match the space usage,
- improve the approximation in  $\Delta$  from a factor 11 to 4,
- asymptotically match the clustering size (whenever  $\ell \in \Omega(\log n / \log k)$ ).

Compared exclusively to deterministic results [13], we instead improve time by a factor near-linear in  $n\ell$ , space by a factor super-linear in  $n\ell$ , and obtain asymptotically equal clustering size for all  $\ell$  (see also Table 1).

**Methodology and contribution.** Our algorithm constructs a clustering iteratively by greedily adding a cluster that covers an approximately-maximum set of uncovered points on  $T$ . The challenge is to compute such a cluster. Previous work [3, 5] presented randomized algorithms for constructing a cluster based on  $\varepsilon$ -net sampling over the set of all candidate clusters. They shatter the set of candidate clusters and show that it has bounded VC dimension, which leads to their asymptotic approximation of  $k$  — the minimum size of an  $(\ell, \Delta)$ -clustering. The algorithm of Conradi and Driemel [13] is more similar to ours. They also simplify the input and iteratively select the cluster with the (exact) maximum coverage to obtain an  $(\ell, \Delta)$ -clustering of size  $\mathcal{O}(k \log n)$ . The key difference lies in finding the next cluster. Conradi and Driemel [13] explicitly consider a set of  $\mathcal{O}(n^3 \ell)$  candidate clusters, which requires  $\mathcal{O}(n^4 \ell)$  time and space to construct.

We make two key contributions that distinguish us from prior works: First, we present a novel simplification algorithm that computes a curve  $S$  such that we may restrict potential

reference curves of clusters to be subcurves of  $S$ . This new curve simplification technique allows us to create a clustering where clusters have radius at most  $4\Delta$  as opposed to  $11\Delta$ . Second, we prove that we may restrict the reference curves to be one of two types:

- vertex-subcurves of  $S$ , which are subcurves that start and end at a vertex of  $S$ ,  
(we may furthermore only consider subcurves whose complexity is a power of 2)
- and subedges of  $S$ , which are subcurves that are a subsegment of a single edge of  $S$ .

We prove that a greedy algorithm that exclusively adds maximal clusters where the reference curve is of one of these two types creates a clustering of size  $\mathcal{O}(k \log n)$ . This characterization reduces the set of candidate clusters from  $\tilde{\mathcal{O}}(n^3 \ell)$  to  $\tilde{\mathcal{O}}(n^2)$  which significantly reduces the time spent compared to [13]. The geometric characterization of these subcurves allow us to compute candidate clusters on the fly, significantly reducing space usage.

## 2 Preliminaries

A (*polygonal*) curve with  $n$  vertices is a piecewise-linear map  $P: [1, n] \rightarrow \mathbb{R}^d$  whose breakpoints (called *vertices*) are at each integer parameter, and whose pieces are called *edges*. We denote by  $P[a, b]$  the subcurve of  $P$  that starts at  $P(a)$  and ends at  $P(b)$ . If  $a$  and  $b$  are integers, we call  $P[a, b]$  a *vertex subcurve* of  $P$ . Let  $|P|$  denote the number of vertices of  $P$ .

**Fréchet distance.** A *reparameterization* of  $[1, n]$  is a non-decreasing surjection  $f: [0, 1] \rightarrow [1, n]$ . Two reparameterizations  $f$  and  $g$  of  $[1, m]$  and  $[1, n]$ , respectively, describe a *matching*  $(f, g)$  between two curves  $P$  and  $Q$  with  $n$  and  $m$  vertices, where for any  $t \in [0, 1]$ , point  $P(f(t))$  is matched to  $Q(g(t))$ . A matching  $(f, g)$  is said to have *cost*  $\max_t \|P(f(t)) - Q(g(t))\|$ , where  $\|\cdot\|$  denotes the Euclidean norm. A matching with cost at most  $\Delta$  is called a  $\Delta$ -*matching*. The (continuous) *Fréchet distance*  $d_F(P, Q)$  between  $P$  and  $Q$  is the minimum cost over all matchings.

**Free space diagram.** The *parameter space* of curves  $P$  and  $Q$  with  $m$  and  $n$  vertices, respectively, is given by the orthogonal rectangle  $[1, m] \times [1, n]$ . This parameter space is associated with a regular grid whose cells are the squares  $[i, i+1] \times [j, j+1]$  for integers  $i$  and  $j$ . A point  $(x, y)$  in the parameter space corresponds to the pair of points  $P(x)$  and  $Q(y)$ . We say that  $(x, y)$  is  $\Delta$ -*free* if  $\|P(x) - Q(y)\| \leq \Delta$ . The  $\Delta$ -*free space diagram*  $\Delta\text{-FSD}(P, Q)$  of  $P$  and  $Q$  is the set of  $\Delta$ -free points in the parameter space of  $P$  and  $Q$ . The *obstacles* of  $\Delta\text{-FSD}(P, Q)$  are the connected components of  $([1, m] \times [1, n]) \setminus \Delta\text{-FSD}(P, Q)$ .

Alt and Godau [4] observe that the Fréchet distance between  $P[x_1, x_2]$  and  $Q[y_1, y_2]$  is at most  $\Delta$  if and only if there is a bimonotone path in  $\Delta\text{-FSD}(P, Q)$  from  $(x_1, y_1)$  to  $(x_2, y_2)$  (and  $x_1 \leq x_2$  and  $y_1 \leq y_2$ ).

**Input and output.** Our input is a curve  $T$  with  $n$  vertices, which we will call the *trajectory*, some integer parameter  $\ell \geq 2$ , and some distance parameter  $\Delta \geq 0$ . We consider *clustering* subtrajectories of  $T$  using *pathlets*:

► **Definition 1 (Pathlet).** An  $(\ell, \Delta)$ -pathlet is a tuple  $(P, \mathcal{I})$  where  $P$  is a curve with  $|P| \leq \ell$  and  $\mathcal{I}$  is a set of intervals in  $[1, n]$ , where  $d_F(P, T[a, b]) \leq \Delta$  for all  $[a, b] \in \mathcal{I}$ . We call  $P$  the reference curve of  $(P, \mathcal{I})$ .

We can see a pathlet  $(P, \mathcal{I})$  as a cluster, where the center is  $P$  and all subtrajectories induced by  $\mathcal{I}$  get mapped to  $P$ . See Figure 1. An  $(\ell, \Delta)$ -clustering of  $T$  is defined as follows:

► **Definition 2.** An  $(\ell, \Delta)$ -clustering  $C$  is a set of  $(\ell, \Delta)$ -pathlets with  $\bigcup_{(P, \mathcal{I}) \in C} \bigcup_{I \in \mathcal{I}} I = [1, n]$ .

Throughout this paper, we let  $k_\ell(\Delta)$  denote the smallest integer for which there exists an  $(\ell, \Delta)$ -clustering of size  $k_\ell(\Delta)$ . The goal is to find an  $(\ell, \Delta')$ -clustering  $C$  where  $|C|$  is not too large compared to  $k_\ell(\Delta)$ , and  $\Delta' \in \mathcal{O}(\Delta)$ .

**Weighting a cluster.** We define a *universe*  $\mathcal{U}$  as any set of interior-disjoint closed intervals that together cover  $[1, n]$ . Given a fixed universe  $\mathcal{U}$ , we can weigh each pathlet by what we call its *coverage*:

► **Definition 3.** The coverage over  $\mathcal{U}$  of a pathlet  $(P, \mathcal{I})$  is  $\text{Cov}_{\mathcal{U}}(P, \mathcal{I}) = \{I \in \mathcal{U} \mid I \subseteq \text{Cov}(P, \mathcal{I})\}$ . The coverage of a set  $C$  of pathlets is  $\text{Cov}_{\mathcal{U}}(C) = \sum_{(P, \mathcal{I}) \in C} \text{Cov}_{\mathcal{U}}(P, \mathcal{I})$ .

Whenever  $\mathcal{U}$  is clear from context we omit the subscript  $\mathcal{U}$ .

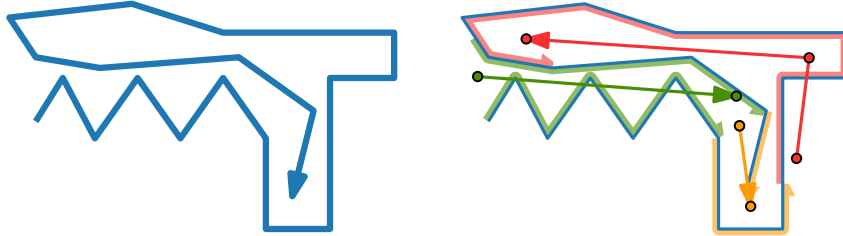
► **Definition 4** (Reference optimal). Let the universe  $\mathcal{U}$  be fixed and let  $C$  be a set of  $(\ell, \Delta)$ -pathlets. An  $(\ell, \Delta)$ -pathlet  $(P, \mathcal{I})$  is reference-optimal if its coverage over  $\mathcal{U} \setminus \text{Cov}(C)$ , i.e.,  $|\text{Cov}(P, \mathcal{I}) \setminus \text{Cov}(C)|$ , is maximum over all  $(\ell, \Delta)$ -pathlets with the same reference curve.

► **Definition 5.** Let the universe  $\mathcal{U}$  be fixed and let  $C$  be a set of  $(\ell, \Delta)$ -pathlets. An  $(\ell, \Delta)$ -pathlet  $(P, \mathcal{I})$  is optimal whenever  $|\text{Cov}(P, \mathcal{I}) \setminus \text{Cov}(C)|$  is maximum over all  $(\ell, \Delta)$ -pathlets.

### 3 Algorithmic outline

Our algorithmic input is a trajectory  $T$ , an integer  $\ell \geq 2$ , and value  $\Delta \geq 0$ . We provide a high-level overview of our algorithm here. Our approach can be decomposed as follows:

1. Reference curves may lie anywhere in the ambient space. Our first step is to restrict where these reference curves may lie. In Section 4 we construct a  $2\Delta$ -simplification  $S$  of  $T$ , and prove that for any  $(\ell, \Delta)$ -pathlet  $(P, \mathcal{I})$ , there exists a subcurve  $S[a, d]$  of  $S$  for which  $(S[a, d], \mathcal{I})$  is an  $(\ell + 2 - |\mathbb{N} \cap \{a, d\}|, \Delta')$ -pathlet, where  $\Delta' = 4\Delta$ . Hence we may restrict our attention to pathlets where the reference curve is a subcurve of  $S$ , if we allow for a slightly higher complexity. This higher complexity is circumvented later on, to still give an  $(\ell, \Delta')$ -clustering.
2. In Section 5, Given  $S$  and  $T$ , we smartly create some universe  $\mathcal{U}$ . We prove, by adapting the argument for greedy set cover, that any algorithm that iteratively computes an optimal  $(\ell, \Delta)$ -pathlet outputs a clustering of size  $\mathcal{O}(k_\ell(\Delta) \log n)$ .



■ **Figure 1** The trajectory  $T$  (blue, left) is covered by three pathlets. Each pathlet is defined by a reference curve (green, red, yellow) and the subcurve(s) of  $T$  the curve covers.

3. In Section 6 we give the general algorithm. We choose some  $\Delta' \in \Theta(\Delta)$ . We iteratively construct an  $(\ell, \Delta')$ -clustering of size  $\mathcal{O}(k_\ell(\Delta) \log n)$ . Our greedy iterative algorithm maintains a set  $C$  of pathlets and adds an  $(\ell, \Delta')$ -pathlet  $(P, \mathcal{I})$  to  $C$  at every iteration. Consider having a set of pathlets  $C = \{(P_i, \mathcal{I}_i)\}$ . We greedily select a pathlet  $(P, \mathcal{I})$  that covers as much of  $\mathcal{U} \setminus \text{Cov}(C)$  as possible, and add it to  $C$ . Formally, we select a  $(\Delta, \frac{1}{17})$ -maximal  $(\ell, \Delta')$ -pathlet: an  $(\ell, \Delta')$ -pathlet  $(P, \mathcal{I})$  such that

$$|\text{Cov}(P, \mathcal{I}) \setminus \text{Cov}(C)| \geq \frac{1}{17} |\text{Cov}(P', \mathcal{I}') \setminus \text{Cov}(C)|$$

for all  $(\ell, \Delta)$ -pathlets  $(P', \mathcal{I}')$ .

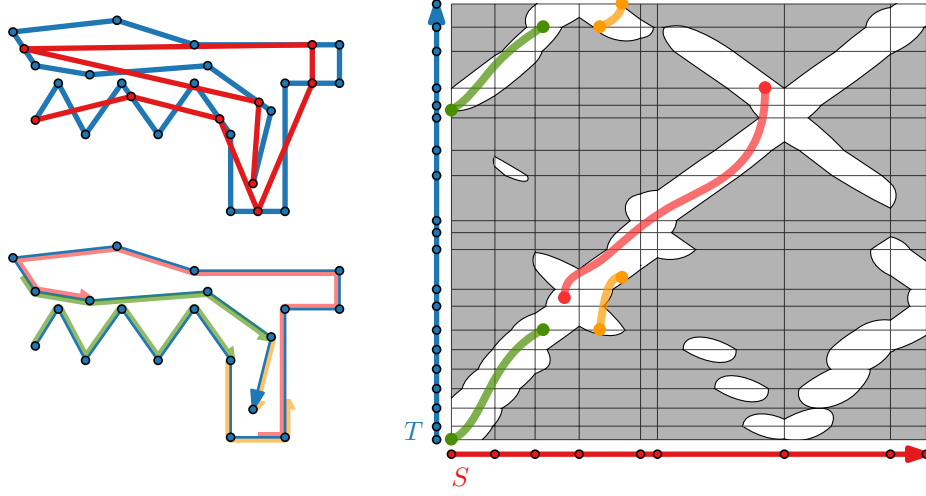
4. The subsequent goal is to compute  $(\Delta, \frac{1}{17})$ -maximal pathlets. We restrict pathlets to two types: those where the reference curve is 1) a vertex subcurve of  $S$ , or 2) a subsegment of an edge of  $S$ . Then we give algorithms for constructing pathlets of these types with a certain quality guarantee, i.e., pathlets that cover at least a constant fraction of what the optimal pathlet of that type covers. These algorithms are given in Sections 8 and 9.

