Exploration of a Graph-based Density-Sensitive Metric

Timothy Chu
CMU
tzchu@andrew.cmu.edu

Gary L. Miller
CMU
glmiller@cs.cmu.edu

Donald Sheehy
University of Connecticut
don.r.sheehy@gmail.com

October 26, 2018

Abstract

We consider a simple graph-based metric on points in Euclidean space known as the edge-squared metric. This metric is defined by squaring the Euclidean distance between points, and taking the shortest paths on the resulting graph. This metric has been studied before in wireless networks and machine learning, and has the **density-sensitive** property: distances between two points in the same cluster are short, even if their Euclidean distance is long. This property is desirable in machine learning.

In this paper, we show that this metric is equal to previously studied geodesic-based metric defined on points in Euclidean space. Previous best work showed that these metrics are 3-approximations of each other. It was not known or suspected that these metrics were equal. We give fast algorithms to compute sparse spanners of this distance, a problem that can be seen as a generalization of both the Euclidean spanner and the Euclidean MST problem. Spanners of the edge-squared metric are faster to compute and sparser than the best known Euclidean spanners in a variety of settings.

1 Introduction

A foundational hypothesis in non-linear dimension reduction and machine learning is that data can be represented as points in Euclidean space, and that graphs on these points can be generated to solve a variety of problems on the data, including classification, regression, and clustering [29, 26, 53, 32, 43, 74, 55]. Such graphs can induce similarity measures or distances on the point set, which are used as the key model when generating a clustering or classification [29, 16, 67]

Although graph algorithms have been extensively studied, the problem of generating metrics and graphs on data sets is an active topic whose study has mostly occurred in machine learning, statistics, and geometry [44, 6, 58, 73, 16, 43, 4, 69, 38, 30, 23, 39, 40, 75, 3, 63, 45]. We believe two important problems are: comparing the various methods of generating distances from point sets, and building data structures to quickly compute these distances.

In this paper, we study a particular distance generated from a graph on points in Euclidean space, called the edge-squared metric. It is defined by taking the Euclidean distance squared between two points, and finding the shortest path on the resulting graph. This metric has the property that two points in a dense cluster are considered close, even if their Euclidean distance is far [16]. This property is known as **density-sensitivity** [73], and is desirable for clustering and classification [58, 4, 45, 28]. The edge-squared metric has been studied before in the context of machine learning [16, 73, 45, 4, 28] and power-efficient wireless networks [50, 51]. Squared Euclidean distances have been examined before as an optimization objective, and occur in natural settings including k-means clustering and RMS matching [54, 61]. We compare the edge-squared metric to another distance on point-sets known as the nearest-neighbor geodesic distance, first introduced in [28] as the nearest neighbor distance. Close variants of this metric have been studied in probability and machine learning [58, 28, 45]. We then give efficient algorithms constructing sparse spanners for both metrics.

In this paper, we show that the edge-squared metric and nearest-neighbor geodesic distance are identical. It was not previously known or suspected that the two might be the same. This is the first work we know of that equates a discrete metric with a continuous geodesic, and gives the first nontrivial example of a so-called **density-based distance** [58] that can be computed exactly. This considerably improves a result in [28], which showed that the two were 3-approximations of each other. Previous works computing geodesics or density-based distances generally use approximate methods whose run-time grows as the approximation quality improves [46, 67, 4, 3, 45, 28], or calculus of variations [15, 60, 66]. We use a different method. Our proof employs the Kirszbraun theorem, also known as the Lipschitz Extension Theorem [47, 18]. This theorem has been widely used in computational geometry, classification, and metric embedding theory [57, 49, 56, 40]. Our result lets us compute the persistent homology of the nearest-neighbor geodesic, a general problem of interest for many continuous metrics in the computational geometry setting [33, 36, 24, 25, 22, 2].

In order to perform clustering or classification with the edge-squared metric or nearest-neighbor geodesic distance, we provide data structures that admit fast, practical computation of these metrics with theoretical guarantees. The edge-squared metric can have high doubling dimension even if the underlying points are in 2 dimensions [28]. This means that many data structures suitable for low-doubling dimension metrics will not immediately work on it. Despite this, if the underlying point set is in low dimension, we compute a sparse $(1+\varepsilon)$-spanner of the edge-squared metric quickly. This spanner is sparser and faster to compute than the best known $(1+\varepsilon)$-spanners for Euclidean metrics, and uses techniques from well-separated pair decompositions [20] and approximate Euclidean MSTs [11, 12, 21].

A foundational assumption of machine learning is that most data points are samples from a well-behaved probability distribution with low intrinsic dimension [67, 38, 72, 43, 55, 45]. We compute a sparse 1-spanner for our metrics in such a setting. Our 1-spanner is a k-nearest neighbor graph (k-NN graph) with edge weights equal to Euclidean distance squared, with $k = O(2^d \log n)$. Here, d is the intrinsic dimension of the probability density. Note that a sparse 1-spanner of Euclidean distance is not possible in this setting. Our result may allow for fast computation of edge-squared spanners in practice, given the breadth of literature on the k-NN graph [19, 32, 26] and its widespread use in practice [67, 53, 32]. If intrinsic dimension d is constant, our k is nearly optimal: it is believed $k = \Omega(\log n)$ is necessary for connectivity of the k-nearest neighbor graph [14, 37, 53].

We also show how spanners of the edge-squared metric can be seen as a generalization of Euclidean spanners [70, 34, 11, 20], approximate Euclidean MSTs [21, 9, 12, 5, 77], and single linkage clustering [41, 77].

1.1 Definitions and Preliminaries

Edge-squared metric: For $x \in \mathbb{R}^d$, let $\|x\|$ denote the Euclidean norm. For a set of points $P \subset \mathbb{R}^d$:

Definition 1.1. *The edge-squared metric for $a, b \in P$ is*

$$\mathbf{d}_2(a,b) = \min_{(p_0,\ldots,p_k)} \sum_{i=1}^{k} \|p_i - p_{i-1}\|^2,$$

where the minimum is over sequences of points $p_0, \ldots, p_k \in P$ with $p_0 = a$ and $p_k = b$.

Nearest-neighbor geodesic distance: Another metric on the points of P is called the nearest-neighbor geodesic distance, and is denoted $\mathbf{d}_N$. This distance was first defined and studied in [28]. Before we can define it, we need a couple other definitions.

Given any finite set $P \subset \mathbb{R}^k$, there is a real-valued function $\mathbf{r}_P : \mathbb{R}^k \to \mathbb{R}$ defined as $\mathbf{r}_P(z) = \min_{x \in P} \|x - z\|$. A path is a continuous mapping $\gamma : [0,1] \to \mathbb{R}^d$. Let path$(a,b)$ denote the set of piecewise-C_1 paths from a to b. We will compute the lengths of paths relative to the distance function $\mathbf{r}_P$ as follows.

$$\ell(\gamma) := \int_0^1 \mathbf{r}_P(\gamma(t))\|\gamma'(t)\|dt.$$

By considering the velocity of γ, this definition is independent of the parameterization of the path.

Definition 1.2. *The nearest-neighbor geodesic distance is defined as:*

$$\mathbf{d}_N(a,b) := 4 \inf_{\gamma \in \text{path}(a,b)} \ell(\gamma).$$

The factor of 4 *normalizes the metrics.*

In particular, when P has only two points a and b, $\mathbf{d}_2(a,b) = \mathbf{d}_N(a,b)$. This reduces to a high school calculus exercise as the minimum path γ will be a straight line between the points and the nearest neighbor geodesic is

$$\mathbf{d}_N(a,b) = 4\int_0^1 \mathbf{r}_P(\gamma(t))\|\gamma'(t)\|dt = 8\int_0^{\frac{1}{2}} t\|a-b\|^2 dt = \|a-b\|^2 = \mathbf{d}_2(a,b).$$

This observation about pairs of points makes it easy to see that the nearest-neighbor geodesic distance is never greater than the edge-squared distance, as proven in the following lemma.

Lemma 1.3. *For all $s,p \in P$, we have $\mathbf{d}_N(s,p) \leq \mathbf{d}_2(s,p)$.*

Proof. Fix any points $s,p \in P$. Let $q_0,\ldots,q_k \in P$ be such that $q_0 = s$, $q_k = p$ and

$$\mathbf{d}_2(s,p) = \sum_{i=1}^{k} \|q_i - q_{i-1}\|^2.$$

Let $\psi_i(t) = tq_i + (1-t)q_{i-1}$ be the straight line segment from q_{i-1} to q_i. Observe that $\ell(\psi_i) = \|q_i - q_{i-1}\|^2/4$, by the same argument as in the two point case. Then, let ψ be the concatenation of the ψ_i and it follows that

$$\mathbf{d}_2(s,p) = 4\ell(\psi) \geq 4 \inf_{\gamma \in \text{path}(s,p)} \ell(\gamma) = \mathbf{d}_N(s,p). \qquad \square$$

Spanners: For real value $t \geq 1$, a t-spanner of a weighted graph G is a subgraph S such that $d_G(x,y) \leq d_S(x,y) \leq t \cdot d_G(x,y)$ where d_G and d_S represent the shortest path distance functions between vertex pairs in G and S. Spanners of Euclidean distances, and general graph distances, have been studied extensively, and their importance as a data structure is well established. [27, 70, 20, 42].

k-nearest neighbor graphs: The k-nearest neighbor graph (k-NN graph) for a set of objects V is a graph with vertex set V and an edge from $v \in V$ to its k most similar objects in V, under a given distance measure. In this paper, the underlying distance measure is Euclidean, and the edge weights are Euclidean distance squared. k-NN graph constructions are a key data structure in machine learning [32, 26], clustering [53], and manifold learning [67].

