COMPRESSIBILITY BARRIERS TO NEIGHBORHOOD-PRESERVING DATA VISUALIZATIONS

Szymon Snoeck
Applied Mathematics Department
Columbia University
New York, NY 10027
sgs2179@columbia.edu

Noah Bergam
Computer Science Department
Columbia University
New York, NY 10027
njb2154@columbia.edu

Nakul Verma
Computer Science Department
Columbia University
New York, NY 10027
verma@cs.columbia.edu

ABSTRACT

To what extent is it possible to visualize high-dimensional datasets in a two- or three-dimensional space? We reframe this question in terms of embedding n-vertex graphs (representing the neighborhood structure of the input points) into metric spaces of low doubling dimension d, in such a way that maintains the separation between neighbors and non-neighbors. This seemingly lax embedding requirement is surprisingly difficult to satisfy. Our investigation shows that an overwhelming fraction of graphs require $d = \Omega(\log n)$. Even when considering sparse regular graphs, the situation does not improve, as an overwhelming fraction of such graphs requires $d = \Omega(\log n / \log\log n)$. The landscape changes dramatically when embedding into normed spaces. In particular, all but a vanishing fraction of graphs demand $d = \Theta(n)$. Finally, we study the implications of these results for visualizing data with intrinsic cluster structure. We find that graphs produced from a planted partition model with k clusters on n points typically require $d = \Omega(\log n)$, even when the cluster structure is salient. These results challenge the aspiration that constant-dimensional visualizations can faithfully preserve neighborhood structure.

1 Introduction

Visualizing a 10,000-point, 1,000-dimensional dataset in a two-dimensional plot is a bold pursuit. It is also a common practice across natural and social scientific research literature, from biology to economics to physics, where colorful UMAP and t-SNE plots have become a standard feature of data analysis [Kobak and Berens, 2019, Dimitriadis et al., 2018, Han et al., 2021]. In light of the well-known impossibility of low-distortion metric embeddings in constant dimensions (see e.g. Chapter 15 of Matoušek [2013]), the putative theoretical justification of these extreme dimension reduction procedures is that they need not preserve all minutiae of the input data. Instead, the argument goes, highlighting only the most basic structures, like local neighborhoods of points, is enough for most exploratory data analysis purposes. We study the conditions under which such structure-preserving embeddings are possible.

We demonstrate that, in many scenarios of practical interest, it is impossible to embed a point cloud in *any* metric space (let alone a Euclidean space) of constant dimension while preserving neighborhood structure.

Our analysis begins by re-framing low-dimensional data visualization in terms of embedding the neighbor graphs of the input point cloud into metric spaces. Let V be a size-n set representing our data points of interest. Let $\mathcal{G}_n(V)$ or simply $\mathcal{G}_n$ for short denote the set of all n-vertex unweighted, undirected, simple graphs on V. Let $(\mathcal{X}, \rho)$ be a target metric space of interest with doubling dimension $d = \dim(\mathcal{X})$ (see Section 3 for a definition). We think of a map $f : \mathcal{G}_n \to \mathcal{X}^n$ as a *data-visualization algorithm* and $f(G)$ as an *embedding* of a specific graph $G = (V, E) \in \mathcal{G}_n$ into $\mathcal{X}$ via the algorithm f. Out of convenience, we write $f_G(v)$ to denote the point in $f(G)$ to which the vertex v of G is being mapped. If there is an edge between u and v in G, we write $u \sim_G v$, or simply $u \sim v$ when clear from context (symmetrically, we write $u \not\sim v$ if there is no direct edge).

Within this framework, we study data visualization algorithms that are faithful to the neighborhood structure of input graphs, in the sense of keeping neighbors close and non-neighbors far. We define this desideratum below, and then study the minimum doubling dimension of the output metric space, $\mathcal{X}$, necessary to accommodate it.

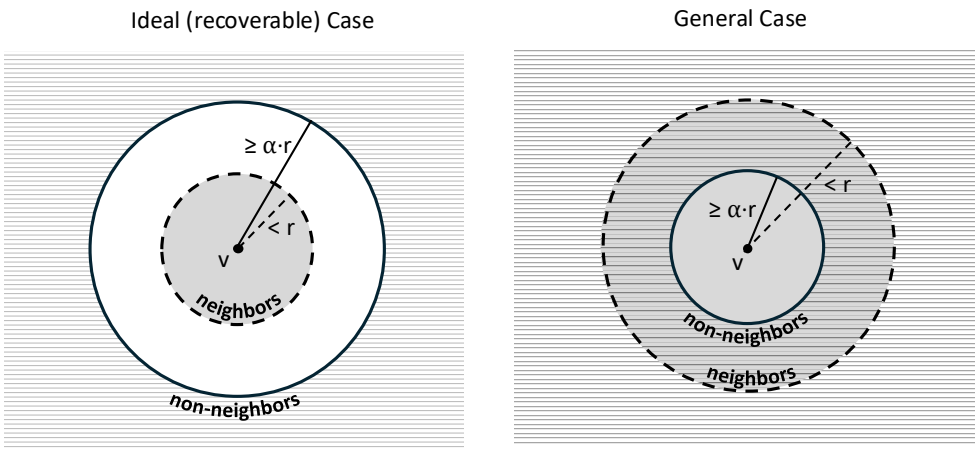

Figure 1: An f-embedding of a vertex $v \in V$. Left: non-neighbor and neighbor relations of v are recoverable by thresholding at r ($\alpha \geq 1$). Right: the more relaxed case of preservation where overlap is allowed ($\alpha < 1$).

Definition 1. *Fix $\alpha \geq 0$. A data visualization algorithm $f : \mathcal{G}_n \to \mathcal{X}^n$ is said to α-preserve $G \in \mathcal{G}_n$ if there exists $r = r_G > 0$ (the neighborhood threshold) such that for all distinct $u, v \in V$,*

$$(1)\ u \sim v \implies \rho(f_G(u), f_G(v)) < r, \qquad \textit{(neighbor proximity)}$$

$$(2)\ u \nsim v \implies \rho(f_G(u), f_G(v)) \geq \alpha \cdot r. \qquad \textit{(non-neighbor separation)}$$

Note that when $\alpha \geq 1$, one can recover the input graph G from the embedding $f(G)$ by simply drawing an edge between any two vertices embedded within a distance of r. We call this special case of preservability *recoverability*, and we notice that, depending on the setting, this distinction can have a sharp impact on the difficulty of preservability. Furthermore, preservability (including the special case of recoverability) stands out from other notions of structure preserving embeddings, such as ordinal embeddings, low-distortion embeddings, etc., due to its strictly local nature; it only requires the neighbor structure to be preserved, while non-neighbor structure can be distorted arbitrarily. Interestingly, while it seems like the locality of α-preservation should make it easier to satisfy from a compressibility perspective, we find that α-preserving a typical graph even in general metrics is as hard as near-isometric embedding of points in ℓ_2 (typically considered very rigid for embedding purposes) in that both require $\Omega(\log n)$ dimensions [Larsen and Nelson, 2017].

We are interested in characterizing the minimum dimension necessary to α-preserve a graph. This is formalized in what we call the α-preservation dimension.

Definition 2 (α-preservation dimension)**.** *Fix $\alpha \geq 0$ and $n \in \mathbb{N}$. Let $\mathbb{X}$ be a collection of metric spaces of interest. The α-preservation dimension of G in $\mathbb{X}$ is given by*[1]

$$\dim_\alpha(G, \mathbb{X}) := \min\{\dim(\mathcal{X}) \mid \mathcal{X} \in \mathbb{X} \textit{ and there exists } f : \mathcal{G}_n \to \mathcal{X}^n \textit{ which } \alpha\textit{-preserves } G\}. \tag{1}$$

In other words, it is the smallest $d \in \mathbb{N}$ such that there exists a metric space $\mathcal{X} \in \mathbb{X}$ of doubling dimension d that α-preserves G. If $\mathbb{X}$ is the collection of all possible metric spaces, we shorten the above to $\dim_\alpha(G)$.

Our notion of preservation dimension can be understood as a natural generalization of a well-studied graph invariant known as *sphericity* [Maehara, 1984]. The sphericity of a graph G is the minimum dimension d such that the vertices of G can be distinctly embedded into a d-dimensional Euclidean space such that the embedded vertices are distance at most 1 if and only if they are edge-connected. Preservation dimension relaxes this notion by (1) parametrizing the separation between neighbors versus non-neighbors via α (see Figure 1) and (2) allowing for embeddings into general metric spaces. In developing this generalization of sphericity, we obtain a more fine-grained understanding of structure-preserving metric embeddings of graphs.

[1] In the degenerate case when no such f exists, $\dim_\alpha(G, \mathbb{X})$ is undefined.

Our main results are as follows.