**Reachability graph.** We introduce the *reachability graph* in Section 7. This graph is defined on a subcurve  $W$  of  $S$  and a set  $Z$  of points in  $\Delta'$ -FSD( $W, T$ ). The reachability graph  $G(W, T, Z)$  is a directed acyclic graph whose vertices are the set of points  $Z$ , together with certain boundary points of the free space  $\Delta'$ -FSD( $W, T$ ) and a collection of *steiner points*. Given two points  $(x, y)$  and  $(x', y')$  in  $Z$ , the graph contains a directed path from  $(x, y)$  to  $(x', y')$  if and only if  $d_F(W[x, x'], T[y, y']) \leq \Delta'$ .

We treat the free space diagram as a rectilinear polygon  $\mathcal{R}$  with rectilinear holes, obtained by reducing all obstacles of  $\Delta'$ -FSD( $W, T$ ) to their intersections with the parameter space grid. We show that a bimonotone path between two points  $p$  and  $q$  exists in  $\Delta'$ -FSD( $W, T$ ) if and only if a rectilinear shortest path between  $p$  and  $q$  in  $\mathcal{R}$  has length  $\|p - q\|_1$ , the  $L_1$ -distance between  $p$  and  $q$ . The reachability graph  $G(W, T, Z)$  is defined as the *shortest paths preserving graph* [20] for the set  $Z$  with respect to  $\mathcal{R}$ , made into a directed graph by directing edges, which are all horizontal or vertical, to the right or top. This graph has  $\mathcal{O}((|W|n + |Z|) \log(n|Z|))$  complexity, and a shortest path in the graph between points in  $Z$  is also a rectilinear shortest path between the corresponding points in  $\mathcal{R}$ .

**Vertex-to-vertex pathlets.** In Section 8 we construct a pathlet where the reference curve is a vertex subcurve of  $S$ . For a given vertex  $S(i)$  of  $S$ , we construct reference-optimal  $(\ell, \Delta')$ -pathlets  $(S[i, i + j], \mathcal{I}_j)$  for all  $j \in [\ell]$ . We first identify a set  $Z$  of  $\mathcal{O}(n\ell)$  *critical points* in  $\Delta'$ -FSD( $S[i, i + \ell], T$ ). We show that for every reference curve  $S[i, i + j]$ , there is a reference-optimal  $(\ell, \Delta')$ -pathlet  $(S[i, i + j], \mathcal{I}_j)$  where for each interval  $[y, y'] \in \mathcal{I}_j$ , the points  $(i, y)$  and  $(i + j, y')$  are critical points. We construct the intervals  $\mathcal{I}_j$  through a sweepline algorithm over the reachability graph  $G(S[i, i + \ell], T, Z)$ , which has  $\mathcal{O}(n\ell \log n)$  complexity. Our sweepline computes, for all  $j \in [\ell]$ , a reference-optimal  $(j, \Delta')$ -pathlet  $(S[i, i + j], \mathcal{I}_j)$  by iterating over all in-edges to critical points  $(i + j, y)$  in  $G(S[i, i + \ell], T, Z)$ . Doing this for all  $i$  (and remembering the optimum) thereby takes  $\mathcal{O}(n^2 \ell \log^2 n)$  time and  $\mathcal{O}(n\ell \log n)$  space.

**Subedge pathlets.** In Section 9 we construct a pathlet where the reference curve is a subsegment of an edge of  $S$ . For a given edge  $e$  of  $S$ , we again first identify a set  $Z$  of  $\mathcal{O}(n^2)$  *critical points* in  $\Delta'$ -FSD( $e, T$ ). However, rather than restricting the intervals in pathlets based on these critical points, we restrict the reference curves based on these critical points. Specifically, there are  $m = \mathcal{O}(n)$  unique  $x$ -coordinates of points in  $Z$ , which we order as



■ **Figure 2** Top left: A simplification  $S$  (red) of the trajectory  $T$  (blue). Right: The diagram  $\Delta'$ -FSD( $S, T$ ) in white. The obstacles of the diagram are colored in gray. The clustering (bottom left) corresponds to a set of colored bimonotone paths, where paths of a given color are horizontally aligned, and the paths together span the entire vertical axis.

$x_1, \dots, x_m$ . We show that by allowing for pathlets to use subsegments of the reversal  $\overleftarrow{e}$  of  $e$  as reference curves, we may restrict reference curves to be of the form  $e[x_i, x_{i'}]$  or  $\overleftarrow{e}[x_i, x_{i'}]$  to not lose much coverage. That is, the optimal  $(2, \Delta')$ -pathlet with such a reference curve covers at least one-fourth of what any other  $(2, \Delta')$ -pathlet using a subsegment of  $e$  as a reference curve covers.

The remainder of our subedge pathlet construction algorithm follows the same procedure as for vertex-to-vertex pathlets, though with the following optimization. We consider every  $x_i$  separately, for  $i \in [m]$ . However, rather than considering all reference curves  $e[x_i, x_{i'}]$ , of which there are  $m - i$ , we consider only  $\mathcal{O}(\log(m - i))$  reference curves. The main observation is that we may split a pathlet  $(e[x_i, x_{i'}], \mathcal{I})$  into two:  $(e[x_i, x_{i+2^j}], \mathcal{I}_1)$  and  $(e[x_{i+2^j}, x_{i'}], \mathcal{I}_2)$ , for some  $j \leq \log(m - i)$ . One of the two pathlets covers at least half of what  $(e[x_i, x_{i'}], \mathcal{I})$  covers, so an optimal  $(2, \Delta')$ -pathlet  $(e[x_i, x_{i+2^j}], \mathcal{I})$  that is defined by critical points covers at least one-eighth of any other subedge  $(2, \Delta')$ -pathlet  $(e[x, x'], \mathcal{I}')$ .

For every  $i \in [m]$ , we let  $Z_i \subseteq Z$  be the subset of critical points with  $x$ -coordinate equal to  $x_i$  or  $x_{i+2^j}$  for some  $j \leq \log(m - i)$ . We construct the reachability graph  $G(e, T, Z_i)$ , which has  $\mathcal{O}(n \log^2 n)$  complexity. We then proceed as with the vertex-to-vertex pathlets, using a sweepline through the reachability graph. Doing this for all  $i$  (and remembering the optimal pathlet) thereby takes  $\mathcal{O}(n^2 \log^3 n)$  total time and  $\mathcal{O}(n \log^2 n)$  space. Taken over all edges of  $S$ , we obtain a subedge pathlet in  $\mathcal{O}(n^3 \log^3 n)$  time and  $\mathcal{O}(n \log^2 n)$  space.

#### 4 Pathlet-preserving simplifications

We first aim to limit our attention to  $(\ell, 4\Delta)$ -pathlets  $(P, \mathcal{I})$  whose reference curves  $P$  are subcurves of some universal curve  $S$ . This way, we may design an algorithm that considers all subcurves of  $S$ , rather than all curves in  $\mathbb{R}^d$ . This has the additional benefit of allowing the use of the free space diagram  $4\Delta$ -FSD( $S, T$ ) to construct pathlets, as seen in Figure 2.

For any  $(\ell, \Delta)$ -pathlet  $(P, \mathcal{I})$  there exists an  $(n, 2\Delta)$ -pathlet  $(P', \mathcal{I})$  where  $P'$  is a subcurve of  $T$ . Indeed, consider any interval  $[a, b] \in \mathcal{I}$  and choose  $P' = T[a, b]$ . However, restricting



the subcurves of  $T$  to have complexity at most  $\ell$  may significantly reduce the maximum coverage, see for example Figure 3. Instead of restricting pathlets to be subcurves of  $T$ , we restrict them to be subcurves of a different curve  $S$ . We enforce the following property:

► **Definition 6.** For a trajectory  $T$  and value  $\Delta \geq 0$ , a *pathlet-preserving simplification* is a curve  $S$  together with a  $2\Delta$ -matching  $(f, g)$ , where for any subtrajectory  $T[a, b]$  of  $T$  and all curves  $P$  with  $d_F(P, T[a, b]) \leq \Delta$ , the subcurve  $S[s, t]$  matched to  $T[a, b]$  by  $(f, g)$  has complexity  $|S[s, t]| \leq |P| + 2 - |\mathbb{N} \cap \{s, t\}|$ .

► **Theorem 7.** Let  $(S, f, g)$  be a pathlet-preserving simplification of  $T$ . For any  $(\ell, \Delta)$ -pathlet  $(P, \mathcal{I})$ , there exists a subcurve  $S[s, t]$  such that  $(S[s, t], \mathcal{I})$  is an  $(\ell + 2 - |\mathbb{N} \cap \{s, t\}|, 4\Delta)$ -pathlet.

**Proof.** Consider any  $(\ell, \Delta)$ -pathlet  $(P, \mathcal{I})$  and choose some interval  $[a, b] \in \mathcal{I}$ . For all  $[c, d] \in \mathcal{I}$ , via the triangle inequality,  $d_F(T[a, b], T[c, d]) \leq 2\Delta$ . Let  $S[s, t]$  be the subcurve of  $S$  matched to  $T[a, b]$  by  $(f, g)$ . Naturally,  $d_F(S[s, t], T[a, b]) \leq 2\Delta$ , and so by the triangle inequality  $d_F(S[s, t], T[c, d]) \leq 4\Delta$ . By the definition of a pathlet-preserving simplification, we obtain that for every curve  $P'$  with  $d_F(P', T[a, b]) \leq \Delta$ , we have  $|P'| \geq |S[s, t]| - 2 + |\mathbb{N} \cap \{s, t\}|$ . In particular, setting  $P' \leftarrow P$  implies that  $|S[s, t]| \leq \ell + 2 - |\mathbb{N} \cap \{s, t\}|$ . Thus  $(S[s, t], \mathcal{I})$  is an  $(\ell + 2 - |\mathbb{N} \cap \{s, t\}|, 4\Delta)$ -pathlet. ◀

**Prior simplifications.** The curve  $S$  that we construct is a *curve-restricted  $\alpha\Delta$ -simplification* of  $T$ ; a curve whose vertices lie on  $T$ , where for every edge  $s = \overline{T(a)T(b)}$  of  $S$  we have  $d_F(s, T[a, b]) \leq \alpha\Delta$ . Various  $\alpha\Delta$ -simplification algorithms have been proposed [2, 14, 17, 19].

If  $T$  is a curve in  $\mathbb{R}^2$ , Guibas *et al.* [17] provide an  $\mathcal{O}(n \log n)$  time algorithm that constructs a  $2\Delta$ -simplification  $S$  for which there is no  $\Delta$ -simplification  $S'$  with  $|S'| < |S|$ . Their algorithm is not efficient in higher dimensions however.

Agarwal *et al.* [2] also construct a  $2\Delta$ -simplification  $S$  of  $T$  in  $\mathcal{O}(n \log n)$  time. This was applied by Akitaya *et al.* [3] for their subtrajectory clustering algorithm under the discrete Fréchet distance. The simplification  $S$  has a similar guarantee as the simplification of [17]: there exists no *vertex-restricted*  $\Delta$ -simplification  $S'$  with  $|S'| < |S|$ . This guarantee is weaker than that of [17], as vertex-restricted simplifications are simplifications formed by taking a subsequence of vertices of  $T$  as the vertices of the simplification. It can, however, be constructed efficiently in higher dimensions.

As we show in Figure 3, the complexity of a vertex-restricted  $\Delta$ -simplification can be arbitrarily bad compared to the (unrestricted)  $\Delta$ -simplification with minimum complexity. Brünig *et al.* [5] note that for the subtrajectory problem under the continuous Fréchet distance, one requires an  $\alpha\Delta$ -simplification whose complexity has guarantees with respect to the optimal (unrestricted) simplification. They present a  $3\Delta$ -simplification  $S$  (whose definition was inspired by de Berg, Gudmundsson and Cook [14]) with the following property:



■ **Figure 3** There exists a segment  $P$  where  $d_F(P, T[a, b]) \leq \Delta$ . In contrast, for any vertex-restricted  $S$  with  $d_F(T[a, b], S) \leq \Delta$ , the complexity of  $S$  is  $\Theta(|T[a, b]|)$ .

for any subcurve  $T[a, b]$  of  $T$  within Fréchet distance  $\Delta$  of some line segment, there exists a subcurve  $S[s, t]$  of  $S$  with complexity at most 4 that has Fréchet distance at most  $3\Delta$  to  $T[a, b]$ . Thus, there exists no  $\Delta$ -simplification  $S'$  with  $|S'| < |S|/2$ .

**Our new curve simplification.** In Definition 6 we presented yet another curve simplification under the Fréchet distance for curves in  $\mathbb{R}^d$ . Our simplification has a stronger property than the one that is realized by Brüning *et al.* [5]: for any subcurve  $T[a, b]$  and *any* curve  $P$  with  $d_F(P, T[a, b]) \leq \Delta$ , we require that there exists a subcurve  $S[s, t]$  with  $d_F(S[s, t], T[a, b]) \leq 2\Delta$  that has at most two more vertices than  $P$ . This implies both the property of Brüning *et al.* [5] and ensures that no  $\Delta$ -simplification  $S'$  exists with  $|S'| < |S| - 2$ .

In Appendix B, we provide an efficient algorithm for constructing pathlet-preserving simplifications. We relegate this section to the appendix to not distract from the main storyline. The algorithm is an extension of the vertex-restricted simplification of Agarwal *et al.* [2] to construct a curve-restricted simplification instead. For this, we use the techniques of Guibas *et al.* [17] to quickly identify if an edge of  $T$  is suitable to place a simplification vertex on. We combine this check with the algorithm of [2] and obtain:

► **Theorem 8.** *For any trajectory  $T$  with  $n$  vertices and any  $\Delta \geq 0$ , we can construct a pathlet-preserving simplification  $S$  in  $\mathcal{O}(n \log n)$  time.*

## 5 The universe $\mathcal{U}$ and greedy set cover

Subtrajectory clustering is closely related to the *set cover* problem. In this problem, we have a discrete universe  $\mathcal{U}$  and a family of sets  $\mathcal{S}$  in this universe, and the goal is to pick a minimum number of sets in  $\mathcal{S}$  such that their union is the whole universe. The decision variant of set cover is NP-complete [18]. However, the following greedy strategy gives an  $\mathcal{O}(\log |\mathcal{U}|)$  approximation of the minimal set cover size [12]. Suppose we have picked a set  $\hat{\mathcal{S}} \subseteq \mathcal{S}$  that does not yet cover all of  $\mathcal{U}$ . The idea is then to add a set  $S \in \mathcal{S}$  that maximizes  $|S \cap (\mathcal{U} \setminus \bigcup \hat{\mathcal{S}})|$ , and to repeat the procedure until  $\mathcal{U}$  is fully covered.