Gabriel Graphs: The Gabriel graph is a graph where two vertices p and q are joined by an edge if and only if the disk with diameter pq has no other points of S in the interior. The Gabriel graph is a subgraph of the Delaunay triangulation [63], and a 1-spanner of the edge-squared metric [63]. Gabriel graphs will be used in the proof of Theorem 1.6.

1.2 Contributions

Our paper has three main theorems.

Theorem 1.4. *Given a point set $P \in \mathbb{R}^d$, the edge-squared metric on P and the nearest-neighbor geodesic on P are always equivalent.*

Theorem 1.5. *For any set of points in $\mathbb{R}^d$ for constant d, there exists a $(1+\varepsilon)$ spanner of the edge-squared metric, with size $O\left(n\varepsilon^{-d/2}\right)$ computable in time $O\left(n \log n + n\varepsilon^{-d/2} \log \frac{1}{\varepsilon}\right)$. The $\log \frac{1}{\varepsilon}$ term goes away given a fast floor function.*

Theorem 1.6. *Suppose points P in Euclidean space are drawn i.i.d from a Lipschitz probability density bounded above and below by a constant, with support on a smooth, connected, compact manifold with intrinsic dimension d, and smooth boundary of bounded curvature. Then w.h.p. the k-NN graph of P for $k = O(2^d \ln n)$ and edges weighted with Euclidean distance squared, is a 1-spanner of the edge-squared metric on P.*

Theorem 1.4 considerably strengthens a result from in [28], which showed $\mathbf{d}_2$ is a 3-approximation of $\mathbf{d}_N$. Our theorem finds $\mathbf{d}_N$ exactly, and lets us compute the persistent homology of $\mathbf{d}_N$. $\mathbf{d}_N$ is defined on all points in space, and is thus a metric extension [56] of the edge-squared metric and of negative type distances [31] to the entire space.

Theorem 1.5 proves that a $(1 + \varepsilon)$ spanner of the edge-squared metric with points in constant dimension is sparser and quicker to compute than the Euclidean spanners of Callahan and Kosaraju [20]. The latter spanners have $O(n\varepsilon^{-d})$ edges and are computable in $O(n \log n + n\varepsilon^{-d})$ time. To the authors' knowledge, these are the sparsest quickly-constructable Euclidean spanners in terms of ε dependence. Later works on spanners have focused on bounding diameter, degree, or total edge weight [11, 34]. We give a size lower bound for $(1+\varepsilon)$-Euclidean spanners, which is close to the sparsity of our $(1+\varepsilon)$ spanner of the edge-squared metric. Previously, sparse spanners of the edge-squared metric were shown to exist in two dimensions via Yao graphs and Gabriel graphs [50].

Theorem 1.6 proves that a 1-spanner of the edge-squared metric can be found assuming points are samples from a probability density, by using a k-NN graph for appropriate k. Our result is tight when d is constant. This is not possible for Euclidean distance, as a 1-spanner is almost surely the complete graph. Without the probability density assumption, there are point sets in $\mathbb{R}^4$ where 1-spanners of the edge-squared metric require $\Omega(n^2)$ edges. Finally, we show that spanners of p-power metrics, which are edge-squared metrics but with powers of p instead of 2, generalize Euclidean spanners and Euclidean MSTs. p-power metrics were considered in [50].

2 Outline

Section 3 contains the proof of Theorem 1.4, equating the edge-squared metric and nearest-neighbor geodesic distance in all cases. We then compute the persistent homology of the nearest-neighbor geodesic distance. Section 4 outlines a proof of Theorem 1.5, and compares our spanner to new lower bounds on the sparsity of $(1+\varepsilon)$-spanners of the Euclidean metric. We outline a proof of Theorem 1.6 in Section 5 and discuss its implications.

Section 6 introduces the p-power metrics. We show that Euclidean spanners and Euclidean MSTs are special cases of p-power spanners. We show how clustering algorithms including k-means, level-set methods, and single linkage clustering, are special cases of clustering with p-power metrics.

Conclusions and open questions are in Section 7. Full proofs for Theorems 1.6, 1.5 are contained in the Appendix.

3 Edge-Squared Metric is Equivalent to the Nearest-Neighbor Geodesic Distance

In this section, we prove Theorem 1.4. By Lemma 1.3, it suffices to show that for a and b on Euclidean point set P, we have $\mathbf{d}_N(a,b) \geq \mathbf{d}_2(a,b)$ for $a, b \in P$.

Let $P \subset \mathbb{R}^d$ be a set of n points. Pick any *source* point $s \in P$. Order the points of P as $p_1, \ldots, p_n$ so that

$$\mathbf{d}_2(s, p_1) \leq \cdots \leq \mathbf{d}_2(s, p_n).$$

This will imply that $p_1 = s$. It will suffice to show that for all $p_i \in P$, we have $\mathbf{d}_2(s, p_i) = \mathbf{d}_N(s, p_i)$. There are three main steps:

1. We first show that when P is a subset of the vertices of an axis-aligned box, $\mathbf{d} = \mathbf{d}_N$. In this case, shortest paths for $\mathbf{d}$ are single edges and shortest paths for $\mathbf{d}_N$ are straight lines.

2. We then show how to lift the points from $\mathbb{R}^d$ to $\mathbb{R}^n$ by a Lipschitz map m that places all the points on the vertices of a box and preserves $\mathbf{d}_2(s,p)$ for all $p \in P$.

3. Finally, we show how the Lipschitz extension of m is also Lipschitz as a function between nearest-neighbor geodesic distances. We combine these pieces to show that $\mathbf{d} \leq \mathbf{d}_N$. As $\mathbf{d} \geq \mathbf{d}_N$ (Lemma 1.3), this will conclude the proof that $\mathbf{d} = \mathbf{d}_N$.

3.0.1 Boxes

Let Q be the vertices of a box in $\mathbb{R}^n$. That is, there exist some positive real numbers $\alpha_1, \ldots, \alpha_n$ such that each $q \in Q$ can be written as $q = \sum_{i \in I} \alpha_i e_i$, for some $I \subseteq [n]$.

Let the source s be the origin. Let $\mathbf{r}_Q : \mathbb{R}^n \to \mathbb{R}$ be the distance function to the set Q. Setting $r_i(x) := \min\{x_i, \alpha_i - x_i\}$ (a lower bound on the difference in the ith coordinate to a vertex of the box), it follows that

$$\mathbf{r}_Q(x) \geq \sqrt{\sum_{i=1}^{n} r_i(x)^2}. \tag{1}$$

Let $\gamma : [0,1] \to \mathbb{R}^n$ be a curve in $\mathbb{R}^n$. Define $\gamma_i(t)$ to be the projection of γ onto its ith coordinate. Thus,

$$r_i(\gamma(t)) = \min\{\gamma_i(t), \alpha_i - \gamma_i(t)\} \tag{2}$$

and

$$\|\gamma'(t)\| = \sqrt{\sum_{i=1}^{n} \gamma_i'(t)^2}. \tag{3}$$

We can bound the length of γ as follows. For simplicity of exposition we only present the case of a path from the origin to the far corner, $p = \sum_{i=1}^n \alpha_i e_i$.

$$\begin{aligned}
\ell(\gamma) &= \int_0^1 \mathbf{r}_Q(\gamma(t))\|\gamma'(t)\|dt && \text{[by definition]}\\
&\geq \int_0^1 \left(\sqrt{\sum_{i=1}^{n} r_i(\gamma(t))^2}\sqrt{\sum_{i=1}^{n} \gamma_i'(t)^2}\right) dt && \text{[by (1) and (3)]}\\
&\geq \sum_{i=1}^{n} \int_0^1 r_i(\gamma(t))\gamma_i'(t)dt && \text{[Cauchy-Schwarz]}\\
&\geq \sum_{i=1}^{n} \left(\int_0^{\ell_i} \gamma_i(t)\gamma_i'(t)dt + \int_{\ell_i'}^1 (\alpha_i - \gamma_i(t))\gamma_i'(t)dt\right)\\
&\text{[by (2) where } \gamma_i(\ell_i) = \alpha_i/2 \text{ for the first time and } \gamma_i(\ell_i') = \alpha_i/2 \text{ for the last time.]}\\
&= \sum_{i=1}^{n} 2\int_0^{\ell_i} \gamma_i(t)\gamma_i'(t)dt && \text{[by symmetry]}\\
&\geq \sum_{i=1}^{n} \frac{\alpha_i^2}{4} && \text{[basic calculus]}
\end{aligned}$$

It follows that if γ is any curve that starts at s and ends at $p = \sum_{i=1}^n \alpha_i e_i$, then $\mathbf{d}_N(s,p) = \mathbf{d}_2(s,p)$.