- **Preservation in General Metrics**. Though certain kinds of graphs are easily α-preserved in constant dimensions (e.g. cliques, cycles, paths, etc.), we show that these "easy" graphs comprise a vanishing fraction of all graphs: an overwhelming fraction of $G \in \mathcal{G}_n$ require $\dim_\alpha(G) = \Omega\big(\log(n)/\log(\frac{8}{\alpha})\big)$, see Theorem 8(i). Even if we consider only constant-degree regular graphs, a natural model for neighborhood connectivity, the situation is similar: an overwhelming fraction of such graphs require $\Omega\big(\log(n)/(\log(\frac{\log(n)}{\alpha}))\big)$ dimensions, see Theorem 8(ii). We conclude with a full characterization of α-preservation for *all* constant-diameter graphs, see Corollary 16, which highlights a key difference between the cases of recoverability and non-recoverability.
- **Preservation in Normed Spaces**. When we insist on embedding graphs into normed spaces, the neighborhood recoverability landscape changes dramatically: an overwhelming fraction of $G \in \mathcal{G}_n$ require $\dim_{(\alpha>1)}(G) = \Omega(n/\log(\frac{8}{\alpha-1}))$ in *any* normed space, see Theorem 17. For Euclidean spaces, we can improve this to $\dim_{(\alpha=1)}(G, \ell_2) = \Omega(n)$, and as α exceeds 1, a phase-change phenomenon occurs: below a certain threshold α-preservation can be achieved in dimension depending on the graph's spectral properties, and beyond this threshold α-preservation may not be possible, see Proposition 20. Meanwhile, k-regular graphs do not suffer from a $\Omega(n)$ recoverability barrier in normed spaces; $O(k^2 \log n)$ dimensions suffice even in ℓ_2, see Proposition 21.
- **Preservation of Clustered Data**. We study the preservation dimension of graphs generated from a planted partition model. We find that, with high probability, α-preservation requires $\Omega\Big(\frac{(1-\xi)\log(n)+\xi\log(k)}{\log(8/\alpha)}\Big)$ dimensions in general metrics, where ξ is a suitable measure of the cluster salience, see Theorem 23. Furthermore, if we insist on $\alpha > 1$, we are hit with a lower bound of $\Omega(\log(n)/\log(\frac{1}{\alpha-1}))$ regardless of cluster salience.

These results present a formidable barrier for existing data visualization algorithms, which typically embed input points in two- or three-dimensional Euclidean space. Per the lower bounds presented in this paper, such visualizations are doomed to misrepresent a portion of the neighborhood structure. This should be of great concern to practitioners who use these algorithms for data analysis and hypothesis generation with the expectation that they reliably reveal the neighborhood structure of high dimensional datasets.

2 Related Work

Our analysis of α-preservation brings insights and techniques from the graph embedding literature to bear on the problem of data visualization.

2.1 Metric Embeddings of Graphs

Representing graphs faithfully in metric spaces is of intense interest in computer science. There are different techniques and limitations for this endeavor depending on what metric one embeds into and what structure the embedding is supposed to preserve. For instance, if one seeks to embed a graph into Euclidean space in such a way that reflects its cluster structure, then spectral clustering is a standard choice [von Luxburg, 2007], and it provably preserves sufficiently prominent clusters in the input [Ng et al., 2001]. If, on the other hand, one seeks a low-distortion embedding (see Definition 31) of some metric induced by a graph (such as the shortest path metric), one can use the famous result by Bourgain [1985] that guarantees an $O(\log n)$-distortion embedding into $O(\log n)$-dimensional Euclidean space. If one insists on arbitrarily good average distortion-D embeddings into normed spaces, Naor [2021] showed that $n^{\Omega(1/D)}$ dimensions is necessary in general, with constant-degree expanders providing the hard instances. The situation improves if one is willing to assume some intrinsic structure; indeed, if the input metric (derived from the graph) has doubling dimension d, then an $O(d)$-dimensional Euclidean embedding exists with $\mathrm{polylog}(n)$-distortion [Abraham et al., 2008] and an $O(d \log\log n)$-dimensional Euclidean embedding exists with $o(\log n)$ distortion [Chan et al., 2010].

One could also seek out an embedding which preserves graph *neighborhoods*, in the sense that edge-connected vertices are mapped as neighbors and non-edge-connected vertices are mapped as non-neighbors (for some suitable sense of neighborhood). One of the first studies in the direction was by Erdős et al. [1965], who define the dimension of a graph as the minimum $d \in \mathbb{N}$ such that there exists an injection $f : V \to \mathbb{R}^d$ for which $u \sim v \implies \|f(u) - f(v)\|_2 = 1$ for all $u, v \in V$. In this setting, the distances between non-edge connected vertices is irrelevant. Maehara [1984] later introduced a threshold-based notion of graph dimension known as sphericity, where the condition on the embedding becomes $u \sim v \iff \|f(u) - f(v)\|_2 < 1$. Reiterman et al. [1989] showed somewhat strikingly that for $n \geq 37$, all but $(1 - 1/n)$-fraction of graphs have sphericity $\geq n/15$. Further results have lower-bounded sphericity in terms of spectral properties of the graph adjacency matrix [Bilu and Linial, 2005]. More recently, Bhattacharjee and Dasgupta

[2020] developed multiple notions of dimension for *directed* graphs, motivated by the recent surge in the interest of sequential data (e.g. natural language). They relate these notions of embeddability to fundamental graph properties like cyclicity and eigenspectra.

Indyk and Naor [2007] study a notion of local structure-preserving embedding that is motivated by approximate nearest neighbor search. They show that for input metrics with constant aspect ratio (i.e. the diameter over the smallest interpoint distance) or constant doubling dimension, one can achieve efficient $(1+\epsilon)$-approximate nearest neighbor Euclidean embedding in $\Omega(1/\epsilon^2)$ dimensions.[2]. In a similar vein, Bartal et al. [2011] develop a local Johnson-Lindenstrauss-type result which embeds into $\Omega(\log k/\epsilon^2)$-dimensional Euclidean space and promise low distortion between any input point and its k-nearest-neighbors.

2.2 Data Visualization and Other Applications

Data visualization is a type of extreme dimension reduction that is focused on producing two- or three-dimensional outputs which emphasize cluster or local neighborhood structure. Standard (linear) dimension reduction methods like classical multi-dimensional scaling (MDS) and random projections are generally not well-suited for this purpose: when forced into ultra-low dimensions, these embeddings tend to destroy salient structures and display "artifacts" unrelated to the intrinsic structure of the data [Dasgupta et al., 2006, Diaconis et al., 2008]. A similar phenomenon can be said of many popular manifold learning methods like Locally Linear Embedding or Laplacian Eigenmaps [Perraul-Joncas and Meilă, 2013, Goldberg et al., 2008, Venna et al., 2010].

t-SNE [van der Maaten and Hinton, 2008] and UMAP [McInnes et al., 2018], on the other hand, have gained widespread popularity across the general scientific literature for their seemingly remarkable ability to visualize salient structure in high-dimensional data. Shaham and Steinerberger [2017], Linderman and Steinerberger [2019], and Arora et al. [2018] were the first works showing that, for sufficiently well-clustered inputs, t-SNE does indeed output the desired cluster visualization. These results corroborate t-SNE's apparent ability to tease out global cluster structure. What about local neighborhood structure? Im et al. [2018] (building on the precision-recall framework of Venna et al. [2010]) and Chari and Pachter [2023] provide some practical evidence that t-SNE and UMAP are less attuned to faithfully revealing neighborhoods.

If one seeks to embed *labelled* data (for downstream prediction as well as visualization), large-margin nearest neighbors is a canonical linear[3] technique from the Mahalonobis metric learning literature [Weinberger and Saul, 2009]. This method aims to alter the original representation of the data such that the nearest neighbor of any point have the same label, while differently-labelled points are separated with a "margin". This margin or gap is akin to our notion of separability between neighbors and non-neighbors for $(\alpha > 1)$-preservation (see Figure 1, left).

A fundamental question in the backdrop of these studies is: what is the minimum embedding dimension necessary to preserve local neighborhoods? Our work addresses this unifying question in a very general setting, demonstrating when and how the embedding dimension must scale with key properties of the input data (e.g. number of data points, connectivity and neighborhood structure, etc.).

3 Preliminaries

For $(\mathcal{X}, \rho)$ a metric space, let $B_r(x) \subseteq \mathcal{X}$ be an open ball of radius r centered at $x \in \mathcal{X}$. The **doubling dimension** of $\mathcal{X}$, denoted $\dim(\mathcal{X})$, is the smallest integer d such that any open ball of radius $r > 0$ in $\mathcal{X}$ can be covered by at most 2^d open balls of radius $r/2$. Let $\mathcal{M}(S, \epsilon)$ (the packing number) denote the size of the smallest packing of points into S such that the distance between any two points in the packing is at least ϵ. Likewise, let $\mathcal{N}(S, \epsilon)$ (the covering number) denote the size of the smallest covering of S by ϵ-radius balls centered in $\mathcal{X}$. When the context is clear, we may abbreviate $\mathcal{N}(B_R(x), \epsilon)$ as $\mathcal{N}(R, \epsilon)$. Observe the following known results about covering and doubling dimension.

Observation 3. *For all $0 < \epsilon < R$ and $x \in \mathcal{X}$: $\mathcal{N}(B_R(x), \epsilon) \leq (2^{\dim(\mathcal{X})})^{\lceil \log_2(R/\epsilon) \rceil} \leq (2R/\epsilon)^{\dim(\mathcal{X})}$.*

Observation 4. *For any $S \subseteq \mathcal{X}$ and $\epsilon > 0$, we have $\mathcal{M}(S, 2\epsilon) \leq \mathcal{N}(S, \epsilon) \leq \mathcal{M}(S, \epsilon)$.*

Observation 5. *Any n-point metric space has doubling dimension at most $\lceil \log_2(n) \rceil$.*

Let $\mathcal{G}_n^{k\text{-reg}} \subseteq \mathcal{G}_n$ denote the set of k-regular graphs on n vertices.

[2] One should note that an analogous statement for $(1 \pm \epsilon)$-isometry for points with constant doubling dimension is known not to hold [Alon, 2003].

[3] Similar nonlinear techniques often used in practice include contrastive learning [Hadsell et al., 2006] and Siamese networks [Bromley et al., 1994].

For $G = (V, E) \in \mathcal{G}_n$, we use the following notation:

- $A(G) = A \in \{0,1\}^{n \times n}$: the adjacency matrix, where $A_{ij} = 1 \iff (i,j) \in E$,
- $\Delta(G)$: graph diameter: the maximum shortest-path length between any two vertices; ∞ if G is disconnected,
- $\iota(G) \subseteq V$: the largest independent set,
- $\kappa(G) \subseteq V$: the largest clique,
- $G|_{V'} = (V', E')$: the subgraph of G induced by $V' \subseteq V$, where $E' = \{(u,v) \in E : u, v \in V'\}$.