**Defining the universe  $\mathcal{U}$ .** We apply this greedy strategy to subtrajectory clustering, putting the focus on constructing a pathlet that covers the most of some universe  $\mathcal{U}$ . For subtrajectory clustering, the universe is, in principle, infinite. We therefore first define a discrete universe  $\mathcal{U}$  consisting of  $\mathcal{O}(n^3)$  intervals that together cover  $[1, n]$ . We choose this universe carefully, as an optimal covering of  $\mathcal{U}$  with pathlets must have roughly the same size as an optimal covering of  $[1, n]$ . We define  $\mathcal{U}$  using the following set of *critical points* in  $\Delta'$ -FSD( $S, T$ ):

► **Definition 9.** *For  $i \in [|S| - 1]$  and  $j \in [n - 1]$ , consider their corresponding cell (the area  $[i, i + 1] \times [j, j + 1]$ ) and the following six extreme points:*

- *A leftmost point of  $\Delta'$ -FSD( $S, T$ )  $\cap ([i, i + 1] \times [j, j + 1])$ ,*
- *A rightmost point of  $\Delta'$ -FSD( $S, T$ )  $\cap ([i, i + 1] \times [j, j + 1])$ ,*
- *The leftmost and rightmost points of  $\Delta'$ -FSD( $S, T$ )  $\cap ([i, i + 1] \times \{j\})$ , and*
- *The leftmost and rightmost points of  $\Delta'$ -FSD( $S, T$ )  $\cap ([i, i + 1] \times \{j + 1\})$ .*

*Let  $X_{i,j}$  be the set of corresponding  $x$ -coordinates and  $X := \bigcup_{i,j} X_{i,j}$ . For each  $x \in X$ , we call every point  $(x, y)$  that is an endpoint of a connected component (vertical segment) of  $\Delta'$ -FSD( $S, T$ )  $\cap (\{x\} \times [1, n])$  a critical point.*

► **Definition 10.** *Let  $Y^*$  be the set of critical points, sorted by their  $y$ -coordinate. We define the set  $\mathcal{U}$  as the set of intervals in  $[1, n]$  between two consecutive  $y$ -coordinates in  $X$ . Since*



there are at most  $6n$  critical points in  $[i, i+1] \times [j, j+1]$  for each  $i \in [|S| - 1]$  and  $j \in [n - 1]$ , it follows that  $|\mathcal{U}| \leq 6n^3 - 1 = \mathcal{O}(n^3)$ .

► **Lemma 11.** *We can construct  $\mathcal{U}$  in  $\mathcal{O}(n^3)$  time.*

**Proof.** Fix an integer  $i \in [|S| - 1]$ . We compute the critical points inside the cells  $[i, i+1] \times [j, j+1]$ , for all  $j \in [n - 1]$ , in  $\mathcal{O}(n^2)$  time altogether. For this, we compute the sets  $X_{i,j}$  of Definition 9 in  $\mathcal{O}(1)$  time each. Let  $X_i = \bigcup_j X_{i,j}$ . Then, we compute the intersections of each vertical line  $\{x\} \times [1, n]$ , for  $x \in X_i$ , in  $\mathcal{O}(n)$  time each. The critical points inside the cells  $[i, i+1] \times [j, j+1]$ , for  $j \in [n - 1]$ , are the endpoints of connected components of these intersections, and can be computed in  $\mathcal{O}(n)$  time per line, totalling  $\mathcal{O}(n^2)$  time. Summing over all integers  $i$  completes the proof. ◀

**Applying greedy set cover.** In the remainder of this paper, we let  $\mathcal{U}$  denote this discrete universe. We generalize the analysis of the greedy set cover argument to pathlets that cover a (constant) fraction of what the optimal pathlet covers. This relaxes the requirements on the pathlets and helps reduce complexity of the problem. For this, we introduce the following:

► **Definition 12 (Maximal pathlets).** *Given a set  $C$  of pathlets, a  $(\Delta, \frac{1}{c})$ -maximal  $(\ell, \Delta')$ -pathlet  $(P', \mathcal{I}')$  is a pathlet such that there exists no  $(\ell, \Delta)$ -pathlet  $(P, \mathcal{I})$  with*

$$\frac{1}{c} |\text{Cov}(P, \mathcal{I}) \setminus \text{Cov}(C)| \geq |\text{Cov}(P', \mathcal{I}') \setminus \text{Cov}(C)|.$$

In Lemma 13, we show that if we keep greedily selecting  $(\Delta, \frac{1}{c})$ -maximal pathlets for our clustering, the size of the clustering stays relatively small compared to the optimum size. The bound closely resembles the bound obtained by the argument for greedy set cover.

► **Lemma 13.** *Iteratively adding  $(\Delta, \frac{1}{c})$ -maximal pathlets yields a clustering of size at most  $3c \cdot k_\ell(\Delta) \ln(6n) + 1$ .*

**Proof.** Let  $C^* = \{(P_i, \mathcal{I}_i)\}_{i=1}^k$  be an  $(\ell, \Delta)$ -clustering of  $T$  of minimal size. Then  $k := |C^*| = k_\ell(\Delta)$ . Consider iteration  $j$  of the algorithm, where we have some set of  $(\ell, \Delta')$ -pathlets  $C_j$ . Denote by  $W_j = |\mathcal{U}| \setminus \text{Cov}(C_j)$  the “size” of the part of the universe that still needs to be covered. Since  $C^*$  covers  $\mathcal{U}$ , it must cover  $\mathcal{U} \setminus \text{Cov}(C_j)$ . It follows via the pigeonhole principle that there is at least one  $(\ell, \Delta)$ -pathlet  $(P_i, \mathcal{I}_i) \in C^*$  that covers at least  $W_j/k$  intervals in  $\text{Cov}(P_i, \mathcal{I}_i) \setminus \text{Cov}(C_j)$ . Per definition of being  $(\Delta, \frac{1}{c})$ -maximal, our greedy algorithm finds a pathlet  $(P_j, \mathcal{I}_j)$  that covers at least  $\frac{W_j}{ck}$  uncovered intervals. Thus:

$$W_{j+1} = |\mathcal{U}| - |\text{Cov}(C_j) \cup \text{Cov}(P_j, \mathcal{I}_j)| \leq W_j - \frac{W_j}{c \cdot k} = W_j \cdot \left(1 - \frac{1}{c \cdot k}\right).$$

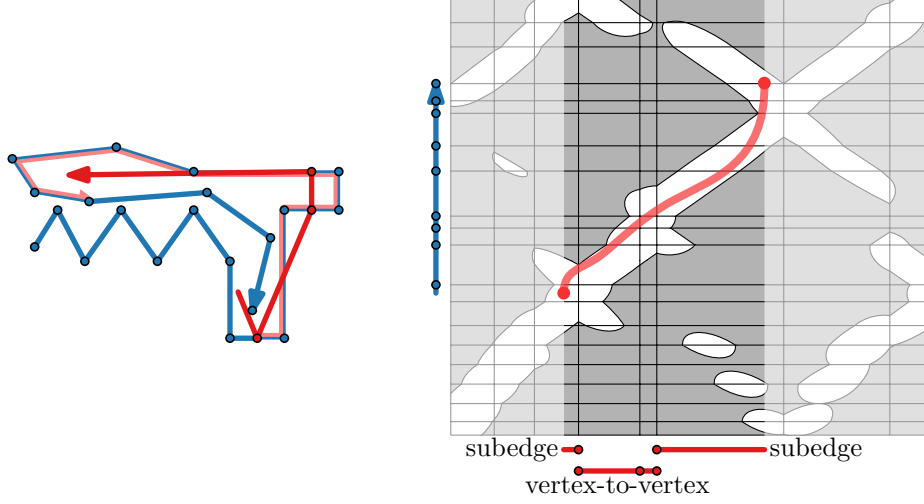
We have that  $W_0 = |\mathcal{U}|$ . Suppose it takes  $k' + 1$  iterations to cover all of  $T'$  with the greedy algorithm. Then before the last iteration, at least one edge of  $T'$  remained uncovered. That is,  $|\mathcal{U}| \cdot \left(1 - \frac{1}{c \cdot k}\right)^{k'} \geq 1$ . We apply that  $e^x \geq 1 + x$  for all real  $x$  to obtain:

$$\frac{1}{e} \geq \left(1 - \frac{1}{c \cdot k}\right)^x$$

for all  $x \geq 1$ . Plugging in  $x \leftarrow c \cdot k$ , it follows that

$$1 \leq |\mathcal{U}| \cdot \left(1 - \frac{1}{c \cdot k}\right)^{k'} = |\mathcal{U}| \cdot \left(1 - \frac{1}{c \cdot k}\right)^{c \cdot k \cdot \frac{k'}{c \cdot k}} \leq |\mathcal{U}| \cdot e^{-\frac{k'}{c \cdot k}}.$$

Hence  $e^{\frac{k'}{c \cdot k}} \leq |\mathcal{U}| - 1$ , showing that  $k' \leq c \cdot k \ln(|\mathcal{U}| - 1)$ . Thus after  $k' + 1 \leq c \cdot k_\ell(\Delta) \ln(|\mathcal{U}| - 1) + 1$  iterations, all of  $T'$ , and therefore  $T$ , is covered. Using that  $|\mathcal{U}| \leq 6n^3 - 1$  completes the proof. ◀



■ **Figure 4** A pathlet (left), corresponding to the red  $\Delta'$ -matching (right), gets split into a vertex-to-vertex and two subedge pathlets. The new pathlets correspond to the parts of the red matching that are vertically above the part of the  $x$ -axis corresponding to the new reference curve.

## 6 Subtrajectory clustering

In this section we present our algorithm for subtrajectory clustering. We first restrict our attention to reference curves of two types.

Recall that using the pathlet-preserving simplification  $S$  of  $T$ , we may already restrict our attention to reference curves that are subcurves of  $S$ . Still, the space of possible reference curves remains infinite. We wish to discretize this space by identifying certain finite classes of reference curves that contain a “good enough” reference curve, i.e., one with which we can construct a pathlet that is  $(\Delta, \frac{1}{c})$ -maximal for some small constant  $c$ .

We distinguish between two types of pathlets, based on their reference curves (note that not all pathlets fit into a class, and that some may fit into both classes):

1. Vertex-to-vertex pathlets: pathlets  $(P, \mathcal{I})$  where  $P$  is a vertex subcurve of  $S$ .
2. Subedge pathlets: pathlets  $(P, \mathcal{I})$  where  $P$  is a subsegment of an edge of  $S$ .

We construct pathlets of the above types, ensuring that they all cover at least some constant fraction of the optimal coverage for pathlets of the same type. Let  $(P_{\text{ver}}, \mathcal{I}_{\text{ver}})$  and  $(P_{\text{sub}}, \mathcal{I}_{\text{sub}})$  respectively be a vertex-to-vertex and subedge  $(\ell, \Delta')$ -pathlet, that respectively cover at least a factor  $\frac{1}{c_{\text{ver}}}$  and  $\frac{1}{c_{\text{sub}}}$  of an optimal pathlet of the same type. We show that one of these two pathlets is a  $(\Delta, \frac{1}{c})$ -maximal pathlet, for  $c = c_{\text{ver}} + 2c_{\text{sub}}$ . For intuition, refer to Figure 4.

► **Lemma 14.** *Given a collection  $C$  of pathlets, let*

$$(P, \mathcal{I}) \in \{(P_{\text{ver}}, \mathcal{I}_{\text{ver}}), (P_{\text{sub}}, \mathcal{I}_{\text{sub}})\}$$

*be a pathlet with maximal coverage among the uncovered points. Then  $(P, \mathcal{I})$  is  $(\Delta, \frac{1}{c})$ -maximal with respect to  $C$ , for  $c = c_{\text{ver}} + 2c_{\text{sub}}$ .*

**Proof.** Let  $(P^*, \mathcal{I}^*)$  be an optimal  $(\ell, \Delta)$ -pathlet. By Theorem 7, there exists a subcurve  $S[x, x']$  of  $S$  such that  $(S[x, x'], \mathcal{I}^*)$  is a  $(\ell + 2 - |\mathbb{N} \cap \{x, x'\}|, \Delta')$ -pathlet. Suppose first that  $S[x, x']$  is a subsegment of an edge of  $S$ , making  $(S[x, x'], \mathcal{I}^*)$  a subedge pathlet with  $|S[x, x']| \leq 2$ . In this case, the coverage of  $(P_{\text{sub}}, \mathcal{I}_{\text{sub}})$  is at least  $\frac{1}{c_{\text{sub}}}$  times the coverage of

$(S[x, x'], \mathcal{I}^*)$  over the uncovered points. Hence  $(P_{\text{sub}}, \mathcal{I}_{\text{sub}})$  is  $(\Delta, \frac{1}{c_{\text{sub}}})$ -maximal. Since the pathlet  $(P, \mathcal{I})$  has at least as much coverage as  $(P_{\text{sub}}, \mathcal{I}_{\text{sub}})$ , it must also be  $(\Delta, \frac{1}{c_{\text{sub}}})$ -maximal.

Next suppose that  $S[x, x']$  is not a subsegment of an edge of  $S$ , meaning  $S[x, x']$  contains at least one vertex of  $S$ . In this case, we split  $S[x, x']$  into three subcurves:

- A suffix  $P_{\text{suf}} = S[x, \lceil x \rceil]$  of an edge,
- A vertex subcurve  $P_{\text{ver}} = S[\lceil x \rceil, \lfloor x' \rfloor]$ , and
- A prefix  $P_{\text{pre}} = S[\lfloor x' \rfloor, x']$  of an edge.

Observe that every subcurve has at most  $\ell$  vertices. The suffix and prefix both trivially have at most  $2 \leq \ell$  vertices. The vertex subcurve has at most the number of vertices of  $S[x, x']$ , but if  $x$ , respectively  $x'$ , is not an integer, then the vertex subcurve loses a vertex compared to  $S[x, x']$ . That is, the vertex subcurve has at most

$$\ell + 2 - |\mathbb{N} \cap \{x, x'\}| - |\{x, x'\} \setminus \mathbb{N}| = \ell + 2 - |\{x, x'\}| = \ell \quad \text{vertices.}$$

Since every interval  $[y, y'] \in \mathcal{I}^*$  corresponds to a  $\Delta'$ -matching  $M$  between  $S[x, x']$  and  $T[y, y']$ , we can decompose  $[y, y']$  into three intervals  $[y, y_1]$ ,  $[y_1, y_2]$  and  $[y_2, y']$ , such that  $M$  decomposes into three  $\Delta'$ -matchings, one between  $P_{\text{suf}}$  and  $T[y, y_1]$ , one between  $P_{\text{ver}}$  and  $T[y_1, y_2]$ , and one between  $P_{\text{pre}}$  and  $T[y_2, y']$ . By decomposing all intervals in  $\mathcal{I}^*$  in this manner, we obtain that there are three  $(\ell, \Delta)$ -pathlets  $(P_{\text{suf}}, \mathcal{I}_{\text{suf}}^*)$ ,  $(P_{\text{ver}}, \mathcal{I}_{\text{ver}}^*)$  and  $(P_{\text{pre}}, \mathcal{I}_{\text{pre}}^*)$  that together have the same coverage as  $(P^*, \mathcal{I}^*)$ .