3.0.2 Lifting the points to $\mathbb{R}^n$

Define a mapping $m : P \to \mathbb{R}^n$. We do this by adding the points $p_1, \ldots, p_n$, as defined above, one point at a time. For each new point we will introduce a new dimension. We start by setting $m(p_1) = 0$ and by induction:

$$m(p_i) = m(p_{i-1}) + \sqrt{\mathbf{d}_2(s,p_i) - \mathbf{d}_2(s,p_{i-1})}e_i, \tag{4}$$

where the vectors e_i are the standard basis vectors in $\mathbb{R}^n$.

Lemma 3.1. *For all $p_i, p_j \in P$, we have*

(i) $\|m(p_j) - m(p_i)\| = \sqrt{|\mathbf{d}_2(s,p_j) - \mathbf{d}_2(s,p_i)|}$, *and*

(ii) $\|m(s) - m(p_j)\|^2 \leq \|m(p_i)\|^2 + \|m(p_i) - m(p_j)\|^2$.

Proof. Proof of (i). Without loss of generality, let $i \leq j$.

$$\begin{aligned}
\|m(p_j) - m(p_i)\| &= \left\| \sum_{k=i+1}^{j} \sqrt{\mathbf{d}_2(s, p_k) - \mathbf{d}_2(s, p_{k-1})}\, e_k \right\| && \text{[from the definition of } m\text{]} \\
&= \sqrt{\sum_{k=i+1}^{j} (\mathbf{d}_2(s, p_k) - \mathbf{d}_2(s, p_{k-1}))} && \text{[expand the norm]} \\
&= \sqrt{\mathbf{d}_2(s, p_j) - \mathbf{d}_2(s, p_i)}. && \text{[telescope the sum]}
\end{aligned}$$

Proof of (ii). As $m(s) = 0$, it suffice to observe that

$$\begin{aligned}
\|m(p_j)\|^2 &= \mathbf{d}_2(s, p_j) && \text{[by } \textit{(i)}\text{]} \\
&\leq \mathbf{d}_2(s, p_i) + |\mathbf{d}_2(s, p_j) - \mathbf{d}_2(s, p_i)| && \text{[basic arithmetic]} \\
&= \|m_(p_i)\|^2 + \|m(p_i) - m(p_j)\|^2 && \text{[by } \textit{(i)}\text{]}
\end{aligned}$$

□

We can now show that m has all of the desired properties.

Proposition 3.2. *Let $P \subset \mathbb{R}^d$ be a set of n points, let $s \in P$ be a designated source point, and let $m : P \to \mathbb{R}^n$ be the map defined as in (4). Let $\mathbf{d}'$ denote the edge squared metric for the point set $m(P)$ in $\mathbb{R}^n$. Then,*

(i) m is 1-Lipschitz as a map between Euclidean metrics,

(ii) m maps the points of P to the vertices of a box, and

(iii) m preserves the edge squared distance to s, i.e. $\mathbf{d}'(m(s), m(p)) = \mathbf{d}_2(s, p)$ for all $p \in P$.

Proof. Proof of (i). To prove the Lipschitz condition, fix any $a, b \in P$ and bound the distance as follows.

$$\begin{aligned}
\|m(a) - m(b)\| &= \sqrt{|\mathbf{d}_2(s, a) - \mathbf{d}_2(s, b)|} && \text{[Lemma 3.1(i)]} \\
&\leq \sqrt{\mathbf{d}_2(a, b)} && \text{[triangle inequality]} \\
&\leq \|a - b\| && \Big[\mathbf{d}_2(a, b) \leq \|a - b\|^2 \text{ by the definition of } \mathbf{d}\Big]
\end{aligned}$$

Proof of (ii). That m maps P to the vertices of a box is immediate from the definition. The box has side lengths $\|m_i - m_{i-1}\|$ for all $i > 1$ and $p_i = \sum_{k=1}^{i} \|m_k - m_{k-1}\| e_k$.
Proof of (iii). We can now show that the edge squared distance to s is preserved. Let $q_0, \ldots, q_k$ be the shortest sequence of points of $m(P)$ that realizes the edge-squared distance from $m(s)$ to $m(p)$, i.e., $q_0 = m(s)$, $q_k = m(p)$, and

$$\mathbf{d}'(m(s), m(p)) = \sum_{i=1}^{k} \|m(q_i) - m(q_{i-1})\|^2.$$

If $k > 1$, then Lemma 3.1(ii) implies that removing q_1 gives a shorter sequence. Thus, we may assume $k = 1$ and therefore, by Lemma 3.1(i),

$$\mathbf{d}'(m(s), m(p)) = \|m(s) - m(p)\|^2 = \mathbf{d}_2(s, p). \qquad \square$$

3.0.3 The Lipschitz Extension

Proposition 3.2 and the Kirszbraun theorem on Lipschitz extensions imply that we can extend m to a 1-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}^n$ such that $f(p) = m(p)$ for all $p \in P$ [47, 71, 18].

Lemma 3.3. *The function f is also 1-Lipschitz as mapping from $\mathbb{R}^d \to \mathbb{R}^n$ with both spaces endowed with the nearest-neighbor geodesic distance.*

Proof. We are interested in two distance functions $\mathbf{r}_P : \mathbb{R}^d \to \mathbb{R}$ and $\mathbf{r}_{f(P)} : \mathbb{R}^n \to \mathbb{R}$. Recall that each is the distance to the nearest point in P or $f(P)$ respectively.

$$\begin{aligned}
\mathbf{r}_{f(P)}(f(x)) &= \min_{q \in f(P)} \|q - f(x)\| && \text{[by definition]}\\
&= \min_{p \in P} \|f(p) - f(x)\| && [q = f(p) \text{ for some } p]\\
&\leq \min_{p \in P} \|p - x\| && [f \text{ is 1-Lipschitz}]\\
&= \mathbf{r}_P(x). && \text{[by definition]}
\end{aligned}$$

For any curve $\gamma : [0, 1] \to \mathbb{R}^d$ and for all $t \in [0, 1]$, we have $\|(f \circ \gamma)'(t)\| \leq \|\gamma'(t)\|$. It then follows that

$$\ell'(f \circ \gamma) = \int_0^1 \mathbf{r}_{f(P)}(f(\gamma(t)))\|(f \circ \gamma)'(t)\|dt \leq \int_0^1 \mathbf{r}_P(\gamma(t))\|\gamma'(t)\|dt = \ell(\gamma), \tag{5}$$

where ℓ' denotes the length with respect to $\mathbf{r}_{f(P)}$. Thus, for all $a, b \in P$,

$$\begin{aligned}
\mathbf{d}_N(a, b) &= 4 \inf_{\gamma \in \text{path}(a,b)} \ell(\gamma) && \text{[by definition]}\\
&\geq 4 \inf_{\gamma \in \text{path}(a,b)} \ell'(f \circ \gamma) && \text{[by (5)]}\\
&\geq 4 \inf_{\gamma' \in \text{path}(f(a),f(b))} \ell'(\gamma') && [\text{because } f \circ \gamma \in \text{path}(f(a), f(b))]\\
&= \mathbf{d}_N(f(a), f(b)). && \text{[by definition]}
\end{aligned}$$

$\square$

We now restate Theorem 1.4 for convenience, and prove it.

Theorem 3.4. *For any point set $P \subset \mathbb{R}^d$, the edge squared metric $\mathbf{d}$ and the nearest-neighbor geodesic distance $\mathbf{d}_N$ are identical.*

Proof. Fix any pair of points s and p in P. Define the Lipschitz mapping m and its extension f as in (4). Let $\mathbf{d}'$ and $\mathbf{d}'_N$ denote the edge-squared and nearest-neighbor geodesic distances on $f(P)$ in $\mathbb{R}^n$.

$$\begin{aligned}\mathbf{d}_2(s,p) &= \mathbf{d}'(m(s), m(p)) && \text{[Proposition 3.2(iii)]}\\ &= \mathbf{d}'_N(m(s), m(p)) && \text{[$f(P)$ are vertices of a box]}\\ &\leq \mathbf{d}_N(s,p) && \text{[Lemma 3.3]}\end{aligned}$$

We have just shown that $\mathbf{d} \leq \mathbf{d}_N$ and Lemma 1.3 states that $\mathbf{d} \geq \mathbf{d}_N$, so we conclude that $\mathbf{d} = \mathbf{d}_N$ as desired. □

3.1 Persistent Homology of the Nearest-neighbor Geodesic Distance

In this section, we show how to compute the so-called persistent homology [33] of the nearest-neighbor geodesic distance in two different ways, one ambient and the other intrinsic. The latter relies on Theorem 1.4 and would be quite surprising without it.

Persistent homology is a popular tool in computational geometry and topology to ascribe quantitative topological invariants to spaces that are stable with respect to perturbation of the input. In particular, it's possible to compare the so-called persistence diagram of a function defined on a sample to that of the complete space [24]. These two aspects of persistence theory—the intrinsic nature of topological invariants and the ability to rigorously compare the discrete and the continuous—are both also present in our theory of nearest-neighbor geodesic distances. Indeed, the primary motivation for studying these metrics was to use them as inputs to persistence computations for problems such as persistence-based clustering [25] or metric graph reconstruction [1].