For an embedding algorithm f (as defined in Introduction), we shall use $\mathrm{diam}(f(G)) := \max_{u,v} \rho(f_G(u), f_G(v))$ to denote the diameter of the f-embedding of G in $\mathcal{X}$.

All the omitted proofs can be found in the Appendix.

3.1 Elementary Observations about α-Preservation

Before presenting our main results, we show that α-preservation is only interesting for the $\alpha \in (0,2)$ case. This is because, aside from some very trivial cases, all graphs simply do not admit $(\alpha \geq 2)$-preservation. In particular:

Proposition 6. *Let $\alpha \geq 2$. For all $G \in \mathcal{G}_n$, either (i) $\dim_\alpha(G)$ is undefined (i.e. α-preservation is not possible), or (ii) all connected components in G are cliques and G can be α-preserved in $\mathbb{R}$.*

Proof. Suppose $G = (V, E) \in \mathcal{G}_n$ contains a connected component that is not a clique. We will show G is not $(\alpha \geq 2)$-preservable in any metric space. Let $V' \subseteq V$ be the non-clique connected component of G. The diameter of $G|_{V'}$ is at least 2. Thus there exist two vertices $u, v \in V'$ such that the shortest path between them is of length exactly 2. Let w be the "connecting" vertex such that $u \sim w \sim v$, and assume towards contradiction there exists an algorithm f that α-preserves G in some metric space $\mathcal{X}$ (with neighborhood threshold of $r > 0$, see Definition 1). Then the maximum distance between u and v in the embedding is:

$$\rho(f_G(u), f_G(v)) \leq \rho(f_G(u), f_G(w)) + \rho(f_G(w), f_G(v)) < 2r,$$

which implies that $u \sim v$ since $\alpha \geq 2$ implies $u \not\sim v \implies \rho(f_G(u), f_G(v)) \geq r \cdot \alpha \geq 2r$. This contradicts the fact that the shortest path between u and v is of length 2. Thus, an f that α-preserves G cannot exist.

Suppose all connected components of G are cliques. Observe that any visualization algorithm $f : \mathcal{G}_n \to \mathbb{R}$ that maps each clique to a unique point in $\mathbb{R}$ α-preserves these graphs for all $\alpha > 0$, in particular $\alpha \geq 2$ (simply choose $r > 0$ small enough). □

We also observe the following intuitive monotonicity properties of α-preservation.

Proposition 7. *For $G \in \mathcal{G}_n$ and a collection of metric spaces $\mathbb{X}$, assume $\dim_\alpha(G, \mathcal{X})$ is well-defined. The visualization dimension satisfies the following properties:*

(i) If $\beta \leq \alpha$ then $\dim_\beta(G, \mathbb{X}) \leq \dim_\alpha(G, \mathbb{X})$.

(ii) If $\mathbb{X} \subseteq \mathbb{Y}$ then $\dim_\alpha(G, \mathbb{Y}) \leq \dim_\alpha(G, \mathbb{X})$.

Throughout the rest of the paper we shall assume $\alpha \in (0,2)$.

4 Preservation in General Metric Spaces

Take any graph $G \in \mathcal{G}_n$. Arguably the most natural realization of this graph in a metric space is induced by its shortest path distances. This immediately provides an α-preservation of G for *all* $\alpha \in (0,2)$ in an n-point metric space, yielding a doubling dimension of $O(\log n)$ (see Observation 5). The key question is whether this $\log(n)$-scaling is necessary. It turns out that there exist graphs which do require this scaling. In fact, these hard instances comprise an overwhelming fraction of $\mathcal{G}_n$ and persist even if we allow for a substantial overlap between neighbors and non-neighbors ($\alpha \ll 1$) Interestingly, the situation does not improve even for graphs with low connectivity, e.g. k-regular graphs.

4.1 Lower Bounds

Theorem 8. *For any $\alpha \in (0, 2)$, we have the following.*

(i) For all $n \geq 82$, at least $1 - 2^{-n/5}$ fraction of $G \in \mathcal{G}_n$:

$$\dim_\alpha(G) \geq \frac{\log(n) - 2\log(2)}{2\log(8/\alpha)} = \Omega\left(\frac{\log(n)}{\log(8/\alpha)}\right).$$

(ii) For all even integers $n \geq 6$ and $k \geq 4$, at least $1 - O(n^{-k+2})$ fraction of $G \in \mathcal{G}_n^{k\text{-reg}}$:

$$\dim_\alpha(G) \geq \frac{\log(n/(k+1))}{\log\left(\frac{4}{\alpha}\left\lceil \frac{\log(n-1)}{\log\left(\frac{k}{2\sqrt{k-1}+1/2}\right)}\right\rceil\right)} = \Omega\left(\frac{\log(n/k)}{\log\frac{\log n}{\log k} + \log(4/\alpha)}\right).$$

This incompressibility result is realized by the prevalence of graphs with high connectivity. Take for instance the star graph on n nodes: it is a diameter-2 graph with $n-1$ edges. Intuitively, a neighborhood preservation (neighbors close, non-neighbors far) of a star graph requires packing $\Omega(n)$ points in a $O(1)$-diameter ball, yielding a $\log(n)$-type lower bound on the dimension of the target metric space. We can apply the same intuition for constant-degree expanders (e.g. Ramanujan graphs [Hoory et al., 2006, Huang et al., 2024]), yielding a similar result for $\mathcal{G}_n^{k\text{-reg}}$.

We now make this intuition precise. A key quantity in our analysis will be the notion of a clique partition of a graph (also known as the *minimum clique cover* in the literature), which is a natural measure of neighborhood connectivity.

Definition 9 (clique partition)**.** *For $G \in \mathcal{G}_n$, define the* clique partition *of G, denoted $P(G)$, as the smallest-sized partition*[4] *of V such that for all $S \in P(G)$, $G|_S$ is a clique.*

One can relate the clique partition to fundamental graph quantities which will be helpful in our discussion later.

Observation 10. *For all $G = (V, E)$, we have $\chi(G^c) = |P(G)| \geq \max\left(|\iota(G)|, \frac{|V|}{|\kappa(G)|}\right)$, where $\chi(G^c)$ is the chromatic number of the complement graph of G.*

We can quantify the difficulty of α-preservation dimension in terms of clique partitions of the input graph G.

Lemma 11. *For all $G = (V, E) \in \mathcal{G}_n$, and all $\alpha \in (0, 2)$,*[5]

$$\dim_\alpha(G) \geq \max_{\substack{U \subseteq V \\ |U| \geq 2}} \frac{\log|P(G|_U)|}{\log(4\,\Delta(G|_U)/\alpha)}.$$

Before proving this, it is instructive to understand the significance of the ratio: it captures our intuition of large packing (numerator) in a small space (denominator). Consider the aforementioned star graph, which has a $\Omega(n)$-sized clique partition and $O(1)$ diameter, recovering the expected $\log(n)$-lower bound. In contrast, the cycle graph——which can clearly be $(\alpha < 2)$-preserved in constant-dimensional ℓ_1 space——has both $\Omega(n)$-sized clique partition and diameter, relaxing our lower bound to accommodate this case.

Proof. Let $f_G : V \to \mathcal{X}^n$ be an α-preservation of G into some target metric $(\mathcal{X}, \rho)$ with neighbor threshold $r > 0$ (see Definition 1). Note that such a map always exists (see Observation 14). Now for any $U \subseteq V$, let $G' := G|_U$ be a vertex-induced subgraph. Without loss of generality, assume G' has finite diameter (otherwise the bound is trivially true). We proceed via a covering argument.

First, we show that $\mathrm{diam}(f_G(U)) < r \cdot \Delta(G')$. Of course, every pair of vertices $u, v \in U$ have a path length of at most $\Delta(G')$. A trivial application of triangle inequality shows $\rho(f_G(u), f_G(v))$ is bounded by $r \cdot \Delta(G')$ since every successive vertex pair in the path from u to v needs to be within distance r.

Next, we show that $\mathcal{N}(f_G(U), \alpha \cdot r/2) \geq |P(G')|$. Consider any $(\alpha \cdot r/2)$-covering $\mathcal{C} = \{c_1, \ldots, c_m\}$ of $f_G(U)$. For each $i \in [m]$, iteratively define

$$P_i := \left\{u \in U \;\middle|\; f_G(u) \in B_{\alpha \cdot r/2}(c_i)\right\} \setminus \bigcup_{k<i} P_k.$$

[4] We break ties in an arbitrary but fixed manner.

[5] We use the convention that the maximum of an empty set in $\mathbb{R}$ is $-\infty$.

Observe that $G'|_{P_i}$ is a clique, since for all distinct $u, v \in P_i$, $\rho(f_G(u), f_G(v)) < \alpha \cdot r$, which by definition of α-preservation implies $u \sim v$. Note that $P' := \{P_i\}_{i\in[m]}$ constitutes a clique partition of U, therefore $|P(G')| \leq |P'| \leq m$.