We have at least one of the following:

- $(P_{\text{suf}}, \mathcal{I}_{\text{suf}}^*)$  covers at least a factor  $\frac{c_{\text{sub}}}{c_{\text{ver}} + 2c_{\text{sub}}}$  of what  $(P^*, \mathcal{I}^*)$  covers, or
- $(P_{\text{ver}}, \mathcal{I}_{\text{ver}}^*)$  covers at least a factor  $\frac{c_{\text{ver}}}{c_{\text{ver}} + 2c_{\text{sub}}}$  of what  $(P^*, \mathcal{I}^*)$  covers, or
- $(P_{\text{pre}}, \mathcal{I}_{\text{pre}}^*)$  covers at least a factor  $\frac{c_{\text{sub}}}{c_{\text{ver}} + 2c_{\text{sub}}}$  of what  $(P^*, \mathcal{I}^*)$  covers.

Regardless of what statement holds, the pathlet  $(P, \mathcal{I})$  covers at least a factor  $\frac{1}{c_{\text{ver}} + 2c_{\text{sub}}}$  of what  $(P^*, \mathcal{I}^*)$  covers. Thus we have that  $(P, \mathcal{I})$  is  $(\Delta, \frac{1}{c_{\text{ver}} + 2c_{\text{sub}}})$ -maximal. ◀

Next we combine the previous ideas on simplification and greedy algorithms and present our algorithm for subtrajectory clustering. The algorithm uses subroutines for constructing the two types of pathlets described above, as well as a data structure for comparing their coverages to select the best pathlet for the clustering.

Our pathlet construction algorithms guarantee that  $c_{\text{ver}} = 1$  and  $c_{\text{sub}} = 8$ . By Lemma 14, the pathlet with the most coverage is therefore  $(\Delta, \frac{1}{c})$ -maximal with respect to the uncovered points, for  $c = 1 + 2 \cdot 8 = 17$ . By Lemma 13, the resulting  $(\ell, \Delta')$ -clustering has a size of at most  $17k_\ell(\Delta) \ln(|\mathcal{U}| - 1) + 1$ . Since our universe  $\mathcal{U}$  has size at most  $6n^3 - 1$ , the clustering has a size of at most  $51k_\ell(\Delta) \ln(6n) + 1$ .

**A data structure for comparing pathlets.** Recall that we fixed some discrete universe  $\mathcal{U}$  of  $\mathcal{O}(n^3)$  intervals, and that we denote  $\text{Cov}(P, \mathcal{I}) = \text{Cov}_{\mathcal{U}}(P, \mathcal{I})$ . In each iteration of our greedy algorithm, we select one of two pathlets whose coverage is the maximum over  $\mathcal{U} \setminus \text{Cov}(C)$ , given the current set of picked pathlets  $C$ . To compare the coverages of pathlets, we make use of binary search trees built on  $\mathcal{U}$  and  $\text{Cov}(C)$ :

► **Lemma 15.** *In  $\mathcal{O}(n^3 \log n)$  time, we can preprocess  $\mathcal{U}$  and  $\text{Cov}(C)$  into a data structure of  $\mathcal{O}(n^3)$  size, such that given a pathlet  $(P, \mathcal{I})$ , the value  $|\text{Cov}(P, \mathcal{I}) \setminus \text{Cov}(C)|$  can be computed in  $\mathcal{O}(|\mathcal{I}| \log n)$  time.*

**Proof.** We make use of a general data structure for storing a set  $\mathcal{I}$  of  $m$  interior-disjoint intervals, such that given a query interval  $I$ , the number of intervals in  $\mathcal{I}$  that are fully contained in  $I$  can be reported efficiently. For the data structure, we store the (multiset of)

endpoints of intervals in  $\mathcal{I}$  in a balanced binary search tree. The tree uses  $\mathcal{O}(m)$  space and is constructed in  $\mathcal{O}(m \log m)$  time.

We report the number of intervals in  $\mathcal{I}$  contained in a query interval  $I$  as follows. An interval  $[a, b] \in \mathcal{I}$  is contained in  $I$  if and only if both  $a$  and  $b$  are. Furthermore, there are  $k' \leq 2$  intervals in  $\mathcal{I}$  that  $I$  intersects but does not contain. Thus, if  $I$  contains  $k$  endpoints stored in the binary search tree, then it contains  $(k - k')/2$  intervals of  $\mathcal{I}$ . We compute  $k'$  by reporting the intervals of  $\mathcal{I}$  containing the endpoints of  $I$  in  $\mathcal{O}(\log m)$  time. Computing  $k$  and then reporting  $(k - k')/2$  takes an additional  $\mathcal{O}(\log m)$  time. Thus we answer a query in  $\mathcal{O}(\log m)$  time.

We use the above data structure to efficiently compute  $|\text{Cov}(P, \mathcal{I}) \setminus \text{Cov}(C)|$  for a query pathlet  $(P, \mathcal{I})$ . For this, we preprocess both  $\mathcal{U}$  and  $\text{Cov}(C)$  into the above data structure. Since  $\text{Cov}(C) \subseteq \mathcal{U}$  and  $|\mathcal{U}| = \mathcal{O}(n^3)$ , this takes  $\mathcal{O}(n^3 \log n)$  time, and the data structures use  $\mathcal{O}(n^3)$  space. With the two data structures, we report the values  $|\text{Cov}_{\mathcal{U}}(P, \mathcal{I}) \cap \mathcal{U}|$  and  $|\text{Cov}(P, \mathcal{I}) \cap \text{Cov}(C)|$  in  $\mathcal{O}(\log n)$  time. We then report

$$|\text{Cov}(P, \mathcal{I}) \setminus \text{Cov}(C)| = |\text{Cov}(P, \mathcal{I}) \cap \mathcal{U}| - |\text{Cov}(P, \mathcal{I}) \cap \text{Cov}(C)|. \quad \blacktriangleleft$$

**Asymptotic complexities.** Our algorithm iteratively constructs a set  $C$  of  $\mathcal{O}(k_\ell(\Delta) \log n)$  pathlets. Before we start constructing pathlets, we compute the universe  $\mathcal{U}$  of  $\mathcal{O}(n^3)$  intervals. This takes  $\mathcal{O}(n^3)$  time (Lemma 11).

In each iteration, we construct the data structure of Lemma 15 on the universe  $\mathcal{U}$  and current set of pathlets  $C$ . This takes  $\mathcal{O}(n^3 \log n)$  time and uses  $\mathcal{O}(n^3)$  space. Constructing the vertex-to-vertex pathlet then takes  $\mathcal{O}(n^2 \ell \log^2 n)$  time and uses  $\mathcal{O}(n \ell \log n)$  space (Theorem 20). The subedge pathlet takes  $\mathcal{O}(n^3 \log^3 n)$  time and  $\mathcal{O}(n \log^2 n)$  space to construct (Theorem 23).

To decide which pathlet to use in the clustering, we make further use of the data structure of Lemma 15. All constructed pathlets  $(P, \mathcal{I})$  have  $|\mathcal{I}| \leq n$ , and so we compute the coverages of the two pathlets in  $\mathcal{O}(n \log n)$  time. By summing up all complexities, we derive our main theorem:

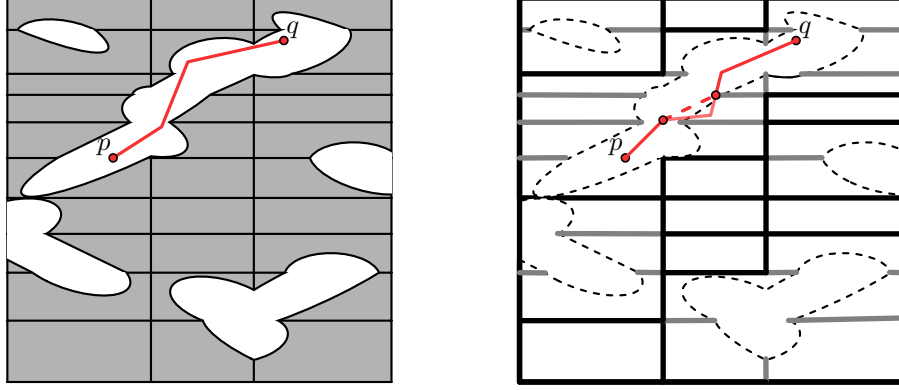
► **Theorem 16.** *Given a trajectory  $T$  with  $n$  vertices, an integer  $\ell \geq 2$ , and a value  $\Delta \geq 0$ , we can construct an  $(\ell, 4\Delta)$ -clustering of size at most  $51k_\ell(\Delta) \ln(6n) + 1$  in  $\mathcal{O}(k_\ell(\Delta) n^3 \log^4 n)$  time and using  $\mathcal{O}(n^3)$  space.*

## 7 The reachability graph

Let  $\Delta' = 4\Delta$ . For any subcurve  $W$  of  $S$  and a set of points  $Z$  in  $\Delta'$ -FSD( $W, T$ ) we define the *reachability graph*  $G(W, T, Z)$ . The vertices of this graph are the set of points  $Z$ , together with some Steiner points in  $[1, |W|] \times [1, |T|]$ . The reachability graph  $G(W, T, Z)$  is a directed graph where, for any two  $\mu_1, \mu_2 \in Z$ , there exists a directed path from  $\mu_1$  to  $\mu_2$  if and only if  $\mu_2$  is reachable from  $\mu_1$  in the free space  $\Delta'$ -FSD( $W, T$ ).

We define a reachability graph with  $\mathcal{O}((n|W| + |Z|) \log(n|Z|))$  vertices and edges, and can be constructed in  $\mathcal{O}((n|W| + |Z|) \log(n|Z|))$  time.

► **Theorem 17.** *Let  $W$  be a subcurve of  $S$  and  $Z$  a set of points in  $\Delta'$ -FSD( $W, T$ ). The reachability graph  $G(W, T, Z)$  has  $\mathcal{O}((n|W| + |Z|) \log(n|Z|))$  vertices and edges, and can be constructed in  $\mathcal{O}((n|W| + |Z|) \log(n|Z|))$  time.*



■ **Figure 5** (left) The  $\Delta'$ -free space diagram of  $W$  and  $T$  with points  $p$  and  $q$  connected by a bimonotone path. (right) The obstacles of  $\mathcal{R}$  are made up of all grid edges that are entirely contained in the obstacles of  $\Delta'$ -FSD( $W, T$ ) (shown in black) plus the gray segments. We may transform any bimonotone path between  $p$  and  $q$  into one that lies in  $\Delta'$ -FSD( $W, T$ ).

**Constructing the graph.** Lemma 18 shows that when focusing on reachability between points in  $\Delta'$ -FSD( $W, T$ ), we can simplify obstacles of the free space diagram to the parameter space grid, minus the free space. See Figure 5.

These simplified obstacles can be represented in  $\mathcal{O}(|W|n)$  time as a set of horizontal and vertical line segments (whose endpoints are not included, except possibly some that meet the boundary of  $[1, |W|] \times [1, |T|]$ ). The complement of these segments in the parameter space  $[1, |W|] \times [1, |T|]$  gives a rectilinear polygon with rectilinear holes  $\mathcal{R}$ .

► **Lemma 18.** *Let  $p$  and  $q$  be two points in  $\Delta'$ -FSD( $W, T$ ). There is a bimonotone path from  $p$  to  $q$  in  $\Delta'$ -FSD( $W, T$ ) if and only if there is a bimonotone path from  $p$  to  $q$  in  $\mathcal{R}$ .*

**Proof.** Since  $\Delta'$ -FSD( $W, T$ ) is completely contained in  $\mathcal{R}$ , any path in  $\Delta'$ -FSD( $W, T$ ) is also a path in  $\mathcal{R}$ . To transform a path from  $p$  to  $q$  in  $\mathcal{R}$  to a path in  $\Delta'$ -FSD( $W, T$ ), replace each maximal subpath  $\pi$  that lies inside a cell of  $\Delta'$ -FSD( $W, T$ ), with the segment connecting its endpoints. The obstacles of  $\mathcal{R}$  agree with the obstacles of  $\Delta'$ -FSD( $W, T$ ) on the boundary of cells, and thus if  $\pi$  starts or ends on the boundary of a cell, the respective endpoint lies in  $\Delta'$ -FSD( $W, T$ ). Additionally, because  $p$  and  $q$  lie in  $\Delta'$ -FSD( $W, T$ ), we have that  $\pi$  must always start and end at points in  $\Delta'$ -FSD( $W, T$ ). By convexity of the free space inside a cell, the line segment connecting the endpoints of  $\pi$  lies in  $\Delta'$ -FSD( $W, T$ ), and so does the resulting path. This replacement preserves bimonotonicity, completing the proof. ◀

To obtain  $G(W, T, Z)$  we first construct an undirected graph  $G(Z)$ . This graph is the *shortest paths preserving graph* by Widmayer [20]. The vertices of  $G(Z)$  are the points in  $Z$ , together with the vertices of  $\mathcal{R}$  and some Steiner points. By weighting each edge by its length, the graph perfectly encodes rectilinear distances between points in  $Z$ . That is, the rectilinear distance in  $\mathcal{R}$  between two points in  $Z$  is equal to their distance in  $G(Z)$ .

The number of vertices of  $\mathcal{R}$  is  $\mathcal{O}(n|W|)$ , giving the graph a complexity of  $\mathcal{O}((n|W| + |Z|) \log(n|Z|))$ . The graph can be constructed in  $\mathcal{O}((n|W| + |Z|) \log(n|Z|))$  time [20].

The edges of  $G(Z)$  are all horizontal or vertical line segments. We set  $G(W, T, Z)$  to be the graph  $G(Z)$ , but with each edge directed towards the right (if horizontal) or top (if vertical). Observe that  $G(W, T, Z)$  perfectly encodes reachability: for two points  $p = (x, y)$  and  $q = (x', y')$  in  $Z$ , if there is a bimonotone rectilinear path from  $p$  to  $q$  in  $\mathcal{R}$ , then any rectilinear shortest path from  $p$  to  $q$  must be bimonotone, and hence there must be a

(bimonotone) path between them in  $G(W, T, Z)$ . Conversely, any path in  $G(W, T, Z)$  is also a path in  $\mathcal{R}$ . Thus  $d_F(P[x, x'], T[y, y']) \leq \Delta'$  if and only if there is a (bimonotone) path from  $(x, y)$  to  $(x', y')$  in  $G(W, T, Z)$ .