The input for persistence computation is a *filtration*—a nested sequence of spaces, usually parameterized by a real number $\alpha \geq 0$. The output is a set of points in the plane called a *persistence diagram* that encodes the birth and death of topological features like connected components, holes, and voids.

The Ambient Persistent Homology Perhaps the most popular filtration to consider on a Euclidean space is the sublevel set filtration of the distance to a sample P. This filtration is $(F_\alpha)_{\alpha \geq 0}$, where

$$F_\alpha := \{x \in \mathbb{R}^d \mid \mathbf{r}_P(x) \leq \alpha\},$$

for all $\alpha \geq 0$. If one wanted to consider instead the nearest-neighbor geodesic distance $\mathbf{d}_N$, one gets instead a filtration $(G_\alpha)_{\alpha \geq 0}$, where

$$G_\alpha := \{x \in \mathbb{R}^d \mid \min_{p \in P} \mathbf{d}_N(x,p) \leq \alpha\},$$

for all $\alpha \geq 0$.

Both the filtrations (F_α) and (G_α) are unions of metric balls. In the former, they are Euclidean. In the latter, they are the metric balls of $\mathbf{d}_N$. These balls can look very different, for example, for $\mathbf{d}_N$, the metric balls are likely not even convex. However, these filtrations are very closely related.

Lemma 3.5. *For all $\alpha \geq 0$, $F_\alpha = G_{2\alpha^2}$.*

Proof. The key to this exercise is to observe that the nearest point $p \in P$ to a point x is also the point that minimizes $\mathbf{d}_N(x,p)$. To prove this, we will show that for any $p \in P$ and any path $\gamma \in \mathrm{path}(x,p)$, we have $\ell(\gamma) \geq \frac{1}{2}\mathbf{r}_P(x)^2$. Consider any such x, p, and γ. The euclidean length of γ must be at least $\mathbf{r}_P(x)$, so we will assume that $\|\gamma'\| = \mathbf{r}_P(x)$ and will prove the lower bound on the subpath starting at x of length exactly $\mathbf{r}_P(x)$. This will imply a lower bound on the whole path. Because $\mathbf{r}_P$ is 1-Lipschitz, we have $\mathbf{r}_P(\gamma(t)) \geq (1-t)\mathbf{r}_P(x)$ for all $t \in [0,1]$. It follows that

$$\ell(\gamma) = \int_0^1 \mathbf{r}_P(\gamma(t))\|\gamma'(t)\|dt \geq \mathbf{r}_P(x)^2 \int_0^1 (1-t)dt = \frac{1}{2}\mathbf{r}_P(x)^2$$

The bound above applies to any path from x to a point $p \in P$, and so,

$$\mathbf{d}_N(x,p) = 4 \inf_{\gamma \in \mathrm{path}(x,p)\ell(\gamma)} \geq 2\mathbf{r}_P(x).$$

If p is the nearest neighbor of x in P, then $\mathbf{d}_N(x,p) = 2\mathbf{r}_P(x)$, by taking the path to be a straight line. It follows that $\min_{p \in P} \mathbf{d}_N(x,p) = 2\mathbf{r}_P(x)$. □

The preceding lemma shows that the two filtrations are equal up to a monotone change in parameters. By standard results in persistent homology, this means that their persistence diagrams are also equal up to the same change in parameters. This means that one could use standard techniques such as α-complexes [33] to compute the persistence diagram of the Euclidean distance and convert it to the nearest-neighbor geodesic distance afterwards. Moreover, one observes that the same equivalence will hold for variants of the nearest-neighbor geodesic distance that take other powers of the distance.

Intrinsic Persistent Homology Recently, several researchers have considered intrinsic nerve complexes on metric data, especially data coming from metric graphs [2, 36]. These complexes are defined in terms of the intersections of metric balls in the input. The vertex set is the input point set. The edges at scale α are pairs of points whose α-radius balls intersect. In the intrinsic Čech complex, triangles are defined for three way intersections, tetrahedra for four-way intersections, etc.

In Euclidean settings, little attention was given to the difference between the intrinsic and the ambient persistence, because a classic result, the Nerve Theorem [17], and its persistent version [24] guaranteed there is no difference. The Nerve theorem, however, requires the common intersections to be contractible, a property easily satisfied by convex sets such as Euclidean balls. However, in many other topological metric spaces, the metric balls might not be so well-behaved. In particular, the nearest-neighbor geodesic distance has metric balls which may take on very strange shapes, depending on the density of the sample. This is similarly true for graph metrics. So, in these cases, there is a difference between the information in the ambient and the intrinsic persistent homology.

Theorem 3.6. *Let $P \subset \mathbb{R}^d$ be finite and let $\mathbf{d}_N$ be the nearest-neighbor geodesic distance with respect to P. The edges of the intrinsic Čech filtration with respect to $\mathbf{d}_N$ can be computed exactly in polynomial time.*

Proof. The statement is equivalent to the claim that $\mathbf{d}_N$ can be computed exactly between pairs of points of P, a corollary of Theorem 3.4. Two radius α balls will intersect if and only of the distance between their centers is at most 2α. The bound on the distance necessarily implies a path and the common intersection will be the midpoint of the path. □

4 Fast, Sparse Spanner for the Edge-Squared Metric

Now we outline a proof for Theorem 1.5, which shows that one can construct a $(1+\varepsilon)$ edge-squared spanner of size $O(n\varepsilon^{-d/2})$ in time $O\left(n\log n + n\varepsilon^{-d/2}\log\left(\frac{1}{\varepsilon}\right)\right)$, for points in constant dimensional space. The full proof is in Appendix A. By Theorem 1.4, this spanner is also a good spanner of the nearest-neighbor geodesic distance. Note that this spanner is sparser and faster in terms of epsilon dependency than the best spanner for Euclidean distances known to the authors , which has $O\left(\varepsilon^{-d}\right)$ edges and runs in $O\left(n\log n + \varepsilon^{-d}n\log\left(\frac{1}{\varepsilon}\right)\right)$ time [20]. We rely extensively on well-separated pair decompositions (WSPDs), and this outline assumes familiarity with that notation. For a comprehensive set of definitions and notations on well separated pairs, refer to any of [21, 12, 20, 11]. Our proof consists of three parts.

1. Showing that connecting a $(1+O(\delta^2))$-approximate shortest edge in a $1/\delta$ well separated pair for all the pairs in the decomposition gives a $1+O(\delta^2)$ edge-squared spanner. The processing for this step takes $O(n\log n + \delta^{-d}n)$ time.

2. Previous work contains an algorithm computing $1+O(\delta^2)$-approximate shortest edge in a $1/\delta$ well separated pair for all the pairs in a WSPD, and takes $O(1)$ time per pair. The pre-processing for this step will be bounded by $O(\delta^{-d}n\log\left(\frac{1}{\delta}\right))$ time. The $\log\left(\frac{1}{\delta}\right)$ factor goes away given a fast floor function. This procedure was first introduced in [21].

3. Putting these two together, and setting $\epsilon = \delta^2$ gives us a $1+\epsilon$ spanner with $O(\epsilon^{-d/2}n)$ edges in $O(n\log n + \epsilon^{-d/2}n)$ time.

Full details of this proof are contained in Appendix A

4.1 Lower Bounds for Sparsity of Euclidean Spanners

Theorem 4.1. *For constant d and any fixed ε, there exists a set of points such that any $(1+\varepsilon)$ Euclidean spanner in $\mathbb{R}^d$ needs $\Omega\left(n\varepsilon^{-\lfloor d/2\rfloor+1}\right)$ edges.*

Here, we show that our edge-squared spanner is about as sparse as the theoretically optimal Euclidean spanner with the same approximation quality. The set of points is chosen adversarially for a given ε.

Proof. (of Theorem 4.1) Take points spaced at least 4ϵ apart on the surface of the unit ball on the first $d/2$ dimensions. Then, take points spaced at least 4ϵ apart on the surface of the unit ball on the remaining $d/2$ dimensions. Let the first set of points be A, and the second set of points be B. You can pack $\Theta(\varepsilon^{-d/2+1})$ points into both A and B this way. Each distance crossing from A to B has Euclidean distance exactly equal to 2. Therefore, any edge from A to B must be in a $(1+\varepsilon)$ spanner of the Euclidean distance. We have constructed a set $P := A \cup B$ with $\Theta(\varepsilon^{-d/2+1})$ points, whose $(1+\varepsilon)$ Euclidean spanner must have at least $\Theta(\varepsilon^{-d+2})$ edges. This construction can have arbitrarily many points n, by duplicating $n\varepsilon^{d/2-1}$ copies of P arbitrarily far away from each other. The result has n vertices, and must have at least $\Omega(n\varepsilon^{-d/2+1})$ edges in any $(1+\epsilon)$ Euclidean spanner. □

By substituting $\sqrt{\varepsilon}$ for ε in the construction, we can additionally show a lower bound for the sparsity of an edge-squared spanner.