Therefore, by Observation 3,

$$|P(G')| \leq \mathcal{N}\left(r \cdot \Delta(G'), \frac{\alpha \cdot r}{2}\right) \leq \left(\frac{4\,\Delta(G')}{\alpha}\right)^{\dim(\mathcal{X})}.$$

Equivalently, for *any* α-preserving $f_G : V \to \mathcal{X}^n$ on a metric space $\mathcal{X}$ and subgraph $G' := G|_U$ induced from *any* $U \subseteq V$, we have $\dim(\mathcal{X}) \geq \Big(\log|P(G')|\Big) \Big/ \Big(\log(4\,\Delta(G')/\alpha)\Big)$ which yields the lemma statement. □

The proof of Theorem 8 proceeds by analyzing the key quantities $|P(G)|$ and $\Delta(G)$ for typical graphs in $\mathcal{G}_n$. We think of a typical graph as being drawn uniformly at random from $\mathcal{G}_n$ (equivalently, produced by the Erdős–Rényi model with parameter $1/2$). With high probability, such a graph has constant diameter and a large clique partition. The latter follows from Observation 10 and a refined analysis of established bounds on clique numbers of Erdős–Rényi graphs [Frieze and Karoński, 2016]. For the $\mathcal{G}_n^{k\text{-reg}}$ lower bound, a similar analysis yields $\Delta(G) = O(\log n)$ and $|P(G)| = \Omega(n)$. The diameter upper bound follows from well-known expander properties of k-regular graphs [Bollobás and de la Vega, 1982, Huang et al., 2024]. The clique partition lower bound follows from analyzing $\iota(G)$ [Wormald, 1995]. Formal details are provided in Appendix B.1.

The clique partition $P(G)$ is a powerful concept, which helps us establish the near-worst-case difficulty of α-preserving typical graphs for *all* $\alpha \in (0, 2)$. In light of Corollary 16 and Lemma 25, it tightly characterizes the α-preservation dimension of such graphs, which constitute $(1 - 2^{-\Omega(n)})$-fraction of $\mathcal{G}_n$.

For the remaining (atypical) fraction of graphs, the situation can be different. Consider the following graph: let $G \in \mathcal{G}_n$ (with n even) consist of two disjoint cliques of size $n/2$ and exactly $n/2$ edges connecting vertices of the two cliques 1-to-1. A straightforward packing argument gives $\dim_\alpha(G) = \Omega(\log(n))$ for all $\alpha = 1 + \Omega(1)$, yet our key Lemma 11 is rendered ineffective since $|P(G)| = 2$ and $\Delta(G) = 2$.

This example highlights how, as α grows past unity, the problem of α-preservation can become significantly harder for certain graphs. The key difficulty of $(\alpha > 1)$-preserving this graph stems from the fact that there exist a large number of vertices in close proximity that have *distinct, yet overlapping* neighborhoods. To formalize this, we introduce the concept of a "neighborhood partition," which we leverage in Lemma 13, strengthening Lemma 11 when $\alpha = 1 + \Omega(1)$.

Definition 12 (neighborhood partition)**.** *For $G = (V, E) \in \mathcal{G}_n$ and $U \subseteq V$, define the* neighborhood partition *of $G|_U$, denoted $\mathrm{C}(G|_U)$, as the smallest-sized partition*[6] *of U such that all u and u' are in the same part if and only if they have identical neighborhoods with respect to G. That is, for all $u, u' \in S \in \mathrm{C}(G|_U)$, we have $N(u) = N(u')$, where $N(u) := \{v \in V : v \sim_G u \text{ or } v = u\}$.*

Lemma 13. *For all $G = (V, E) \in \mathcal{G}_n$ and $\alpha \in (1, 2)$,*

$$\dim_\alpha(G) \geq \max_{\substack{U\subseteq V \\ |U|\geq 2}} \frac{\log\big|\,\mathrm{C}(G|_U)\big|}{\log(\frac{4\,\Delta(G|_U)}{\alpha-1})}.$$

Proof. It suffices to exhibit the lower bound for all $U \subseteq V$ with $G|_U$ connected and $|\,\mathrm{C}(G|_U)| \geq 2$ otherwise the bound is vacuously true. Fix any $\alpha \in (1, 2)$ and consider an α-preservation f_G of G in $(\mathcal{X}, \rho)$. Note that for all $u', u'' \in V$ which are not in the same part of $\mathrm{C}(G|_U)$, there exists $v \in V$ that is in either $N(u')$ or $N(u'')$ but not both since $N(u') \neq N(u'')$. WLOG assume $v \in N(u') \setminus N(u'')$. By definition of α-preservation, this tells us:

$$\alpha r \leq \rho(f_G(u''), f_G(v)) \leq \rho(f_G(u'), f_G(v)) + \rho(f_G(u'), f_G(u'')) < \rho(f_G(u'), f_G(u'')) + r$$
$$\implies \forall u', u'' \text{ in distinct parts of } C(G|_U) \qquad \rho(f_G(u'), f_G(u'')) > r(\alpha - 1),$$

where r is the neighborhood threshold of f_G (cf. Definition 1). Since $f_G(U)$ has diameter at most $r \cdot \Delta(G|_U)$, constitutes an $r(\alpha-1)$-packing of a ball of diameter $r \cdot \Delta(G|_U)$ of size $|C(G|_U)|$. Due to maximal packing estimates (cf. Observation 4), we have:

$$|\,\mathrm{C}(G|_U)| \leq \mathcal{M}\left(r \cdot \Delta(G|_U), r(\alpha-1)\right) \leq \mathcal{N}\left(r \cdot \Delta(G|_U), \frac{r(\alpha-1)}{2}\right) \leq \left(\frac{4r \cdot \Delta(G|_U)}{r \cdot (\alpha-1)}\right)^{\dim(\mathcal{X})}.$$

Rearranging the terms yields the lower bound. □

[6] We break ties in an arbitrary but fixed manner.

4.2 Upper Bounds

As discussed at the beginning of the section, one trivially has an upper bound on the α-preservation dimension of a graph G in terms of the doubling dimension of the graph shortest path metric ρ_G.

Observation 14. *For all $G \in \mathcal{G}_n$ and $\alpha \in (0,2)$, $\dim_\alpha(G) \leq \dim(\mathcal{X}_{\rho_G}) \leq \lceil \log_2 n \rceil$, where $\mathcal{X}_{\rho_G}$ is the n-point shortest path metric derived from G.*

Observe that this upper bound in terms of the shortest path metric is clearly not tight. Consider for instance the complete graph on n vertices. The doubling dimension (with respect to the shortest path metric) of this graph is $\Theta(\log n)$, yet it can be α-preserved in constant dimensions, by simply sending all points to a one-point metric space. It is natural to ask whether an α-dependent bound is possible.

Proposition 15. *For all $G \in \mathcal{G}_n$,*

$$\dim_\alpha(G) \leq \begin{cases} \left\lceil \log_2(3) \left\lceil \frac{\log |P(G)|}{\log \lceil 1/\alpha \rceil} \right\rceil \right\rceil & \alpha \in (0,1) \\ \left\lceil \log_2 |P(G)| + \log_2(3) \right\rceil & \alpha = 1 \\ \left\lceil \log_2 |P(G)| + \log_2(3) \left\lceil \max_{S \in P(G)} \frac{\log |C(G|_S)|}{\log \lceil \frac{1}{\alpha - 1} \rceil} \right\rceil \right\rceil & \alpha \in (1,2) \end{cases}.$$

The proof follows an intuitive construction which hinges on the clique partition. For $\alpha < 1$, each part of the clique partition can be collapsed to a point yielding a $|P(G)|$-sized metric space of doubling dimension $O(\log |P(G)|)$. For $\alpha \geq 1$, each part S is embedded as a sub-metric which necessarily contains some $\alpha - 1$ packing depending on its neighborhood partition $C(G|_S)$. Compare this with our lower bounds: Lemma 11 for the $\alpha \leq 1$ case and Lemma 13 for the $\alpha > 1$ case.

This enables us to give a full characterization of α-preservation dimension for constant-diameter graphs, which by Lemma 25 constitute an overwhelming fraction of $\mathcal{G}_n$.

Corollary 16. *For $\alpha \in (0,2)$ and $G \in \mathcal{G}_n$ such that $\Delta(G) = O(1)$,*

$$\dim_\alpha(G) = \Theta\Big(\frac{\log |P(G)|}{\log(8/\alpha)} + \mathbb{1}[\alpha > 1] \cdot \frac{\log |C(G)|}{\log(\frac{4}{\alpha - 1})} \Big).$$

5 Preservation in Normed Spaces

The designation of metric space is incredibly general. Practitioners are usually interested in producing faithful data visualizations into more structured spaces. For instance, it is often desirable to work with data in a normed space due to the ability to add and scale the points as vectors. The significance of the norm structure is not just mathematical convenience but also its direct interpretability for data visualization: we tend to think in terms of normed (especially Euclidean) space because it obeys similar principles as physical space. This motivates us to study α-preservation into such spaces. We find that this restriction comes at a steep cost in terms of the embedding dimension.

5.1 Lower Bounds

It turns out that $(\alpha \geq 1)$-preservation (that is neighborhood *recoverability*) in normed spaces is exponentially harder than it is for general metrics: an overwhelming fraction of graphs in $\mathcal{G}_n$ require dimension that scales *linearly* in n, in contrast to the logarithmic scaling in the case of general metrics (cf. Theorem 8).

Theorem 17. *We have the following.*

(i) ***(General normed spaces)*** *Let $\mathbb{L}$ be the collection of all normed spaces. For all $\alpha \in (1,2)$ and $n \geq 82$, we have that for at least $1 - 2^{-n/6}$ fraction of $G \in \mathcal{G}_n$:*

$$\dim_\alpha(G, \mathbb{L}) \geq \frac{n}{3 \log_2(\frac{16}{\alpha - 1})} = \Omega\left(\frac{n}{\log(\frac{16}{\alpha - 1})} \right).$$

(ii) ***(Euclidean spaces)*** *For $\alpha = 1$ and $n \geq 0$, we have that for at least $1 - 2^{-n}$ fraction of $G \in \mathcal{G}_n$,*

$$\dim_{(\alpha=1)}(G, \ell_2) \geq \frac{n}{15} - \frac{1}{4}.$$

Before proceeding to the discussion of our key theorem, a few remarks are in order.