► **Theorem 17.** *Let  $W$  be a subcurve of  $S$  and  $Z$  a set of points in  $\Delta'$ -FSD( $W, T$ ). The reachability graph  $G(W, T, Z)$  has  $\mathcal{O}((n|W| + |Z|) \log(n|Z|))$  vertices and edges, and can be constructed in  $\mathcal{O}((n|W| + |Z|) \log(n|Z|))$  time.*

## 8 Vertex-to-vertex pathlets

Let  $\Delta' = 4\Delta$ , and let  $C$  be a set of pathlets. Recall that we can compute  $|\text{Cov}(P, \mathcal{I}) \setminus \text{Cov}(C)|$ , for a given pathlet  $(P, \mathcal{I})$ , in  $\mathcal{O}(|\mathcal{I}| \log n)$  time (Lemma 15). We fix some integer  $i$ . We then give an algorithm for constructing a vertex-to-vertex  $(\ell, \Delta')$ -pathlet  $(P, \mathcal{I})$  where  $P$  starts at the  $i$ 'th vertex of  $S$ , and its coverage over  $\mathcal{U} \setminus \text{Cov}(C)$  is maximum.

We find for each subcurve  $S'$  of  $S$  of length at most  $\ell$  a reference-optimal  $(\ell, \Delta')$ -pathlet. To this end, we consider each vertex  $S(i)$  of  $S$  separately. We construct a set of reference-optimal pathlets  $(S[i, i+1], \mathcal{I}_1), \dots, (S[i, i+j], \mathcal{I}_j), \dots, (S[i, i+\ell], \mathcal{I}_\ell)$ . We let each interval  $\mathcal{I}_j$  contain all maximal intervals  $[y, y']$  for which  $d_F(S[i, i+j], T[y, y']) \leq \Delta'$ , and thus all maximal intervals for which  $(i, y)$  can reach  $(i+j, y')$  by a bimonotone path in  $\Delta'$ -FSD( $S, T$ ).

Recall that in Definition 9 we defined a set of *critical points*. Let  $Z$  denote all critical points in  $\Delta'$ -FSD( $S[i, i+\ell], T$ ) of the form  $(i+j, y)$ , for integers  $j \in [\ell]$ . That is,  $Z$  contains for all  $j \in [\ell]$  the endpoints of all connected components (vertical line segments) of  $\Delta'$ -FSD( $S, T$ )  $\cap (\{i+j\} \times [1, n])$ . Since each cell has at most  $\mathcal{O}(1)$  such critical points, it follows that  $|Z| \in \mathcal{O}(n\ell)$ . Observe that for any  $\Delta'$ -matching between  $T$  and a vertex-to-vertex subcurve, we can always extend each curve in the matching to start and end at a point in  $Z$ :

► **Observation 19.** *Let  $(P, \mathcal{I})$  be a vertex-to-vertex  $(\ell, \Delta')$ -pathlet where  $P$  starts at the  $i$ 'th vertex. Then there exists an  $(\ell, \Delta')$ -pathlet  $(P, \mathcal{I}')$  with  $\text{Cov}(P, \mathcal{I}) \subseteq \text{Cov}(P, \mathcal{I}')$  such that for each interval in  $\mathcal{I}'$ , the corresponding bimonotone path in  $\Delta'$ -FSD( $S, T$ ) starts and ends at a point in  $Z$ .*

We create a sweepline algorithm that, for each  $j \in [\ell]$ , constructs a reference-optimal  $(\ell, \Delta')$ -pathlet  $(S[i, j], \mathcal{I}_j)$ . We let each interval  $\mathcal{I}_j$  contain all maximal intervals  $[y, y']$  for which  $d_F(S[i, i+j], T[y, y']) \leq \Delta'$ , and thus all maximal intervals for which  $(i, y)$  can reach  $(i+j, y')$  by a bimonotone path in  $\Delta'$ -FSD( $S, T$ ). Note that both  $(i, y)$  and  $(i+j, y')$  are critical points. Thus we aim to find all maximal intervals  $[y, y']$  for which  $\Delta'$ -FSD( $S, T$ ) contains a bimonotone path between critical points  $(i, y)$  and  $(i+j, y')$ .

To this end, we construct, for each  $i \in [n]$ , the reachability graph  $G(S[i, i+\ell], T, Z)$  from Section 7, which encodes reachability between all critical points. This graph takes  $\mathcal{O}((n\ell + |Z|) \log(n|Z|)) = \mathcal{O}(n\ell \log n)$  time to construct and has complexity  $\mathcal{O}(n\ell \log n)$  (see Theorem 17). We aim to annotate each vertex  $\mu$  (which does not necessarily have to be a critical point) in  $G(S[i, i+\ell], T, Z)$  with the minimum  $y$ , such that there exists a critical point  $(i, y)$  that can reach  $\mu$ . We annotate  $\mu$  with  $\infty$  if no such value  $y$  exists.

**Annotating vertices.** We begin by annotating the vertices  $(i, y)$  in  $\mathcal{O}(n)$  time, by scanning over them in order of increasing  $y$ -coordinate. We go over the remaining vertices in  $yx$ -lexicographical order, where we go over the vertices based on increasing  $y$ -coordinate, and increasing  $x$ -coordinate when ties arise. Each vertex  $\mu$  that we examine has only incoming arcs originating from vertices below and left of  $\mu$ . By our lexicographical ordering, each of these vertices are already annotated. The minimal  $y$  for which there exists a path from  $(i, y)$



to  $\mu$ , must be the minimum over all its incoming arcs which we compute in time proportional to the in-degree of  $\mu$ . If  $\mu$  has no incoming arcs, we annotate it with  $\infty$ .

Let  $V$  and  $A$  be the sets of  $\mathcal{O}(n\ell \log n)$  vertices and arcs of  $G(S[i, i + \ell], T, Z)$ . For the above annotation procedure, we first compute the  $yx$ -lexicographical ordering of the vertices, based on their corresponding points in the parameter space. This takes  $\mathcal{O}(|V| \log |V|)$  time. Afterwards, we go over each vertex and each incoming arc exactly once, which take an additional  $\mathcal{O}(|V| + |A|)$  time. In total, we annotate all vertices in  $\mathcal{O}(n\ell \log^2 n)$  time.

**Constructing the pathlets.** With the annotations, constructing the pathlets becomes straightforward. For each  $j \in [\ell]$ , we construct  $\mathcal{I}_j$  as follows. We iterate over all critical point  $(i + j, y')$  in the graph  $G(S[i, i + \ell], T, Z)$ . For each critical point  $(i + j, y')$  with a finite annotation  $y$ , we add the interval  $[y, y']$  to  $\mathcal{I}_j$ . This procedure ensures that  $\mathcal{I}_j$  contains all maximal intervals  $[y, y']$  for which  $d_F(S[i, i + j], T[y, y']) \leq \Delta'$ , creating an optimal pathlet  $(S[i, i + j], \mathcal{I}_j)$  with respect to its reference curve. Since there are  $\mathcal{O}(n)$  critical points per  $j$ , this algorithm uses  $\mathcal{O}(n\ell)$  time. Storing the pathlets takes  $\mathcal{O}(n\ell)$  space. We conclude:

► **Theorem 20.** *Suppose  $\text{Cov}(C)$  is preprocessed by Lemma 15. In  $\mathcal{O}(n^2 \ell \log^2 n)$  time and using  $\mathcal{O}(n\ell \log n)$  space, we can construct an optimal vertex-to-vertex  $(\ell, \Delta')$ -pathlet  $(P, \mathcal{I})$ .*

**Proof.** For a given vertex  $S(i)$ , we compute optimal pathlets  $(S[i, i + j], \mathcal{I}_j)$  with respect to their reference curves for  $j \in [\ell]$  in  $\mathcal{O}(n\ell \log^2 n)$  time, using  $\mathcal{O}(n\ell \log n)$  space. Using the data structure of Lemma 15, we subsequently compute the coverage of one of these pathlets  $\mathcal{O}(n \log n)$  time, so  $\mathcal{O}(n\ell \log n)$  time for all. We pick the best pathlet and remember its coverage. Doing so for all vertices  $S(i)$  of  $S$ , we obtain  $|S|$  pathlets, of which we report the best. By only keeping the best pathlet in memory, rather than all  $|S|$ , the space used by these pathlets is lowered from  $\mathcal{O}(n^2)$  to  $\mathcal{O}(n)$ . ◀

## 9 Subedge pathlets

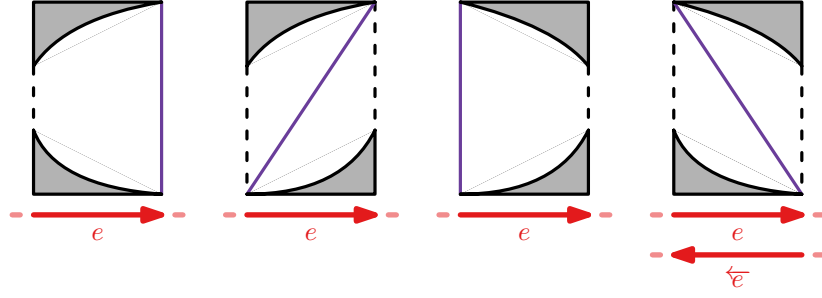
Let  $\Delta' = 4\Delta$ , and let  $C$  be a set of pathlets. We assume that  $\text{Cov}(C)$  has at most  $\mathcal{O}(n^2 \log n)$  connected components. We provide an algorithm for constructing a subedge  $(2, \Delta')$ -pathlet  $(P, \mathcal{I})$  whose coverage – (the sum of lengths in  $\text{Cov}(P, \mathcal{I}) \setminus \text{Cov}(C)$ ) – is at least one-eighth the optimum.

Recall that a subedge pathlet  $(P, \mathcal{I})$  is a pathlet where  $P = e[x, x']$  is a subsegment of some edge  $e$  of  $S$ . We construct a subedge pathlet given the edge  $e$ . We first discretize the problem, identifying a set of  $\mathcal{O}(n^2)$  critical points in  $\Delta'$ -FSD( $e, T$ ). This set ensures that there exists a subedge pathlet  $(e[x, x'], \mathcal{I})$  with at least one-fourth the coverage of any pathlet using a subedge of  $e$  as a reference curve, where for all  $[y, y'] \in \mathcal{I}$ , the points  $(x, y)$  and  $(x', y')$  are both critical points.

For  $j \in [n - 1]$ , consider the following six extreme points of  $\Delta'$ -FSD( $e, T$ )  $\cap ([1, 2] \times [j, j + 1])$  (where some points may not exist):

- A leftmost point of  $\Delta'$ -FSD( $e, T$ )  $\cap ([1, 2] \times [j, j + 1])$ ,
- A rightmost point of  $\Delta'$ -FSD( $e, T$ )  $\cap ([1, 2] \times [j, j + 1])$ ,
- The leftmost and rightmost points of  $\Delta'$ -FSD( $e, T$ )  $\cap ([1, 2] \times \{j\})$ , and
- The leftmost and rightmost points of  $\Delta'$ -FSD( $e, T$ )  $\cap ([1, 2] \times \{j + 1\})$ .

Let  $X_j$  be the set of  $x$ -coordinates of these points, and let  $X = \bigcup X_j$  be the set of all these coordinates. Let  $x_1, \dots, x_m$  be the set of values in  $X$ , sorted in increasing order. We call every point  $(x_i, y)$  that is an endpoint of a connected component (vertical segment) of  $\Delta'$ -FSD( $e, T$ )  $\cap (\{x_i\} \times [1, n])$  a critical point. Let  $Z$  be the set of at most  $4n^2 = \mathcal{O}(n^2)$  critical points.



■ **Figure 6** The connected components of  $\Delta'$ -FSD( $e', T$ ) fall into these four cases, based on where the minima and maxima of the bottom and top parabolic arcs lie. In the first three cases, there is a clear matching with optimal coverage (purple). In the fourth case, the matching is only valid when mirroring the free space, achieved by using  $\overleftarrow{e}$  instead of  $e$ .

Before we restrict pathlets to be defined by these critical points, we first allow a broader range of pathlets. We consider the edge  $\overleftarrow{e}$ , obtained by reversing the direction of  $e$ , and look at constructing a pathlet that is a subedge of either  $e$  or  $\overleftarrow{e}$ . We show that by restricting pathlets to be defined by  $Z$ , while allowing for reference curves that are subcurves of  $\overleftarrow{e}$ , results in losing only a factor four in the maximum coverage.

► **Lemma 21.** *Let  $C$  be a set of pathlets. For any subedge  $(2, \Delta')$ -pathlet  $(e[x, x'], \mathcal{I})$ , there exists a subedge  $(2, \Delta')$ -pathlet  $(P, \mathcal{I}')$  with*

$$|\text{Cov}(P, \mathcal{I}') \setminus \text{Cov}(C)| \geq \frac{1}{4} |\text{Cov}(e[x, x'], \mathcal{I}) \setminus \text{Cov}(C)|,$$

where  $P$  is equal to  $e[x_i, x_j]$  or  $\overleftarrow{e}[x_i, x_j]$  for some  $i$  and  $j$ , and for every interval  $[y, y'] \in \mathcal{I}'$ , the points  $(x_i, y)$  and  $(x_j, y')$  are contained in  $Z$ .

**Proof.** Consider a subedge  $(2, \Delta')$ -pathlet  $(e[x, x'], \mathcal{I})$ . Any interval  $[a, b] \in \mathcal{I}$  corresponds to a bimonotone path from  $(x, a)$  to  $(x', b)$  in  $\Delta'$ -FSD( $e, T$ ). Consider such an interval  $[a, b]$  and a corresponding path  $\pi$ .

Suppose first that  $x_i \leq x \leq x' \leq x_{i+1}$  for some  $i$ . Observe that every connected component of  $\Delta'$ -FSD( $e, T$ )  $\cap ([x_i, x_{i+1}] \times [1, n])$  is bounded on the left and right by (possibly empty) vertical line segments, and that the bottom and top chains are parabolic curves whose extrema are the endpoints of these segments. In particular, these connected components are convex. Thus there is a straight line segment  $e'$  from  $(x, a)$  to  $(x', b)$  in the free space. The line segment  $e^*$  connecting the extrema of the parabolic curves bounding the connected component containing  $(x, a)$  and  $(x', b)$  is longer than  $e'$ . The endpoints of  $e^*$  are both critical points, and  $e^*$  describes a valid matching between a subcurve of  $T$  and either a subcurve of  $e$ , or a subcurve of  $e'$ . See Figure 6. As there are four different reference curves we choose from, the resulting intervals are spread over four different pathlets. Therefore, one of the pathlets must have at least one-fourth the coverage of any subedge pathlet.