Lemma 4.2. *For constant d and any fixed ε, there exists a point set where a $(1+\varepsilon)$ edge-squared spanner must have at least $\Omega\left(n\varepsilon^{-\lfloor d/4\rfloor+1}\right)$ edges.*

The point set is chosen adversarially for a given ε. By setting $d = 4$ and $\varepsilon = \frac{1}{n}$, our construction gives:

Lemma 4.3. *There exists a 4-dimensional set of points, such that any 1-spanner of the edge-squared metric has $\Omega(n^2)$ edges.*

5 Exact-spanners of Edge-Squared Metrics in the Probability Density Setting

Theorem 1.6 states that for $k = O(2^d \log n)$, the k-NN graph of n points drawn i.i.d from a nicely behaved probability distribution is a 1-spanner of the edge-squared metric. Theorem 1.6 is impossible for general point sets: Lemma 4.3 gives an example where a 1-spanner of the edge-squared metric in 4 dimensions requires $\Omega(n^2)$ edges. This result is also impossible for Euclidean distances, whose 1-spanner is the complete graph almost surely. Our theorem implies any off-the-shelf k-nearest neighbor graph generator can compute edge-squared metric. In this section, we outline a proof and defer the analytical details to Appendix B.

First, let us assume that the support of our probability density D has the same dimension as our ambient space. This simplifies our calculations without changing the problem much. Then, we note that as our number of sample points get large, the density inside a k-NN ball around any point x (the ball with radius k^{th}-NN distance, center at x) looks like the uniform distribution on that ball, possibly intersected with a halfspace. The bounding plane of our halfspace represents the boundary of our density D.

For simplicity in the outline, let's suppose that D is convex. If we condition on the radius of the k-NN ball, then the $k-1^{st}$ nearest neighbors of x are distributed roughly according to the above distribution, described by the ball intersected with a halfspace. For any other point p in D, we project p onto the k-NN ball to point p', and show that the ball $p'x$ contains a k^{th}

nearest neighbor w.h.p, when $k = O(2^d \log n)$. This implies ball with diameter px contains a k^{th} nearest neighbor of x, and thus px is not necessary in any 1-spanner of the edge-squared metric. Then we take union bound over all x. A rigorous proof of Theorem 1.6 requires careful analysis, and is contained in Section B. Our proof can be tweaked to show:

Theorem 5.1. *Given a Lipschitz distribution bounded above and below with support on convex set $C \subset \mathbb{R}^d$, the k-NN graph is Gabriel w.h.p. for $k = O(2^d \log n)$.*

6 Relating the Edge-Squared Metric to Euclidean MSTs, Euclidean Spanners, and More

The edge-squared metric on a Euclidean point set, as we recall, is defined by taking the Euclidean distances squared and finding the shortest paths. We could have taken any such power p of the Euclidean distances. We will soon see that taking $p = 1$ gives us the Euclidean distance, and finding spanners of the graph as $\lim p \to \infty$ is the Euclidean MST problem. Let the p-power metric be defined on a Euclidean point set by taking Euclidean distances to the power of p, and performing all-pairs shortest path on the resulting distance graph.

Theorem 6.1. *For all $q > p$, any 1-spanner of the p-power metric is a 1-spanner of the q-power metric on the same point set*

Proof. A 1-spanner of the q-power metric can be made by taking edges uv where

$$\min_{p_0=u,\ldots p_k=v,k\neq 1} \sum_{k} ||p_i - p_{i-1}||^q > ||u - v||^q. \tag{6}$$

If $\sum_{i=1}^{k} ||p_i - p_{i-1}||^q > ||u - v||^q$ for any points $p_1, \ldots p_k$, then $\sum_{i=1}^{k} ||p_i - p_{i-1}||^p > ||u - v||^p$ for any $q > p$. Thus, for all such edges uv satisfying Equation 6:

$$\min_{p_0=u,\ldots p_k=v,k\neq 1} \sum_{k} ||p_i - p_{i-1}||^p > ||u - v||^p.$$

Such edges uv must be included in any 1-spanner of the p-power metric. □

Corollary 6.1.1. *Let P be a set of points in Euclidean space drawn i.i.d. from a Lipschitz probability density bounded above and below, with support on a smooth, compact manifold with intrinsic dimension d, bounded curvature, and smooth boundary of bounded curvature. Then the k-NN graph on P when $k = O(2^d \log n)$ is a 1-spanner of the p-power metric for every $p \geq 2$, w.h.p.*

This follows from combining Theorem 1.6 and Theorem 6.1.

6.1 Relation to the Euclidean MST problem

Definition 6.2. *Let the* ***normalized*** p***-power metric*** *between two points in* $\mathbb{R}^d$ *be the* p*-power metric between the two points, raised to the* $\frac{1}{p}$ *power. Define the normalized* ∞*-power metric as the limit of the normalized* p*-power metric as* $p \to \infty$*.*

Lemma 6.3. *The Euclidean MST is a* 1*-spanner for the normalized* ∞*-power metric.*

This lemma follows from basic properties of the MST. The normalized p-power metrics give us a suite of metrics such that $p = 1$ is the Euclidean distance and $p = \infty$ gives us the distance of the longest edge on the unique MST-path. Setting $p = 2$ gives the edge-squared metric, which sits between the Euclidean and max-edge-on-MST-path distance. Theorem 6.1 establishes that minimal 1-spanners of the (normalized) p-power metric are contained in each other, as p varies from 1 to ∞. The minimal spanner for a general point set when $p = 1$ is the complete graph, and the Euclidean MST is the minimal spanner for $p = \infty$. Thus:

Theorem 6.4. *For points in* $\mathbb{R}^d$*, every* 1*-spanner of the* p*-power metric on that set of points contains every Euclidean MST.*

Corollary 6.4.1. *Every* 1*-spanner for the edge-squared metric and/or Nearest Neighbor Geodesic contains every Euclidean MST.*

6.2 Generalizing Single Linkage Clustering, Level Sets, and k-Centers clustering

If our point set is drawn from a well-behaved probability density, then the normalized edge-power metrics converge to a nice geodesic distance detailed in [45]. When $p = 1$, clustering with this metric is the same as Euclidean metric clustering (k-means, k-medians, k-centers), and when $p = \infty$, clustering with this metric is the same as the widely used level-set method [76, 41, 35, 10]. Thus, clustering with normalized edge-power metrics generalizes these two very popular methods, and interpolates between their advantages. Definitions of the level-set method and a full discussion are contained in Appendix C

7 Conclusions and Open Questions

We examined a graph-based distance on Euclidean point sets, showed it equaled a special density-based distance, and built sparser and faster spanners on this metric than is known for Euclidean distances. Such sparse data structures may be surprising given that the metric can have high doubling dimension. Many problems remain open.

Is there a generalization of Theorem 1.4 to p-power metrics? This would require defining a new version of the nearest-neighbor geodesic distance. Separately, are the proof techniques for Theorem 1.4 of use for computing or approximating other density-based distances? Can non-spanner data structures for clustering with the edge-squared metric be computed efficiently? Such data structures include core-sets and distance oracles [13, 62, 68].

Can we efficiently compute $o(\log n)$-spanners of the p-power metric in high dimension with a nearly linear number of edges? The existence of such spanners has been studied for Euclidean metrics in [42], where the stretch obtained is $\sqrt{\log n}$. Good constructions for $(1+\varepsilon)$-spanners of the normalized ∞-power metric are known: many (but not all) approximate Euclidean MST constructions are $(1+\varepsilon)$-spanners of this metric [21, 77]. Can high-dimensional approximate Euclidean MST algorithms [5, 77, 9] be adapted to create efficient p-power spanners? Any spanner for high-dimensional edge-squared metrics must give the same quality spanner for negative type distances [59, 31], which include l_2 and l_1.

Does computing k-NN graphs with approximate nearest neighbor methods give 1-spanners of the edge-squared metric with high probability? Approximate nearest neighbors have been studied extensively [52, 26, 32], including locality-sensitive hashing for high dimensional point sets [7] and more [48]. Recent work by Andoni et. al. [8] showed how to compute approximate nearest neighbors for any non-Euclidean norm. Perhaps there is a rigorous theory about density-sensitive metrics generated from any such norm? Similar to how the edge-squared metric is generated from the Euclidean distance.

It remains an open question how well clustering or classification with edge-squared metrics and nearest-neighbor geodesic distances performs on real-world data. Experiments have been done by Bijral, Ratliff, and Srebro in [16]. Theorem 1.6 implies that future experiments can be done using any k-nearest-neighbor graph.

References

[1] Mridul Aanjaneya, Frédéric Chazal, Daniel Chen, Marc Glisse, Leonidas Guibas, and Dmitriy Morozov. Metric graph reconstruction from noisy data. *International Journal of Computational Geometry and Applications (IJCGA)*, 22(04):305–325, 2012.

[2] Michal Adamszek, Henry Adams, Florian Frick, Chris Peterson, and Corrine Previte-Johnson. Nerve complexes of circular arcs. *Discrete & Computational Geometry*, 56(2):251–273, 2016.

[3] Pankaj K. Agarwal, Kyle Fox, and Oren Salzman. An efficient algorithm for computing high quality paths amid polygonal obstacles. In *Proceedings of the 27th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1179–1192, 2016.