- It is straightforward to extend Theorem 17(ii) to $\alpha > 1$ for ℓ_2 (see Proposition 7), so long as $\dim_\alpha(G, \ell_2)$ exists (see Proposition 20).
- Unlike in the case of embeddings into general metric spaces, an analogous $\Omega(n)$ lower bound for $\mathcal{G}_n^{k\text{-reg}}$ is impossible. In particular, one can show that an $O(k^2 \log n)$-dimensional ℓ_2 embedding exists for all such graphs (see Proposition 21) which nearly matches the lower bounds of Theorem 8(ii).
- Similarly one cannot hope for an $\Omega(n)$ bound for the case when $\alpha < 1$. It turns out that *all* graphs $G \in \mathcal{G}_n$ can be $(\alpha < 1)$-preserved in ℓ_∞^d for $d = O(\log |P(G)|) = O(\log n)$ (see Proposition 19).

The proof of incompressibility for general normed spaces proceeds via an elegant application of the volume argument. Specifically, the vector space structure of normed spaces enables us to superimpose the sheer plenitude of low-diameter graphs in a small enough region. Noting that disparate graphs (with disparate neighborhood structures) require disparate neighborhood-preserving embeddings in the space, one must require the target dimension to scale with n to accommodate these embeddings. Formal details are provided in Appendix B.2.

The highly regular structure of the Euclidean space enables us to extend our lower bound to the $\alpha = 1$ case in ℓ_2. This requires an alternate analysis. In particular, we leverage the fact that since the (squared) 2-norm distances can be represented by a (quadratic) polynomial, any graph can be distinctly recognized from an $(\alpha = 1)$-preserving embedding in ℓ_2 by a series of polynomial threshold tests. By a theorem due to Warren [1968], we know that the polynomial threshold tests are expressive enough to distinguish the graphs only if the embedding dimension is $\Omega(n)$, yielding the desired extension.[7] Formal details are provided in Appendix B.2.

5.2 Upper Bounds for Preservation in Normed Spaces

A straightforward application of Fréchet's embedding [Fréchet, 1910, Matoušek, 2013] yields $\dim_\alpha(\mathcal{G}_n, \mathbb{L}) = O(n)$ by considering an isometric embedding of the shortest path metric (and hence an $(\alpha \leq 2)$-preservation) of the input graph into ℓ_∞^{n-1}. Specifically:

Observation 18. *For all $\alpha \in (0, 2)$ and $G \in \mathcal{G}_n$, we have $\dim_\alpha(G, \ell_\infty) \leq \lceil \log_2(3) \cdot (n-1) \rceil = O(n)$.*

This trivial result can be refined as follows.

Proposition 19. *For all $G \in \mathcal{G}_n$,*

$$\dim_\alpha(G, \ell_\infty) \leq \begin{cases} \lceil \log_2(3) \left\lceil \frac{\log |P(G)|}{\log \lceil 1/\alpha \rceil} \right\rceil \rceil = O\left(\frac{\log |P(G)|}{\log \lceil 1/\alpha \rceil}\right) & \alpha \in (0, 1) \\ \lceil \log_2(3) \cdot |C(G)| \rceil = O(|C(G)|) & \alpha \in [1, 2) \end{cases}.$$

Proof. For any $G = (V, E) \in \mathcal{G}_n$, let $P(G) = \{P_1, \ldots, P_m\}$ be the m parts of the clique partition.

Case $\alpha \in (0, 1)$. By Observation 29(i) (take $n = m$, $r = 1$ and $\epsilon = \alpha$), m points can be embedded in (open) unit ball with interpoint distances at least α in ℓ_∞^d with $d = \left\lceil \frac{\log m}{\log \lceil 1/\alpha \rceil} \right\rceil$. Thus, the mapping where vertices belonging to each partition P_i of the input graph G is mapped to the i-th point yields an α-preserving embedding in ℓ_∞, with (doubling) dimension at most $\left\lceil \log_2(3) \left\lceil \frac{\log |P(G)|}{\log \lceil 1/\alpha \rceil} \right\rceil \right\rceil = O\left(\frac{\log |P(G)|}{\log \lceil 1/\alpha \rceil}\right)$.

Case $\alpha \in [1, 2)$. Let $G/C(G)$ be the (simple) graph produced by contracting all nodes that are in the same part of $C(G)$ and preserving the edge connectivity. Since any $u, v \in V$ with $N(v) = N(u)$ (see Definition 12) can be embedded identically with no effect on α-preservation, an α-preservation of $G/C(G)$ is an α-preservation of G. Hence we re-apply Observation 18 to $G/C(G)$ to get the bound[8] $\lceil \log_2(3) \cdot |C(G)| \rceil$. □

For the $\alpha < 1$ case, this upper bound improves considerably on Fréchet-style embedding (Observation 18) and nearly matches the general metric space lower bound given by Lemma 11. On the other hand, in the recoverable case, one cannot hope for more than a constant-factor improvement on this straightforward upper bound due to a result by Roberts [1969] who proved that there exist $G \in \mathcal{G}_n$ with $\dim_{\alpha \geq 1}(G, \ell_\infty) \geq \left\lfloor \frac{2}{3} |C(G)| \right\rfloor$.

[7] This method of proof extends to other p-norms (for p even), and may be of independent interest.

[8] Note that this embedding works for $\alpha \in (0, 2)$, but since $|P(G)| \leq |C(G)|$ it offers no improvement.

5.2.1 Preservation in Euclidean Space (ℓ_2)

Theorem 17(ii) establishes a formidable $\Omega(n)$ lower bound on ℓ_2 recoverability for most $G \in \mathcal{G}_n$. A matching $O(n)$ upper bound is trivial: if an α-preservation of n points exists, it is necessarily an α-preservation on the $(n-1)$-dimensional subspace spanned by the points. Here we investigate more refined upper bounds depending on α as well as the structure of the input graph.

We find that $(\alpha < 1)$-preservation dimension in ℓ_2 follows a familiar $\log|P(G)|$ scaling. In contrast, the recoverability case exhibits an interesting phase shift. For α up to a graph-dependent threshold, the graph is recoverable in dimension depending on its spectrum. However, past this threshold, α-preservation is not always possible. In particular:

Proposition 20. *For all $G \in \mathcal{G}_n$, we have*

$$\dim_\alpha(G, \ell_2) \leq \begin{cases} \left\lceil \log_2(5) \left\lceil \frac{4\log(|P(G)|+1)}{2-4\alpha^2} \right\rceil \right\rceil & \alpha \in (0, \frac{1}{\sqrt{2}}) \\ \left\lceil \log_2(5) \left\lceil 12\left(\frac{1+\alpha^2}{1-\alpha^2}\right)^2 \log|P(G)| \right\rceil \right\rceil & \alpha \in \left(\frac{1}{\sqrt{3}}, 1\right) \\ \left\lceil \log_2(5) \cdot \min\left(\lceil 192\lambda_G^2 \log|C(G)| \rceil,\ |C(G)|-1\right) \right\rceil & \alpha \in \left[1, \frac{1}{\sqrt{1-\frac{1}{4\lambda_G}}}\right) \end{cases},$$

where λ_G denotes the maximum eigenvalue of $A(G/C(G))$.

Moreover, for all $n \geq 4$ there exists $G \in \mathcal{G}_n$ which cannot be α-preserved in ℓ_2 for $\alpha > (1-\frac{1}{\lambda_G})^{-1/2}$.

In light of Theorem 17(ii) it is worth noting that only a negligible fraction of $G \in \mathcal{G}_n$ have $\lambda_G = o(n)$. Since λ_G is upper-bounded by the maximum vertex degree, Proposition 20 immediately provides us a recoverability upper bound k-regular graphs.

Corollary 21. *Fix any $0 < k < n$. For all $G \in \mathcal{G}_n^{k\text{-reg}}$ and $\alpha \in (0, \sqrt{1+\frac{1}{4k}})$, we have, $\dim_\alpha(G, \ell_2) \leq \lceil 192k^2 \log n \rceil$.*

6 Preservation of Clustered Data

Most data visualization algorithms seek to visualize clusters in a dataset. Ideally, such algorithms will be able to detect and represent the latent cluster structure in a constant-dimensional metric space. We show, for a general model of clustering, this is impossible, even when the clusters are well-separated. The standard graph-based model for clusters is the *planted partition model*:

Definition 22. *Fix integers $n \geq k \geq 1$. For a partition $\{S_1, \ldots, S_k\}$ of V (of size n) and $0 \leq q \leq p \leq 1$, define the planted partition distribution, $\mathtt{PP}_{p,q}(S_1, \ldots, S_k)$, as the distribution over $\mathcal{G}_n(V)$ such that for all $i, j \in [k]$ distinct, edges within S_i occur independently with probability p and edges between S_i and S_j occur independently with probability q.*

The key parameters p and q capture intra- and inter-cluster connectivity. Naturally, the higher the gap between these two parameters, the more salient the cluster structure of the model. We provide a lower bound on the α-preservation dimension with respect to these parameters.

Theorem 23. *Fix $n \geq k \geq 1$ and let $\{S_1, \ldots, S_k\}$ be a partition of V (of size n). Let $c \geq 1$ be such that $\max_{i\in[k]} |S_i| \leq \frac{cn}{k}$, and $0 < q \leq p \leq 1$ with $q < 1$. Then for all $\alpha \in (0, 2)$ with probability at least $1 - \exp(-\Omega(n^{\min(2(p+q-pq),1)}))$ over $G \sim \mathtt{PP}_{p,q}(S_1, \ldots, S_k)$:*

$$\dim_\alpha(G) \geq \frac{1}{\log(8/\alpha)} \Big((1-\xi_{p,q}) \log n + \xi_{p,q} \log(k/2c) \Big),$$

where $\xi_{p,q} := p - q + pq$ encodes the cluster saliency *of the planted partition model.*

The proof proceeds by estimating clique number and the diameter of typical graphs generated from a planted partition model and applying our key Lemma 11. Formal details are provided in Appendix B.3.