Next suppose that  $x_i \leq x \leq x_{i+1} < x'$  for some  $i$ . At some point,  $\pi$  reaches a point  $(x_{i+1}, a')$ . Let  $(x^*, y^*)$  be the lowest point in the connected component containing  $(x, a)$ . This point is a critical point. By convexity, the segment  $e^*$  from  $(x^*, y^*)$  to  $(x_{i+1}, a')$  lies in the free space. Because  $y^* \leq a$  by our choice of  $(x^*, y^*)$ , we may connect  $e^*$  to the suffix of  $\pi$  that starts at  $(x_{i+1}, a')$  to obtain a matching that starts at a critical point and that covers at least as much of  $T$  as the original matching. Applying a symmetric procedure to the end  $(x', b)$  of  $\pi$  yields a matching that starts and ends at critical points without losing coverage.

Again, since there are four different reference curves to choose from for the intervals in  $\mathcal{I}$ , the resulting intervals are spread over four different pathlets. One of the pathlets must therefore have at least one-fourth the coverage of any subedge pathlet.  $\blacktriangleleft$

We find for each point  $e(x_i)$  of  $e$  corresponding to a critical point a subedge pathlet whose reference curve starts at  $e(x_i)$  and ends at some point  $e(x_j)$  that also corresponds to a critical point. To this end, we consider each point  $e(x_i)$  separately. We proceed akin to the construction for vertex-to-vertex pathlets Section 8, with some optimization steps.

It proves too costly to consider each reference curve  $e[x_i, x_{i'}]$  for every  $x_i$  we consider. By sacrificing the quality of the pathlet slightly, settling for a pathlet with at least one-eighth the coverage of any subedge pathlet rather than one-fourth, we can reduce the number of reference curves we have to consider from  $\Theta(m^2) = \mathcal{O}(n^2)$  to  $\mathcal{O}(m \log m)$ . Let  $(e[x_i, x_{i'}], \mathcal{I})$  be a subedge pathlet. We can split  $e[x_i, x_{i'}]$  into two subedges  $e[x_i, x_{i+2j}]$  and  $e[x_{i'-2j}, x_{i'}]$ . The matchings corresponding to  $\mathcal{I}$  naturally decompose into two sets of matchings (whose matched subcurves may overlap), giving rise to two pathlets  $(e[x_i, x_{i+2j}], \mathcal{I}_1)$  and  $(e[x_{i'-2j}, x_{i'}], \mathcal{I}_2)$  with  $\mathcal{I}_1 \cup \mathcal{I}_2 = \mathcal{I}$ . Thus at least one of these pathlets has at least half the coverage that  $(e[x_i, x_{i'}], \mathcal{I})$  has. By Lemma 21, a pathlet  $(e[x_i, x_{i+2j}], \mathcal{I})$  that has maximum coverage out of all such pathlets then covers at least one-eighth of what any other subedge pathlet  $(e[x, x'], \mathcal{I}')$  covers.

We create a sweepline algorithm that, for each  $e[x_i, x_{i+2j}]$  (with  $j \leq \log(m-i)$ ), constructs a reference-optimal  $(\ell, \Delta')$ -pathlet  $(e[x_i, x_{i+2j}], \mathcal{I}_j)$ . We let each interval  $\mathcal{I}_j$  contain all maximal intervals  $[y, y']$  for which  $d_F(e[x_i, x_{i+2j}], T[y, y']) \leq \Delta'$ , and thus all maximal intervals for which  $(x_i, y)$  can reach  $(x_{i+2j}, y')$  by a bimonotone path in  $\Delta'$ -FSD( $S, T$ ). Note that both  $(x_i, y)$  and  $(x_{i+2j}, y')$  are critical points. Thus we aim to find all maximal intervals  $[y, y']$  for which the critical point  $(x_i, y)$  can reach the critical point  $(x_{i+2j}, y')$  by a bimonotone path in  $\Delta'$ -FSD( $S, T$ ).

Let  $Z_i$  be the subset of  $\mathcal{O}(n \log n)$  critical points with  $x$ -coordinate equal to  $x_i$  or  $x_{i+2j}$  for some  $j \leq \log(m-i)$ . We construct, for each  $i \in [n]$ , the reachability graph  $G(e, T, Z_i)$  from Section 7, which encodes reachability between the critical points in  $Z_i$ . This graph takes  $\mathcal{O}((n + |Z_i|) \log(n|Z_i|)) = \mathcal{O}(n \log^2 n)$  time to construct and has complexity  $\mathcal{O}(n \log^2 n)$  (see Theorem 17). We aim to annotate each vertex  $\mu$  (note that  $\mu$  does not have to be a critical point) in  $G(e, T, Z_i)$  with the minimum  $y$ , such that there exists a critical point  $(x_i, y)$  that can reach  $\mu$ . We annotate  $\mu$  with  $\infty$  if no such  $y$  exists.

**Annotating vertices and asymptotic analysis.** We first annotate the vertices  $(x_i, y)$  in  $\mathcal{O}(n)$  time by scanning over them in order of increasing  $y$ -coordinate. We process the remaining vertices in  $yx$ -lexicographical order, first by increasing  $y$ -coordinate, and by increasing  $x$ -coordinate when ties arise. Each vertex  $\mu$  that we consider has only incoming arcs that originate from vertices below and left of  $\mu$ . By our lexicographical ordering, each of these vertices are already annotated. The minimal  $y$  for which there exists a path from  $(x_i, y)$  to  $\mu$ , must be the minimum over all its incoming arcs which we compute in time proportional to the in-degree of  $\mu$ . If  $\mu$  has no incoming arcs, we annotate it with  $\infty$ .

Let  $V$  and  $A$  be the sets of  $\mathcal{O}(n \log^2 n)$  vertices and arcs of  $G(e, T, Z_i)$ . For the above annotation procedure, we first compute the  $yx$ -lexicographical ordering of the vertices, based on their corresponding points in the parameter space. This takes  $\mathcal{O}(|V| \log |V|)$  time. Afterwards, we go over each vertex and each incoming arc exactly once, which take an additional  $\mathcal{O}(|V| + |A|)$  time. In total, we annotate all vertices in  $\mathcal{O}(n \log^3 n)$  time.

**Constructing the pathlets.** Using the annotations, constructing the pathlets becomes straightforward. For each  $j \in [\log(m - i)]$ , we construct  $\mathcal{I}_j$  as follows. We iterate over all critical point  $(x_{i+2^j}, y')$  in the graph  $G(e, T, Z_i)$ . For each critical point  $(x_{i+2^j}, y')$  with a finite annotation  $y$ , we add the interval  $[y, y']$  to  $\mathcal{I}_j$ . This procedure ensures that  $\mathcal{I}_j$  contains all maximal intervals  $[y, y']$  for which  $d_F(e[x_i, x_{i+2^j}], T[y, y']) \leq \Delta'$ , making an optimal pathlet  $(e[x_i, x_{i+2^j}], \mathcal{I}_j)$  with respect to its reference curve. As there are  $\mathcal{O}(n)$  critical points per  $j$ , this algorithm uses  $\mathcal{O}(n \log n)$  time. Storing the pathlets takes  $\mathcal{O}(n \log n)$  space. Thus, we conclude the following:

► **Lemma 22.** *Let  $C$  be a set of pathlets where  $\text{Cov}(C)$  has  $\mathcal{O}(n^2 \log n)$  connected components. Suppose  $\text{Cov}(C)$  is preprocessed into the data structure of Lemma 15. In  $\mathcal{O}(n^2 \log^3 n)$  time and using  $\mathcal{O}(n \log^2 n)$  space, we can construct a  $(2, \Delta')$ -pathlet  $(P, \mathcal{I})$  with*

$$\|\text{Cov}(P, \mathcal{I}) \setminus \text{Cov}(C)\| \geq \frac{1}{8} \|\text{Cov}(P', \mathcal{I}') \setminus \text{Cov}(C)\|$$

for any  $(2, \Delta')$ -pathlet  $(P', \mathcal{I}')$  where  $P'$  is a subsegment of a given directed line segment  $e$ . The intervals in  $\mathcal{I}$  all have endpoints that come from a set of at most  $4n^2$  values.

**Proof.** For a given point  $e(x_i)$ , we compute optimal pathlets  $(e[x_i, x_{i+2^j}], \mathcal{I}_j)$  with respect to their reference curves for  $j \in [\log(m - i)]$  in  $\mathcal{O}(n \log^3 n)$  time, using  $\mathcal{O}(n \log n)$  space. Using the data structure of Lemma 15, we subsequently compute the coverage of one of these pathlets  $\mathcal{O}(n \log n)$  time, so  $\mathcal{O}(n \log^2 n)$  time for all. We pick the best pathlet and remember its coverage. Doing so for all points  $e(x_i)$ , we obtain  $m$  pathlets, of which we report the best. This pathlet has at least one-eighth the coverage of any other subedge  $(2, \Delta')$ -pathlet  $(e[x, x'], \mathcal{I})$ . By only keeping the best pathlet in memory, rather than all  $m$ , the space used by these pathlets is lowered from  $\mathcal{O}(mn)$  to  $\mathcal{O}(n)$ . ◀

► **Theorem 23.** *Suppose that the universe  $\mathcal{U}$  and the coverage  $\text{Cov}(C)$  is preprocessed into the data structure of Lemma 15. In  $\mathcal{O}(n^3 \log^3 n)$  time and using  $\mathcal{O}(n \log^2 n)$  space, we can construct a  $(2, \Delta')$ -pathlet  $(P, \mathcal{I})$  with*

$$|\text{Cov}(P, \mathcal{I}) \setminus \text{Cov}(C)| \geq \frac{1}{8} |\text{Cov}(P', \mathcal{I}') \setminus \text{Cov}(C)|$$

for any subedge  $(2, \Delta')$ -pathlet  $(P', \mathcal{I}')$ .

## 10 Conclusion

In this work, we presented an improved approximation algorithm for subtrajectory clustering. We discuss our technical contribution, and how it differs from previous works, our asymptotic improvements and finally interesting directions for future work.

**Technical contribution.** Our technical contributions are threefold:

First, we introduced a new type of curve simplification in Section 4. This simplification allows us to construct a curve  $S$ , such that our clustering needs to consider only pathlets whose reference curve is a subcurve of  $S$ . Although numerous similar curve-simplification algorithms exist, our method distinguishes itself by lying significantly closer to the input curve  $T$ . Consequently, our approximation algorithm is a 4-approximation in  $\Delta$ , compared to the 11-approximations of prior works. We consider this simplification to be of independent interest, as future works may immediately use our simplification method to obtain 4-approximations in  $\Delta$  also.

Secondly, we considered in Section 6 an extension to the greedy set cover algorithm wherein each iteration adds an approximately-maximum covering element, rather than a maximum one. Observe that  $P$  can always be divided into at most three subcurves, where at most one of them starts and ends at a vertex of  $S$  (a vertex-subcurve) and at most two of them are subcurves of an edge of  $S$  (a subedge of  $S$ ). We design a greedy meta-algorithm, that in each iteration computes an  $(\ell, \Delta)$ -pathlet  $(P, \mathcal{I})$  with approximately-maximum coverage, whose reference curve is a vertex-subcurve or subedge of  $S$ . Our approximately greedy set cover analysis shows that our meta-algorithm computes a clustering of size  $\mathcal{O}(k \log n)$ . A key takeaway from our construction is that by restricting our attention to vertex-subcurves and subedges of  $S$ , we significantly reduce the set of candidate pathlets from  $\tilde{\mathcal{O}}(n^3 \ell)$  to  $\tilde{\mathcal{O}}(n^2)$ . We consider this fact to also be of independent interest. Indeed, our subsequent algorithm spends near-linear time per candidate pathlet but future works may discover more efficient algorithms over the same smaller candidate set.

Finally, we presented algorithms in Sections 8 and 9 that compute the corresponding candidate pathlet for a candidate reference curve in near-linear time and near-linear space. The key observation to this contribution is that we show that it suffices to compute all candidate pathlets on the fly, significantly reducing the space.

**Asymptotic improvements.** Compared to the best prior deterministic work [13], our algorithm improves the running time by a factor near-linear in  $n\ell$ , improves the space by a factor near-linear in  $n^2\ell$ , and improves the approximation in  $\Delta$  from a factor 11 to 4, all whilst asymptotically matching the size of the clustering. We consider this a significant improvement over the state-of-the-art.

When we compare to previous randomized work [3, 5] we improve the running time by a factor near-linear in  $\ell$ , improve space by a factor  $n$ , and improve the approximation in  $\Delta$  from a factor 11 to 4. A downside of our approach is that, compared to randomised works, we only asymptotically match the clustering size whenever  $\ell$  is relatively large (i.e.,  $\ell \in \Omega(\log n / \log k)$ ). However, we note that on all other algorithmic quality measures, we still offer a substantial improvement whilst also being deterministic. In addition, when considering algorithmic performance in practice, we note that these previous randomized results [3, 5] use  $\varepsilon$ -net sampling. Such a sampling procedure leads to very high hidden constants in the asymptotic clustering size which makes such an approach impractical.

**Future work.** We think it remains an interesting open problem whether one can obtain a clustering size of  $\mathcal{O}(k\ell \log k)$  in a deterministic manner. We also note that, currently, our algorithm considers a set of  $\tilde{\mathcal{O}}(n^2)$  reference curves  $P$ , and computes an  $(\ell, \Delta)$ -pathlet  $(P, \mathcal{I})$  with approximately-maximum coverage for each reference curve independently, in near-linear time. We think it is an interesting open problem whether one can present an algorithm that is overall more efficient whenever these maximum pathlets are considered simultaneously rather than independently.

---

## References

- 1 Pankaj K. Agarwal, Kyle Fox, Kamesh Munagala, Abhinandan Nath, Jiangwei Pan, and Erin Taylor. Subtrajectory clustering: Models and algorithms. In *proc. 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS)*, pages 75–87, 2018. doi:10.1145/3196959.3196972.