[4] Morteza Alamgir and Ulrike von Luxburg. Shortest path distance in random k-nearest neighbor graphs. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

[5] Joshua Alman, Timothy M. Chan, and Ryan Williams. Polynomial representations of threshold functions and algorithmic applications. In *57th Annual Symposium on Foundations of Computer Science (FOCS 2016)*, FOCS 2016, 2016. URL: `https://arxiv.org/abs/1608.04355`.

[6] Nina Amenta and Marshall Bern. Surface reconstruction by Voronoi filtering. *Discrete & Computational Geometry*, 22:481–504, 1999.

[7] Alexandr Andoni, Piotr Indyk, Thijs Laarhovn, Ilya Razenshteyn, and Ludwig Schmidt. Practical and optimal lsh for angular distance. 29th Annual Conference on Neural Information Processing Systems (NIPS), 2015. URL: `https://arxiv.org/abs/1509.02897`.

[8] Alexandr Andoni, Assaf Naor, Aleksandar Nikolov, Ilya Razenshteyn, and Erik Waingarten. Navigating nets: Simple algorithms for proximity search. In *59th Annual Symposium on Foundations of Computer Science (FOCS)*, 2018. URL: `https://ilyaraz.org/static/papers/daher.pdf`.

[9] Alexandr Andoni, Aleksandar Nikolov, Krzysztof Onak, and Grigory Yaroslavtsev. Parallel algorithms for geometric graph problems. STOC 2014. ACM. URL: `https://arxiv.org/abs/1401.0042`.

[10] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, and Jorg Sander. Optics: Ordering points to identify cluster structure. ACM SIGMOD International Conference on Management of Data, 1999. URL: `http://www.dbs.ifi.lmu.de/Publikationen/Papers/OPTICS.pdf`.

[11] Sunil Arya, Gautam Das, David M. Mount, Jeffrey S. Salowe, and Michiel Smid. Euclidean spanners: Short, thin, and lanky. In *Proceedings of the Twenty-seventh*

Annual ACM Symposium on Theory of Computing, STOC '95, pages 489–498, New York, NY, USA, 1995. ACM. URL: `http://doi.acm.org/10.1145/225058.225191`, `doi:10.1145/225058.225191`.

[12] Sunil Arya and David M. Mount. A fast and simple algorithm for computing approximate euclidean minimum spanning trees. In *Proceedings of the Twenty-seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '16, pages 1220–1233, Philadelphia, PA, USA, 2016. Society for Industrial and Applied Mathematics. URL: `http://dl.acm.org/citation.cfm?id=2884435.2884520`.

[13] Mihai Bādoiu, Sariel Har-Peled, and Piotr Indyk. Approximate clustering via coresets. In *Proceedings of the Thiry-fourth Annual ACM Symposium on Theory of Computing*, STOC '02, pages 250–257, New York, NY, USA, 2002. ACM. URL: `http://doi.acm.org/10.1145/509907.509947`, `doi:10.1145/509907.509947`.

[14] Paul Balister, Bela Bollobas, Amites Sarkar, and Mark Walters. Connectivity of random k-nearest-neighbour graphs. *Advances in Applied Probability*, 37(1):1–24, 2005. URL: `http://www.jstor.org/stable/30037313`.

[15] Johann Bernoulli. Brachistochrone problem. *Acta Eruditorum*, June 1696.

[16] Avleen Singh Bijral, Nathan D. Ratliff, and Nathan Srebro. Semi-supervised learning with density based distances. In Fabio Gagliardi Cozman and Avi Pfeffer, editors, *UAI*, pages 43–50. AUAI Press, 2011.

[17] Karol Borsuk. On the imbedding of systems of compacta in simplicial complexes. *Fund. Math.*, 35:217–234, 1948.

[18] Ulrich Brehm. Extensions of distance reducing mappings to piecewise congruent mappings on ℓ_m. *Journal of Geometry*, 16(1):187–193, 1981.

[19] M.R. Brito, E.L. Chvez, A.J. Quiroz, and J.E. Yukich. Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection. *Statistics and Probability Letters*, 35(1):33 – 42, 1997. URL: `http://www.sciencedirect.com/science/article/pii/S0167715296002131`, `doi:https://doi.org/10.1016/S0167-7152(96)00213-1`.

[20] Paul B. Callahan and S. Rao Kosaraju. Faster algorithms for some geometric graph problems in higher dimensions. In *Proceedings of the Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '93, pages 291–300, Philadelphia, PA, USA, 1993. Society for Industrial and Applied Mathematics. URL: `http://dl.acm.org/citation.cfm?id=313559.313777`.

[21] Paul B. Callahan and S. Rao Kosaraju. A decomposition of multidimensional point sets with applications to k-nearest-neighbors and n-body potential fields. *J. ACM*, 42(1):67–90, January 1995. URL: `http://doi.acm.org/10.1145/200836.200853`, `doi:10.1145/200836.200853`.

[22] Gunnar Carlsson. Topology and data. *Bull. Amer. Math. Soc.*, 46:255–308, 2009.

[23] Nicholas J. Cavanna and Donald R. Sheehy. Adaptive metrics for adaptive samples. In *Proceedings of the Canadian Conference on Computational Geometry*, 2016.

[24] Frédéric Chazal and Steve Y. Oudot. Towards persistence-based reconstruction in Euclidean spaces. In *Proceedings of the 24th ACM Symposium on Computational Geometry*, pages 232–241, 2008.

[25] Frédéric Chazal Chazal, Leonidas J. Guibas, Steve Y. Oudot, and Primoz Skraba. Persistence-based clustering in riemannian manifolds. *J. ACM*, 60(6:41):97–106, 2013. URL: `http://doi.acm.org/10.1145/2535927`, `doi:10.1145/2535927`.

[26] Jie Chen, Hawren Fang, and Yousef Saad. Fast approximate knn graph construction for high dimensional data via recursive lanczos bisection. Journal of Machine Learning Research, 2009.

[27] P Chew. There is a planar graph almost as good as the complete graph. In *Proceedings of the Second Annual Symposium on Computational Geometry*, SCG '86, pages 169–177, New York, NY, USA, 1986. ACM. URL: `http://doi.acm.org/10.1145/10515.10534`, `doi:10.1145/10515.10534`.

[28] Michael B. Cohen, Brittany Terese Fasy, Gary L. Miller, Amir Nayyeri, Donald R. Sheehy, and Ameya Velingker. Approximating nearest neighbor distances. In *Proceedings of the Algorithms and Data Structures Symposium*, 2015.

[29] Samuel I. Daitch, Jonathan A. Kelner, and Daniel A. Spielman. Fitting a graph to vector data. ICML '09, 2009.

[30] T. K. Dey. *Curve and Surface Reconstruction : Algorithms with Mathematical Analysis.* Cambridge University Press, 2007.

[31] Michel Marie Deza and Monique Laurent. *Geometry of Cuts and Metrics*. Number ISBN 978-3-642-04295-9. Springer, 1997.

[32] Wei Dong, Moses Charikar, and Kai Li. Efficient k-nearest neighbor graph construction for generic similarity measures. Proceeding of the International Conference on World Wide Web, pages 577-586, 2011.

[33] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. *Discrete & Computational Geometry*, 4(28):511–533, 2002.

[34] Michael Elkin and Shay Solomon. Optimal euclidean spanners: Really short, thin and lanky. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, STOC '13, pages 645–654, New York, NY, USA, 2013. ACM. URL: `http://doi.acm.org/10.1145/2488608.2488691`, `doi:10.1145/2488608.2488691`.

[35] Martin Ester, Hans-Peter Kriegel, Jorg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pages 226–231. AAAI Press, 1996. URL: `http://dl.acm.org/citation.cfm?id=3001460.3001507`.

[36] Ellen Gasparovic, Maria Gommel, Emilie Purvine, Bei Wang, Yusu Wang, and Lori Ziegelmeier. A complete characterization of the 1-dimensional intrinsic cech persistence diagrams for metric graphs. https://arxiv.org/abs/1702.07379, 2017.

[37] Jose Maria Gonzalez-Barrios and Aldofo J. Quiroz. A clustering procedure based on the comparison between the k nearest neighbors graph and the minimal spanning tree. Statistics and Probability Letters. Elsevier, 2003. URL: `https://www.sciencedirect.com/science/article/pii/S0167715202004212`.

[38] A.N. Gorban, B. Kégl, D.C. Wunsch, and A. Zinovyev, editors. *Principal Manifolds for Data Visualization and Dimension Reduction*, volume 58 of *LNCSE*. Springer, 2007.

[39] Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Efficient classification for metric data. IEEE Trans. Information Theory 60(9), 2014.

[40] Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Efficient regression in metric spaces via approximate lipschitz extension. IEEE Trans. Information Theory 63(8) 4348-4849, 2017.

[41] J. C. Gower and G. J. S. Ross. Minimum spanning trees and single linkage cluster analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 18(1):54–64, 1969. URL: `http://www.jstor.org/stable/2346439`.

[42] Sariel Har-Peled, Piotr Indyk, and Anastasios Sidiropoulos. Euclidean spanners in high dimensions. In *SODA*, 2013.

[43] Tatsunori B. Hashimoto, Yi Sun, and Tommi S. Jaakkola. From random walks to distances on unweighted graphs. NIPS 2015, 2015.