We can think about this result as interpolating between two extremes:

- *Clustered data.* When $0 < q < p = 1$ (with q constant), we have $\xi_{p,q} = 1$, so the lower bound becomes $\Omega(\frac{\log(k/3c))}{\log(8/\alpha)})$. This essentially matches our upper bound in Proposition 15 for $\alpha \leq 1$, since $|P(G)| = k$.
- *Unclustered data.* When $p = q = \frac{1}{2}$, the lower bound becomes $\Omega(\frac{(3/4)\log n + (1/4)\log(k/3c)}{\log(8/\alpha)}) = \Omega(\frac{\log(n)}{\log(8/\alpha)})$, which matches the upper bound in Proposition 15.

For $\alpha > 1$, we have a stronger lower bound: even when $p = 1$ (the fully connected clusters case), the α-preservation dimension is $\Omega(\log n)$ with overwhelming probability.

Proposition 24. *Fix any $\alpha > 1$. Pick $p = 1$ and $0 < q < 1$. Then if $\max_{i\in[k]} |S_i| \leq cn/k$, $G \sim \mathtt{PP}_{p,q}(S_1, ..., S_k)$ with probability at least $1 - n^2\left(\max(q, 1-q)^{2n(1-c/k)} + e^{-q^2(n-1)}\right)$, we have $\dim_\alpha(G) \geq \frac{\log(n)}{\log(\frac{8}{\alpha-1})}$.*

7 Discussion

Given the pervasive use of data visualization techniques like t-SNE and UMAP which "just seem to work" in practice, it is tempting to believe that data visualization is essentially a solved problem. Our analysis of α-preservation reveals some evidence to the contrary: in many situations of practical interest, including when data has extremely pronounced cluster structure, visualizing neighbor and non-neighbor relationships fundamentally requires high dimensions.

One may wonder how the incompressibility results presented in this work square with more "positive" findings regarding the possibility of constant-dimensional data visualizations. Arora et al. [2018], for instance, established that two-dimensional t-SNE plots successfully visualize well-clustered data. In effect, their result speaks to the fact that α-preservation of extremely clustered neighborhood graphs (i.e. $q = 0$ and $p = 1$, in the context of Definition 22) can be done in constant dimensions. Our results show that any introduction of noise, between or within clusters, requires strictly greater than constant dimension (specifically $\Omega(\log k)$ for $q > 0$ and $\Omega(\log n)$ for $p < 1$) for neighborhood preservability. Similarly, Sarkar [2011] gave a construction on how to embed trees arbitrarily well in two-dimensional hyperbolic spaces, again implying good neighbor preservation in constant dimensions. At first glance, this seems to contradict the findings of Lemma 11, which would imply that $\Omega(\log n)$ doubling dimensions are necessary to α-preserve balanced, constant-degree trees. Note, however, that a constant-dimensional hyperbolic space does not have a constant *doubling* dimension.

One can think of $(\alpha \leq 1)$-preservation as a relaxation of $(1/\alpha)$-distortion embedding (indeed, a bounded distortion embedding has correspondingly bounded preservation, see Observation 32). Since neighborhood preservation is fundamentally a *local* concept, it is particularly well-suited to embedding geometric objects such as manifolds that are characterized by their local structure (indeed, our results can be extended and sharpened for such data). In the grand scheme of metric embedding desiderata, preservation is refreshingly lenient compared to low-distortion. For instance, 1-preservation (the hardest form of $\alpha \leq 1$ preservation) is possible for any graph in ℓ_2, whereas 1-distortion (the hardest form of low-distortion embedding, of course) is very much not possible: even some graph metrics of constant doubling dimension require $\sqrt{\log n}$-distortion in ℓ_2 [Gupta et al., 2003]!

There are many avenues of further investigation. For instance, it is natural to consider *approximate α-preservation*: given some neighborhood graph and a fixed target dimension d, what fraction of neighborhoods can be α-preserved in d? Does this relaxation dramatically change the stringent lower bounds of this work? If so, then answering this question for real-world datasets would allow us calibrate our expectations for data visualization on a case-by-case basis.

More generally, it is tempting to pursue an *algorithmic realization of α-preservation*. One can think of the construction of an optimal α-preservation as the minimization of a very difficult objective—the doubling dimension—with respect to basic linear constraints—dictating the metric structure and the α-preservation thresholds. For preservation in ℓ_2 this optimization (as well as its approximate counterpart) can be approximated by a semidefinite program. Are there alternative, less obvious algorithms? Our study of α-preservation suggests that certain graph statistics, including $C(G)$, $P(G)$, and the graph spectrum, can help to quantify α-preservation.

Pursuing algorithms for α-preservation brings us full circle. Our study was motivated by the idea that popular data visualization methods boil down to extracting a useful neighborhood graph and optimizing some sort of relaxed α-preservation-esque objective in a fixed dimension. The jury is still out: does this practice constitute a provably effective approximation algorithm for a well-defined problem? Or is the problem, even with all the standard relaxations, simply too hard?

References

I. Abraham, Y. Bartal, and O. Neiman. Embedding metric spaces in their intrinsic dimension. *Symposium on Discrete Algorithms (SODA)*, pages 363–372, 2008.

N. Alon. Problems and results in extremal combinatorics, part I. *Discrete Mathematics*, 273(1-3):31–53, 2003.

S. Arora, W. Hu, and P. K. Kothari. An analysis of the t-SNE algorithm for data visualization. In *Conference on learning theory (COLT)*, pages 1455–1462, 2018.

Y. Bartal, B. Recht, and L. J. Schulman. Dimensionality reduction: beyond the Johnson-Lindenstrauss bound. In *Symposium on Discrete Algorithms (SODA)*, pages 868–887, 2011.

F. Bavaud. On the Schoenberg transformations in data analysis: theory and illustrations. *Journal of Classification*, 28: 297–314, 2011.

R. Bhattacharjee and S. Dasgupta. What relations are reliably embeddable in Euclidean space? In *Algorithmic Learning Theory (ALT)*, pages 174–195, 2020.

Y. Bilu and N. Linial. Monotone maps, sphericity and bounded second eigenvalue. *Journal of Combinatorial Theory, Series B*, 95(2):283–299, 2005.

B. Bollobás and W. F. de la Vega. The diameter of random regular graphs. *Combinatorica*, 2:125–134, 1982.

J. Bourgain. On Lipschitz embedding of finite metric spaces in Hilbert space. *Israel Journal of Math*, pages 46–52, 1985.

J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah. Signature verification using a “Siamese" time delay neural network. *Advances in Neural Information Processing Systems (NIPS)*, 1994.

T.-H. H. Chan, A. Gupta, and K. Talwar. Ultra-low-dimensional embeddings for doubling metrics. *Journal of the ACM (JACM)*, 57(4):1–26, 2010.

T. Chari and L. Pachter. The specious art of single-cell genomics. *PLOS Computational Biology*, 19(8), 2023.

F. R. Chung. Diameters and eigenvalues. *Journal of the American Mathematical Society*, 2(2):187–196, 1989.

S. Dasgupta and A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

S. Dasgupta, D. Hsu, and N. Verma. A concentration theorem for projections. *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2006.

P. Diaconis, S. Goel, and S. Holmes. Horseshoes in multidimensional scaling and local kernel methods. *The Annals of Applied Statistics*, 2(3), 2008.

G. Dimitriadis, J. P. Neto, and A. R. Kampff. t-SNE visualization of large-scale neural recordings. *Neural computation*, 30(7):1750–1774, 2018.

P. Erdős, F. Harary, and W. T. Tutte. On the dimension of a graph. *Mathematika*, 12(2):118–122, 1965.

J. Friedman. *A proof of Alon's second eigenvalue conjecture and related problems*. American Mathematical Society (AMS), 2008.

A. Frieze and M. Karoński. *Introduction to Random Graphs*. Cambridge University Press, 2016.

M. Fréchet. Les dimensions d'un ensemble abstrait. *Mathematische Annalen*, 68:145–168, 1910.

Y. Goldberg, A. Zakai, D. Kushnir, and Y. Ritov. Manifold learning: The price of normalization. *Journal of Machine Learning Research (JMLR)*, 9(8), 2008.

A. Gupta, R. Krauthgamer, and J. R. Lee. Bounded geometries, fractals, and low-distortion embeddings. *Symposium on Foundations of Computer Science (FOCS)*, pages 534–543, 2003.

R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

Y. Han, S. Liu, D. Cong, Z. Geng, J. Fan, J. Gao, and T. Pan. Resource optimization model using novel extreme learning machine with t-distributed stochastic neighbor embedding: application to complex industrial processes. *Energy*, 225, 2021.

S. Hoory, N. Linial, and A. Wigderson. Expander graphs and their applications. *Bulletin of the American Mathematical Society*, 43(4):439–561, 2006.

J. Huang, T. Mckenzie, and H.-T. Yau. Ramanujan property and edge universality of random regular graphs. *Computing Research Repository (CoRR)*, abs/2412.20263, 2024.

D. J. Im, N. Verma, and K. Branson. Stochastic neighbor embedding under f-divergences. *Computing Research Repository (CoRR)*, abs/1811.01247, 2018.