- 2 Pankaj K. Agarwal, Sarel Har-Peled, Nabil H. Mustafa, and Yusu Wang. Near-linear time approximation algorithms for curve simplification. *Algorithmica*, 42(3):203–219, 2005. doi:10.1007/s00453-005-1165-y.
- 3 Hugo A. Akitaya, Frederik Brünig, Erin Chambers, and Anne Driemel. Subtrajectory clustering: Finding set covers for set systems of subcurves. *Computing in Geometry and Topology*, 2(1):1:1–1:48, 2023. doi:10.57717/cgt.v2i1.7.
- 4 Helmut Alt and Michael Godau. Computing the Fréchet distance between two polygonal curves. *International Journal of Computational Geometry & Applications*, 5:75–91, 1995. doi:10.1142/S0218195995000064.
- 5 Frederik Brünig, Jacobus Conradi, and Anne Driemel. Faster approximate covering of subcurves under the Fréchet distance. In *proc. 30th Annual European Symposium on Algorithms (ESA)*, pages 28:1–28:16, Dagstuhl, Germany, 2022. doi:10.4230/LIPIcs.ESA.2022.28.
- 6 Kevin Buchin, Maike Buchin, David Duran, Brittany Terese Fasy, Roel Jacobs, Vera Sacristan, Rodrigo I. Silveira, Frank Staals, and Carola Wenk. Clustering trajectories for map construction. In *proc. 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 1–10, 2017. doi:10.1145/3139958.3139964.
- 7 Kevin Buchin, Maike Buchin, Joachim Gudmundsson, Jorren Hendriks, Erfan Hosseini Sereshgi, Vera Sacristán, Rodrigo I. Silveira, Jorrick Sleijster, Frank Staals, and Carola Wenk. Improved map construction using subtrajectory clustering. In *proc. 4th ACM SIGSPATIAL Workshop on Location-Based Recommendations, Geosocial Networks, and Geoadvertising*, pages 1–4, 2020. doi:10.1145/3423334.3431451.
- 8 Kevin Buchin, Maike Buchin, Joachim Gudmundsson, Maarten Löffler, and Jun Luo. Detecting commuting patterns by clustering subtrajectories. *International Journal of Computational Geometry & Applications*, 21(03):253–282, 2011. doi:10.1142/S0218195911003652.
- 9 Kevin Buchin, Anne Driemel, Joachim Gudmundsson, Michael Horton, Irina Kostitsyna, Maarten Löffler, and Martijn Struijs. Approximating  $(k, \ell)$ -center clustering for curves. In *proc. Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2922–2938, 2019. doi:10.1137/1.9781611975482.181.
- 10 Maike Buchin and Dennis Rohde. Coresets for  $(k, \ell)$ -median clustering under the Fréchet distance. In *proc. 8th International Conference on Algorithms and Discrete Applied Mathematics (CALDAM)*, pages 167–180, 2022. doi:10.1007/978-3-030-95018-7\_14.
- 11 Siu-Wing Cheng and Haoqiang Huang. Curve simplification and clustering under Fréchet distance. In *proc. 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1414–1432, 2023. doi:10.1137/1.9781611977554.ch51.
- 12 Vasek Chvátal. A greedy heuristic for the set-covering problem. *Mathematics of Operations Research*, 4(3):233–235, 1979. doi:10.1287/MOOR.4.3.233.
- 13 Jacobus Conradi and Anne Driemel. Finding complex patterns in trajectory data via geometric set cover. *arXiv preprint arXiv:2308.14865*, 2023.
- 14 Mark de Berg, Atlas F. Cook, and Joachim Gudmundsson. Fast Fréchet queries. *Computational Geometry*, 46(6):747–755, 2013. doi:10.1016/j.comgeo.2012.11.006.
- 15 Anne Driemel, Amer Krivošija, and Christian Sohler. Clustering time series under the Fréchet distance. In *proc. twenty-seventh annual ACM-SIAM symposium on Discrete algorithms (SODA)*, pages 766–785, 2016. doi:10.1137/1.9781611974331.ch5.
- 16 Joachim Gudmundsson and Sampson Wong. Cubic upper and lower bounds for subtrajectory clustering under the continuous Fréchet distance. In *proc. 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 173–189, 2022. doi:10.1137/1.9781611977073.9.
- 17 Leonidas J. Guibas, John Hersherberger, Joseph S. B. Mitchell, and Jack Snoeyink. Approximating polygons and subdivisions with minimum link paths. *International Journal of Computational Geometry & Applications*, 3(4):383–415, 1993. doi:10.1142/S0218195993000257.
- 18 Richard M. Karp. Reducibility among combinatorial problems. In *proc. Symposium on the Complexity of Computer Computations*, The IBM Research Symposia Series, pages 85–103, 1972. doi:10.1007/978-1-4684-2001-2\_9.



- 19 Mees van de Kerkhof, Irina Kostitsyna, Maarten Löffler, Majid Mirzanezhad, and Carola Wenk. Global curve simplification. *European Symposium on Algorithms (ESA)*, 2019.
- 20 Peter Widmayer. On graphs preserving rectilinear shortest paths in the presence of obstacles. *Annals of Operations Research*, 33(7):557–575, 1991. doi:10.1007/BF02067242.

## A The interior-disjoint setting

Previous definitions of subtrajectory clustering have imposed various restrictions on the pathlets in the clustering. For example, in [6,7,8,16] the pathlets must be *interior-disjoint*. A pathlet  $(P, \mathcal{I})$  is interior-disjoint whenever the intervals in  $\mathcal{I}$  are pairwise interior-disjoint. While we do not give dedicated algorithms for the interior-disjoint setting, we show in Lemma 25 that we can efficiently convert any pathlet into two interior-disjoint pathlets with the same coverage. This gives a post-processing algorithm for converting a clustering  $C$  into an interior-disjoint clustering  $C'$  with at most twice the number of pathlets. We first show the following auxiliary lemma.

► **Lemma 24.** *Given a set of intervals  $\mathcal{I}$ , we can compute a subset  $\mathcal{I}' \subseteq \mathcal{I}$  with ply<sup>1</sup> at most two and with  $\bigcup \mathcal{I}' = \bigcup \mathcal{I}$  in  $\mathcal{O}(|\mathcal{I}| \log |\mathcal{I}|)$  time.*

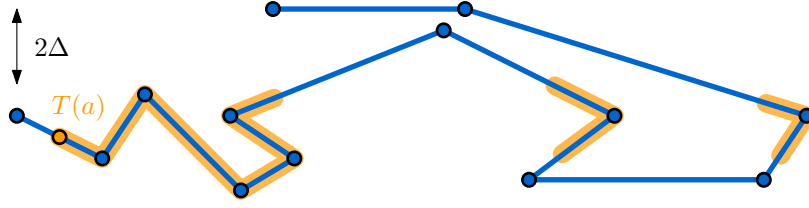
**Proof.** We first sort the intervals of  $\mathcal{I}$  based on increasing lower bound. We then remove all intervals in  $\mathcal{I}$  that are contained in some other interval in  $\mathcal{I}$ , which can be done in a single scan over  $\mathcal{I}$  by keeping track of the largest endpoint of an interval encountered so far. We initially set  $\hat{\mathcal{I}} = \emptyset$  and iterate over the remaining intervals in order of increasing lower bound. During iteration, we maintain the invariant that  $\hat{\mathcal{I}}$  has ply at most two. Let  $I_1, \dots, I_k$  be the intervals in  $\hat{\mathcal{I}}$  in order of increasing lower bound. Suppose we consider adding an interval  $I \in \mathcal{I}$  to  $\hat{\mathcal{I}}$ . If  $I \subseteq \hat{\mathcal{I}}$ , then we ignore  $I$ , since it does not add anything to the coverage of  $(P, \hat{\mathcal{I}})$ . Otherwise, we set  $\hat{\mathcal{I}} \leftarrow \hat{\mathcal{I}} \cup \{I\}$ . This may have increased the ply of  $\hat{\mathcal{I}}$  to three, however. We next show that in this case, we can remove an interval from  $\hat{\mathcal{I}}$  to decrease the ply back to two, without altering  $\bigcup \hat{\mathcal{I}}$ .

Observe that if the ply of  $\hat{\mathcal{I}}$  increases to three, then  $I_{k-1}$ ,  $I_k$  and  $I$  must intersect. Indeed,  $I$  must have a common intersection with two other intervals in  $\hat{\mathcal{I}}$ . Suppose for sake of contradiction that there is some  $I_i \in \hat{\mathcal{I}}$  that intersects  $I_i$  for some  $i < k - 1$ . Then  $I_i$  must contain the lower bounds of  $I_{k-1}$  and  $I_k$ . However,  $I_{k-1}$  must then also contain the lower bound of  $I_k$ , as otherwise  $I_{k-1} \subset I_i$ , which means that  $I_{k-1}$  was already filtered out at the beginning of the algorithm. Thus,  $I_i$ ,  $I_{k-1}$  and  $I_k$  have a common intersection (the lower bound of  $I_k$ ), which contradicts our invariant that  $\hat{\mathcal{I}}$  has ply at most two. Now that we know that  $I_{k-1}$ ,  $I_k$  and  $I$  intersect, note that  $I_k \subseteq I_{k-1} \cup I$ , since the lower bound of  $I_k$  lies between those of  $I_{k-1}$  and  $I$ , and  $I \not\subseteq I_k$ , so the upper bound of  $I_k$  lies between those of  $I_{k-1}$  and  $I$  as well. Hence we can set  $\hat{\mathcal{I}} \leftarrow \hat{\mathcal{I}} \setminus \{I_k\}$  to reduce the ply back to two, while keeping  $\bigcup \hat{\mathcal{I}}$  the same. After sorting  $\mathcal{I}$ , the above algorithm constructs  $\hat{\mathcal{I}}$  in  $\mathcal{O}(|\mathcal{I}|)$  time. This gives a total running time of  $\mathcal{O}(|\mathcal{I}| \log |\mathcal{I}|)$ . ◀

► **Lemma 25.** *Given an  $(\ell, \Delta)$ -pathlet  $(P, \mathcal{I})$ , we can construct two interior-disjoint  $(\ell, \Delta)$ -pathlets  $(P_1, \mathcal{I}_1)$  and  $(P_2, \mathcal{I}_2)$  with  $\mathcal{I}_1 \cup \mathcal{I}_2 = \mathcal{I}$  in  $\mathcal{O}(|\mathcal{I}| \log |\mathcal{I}|)$  time.*

**Proof.** First construct a subset  $\mathcal{I}' \subseteq \mathcal{I}$  with ply at most two and  $\bigcup \mathcal{I}' = \bigcup \mathcal{I}$  using Lemma 24. Then sort  $\mathcal{I}'$  based on increasing lower bound. Construct  $\mathcal{I}_1$  by iterating over  $\mathcal{I}'$  and greedily taking any interval that is interior-disjoint from the already picked intervals. Finally, set  $\mathcal{I}_2 \leftarrow \mathcal{I}' \setminus \mathcal{I}_1$ . ◀

<sup>1</sup> The ply of a set of intervals is the maximum number of intervals with a common intersection.



■ **Figure 7** Consider the trajectory  $T$  and some point  $T(a)$ . For some  $\Delta$ , we can indicate all  $T(b) \in T$  with  $a \leq b$  where for line  $s = \overline{T(a)T(b)}$ ,  $d_F(s, T[a, b]) \leq 2\Delta$ . Note that this set  $\mathbb{B}(a)$  is not a connected set of intervals on  $[1, n]$ .

## B Constructing a pathlet-preserving simplification

### B.1 Defining our $2\Delta$ -simplification $S$

We consider the vertex-restricted simplification defined by Agarwal *et al.* [2] and generalize their  $2\Delta$ -simplification definition, allowing vertices to lie anywhere on  $T$  (whilst still appearing in order along  $T$ ). This way, we obtain a simplification with at most as many vertices as the optimal unrestricted  $\Delta$ -simplification (see Figure 7).

► **Definition 26.** Let  $T$  be a trajectory with  $n$  vertices,  $\Delta \geq 0$  and  $a \in [1, n]$ . We define the set  $\mathbb{B}(a) = \{b \geq a \mid d_F(\overline{T(a)T(b)}, T[a, b]) \leq 2\Delta\}$ .

► **Definition 27.** Let  $T$  be a trajectory with  $n$  vertices and  $\Delta \geq 0$ . We define our  $2\Delta$ -simplified curve  $S$  as follows: the first vertex of  $S$  is  $T(1)$ . The second vertex of  $S$  may be any point  $T(a)$  with  $a$  as a rightmost endpoint of an interval in  $\mathbb{B}(1)$ . The third vertex of  $S$  may be any point  $T(b)$  with  $b$  as a rightmost endpoint of an interval in  $\mathbb{B}(a)$ , and so forth.

Per definition of the set  $\mathbb{B}(a)$ , the resulting curve  $S$  is a  $2\Delta$ -simplified curve. Let  $(f, g)$  be any  $2\Delta$ -matching between  $S$  and  $T$  that matches the vertices of  $S$  to the points on  $T$  that define them. We prove that  $(S, f, g)$  is a pathlet-preserving simplification.

► **Lemma 28.** The curve  $S$  as defined above, together with the matching  $(f, g)$ , forms a pathlet-preserving simplification.

**Proof.** We show that for any subcurve  $T[a, b]$  and all curves  $P$  with  $d_F(P, T[a, b]) \leq \Delta$ , the subcurve  $S[s, t]$  matched to  $T[a, b]$  by  $(f, g)$  has complexity  $|S[s, t]| \leq |P| + 2 - |\mathbb{N} \cap \{s, t\}|$ . For brevity, we write  $X = T[a, b]$ .

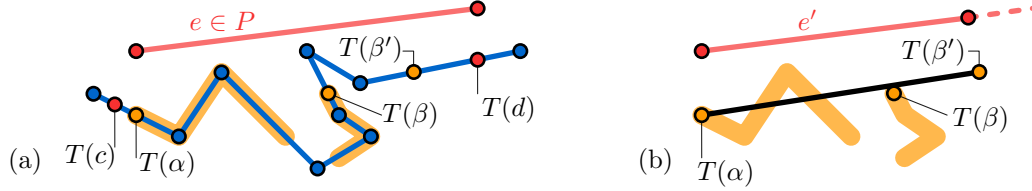
Fix any curve  $P$  with  $d_F(P, X) \leq \Delta$ . There exists a  $\Delta$ -matchings  $(f', g')$  between  $P$  and  $X$ . Per construction, the subcurve  $S[s, t]$  has Fréchet distance  $d_F(S[s, t], X) \leq 2\Delta$  to  $X$ . Any vertex of  $T$  that is a vertex of  $S[s, t]$  is also a vertex of  $T[a, b]$ . Let  $S[x, y]$  be the maximal vertex subcurve of  $S[s, t]$ . This curve naturally has  $|S[s, t]| - 2 + |\mathbb{N} \cap \{s, t\}|$  vertices. We argue that  $|P| \geq |S[x, y]|$ .