[44] Matthias Hein, Jean-Yves Audibert, and Ulrike von Luxburg. Graph laplacians and their convergence on random neighborhood graphs. *Journal of Machine Learning Research*, 8:1325–1368, 2007.

[45] Sung Jin Hwang, Steven B. Damelin, and Alfred O. Hero III. Shortest path through random points. *Ann. Appl. Probab.*, 26(5):2791–2823, 10 2016. URL: `http://dx.doi.org/10.1214/15-AAP1162`, `doi:10.1214/15-AAP1162`.

[46] R. Kimmel and J. A. Sethian. Computing geodesic paths on manifolds. *Proceedings of the National Academy of Sciences (PNAS)*, 17, 1998. `doi:10.1073/pnas.95.15.8431`.

[47] M. Kirszbraun. Über die zusammenziehende und lipschitzsche transformationen. *Fundamenta Mathematicae*, 22(1):77–108, 1934. URL: `http://eudml.org/doc/212681`.

[48] Thijs Laarhoven. Graph-based time-space trade-offs for approximate near neighbors. In *Symposium on Computational Geometry*, 2018.

[49] James R. Lee and Assaf Naor. Extending lipschitz functions via random metric partitions. Invent. Math 160, 1, 59–95, 2005.

[50] Xiang-Yang Li, Peng-Jun Wan, and Yu Wang. Power efficient and sparse spanner for wireless ad hoc networks. In *Proceedings Tenth International Conference on Computer Communications and Networks*, 2001.

[51] Xiang-Yang Li, Peng-Jun Wan, Yu Wang, and O. Frieder. Sparse power efficient topology for wireless networks. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences*, 2002.

[52] Ting Liu, Andrew W. Moore, Alexander Gray, and Ke Yang. An investigation of practical approximate nearest neighbor algorithms. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, NIPS'04, pages 825–832, Cambridge, MA, USA, 2004. MIT Press. URL: `http://dl.acm.org/citation.cfm?id=2976040.2976144`.

[53] Ulrike Von Luxburg. Tutorial on spectral clustering. Statistics and Computing 17(4), 2007.

[54] J.B. MacQueen. Some methods for classification and analysis of multivariate observations. pages 281–297, 1967.

[55] James McQueen, Marina Meilă, Jacob VanderPlas, and Zhongyue Zhang. Megaman: Scalable manifold learning in python. *Journal of Machine Learning Research*, 17:1–5, 2016.

[56] Assaf Naor. *Metric Embeddings and Lipschitz Extensions.* Princeton University, Department of Mathematics, 2015.

[57] Assaf Naor, Yuval Peres, Oded Schramm, and Scott Sheffield. Markov chains in smooth banach spaces and gromov-hyperbolic metric spaces. *Duke Mathematical Journal*, 134:165–197, 2006. URL: `https://projecteuclid.org/euclid.dmj/1152018507`, `doi:doi:10.1215/S0012-7094-06-13415-4`.

[58] Sajama and Alon Orlitsky. Estimating and computing density based distance metrics. In *ICML '05*, pages 760–767, New York, NY, USA, 2005. ACM. URL: `http://doi.acm.org/10.1145/1102351.1102447`, `doi:10.1145/1102351.1102447`.

[59] I. J. Schoenberg. On certain metric spaces arising from euclidean spaces by a change of metric and their imbedding in hilbert space. *Annals of Mathematics*, 38(4):787–793, 1937. URL: `http://www.jstor.org/stable/1968835`.

[60] Karl Schwarzschild. Uber das gravitationsfeld eines massenpunktes nach der einsteinschen theorie. *Sitzungsberichte der Deutschen Akademie der Wissenschaften zu Berlin*, 1916.

[61] R. Sharathkumar and Pankaj K. Agarwal. Algorithms for the transportation problem in geometric settings. In *Proceedings of the Twenty-third Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '12, pages 306–317, Philadelphia, PA, USA, 2012. Society for Industrial and Applied Mathematics. URL: `http://dl.acm.org/citation.cfm?id=2095116.2095145`.

[62] Christian Sohler and David Woodruff. Strong coresets for k-median and subspace approximation, goodbye dimension. In *Foundations of Computer Science (FOCS)*, 2018.

[63] Prashant Sridhar. An experimental study into spectral and geometric approaches to data clustering. Master's thesis, Carnegie Mellon University, October 2015. CMU CS Tech Report CMU-CS-15-149.

[64] Werner Stuetzle. Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. *Journal of Classification*, 20(1):025–047, May 2003. URL: `https://doi.org/10.1007/s00357-003-0004-6`, `doi:10.1007/s00357-003-0004-6`.

[65] Werner Stuetzle. A generalized single linkage method for estimating the cluster tree of a density. 2007.

[66] H. J. Sussmann and J. C. Willems. 300 years of optimal control: from the brachystochrone to the maximum principle. *IEEE Control Systems Magazine*, 17(3):32–44, June 1997. `doi:10.1109/37.588098`.

[67] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[68] Mikkel Thorup and Uri Zwick. Approximate distance oracles. *J. ACM*, 52(1):1–24, January 2005. URL: `http://doi.acm.org/10.1145/1044731.1044732`, `doi:10.1145/1044731.1044732`.

[69] Daniel Ting, Ling Huang, and Michael Jordan. An analysis of the convergence of graph laplacians. In *Proceedings of the 27th International Conference on Machine Learning*, pages 1079–1086, 2010.

[70] Pravin M. Vaidya. A sparse graph almost as good as the complete graph on points in k dimensions. *Discrete & Computational Geometry*, 6(3):369–381, Sep 1991. URL: `https://doi.org/10.1007/BF02574695`, `doi:10.1007/BF02574695`.

[71] F. A. Valentine. A Lipschitz condition preserving extension for a vector function. *American Journal of Mathematics*, 67(1):83–93, 1945. URL: `http://www.jstor.org/stable/2371917`.

[72] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[73] Pascal Vincent and Yoshua Bengio. Density sensitive metrics and kernels. In *Snowbird Workshop*, 2003.

[74] Ulrike von Luxburg and Morteza Alamgir. Density estimation from unweighted k-nearest neighbor graphs: A roadmap. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'13, pages 225–233, USA, 2013. Curran Associates Inc. URL: `http://dl.acm.org/citation.cfm?id=2999611.2999637`.

[75] Ulrike von Luxburg and Olivier Bousquet. Distance based classification with lipschitz functions. Journal of Machine Learning Research 5: 669-695, 2004.

[76] D. Wishart. Mode analysis: A generalization of the nearest neighor which reduces chaining effects. 1969.

[77] Grigory Yaroslavtsev and Adithya Vadapalli. Massively parallel algorithms and hardness for single linkage clustering under l_p distances. Arxiv. ACM, 2017. URL: `https://arxiv.org/pdf/1710.01431`.

A Proving Faster and Sparser-than-Euclidean Approximate Spanners

In this appendix, we finish the proof of Theorem 1.5 based on the outline given in Section 4.

A.1 $1 + O(\delta^2)$ spanners can be generated from a $1/\delta$ WSPD

Definition A.1. *Let e be a **critical** edge in a shortest path metric on any graph if the (possibly-not-unique) shortest path between the endpoints of e is the edge e.*

Lemma A.2. *The set of critical edges on any graph forms a* 1*-spanner of the shortest path metric.*

The above lemma is known in the literature.

To check that any graph H is a $(1 + O(\delta^2)$ spanner of any graph G, it suffices to prove that all critical edges in the edge-squared metric have a stretch no larger than $1 + O(\delta^2)$. Let G be the edge-squared graph arising from points $P \subset \mathbb{R}^d$. Build a well-separated pair decomposition on P, with pairs given as $\{A_1, B_1\}, \{A_2, B_2\}, \ldots \{A_m, B_m\}$. Create a spanner

H as follows: for each pair $\{A_i, B_i\}$, connect an edge $\{a, b\}, a \in A_i, b \in B_i$ such that the Euclidean distance between a and b is a $(1 + c\delta^2)$ approximation of the shortest distance between point sets A_i and B_i, for some constant c independent of i. This can be accomplished in $O(1)$ time assuming a preprocessing step of $O(\delta^{-d} \log\left(\frac{1}{\delta}\right)$ time, as noted in Callahan's paper on constructing a Euclidean MST [21]. Do this for all $1 \leq i \leq m$.

For each critical edge (s, t), consider the well-separated pair $\{A, B\}$ that (s, t) is part of. Let $s \in A$ and $t \in B$. Let (a, b) be a $(1 + c\delta^2)$-approximate shortest edge between A and B $(a \in A, b \in B)$. Scale $||a - b||_2$ to be 1. A and B have Euclidean radius at most δ, by the definition of a well separated pair. By induction on Euclidean distance, H is an edge-squared 2-spanner of the edge-squared metric for all points in A and B and all points in B (assuming sufficiently small δ).

Lemma A.3.

$$dist_H(s,t) \leq dist_H(s,a) + dist_H(a,b) + dist_H(b,t) \leq 1 + O(\delta^2)$$

Proof. We know $dist_H(a, b) = 1$ by our scaling, and

$$dist_H(s,a) \leq 2 \cdot (dist_G(s,a)) \leq 2 \cdot ||s-a||^2 \leq 8\delta^2$$

The first inequality follows by the inductive hypothesis that H is a 2-spanner of G in A. The third inequality follows since both s and a are contained in a ball of radius δ.