P. Indyk and A. Naor. Nearest-neighbor-preserving embeddings. *ACM Transactions on Algorithms (TALG)*, 3(3), 2007.

D. Kobak and P. Berens. The art of using t-SNE for single-cell transcriptomics. *Nature Communications*, 10(1):5416, 2019.

K. G. Larsen and J. Nelson. Optimality of the Johnson-Lindenstrauss lemma. pages 633–638, 2017.

G. C. Linderman and S. Steinerberger. Clustering with t-SNE, provably. *SIAM Journal on Mathematics of Data Science*, 1(2):313–332, 2019.

H. Maehara. Space graphs and sphericity. *Discrete Applied Mathematics*, 7(1):55–64, 1984.

J. Matoušek. *Lectures on Discrete Geometry*, volume 212. Springer Science & Business Media, 2013.

L. McInnes, J. Healy, and J. Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *Computing Research Repository (CoRR)*, abs/1802.03426, 2018.

A. Naor. An average John theorem. *Geometry & Topology*, 25(4):1631–1717, 2021.

A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems (NIPS)*, 14, 2001.

D. Perraul-Joncas and M. Meilă. Non-linear dimensionality reduction: Riemannian metric estimation and the problem of geometric discovery. *Computing Research Repository (CoRR)*, abs/1305.7255, 2013.

J. Reiterman, V. Rödl, and E. Šiňajová. Embeddings of graphs in Euclidean spaces. *Discrete & Computational Geometry (DCG)*, 4:349–364, 1989.

F. S. Roberts. On the boxicity and cubicity of a graph. *Recent Progress in Combinatorics*, pages 301–310, 1969.

R. Sarkar. Low distortion delaunay embedding of trees in hyperbolic plane. *International symposium on graph drawing*, pages 355–366, 2011.

U. Shaham and S. Steinerberger. Stochastic neighbor embedding separates well-separated clusters. *Computing Research Repository (CoRR)*, abs/1702.02670, 2017.

T. Tkocz. An upper bound for spherical caps. *The American Mathematical Monthly*, 119(7):606–607, 2012.

L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research (JMLR)*, 9(11), 2008.

J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research (JMLR)*, 11(2), 2010.

U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17:395–416, 2007.

H. Warren. Lower bounds for approximation by nonlinear manifolds. *Transactions of the American Mathematical Society (AMS)*, pages 167–178, 1968.

K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research (JMLR)*, 10(2), 2009.

N. C. Wormald. Differential Equations for Random Processes and Random Graphs. *The Annals of Applied Probability*, 5(4):1217–1235, 1995.

A Useful Supporting Results

Lemma 25. *Fix any $q \in (0, 1]$. Let $\mathcal{D}_G$ denote a distribution over $\mathcal{G}_n(V)$ where the each edge appears independently with probability at least q. Then $\mathbb{P}_{G\sim\mathcal{D}_G}\Big[\Delta(G) \leq 2\Big] \geq 1 - n^2e^{-q^2(n-1)} \geq 1 - 2^{-\Omega(n)}$.*

Proof. The probability that any distinct $u, v \in V$ have distance strictly greater than 2 is the probability that they do not share an edge and have no common neighbor, which (by independence of edges) is $\leq (1-q)\cdot(1-q^2)^{n-2}$. By union bounding over all pairs of vertices, $\mathbb{P}[\Delta(G) > 2] \leq \binom{n}{2}(1-q)\cdot(1-q^2)^{n-2} \leq n^2e^{-q^2(n-1)}$. □

Corollary 26. *Let $S \subset \mathcal{G}_n$ be the subset of all graphs with diameter at most 2, then $|S| \geq (1 - n^2e^{-(n-1)/4})|\mathcal{G}_n|$. In particular, if $n \geq 82$, $|S| \geq (1 - 2^{-n/5})|\mathcal{G}_n| = (1 - 2^{-\Omega(n)})|\mathcal{G}_n|$.*

Proof. Recall that the uniform draw over $\mathcal{G}_n$ is equal in distribution to an Erdős–Rényi model over n nodes, $\mathcal{G}(n, \frac{1}{2})$, where each edge appears independently with probability $\frac{1}{2}$.

Invoking Lemma 25 (with $q = 1/2$) gives us that with probability at least $1 - n^2e^{-(n-1)/4}$ over a graph G drawn uniformly from $\mathcal{G}_n$, we have that $\Delta(G) \leq 2$. □

Lemma 27. *For $k \geq 4$ and $n \geq 6$ even integers, there exists a universal constant $c > 0$ such that*

$$\mathbb{P}_{G\sim\mathrm{unif}(\mathcal{G}_n^{k\text{-}reg})}\left[\Delta(G) \leq \left\lceil \frac{\log(n-1)}{\log\left(\frac{k}{2\sqrt{k-1}+1/2}\right)}\right\rceil\right] \geq 1 - cn^{-k+2}.$$

Proof. Let $\lambda_2(G)$ be the second largest eigenvalue of G's adjacency matrix. By Chung [1989] we have $\Delta(G) \leq \lceil \frac{\log(n-1)}{\log(k/\lambda_2(G))}\rceil$. By Friedman [2008], for $k \geq 4$ and $n \geq 6$ even integers, there exists a constant $c > 0$ such that with probability $1 - cn^{-\lceil\sqrt{k-1}\rceil+1} \geq 1 - cn^{-k+2}$, a random regular graph has $\lambda_2(G) \leq 2\sqrt{k-1} + 1/2$. Combining these results, we have with high probability that $\Delta(G) \leq \left\lceil \frac{\log(n-1)}{\log\left(\frac{k}{2\sqrt{k-1}+1/2}\right)}\right\rceil$. □

Lemma 28 (Johnson-Lindenstrauss; implied by Theorem 1 of Dasgupta and Gupta [2003])**.** *For any $0 < \epsilon \leq \frac{1}{2}$ and any integer n, let k be a positive integer such that $k \geq \frac{12\log n}{\epsilon^2}$. Then for any set V of n points in $\mathbb{R}^d$, there is a map $f : \mathbb{R}^d \to \mathbb{R}^k$ such that for all $u, v \in V$,*

$$(1-\epsilon)\|u-v\|_2^2 \leq \|f(u) - f(v)\|_2^2 \leq (1+\epsilon)\|u-v\|_2^2.$$

Observation 29. *For any $0 < \epsilon < r$, we have the following.*

(i) One can always ϵ-pack n points in a diameter r (open) ball in ℓ_∞^d with $d = \left\lceil \frac{\log n}{\log\lceil r/\epsilon\rceil}\right\rceil$.

(ii) One can always ϵ-pack n points in a diameter r (open) ball in ℓ_2^d with $d = \left\lceil \frac{4\log(n+1)}{2-(2\epsilon/r)^2}\right\rceil$.

Proof of Observation 29(i). Consider an ϵ-resolution grid of an r diameter open ball in ℓ_∞^d for $d = \lceil \frac{\log n}{\log\lceil r/\epsilon\rceil}\rceil$. Specifically, consider the grid points $\left\{\epsilon\cdot j : j \in \mathbb{Z}, 0 \leq j < r/\epsilon\right\}^d$. Note that by construction, diameter is strictly less than r and the number of grid points is at least n. Any subset of n points from this grid is an ϵ-packing (that is, for any p, p' distinct from the grid, $\|p - p'\|_\infty \geq \epsilon$). □

Proof of Observation 29(ii). We'll show that n points can be γ-packed in a radius 1 (i.e. diameter 2) open ball of in ℓ_2^d for d at most $\lceil \frac{4\log(n+1)}{2-\gamma^2}\rceil$, this immediately yields the desired result.

Fix any $\delta > 0$ and consider the open ball of radius $1 + \delta$ in ℓ_2^d. It suffices to show that there exists a γ-packing of at least n points on the unit sphere S^{d-1} for d sufficiently large. Let σ be the uniform probability measure over S^{d-1} and $\{p_1, \ldots, p_m\}$ be a maximally sized γ-packing of S^{d-1}. For any

$p \in S^{d-1}$, define $C_\gamma(p) := \{q \in S^{d-1} : \|p-q\|_2 < \gamma\}$ as the γ-spherical cap of S^{d-1} at p. Observe that $C_\gamma(p) = \left\{q \in S^{d-1} : p \cdot q > 1 - \frac{\gamma^2}{2}\right\}$ and that $\bigcup_{i\in[m]} C_\gamma(p_i) = S^{d-1}$. Hence:

$$m \cdot \sigma(C_\gamma(p_1)) \geq \sigma\left(\bigcup_{i\in[m]} C_\gamma(p_i)\right) = \sigma(S^{d-1}) = 1.$$

Noting that the volume of a γ-spherical cap $\sigma(C_\gamma(p_1)) \leq \exp\left(-\frac{d}{4}(2-\gamma^2)\right)$ [Tkocz, 2012], and by letting $\delta \to 0$, we have that having $d = \left\lceil \frac{4\log(n+1)}{2-\gamma^2} \right\rceil$ is sufficient to γ-pack n points in an open unit ball in ℓ_2^d. □

Low-Distortion Implies Good Preservation

In order to directly compare α-preservation and $\frac{1}{\alpha}$-distortion embeddings, we broaden the definition of α-preservation to apply to metric spaces:

Definition 30. *For a metric space $(\mathcal{Z}, \sigma)$, let $G_R(\mathcal{Z}) = (\mathcal{Z}, E)$ denote the graph where $(z,w) \in E \iff \sigma(z,w) < R$. We say a map $f : (\mathcal{Z}, \sigma) \to (\mathcal{X}, \rho)$ is an (α, R)-preservation of $(\mathcal{Z}, \rho)$ for $\alpha \in (0,2)$ and $R > 0$ if its output is an α-preservation of $G_R(\mathcal{Z})$.*

Note that $(\alpha, 2)$-preserving $G \in \mathcal{G}_n$ with respect to the shortest path metric coincides precisely with the regular definition of α-preserving G. Consider the classical notion of structure retaining embedding:

Definition 31 (Matoušek [2013])**.** *Fix any $\alpha \leq 1$. A mapping $f : (\mathcal{Z}, \sigma) \to (\mathcal{X}, \rho)$ is called a $\frac{1}{\alpha}$-*distortion*, if there exist $c > 0$ such that for all $u, v \in \mathcal{Z}$:*

$$\alpha c \cdot \sigma(u,v) \leq \rho(f(u), f(v)) \leq c \cdot \sigma(u,v).$$

It is not hard to see that α-preservation is strictly weaker than $\frac{1}{\alpha}$-distortion.