Suppose for sake of contradiction that  $|P| < |S[x, y]|$ . By the pigeonhole principle, there must exist an edge  $e_S = \overline{T(\alpha)T(\beta)}$  of  $S[x, y]$ , as well as an edge  $e_P$  of  $P$  matched to some subcurve  $T[c, d]$  of  $T$  by  $(f', g')$ , such that  $c \leq \alpha \leq \beta < d$ . We claim that  $\beta' \in \mathbb{B}(\alpha)$  for all  $\beta' \leq d$ .

The proof is illustrated in Figure 8. For any  $\beta' \leq d$  there exists a subedge  $e = e_P[x_1, x_2]$  of  $e_P$  that is matched to  $T[\alpha, \beta']$  by  $(f, g)$ . Per definition of a  $\Delta$ -matching we have that  $d_F(T[\alpha, \beta'], e) \leq \Delta$ . The  $2\Delta$ -matching implies that  $\|e(1) - T(\alpha)\| \leq 2\Delta$  and  $\|e(2) - T(\beta')\| \leq 2\Delta$ . We use this fact to apply [2, Lemma 3.1] and note that  $d_F(e, \overline{T(\alpha)T(\beta')}) \leq \Delta$ . Applying

the triangle inequality, we conclude that  $d_F(T[\alpha, \beta'], \overline{T(a)T(b')}) \leq 2\Delta$ . It follows that  $\beta' \in \mathbb{B}(\alpha)$ .

We obtain that  $[\alpha, d]$  is contained in the first connected component of  $\mathbb{B}(\alpha)$ . However, per construction of  $S$ ,  $\beta$  is a rightmost endpoint of a connected component in  $\mathbb{B}(\beta)$ . This contradicts the fact that  $\beta \in [\alpha, d]$ .  $\blacktriangleleft$



**Figure 8** (a) The construction in the proof of Lemma 28. We have an edge  $e$  with  $d_F(e, T[c, d]) \leq \Delta$ . Moreover, for some  $\alpha \in [c, d]$  we show  $\mathbb{B}(\alpha)$  in orange where  $\beta$  is the last value in some connected component of  $\mathbb{B}(\alpha)$ . (b) For any  $\beta' \in [\alpha, d]$ , there exists a subedge  $e'$  of  $e$  with  $d_F(e', T[a, \beta']) \leq \Delta$ .

## B.2 Constructing the simplification

We give an  $\mathcal{O}(n \log n)$  time algorithm for constructing our greedy simplification  $S$  (Definition 27). Given any point  $T(a)$  on  $T$ , we decompose the problem of finding a  $b \in \mathbb{B}(a)$  over all edges of  $T$ . That is, given an  $a \in [1, n]$ , we consider an individual edge  $T[i, i+1]$  of  $T$ . We show how to report the maximal  $b \in [i, i+1] \cap \mathbb{B}(a)$ . Our procedure is based off of the work by Guibas *et al.* [17] on ordered stabbing of disks in  $\mathbb{R}^2$ , and takes  $\mathcal{O}(i - a)$  time. We fix a plane  $H$  in  $\mathbb{R}^d$  that contains  $T(a)$  and  $T[i, i+1]$ . On a high-level, we apply the argument by Guibas *et al.* in  $\mathbb{R}^d$  by restricting the disks to their intersection with  $H$ :

► **Definition 29.** Let  $H$  be some fixed two-dimensional plane in  $\mathbb{R}^d$ . For any  $x \in [1, n]$  denote by  $B_x$  be the ball in  $H$  that is obtained by intersecting a ball of radius  $2\Delta$  centered at  $T(x)$  with  $H$ . For  $a \leq b$  we say that a directed line segment  $e$  in  $H$  stabs all balls in  $[a, b]$  in order if for all  $k \in \{a\} \cup ([a, b] \cap \mathbb{N}) \cup \{b\}$  there are points  $p_k \in e \cap B_k$  such that  $p_k$  comes before  $p_{k'}$  on  $e$  whenever  $k \leq k'$  (see Figure 9 (a)).

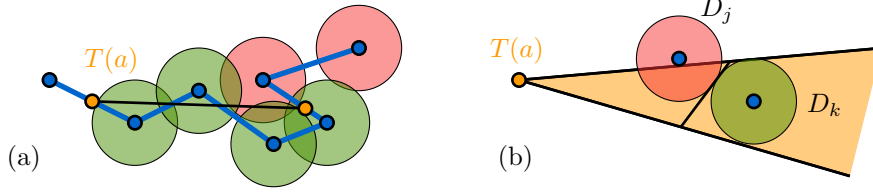
► **Lemma 30** ([17, Theorem 14]). A line segment  $e$  is within Fréchet distance  $2\Delta$  of a subcurve  $T[a, b]$  of  $T$  if and only if the following conditions are met:

1.  $e$  starts within distance  $2\Delta$  of  $T(a)$ ,
2.  $e$  ends within distance  $2\Delta$  of  $T(b)$ , and
3.  $e$  stabs all balls in  $[a, b]$  in order.

**Computing the maximal  $b \in [i, i+1] \cap \mathbb{B}(a)$ .** For any edge  $e = \overline{T(a)T(b)}$  of  $Z$ , the endpoints lie on  $T$  and thus  $e$  trivially satisfies the first two criteria. It follows that if we fix some  $T(a)$  on  $S$  and some edge  $T[i, i+1]$ , then the maximal  $b \in [i, i+1]$  (with  $b \geq a$ ) for which  $\overline{T(a)T(b)}$  stabs balls  $[a, b]$  in order is also the maximal  $b \in [i, i+1] \cap \mathbb{B}(a)$ . We consider the following (slightly reformulated) lemma by Guibas *et al.* [17]:

► **Lemma 31** ([17, Lemma 8]). Let  $[a_j, b_j]$  be a sequence of intervals. There exist  $p_j \in [a_j, b_j]$  with  $p_j \leq p_k$  for all  $j \leq k$ , if and only if there is no pair  $j \leq k$  with  $b_k < a_j$ .

The above lemma is applicable to segments in  $H$  stabbing balls in  $H$ . Indeed, consider all integers  $j \in [a, i]$ . We may view any directed line segment  $e$  in  $H$  as (part of) the real



■ **Figure 9** (a) For two points  $T(a)$  and  $T(b)$  on  $T$ , we consider all  $j \in [a, b] \cap N$  and draw a ball with radius  $2\Delta$  around them in green. (b) For all  $k \in [a, j-1] \cap \mathbb{N}$ , the area  $SW_{j-1}$  must be contained in the orange cone.

number line, and view the intersections between  $e$  and the disks  $D_j$  as intervals. Lemma 31 then implies that  $e$  stabs  $[a, i]$  in order, if and only if no integers  $j, k \in [a, i]$  exist with  $j \leq k$  such that  $e$  leaves  $D_k$  before it enters  $D_j$  (assuming  $e$  intersects all disks).

For all integers  $j \in [a, i]$ , let  $W_j := \{p \in H \mid \overline{T(a)p} \text{ intersects } D_j\}$ . We define the *stabbing wedge*  $SW_j := \{p \in H \mid \overline{T(a)p} \text{ intersects } [a, j] \text{ in order}\}$ . We prove the following:

► **Lemma 32.** *Either  $SW_j = \bigcap_{k \in [a, j] \cap \mathbb{N}} W_k$ , or  $SW_j = \emptyset$ .*

**Proof.** The proof is by induction. The base case is that, trivially,  $SW_{[a]} = W_{[a]}$ . Any line  $\overline{T(a)p}$  that intersects  $[a, j]$  also intersects  $[a, j-1]$ , thus whenever  $SW_{j-1}$  is empty, then  $SW_j$  must also be empty. Suppose now that  $SW_{j-1} = \bigcap_{k \in [a, j-1] \cap \mathbb{N}} W_k$ . We show that  $SW_j$  is either equal to  $SW_{j-1} \cap W_j$ , or it is empty (which, by induction, shows the lemma).

First we show that  $SW_j \subseteq SW_{j-1} \cap W_j$ . Suppose  $SW_j$  is non-empty, and take a point  $p \in SW_j$ . By definition of stabbing wedge  $SW_j$ , the segment  $\overline{T(a)p}$  stabs  $[a, j]$  in order, and thus  $[a, j-1]$  in order. Furthermore,  $p$  must lie in  $W_j$  for  $\overline{T(a)p}$  to be able to stab disk  $D_j$ .

Next we show that  $SW_{j-1} \cap W_j \subseteq SW_j$ . By Lemma 31,  $p \in SW_j$  if and only if  $\overline{T(a)p}$  first enters all disks  $D_{[a]}, \dots, D_{j-1}$  before exiting disk  $D_j$ . Fix some  $p \in SW_{j-1} \cap W_j$ , for all  $k < j$  the line  $\overline{T(a)p}$  intersects  $D_k$ . If  $p \notin SW_j$  then there must exist a  $k < j$  where  $\overline{T(a)p}$  exists  $D_j$  before it enters  $D_k$ . The area  $SW_{j-1}$  must be contained in the cone  $C \subset H$  given by  $T(a)$  and the two tangents of  $D_k$  to  $T(a)$  (Figure 9 (b)) and thus  $\overline{T(a)p}$  is contained in  $C$ . Since  $\overline{T(a)p}$  intersects  $D_k$  in  $C$  after  $D_j$ , it must be that  $D_j \cap C$  is contained in a triangle  $C^*$  formed by the boundary of  $C$  and another tangent of  $D_k$ . However, this means that any segment  $\overline{T(a)q}$  (for  $q \in SW_{j-1} \cap W_j$ ) that stabs the disks  $[a, j-1]$  in order must intersect  $C^*$  before it intersects  $D_k$ . Thus, by Lemma 31, there is no segment  $\overline{T(a)q}$  that stab the disks  $[a, j]$  in order and  $SW_j$  is empty. Thus, either  $SW_j$  is empty or  $SW_j = \bigcap_{[a] \leq k \leq j} W_k$ . ◀

► **Lemma 33.** *Given  $a \in [1, n]$  and edge  $T[i, i+1]$  of  $T$ , we can compute the maximum  $b \in [i, i+1] \cap \mathbb{B}(a)$ , or report that no such  $b$  exists, in  $\mathcal{O}(1 + i - a)$  time.*

**Proof.** By Lemma 30 (and the fact that  $T(a)$  and  $T(b)$  always lie on  $T$ )  $b \in [i, i+1] \cap \mathbb{B}(a)$  if and only if  $T(b) \in SW_i$ . For any edge  $T[i, i+1]$ , we compute the maximal  $b \in [i, i+1] \cap \mathbb{B}(a)$  by first assuming that  $SW_i$  is non-empty. We check afterwards whether this assumption was correct, and if not, we know that no  $b \in [i, i+1]$  exists with  $d_F(T[a, b], \overline{T(a)T(b)}) \leq 2\Delta$ .

For all  $j \in [a, i] \cap N$ , we compute the last value  $b_j \in [i, j]$  such that  $T(b_j) \in W_j$ . This can be done in  $\mathcal{O}(1)$  time per integer  $j$ , as wedges in the plane are formed by two rays and a circular arc in  $H$ , which we can intersect in  $\mathcal{O}(1)$  time. Then we set  $b = \min_j b_j$ .

To check whether  $SW_i$  is non-empty, we determine if  $d_F(T[a, b], \overline{T(a)T(b)}) \leq 2\Delta$ . This takes  $\mathcal{O}(1 + i - a)$  time with the algorithm of Alt and Godau [4]. The assumption that  $SW_i$  is non-empty is correct precisely if  $d_F(T[a, b], \overline{T(a)T(b)}) \leq 2\Delta$ . If  $SW_i$  is empty,  $[i, i+1] \cap \mathbb{B}(a)$

is empty and no output exists. If  $SW_i$  is non-empty, then by Lemma 32, the  $b$  we choose is the maximal  $b \in [i, i+1] \cap \mathbb{B}(a)$ .  $\blacktriangleleft$

► **Lemma 34.** *Given  $a \in [1, n]$ , we can compute a value  $b^*$  that is the maximum of some connected component of  $\mathbb{B}(a)$  in  $\mathcal{O}((1 + b^* - a) \log n)$  time.*

**Proof.** We use Lemma 33 in conjunction with exponential and binary search to compute the maximum  $b^*$  of some connected component of  $\mathbb{B}_{2\Delta}(a)$ :

We search over the edges  $T[i, i+1]$  of  $T$ . For each considered edge we apply Lemma 33 which returns some  $b \in [i, i+1] \cap \mathbb{B}(a)$  (if the set is non-empty). We consider three cases:

If  $b \in [i, i+1)$ , then this value is the maximum of some connected component of  $\mathbb{B}(a)$ . We stop the search and output  $b$ .

If the procedure reports the value  $b = i+1$  then this value may not necessarily be the maximum of a connected component. However, there is sure to be a maximum of at least  $b$ . Hence we continue the search among later edges of  $T$  and discard all edges before, and including,  $T[i, i+1]$ .

If the procedure reports no value then  $[i, i+1] \cap \mathbb{B}(a) = \emptyset$ . Since trivially  $a \in \mathbb{B}(a)$ , it must be that there is a connected component whose maximum is strictly smaller than  $i$ . We continue the search among earlier edges of  $T$  and discard all edges after, and including,  $T[i, i+1]$ .

The above algorithm returns the maximum  $b^* \in [i^*, i^*+1)$  of some connected component of  $\mathbb{B}(a)$ . By applying exponential search first, the edges  $T[i, i+1]$  considered all have  $i \leq 2i^* - a$ . Hence we compute  $b^*$  in  $\mathcal{O}((1 + b^* - a) \log n)$  time.  $\blacktriangleleft$

We now iteratively apply Lemma 34 to construct our curve  $S$ . We obtain a  $2\Delta$ -matching  $(f, g)$  by constructing separate matchings between the edges  $\overline{T(a)T(b)}$  of  $S$  and the subcurves  $T[a, b]$  that they simplify. By Lemma 28 this gives a pathlet-preserving simplification  $(S, f, g)$ .

► **Theorem 8.** *For any trajectory  $T$  with  $n$  vertices and any  $\Delta \geq 0$ , we can construct a pathlet-preserving simplification  $S$  in  $\mathcal{O}(n \log n)$  time.*