The same bound applies for $dist_H(b, t)$. □

Lemma A.4.

$$(1 + c\delta^2)(dist_G(s,t)) \geq dist_G(a,b) = 1$$
$$\Rightarrow dist_G(s,t) \geq \frac{1}{1 + c\delta^2}$$

Lemma A.4 follows from the fact that (a, b) is a $(1 + c\delta^2)$ approximate shortest distance between A and B.

Therefore

$$stretch_H(s,t) \leq \frac{dist_H(s,t)}{dist_G(s,t)} \leq (1 + 16\delta^2)(1 + c\delta^2) = 1 + O(\delta^2) \tag{7}$$

Thus we have proven that H is a $1 + 16\delta^2$ spanner. Now set $\epsilon = \delta^2$, which completes proof of Theorem 1.5.

B Spanners in the Probability Density Setting: Full Proof

Through this section, we assume that D is a probability density function with support on smooth connected compact manifold with intrinsic dimension d embedded in ambient space

$\mathbb{R}^s$, with smooth boundary of bounded curvature. This probability density function is further assumed to be bounded above and below, and to be Lipschitz. For simplicity, we assume that $s = d$, and we can prove all our results when $s > d$ by taking coordinate charts from the manifold into Euclidean space. We will show at the end of the section that if the distribution is supported on a convex set of full dimension in the ambient space, then the k-NN graph is Gabriel for the same k. It is not difficult to see that Gabriel graphs are 1-spanners of the edge-squared metric [63].

Lemma B.1. *Let M be a compact object in $\mathbb{R}^d$, whose boundary is a smooth manifold of dimension $d-1$ with bounded curvature. Let B be any ball with sufficiently small radius r_B with center in M, that intersects the boundary of D at some point x. Let H be the halfspace tangent to M at x containing the center of the ball.*

For any point $Q \in M$, let Q' be the point in B closest to Q. If $d(Q', H)/r_B > c$ for arbitrary constant c, then $d(Q, H) \geq c'$ for some constant c'.

This is a basic fact about the smoothness and bounded curvature of the boundary.

Lemma B.2. *Pick n points from D. W.h.p, any two points in Support(D) with Euclidean distance $\geq \Omega(1)$ have nearest-neighbor geodesic distance of $o(1)$.*

This is implicit in [45].

Lemma B.3. *For any ball B with center O and any point Q' on the boundary of B, let $B_{Q'O}$ be the ball with diameter $Q'O$. Let H be any halfspace containing O. If $d(Q', H)/r_B \leq c$ for some constant c possibly depending on the dimension d, then $\text{Vol}(B_{Q'O} \cap H) \geq \frac{1-c'}{2^d}\text{Vol}(B \cap H)$ for some constant c', where c' goes to 0 as c goes to 0.*

Proof. First, let us consider the case where $d(Q'H) = 0$, that is, Q' is contained in halfspace H'. In this case, dilating $B_{Q'O} \cap H$ by a factor of 2 about point Q' gives a superset of $B \cap H$, as $B_{Q'O}$ maps to B and H maps to a halfspace strictly containing H. In this case, $\text{Vol}(B_{Q'O} \cap H) \geq \frac{1}{2^d}\text{Vol}(B \cap H)$ as desired. The case when $d(Q', H)/r_B$ is bounded follows in a straightforward manner. □

This leads us to our following theorem:

Theorem B.4. *For any n point set P picked i.i.d from D, consider any point O. Let B be the k-NN ball of O. Let $Q \in$ Support(D) be any point outside B, and let the closest point to Q in B be Q'. For a point x inside B on the boundary of D (assuming such a point exists), let H be the tangent halfplane containing the center of B.*

Then: either $d(Q', H)/r_B \leq c'$ for some constant c' or there exists a constant c where $|QO| > c$. Here, c and c' are independent of the number of points chosen, and c' can be set arbitrarily small.

In the latter case, w.h.p. QO is not in the edge-squared 1-spanner. In the former case, setting c' to be a very small constant ϵ lets us say:

$$\mathrm{Vol}(B_{Q'O} \cap H) \geq \frac{1-\epsilon}{2^d}\mathrm{Vol}(B \cap H), \tag{8}$$

or equivalently:

$$\begin{aligned} &\mathop{\mathbb{P}}_{x \sim D}[x \in B_{QO} | x \in B] && (9)\\ \geq &\mathop{\mathbb{P}}_{x \sim D}[x \in B_{Q'O} | x \in B] && (10)\\ \geq &\frac{1-\varepsilon-o(1)}{2^d} && (11) \end{aligned}$$

Expression 10 $>$ Expression 11 follows from Equation 8, and the fact that the radius of the k-NN ball goes to 0 as n gets large, and thus the probability density of sampling x from D conditioned on x being in B approaches the uniform density in $B \cap \mathit{Support}(D)$. Also, $B \cap H$ approaches $B \cap \mathit{Support}(D)$ as the radius of B goes to 0.

Expression 9 $>$ Expression 10 since $B_{QO} \supset B_{Q'O}$. (Here, the k-NN ball B w.r.t. point O is defined as the ball centered at O with radius equal to the distance of the k^{th} nearest neighbor to O).

Note that the $k-1$ nearest neighbors of O, conditioned only on the radius of B, are distributed equivalently to $k-1$ i.i.d samples of D conditioned on containment in B. It follows that for any point Q outside B and in the support of D, where $|QO| < c$:

$$\mathop{\mathbb{P}}_{P \sim D^k}[QO \text{ is not Gabriel w.r.t. } P | Q \notin B] \geq 1 - \left(1 - \frac{1-\varepsilon-o(1)}{2^d}\right)^k \tag{12}$$

Thus, setting $\epsilon = 0.1$ and $k > O(\log n/2^d)$, and factoring in the case where $|QO| > c$, then w.h.p.:

$$\mathop{\mathbb{P}}_{P \sim D^k}[QO \text{ is not critical w.r.t. } P | Q \notin B]$$

Here, we recall that an edge AB is Gabriel with respect to a point set P if and only if B_{AB} does not contain any points in P. Note that every non-Gabriel edge is non-critical, where a critical edge is an edge that must be in the 1-spanner (as in Definition A.1). Thus taking the union bound over $Q, O \in P$ gives us that no edge outside the k-NN graph is critical w.h.p, and thus the k-NN graph contains all critical edges and is a 1-spanner w.h.p.

This proves Theorem 1.6 when the support of D has the same intrinsic dimension as the ambient space. If the support of D has dimension $d < d'$ (where d' is the ambient dimension of the space), simply take coordinate charts from D onto $\mathbb{R}^d$ and the previous arguments will still carry through . We should note that if no point x inside B on the boundary of D exists, then we can ignore H and all the steps of the proof still follow.

C Edge-Power Metrics relate to Single Linkage Clustering, Level Sets, and k-Centers clustering

Many popular clustering algorithms, including k-centers, k-means, and k-medians clustering, use Euclidean distance as a measure of distance between points in $\mathbb{R}^d$. These methods are useful when clusters are spherical and well-separated. However, it is believed by practitioners that density-sensitive distances more accurately capture intrinsic distances between data [4].

The celebrated single-linkage clustering algorithm [41, 77], which is clustering based on an MST, is a widely used tool in machine learning, and gets around many of the problems of the Euclidean distance clustering. In single-linkage clustering, two points are considered similar if the maximum length edge on the path between them in the MST is small. This turns out to be equivalent to computing the normalized ∞-power metric between the two points. Therefore, single linkage clustering can be seen as clustering using the normalized ∞-power metric. Generally, normalized p-power metrics can be seen as an intermediary between Euclidean distances (1-power metrics) and Euclidean MST-based clustering.

Clustering with p-power metric relates to another popular clustering method in machine learning, known as level-set clustering. Loosely speaking, level set clustering involves finding an estimate for the probability density that points are drawn from, finding a cut threshold t, and then taking as clusters all regions with probability density $> t$. Level set clustering has appeared in many incarnations [76, 64, 65], including the celebrated and widely used DBScan method [35] and its considerable number of variations [10]. It is known that level-set clustering is related to single-linkage clustering, as the latter is an approximation of the former [76, 65]. Level-set methods have the advantage that they can find arbitrarily shaped clusters [35], but can cause two points that are very close in Euclidean distance to be considered far apart.

Clustering with the p-power metric incorporates the advantages of both Euclidean distance clustering and level set clustering, as it is both density-sensitive and takes into account overall Euclidean distance between two points. Here, p can be toggled to change the sensitivity of the metric to the underlying density. As the number of samples drawn from our probability density grows large, it has been proven that the behavior of normalized p-power metrics converges to a natural geodesic distance on the underlying probability density [45]. Clustering with this geodesic distance for $p = 1$ is exactly Euclidean clustering, and for $p = \infty$ is exactly the level set method. Thus, clustering with p-power metric converges to a clustering method that smoothly interpolates between Euclidean-distance clustering and level set clustering.