Observation 32. *For any $\alpha \leq 1$. If f constitutes an $\frac{1}{\alpha}$-distortion[9] of $\mathcal{Z}$, then for all $R > 0$, f is an (α, R)-preservation of $\mathcal{Z}$. In particular, if $\mathcal{Z} = G \in \mathcal{G}_n$ is equipped with the shortest path metric, then f constitutes an α-preservation of G.*

Proof. Let f be a $(1/\alpha)$-distortion embedding. Then there exists $c > 0$ s.t. for all $x, y \in \mathcal{Z}$: $c \cdot \sigma(x,y) \leq \rho(f(x), f(y)) \leq \frac{c}{\alpha} \cdot \sigma(x,y)$. Thus setting $r := cR/\alpha$:

$$\sigma(u,v) < R \implies \rho(f_G(v), f_G(w)) < cR/\alpha = r$$
$$\sigma(u,v) \geq R \implies \rho(f_G(v), f_G(w)) \geq cR = \alpha \cdot r.$$

□

To see that this weakness is strict, consider $\frac{1}{\alpha}$-distortion an n-point unit simplex in $\mathbb{R}^n$. The doubling dimension of the embedding will be $\Omega\left(\frac{\log(n)}{\log(\alpha)}\right)$, whereas α-preservation takes 0 dimensions for $R > 1$ and 1 dimension for $R \leq 1$.

B Omitted Proofs

B.1 Proofs from Preservation in General Metrics

Theorem 8. *For any $\alpha \in (0,2)$, we have the following.*

(i) For all $n \geq 82$, at least $1 - 2^{-n/5}$ fraction of $G \in \mathcal{G}_n$:

$$\dim_\alpha(G) \geq \frac{\log(n) - 2\log(2)}{2\log(8/\alpha)} = \Omega\left(\frac{\log(n)}{\log(8/\alpha)}\right).$$

(ii) For all even integers $n \geq 6$ and $k \geq 4$, at least $1 - O(n^{-k+2})$ fraction of $G \in \mathcal{G}_n^{k\text{-reg}}$:

$$\dim_\alpha(G) \geq \frac{\log(n/(k+1))}{\log\left(\frac{4}{\alpha}\left\lceil \frac{\log(n-1)}{\log\left(\frac{k}{2\sqrt{k-1}+1/2}\right)}\right\rceil\right)} = \Omega\left(\frac{\log(n/k)}{\log\frac{\log n}{\log k} + \log(4/\alpha)}\right).$$

[9] Note that $(\frac{1}{\alpha} < 1)$-distortion is impossible by definition.

Proof of Theorem 8(i). The proof proceeds by showing that $1 - 2^{-\Omega(n)}$ fraction of graphs in $\mathcal{G}_n$ have diameter at most 2 and clique number at most $2\sqrt{n}$. Then, applying Observation 10 and Lemma 11 yields the result.

Bounding $\Delta(G)$. Corollary 26 gives us that with probability at least $1 - n^2 e^{-(n-1)/4}$ over a graph G drawn uniformly from $\mathcal{G}_n$, we have that $\Delta(G) \leq 2$.

Bounding $|\kappa(G)|$. For any fixed $m \in [n]$ and $S \subseteq V$, define the random variables: (i) X_m as the number of (not necessarily maximal) cliques of size m in a graph G drawn uniformly from $\mathcal{G}_n$, and (ii) $Y_S = \mathbb{1}\big[G|_S \text{ is a clique}\big]$. Then by Markov's inequality:

$$\begin{aligned}\mathbb{P}_{G\sim\text{unif}(\mathcal{G}_n)}\Big[|\kappa(G)| \geq m\Big] &= \mathbb{P}_{G\sim\mathcal{G}(n,\frac{1}{2})}\Big[|\kappa(G)| \geq m\Big] = \mathbb{P}_{G\sim\mathcal{G}(n,\frac{1}{2})}\Big[X_m \geq 1\Big] \leq \mathbb{E}[X_m] \\ &= \sum_{S\subseteq V:|S|=m} \mathbb{E}[Y_S] = \sum_{S\subseteq V:|S|=m} \left(\frac{1}{2}\right)^{\binom{m}{2}} = \binom{n}{m}\left(\frac{1}{2}\right)^{\binom{m}{2}} \leq n^m 2^{-\binom{m}{2}}.\end{aligned}$$

Thus by picking $m = \lceil 2\sqrt{n} \rceil$ gives that with probability at least $1 - n^{(2\sqrt{n})} \cdot 2^{-\binom{2\sqrt{n}}{2}}$, we have that $|\kappa(G)| \leq 2\sqrt{n}$.

By combining these observations we have that (when $n \geq 82$) for at least $1 - 2^{-n/5}$ fraction of graphs in $G \in \mathcal{G}_n$,

$$\dim_\alpha(G) \geq \frac{\log(|P(G)|)}{\log(4\,\Delta(G)/\alpha)} \geq \frac{\log(n/|\kappa(G)|)}{\log(8/\alpha)} \geq \frac{\log\left(\sqrt{n}/2\right)}{\log(8/\alpha)} = \frac{\log\left(n\right) - 2\log(2)}{2\log(8/\alpha)}.$$

□

Proof of Theorem 8(ii). We will upper bound the diameter and the size of the largest clique of most graphs in $\mathcal{G}_n^{k\text{-reg}}$. Then, again applying Observation 10 and Lemma 11 yields the result.

Bounding $\Delta(G)$. Invoking Lemma 27 gives us that with probability at least $1 - O(n^{-k+2})$ over a graph G drawn uniformly from $\mathcal{G}_n^{k\text{-reg}}$, we have that $\Delta(G) \leq \left\lceil \frac{\log(n-1)}{\log\left(\frac{k}{2\sqrt{k-1}+1/2}\right)} \right\rceil$.

Bounding $|\kappa(G)|$. Since no vertex has degree larger than k, all $G \in \mathcal{G}_n^{k\text{-reg}}$ has $|\kappa(G)| \leq k+1$.

By combining these observations, for any $n \geq 6$ and $k \geq 4$ even integers, we have that for at least $1 - O(n^{-k+2})$ fraction of graphs in $G \in \mathcal{G}_n^{k\text{-reg}}$,

$$\dim_\alpha(G) \geq \frac{\log(n/|\kappa(G)|)}{\log(4\,\Delta(G)/\alpha)} \geq \frac{\log(n/(k+1))}{\log\left(\frac{4}{\alpha}\left\lceil \frac{\log(n-1)}{\log\left(\frac{k}{2\sqrt{k-1}+1/2}\right)} \right\rceil\right)} = \Omega\left(\frac{\log(n/k)}{\log\log_k(n) + \log(4/\alpha)}\right).$$

□

Proof of Proposition 15. For any $G = (V, E) \in \mathcal{G}_n$, let $P(G) = \{P_1, \ldots, P_m\}$ be the m parts of the clique partition.

Case $\alpha < 1$. Proposition 19 realizes $(\alpha < 1)$-preservation in ℓ_∞ in at most $\left\lceil \log_2(3)\left\lceil \frac{\log|P(G)|}{\log\lceil 1/\alpha \rceil} \right\rceil \right\rceil$ (doubling) dimensions.

Case $\alpha > 1$. Fix any $\epsilon > 0$ small enough such that $\lceil \frac{1-\epsilon}{\alpha-1+\epsilon} \rceil = \lceil \frac{1}{\alpha-1} \rceil$. Consider an n-point space $\mathcal{X}$ (each point corresponding to a node in V), with distance ρ defined as (for any $v, v' \in V$)

$$\rho(v,v') := \begin{cases} 0 & v = v' \\ \alpha & v \not\sim_G v' \\ 1-\epsilon & v \sim_G v',\ \{v,v'\} \not\subseteq P_i \\ \sigma_i(v,v') & \{v,v'\} \subseteq P_i \end{cases},$$

where $\sigma_i(v, v')$ is defined as follows. Let $\mathrm{C}(G|_{P_i}) = \{S_{i_1}, \ldots, S_{i_k}\}$ be the i_k parts of the neighborhood partition of $G|_{P_i}$. Recall that (see e.g. Observation 29) i_k points can be $(\alpha - 1 + \epsilon)$-packed in a $(1-\epsilon)$-diameter open ball ℓ_∞^d (for $d = \lceil \log(i_k)/\log\lceil \frac{1-\epsilon}{\alpha-1+\epsilon} \rceil \rceil$). Thus mapping all vertices in P_i to i_k points such that all elements in the same neighborhood part S_{i_j} are mapped to a single point and of different parts distance at least $(\alpha - 1 + \epsilon)$ apart (but contained within a $(1-\epsilon)$ diameter open ball), we can define $\sigma_i(v, v')$ as per the distances induced by this mapping (in ℓ_∞). It is instructive to note that either $\sigma_i(v, v') = 0$ or $\alpha - 1 + \epsilon \leq \sigma_i(v, v') < 1 - \epsilon$.

It is not hard to see that ρ is a valid *pseudo*-metric on $\mathcal{X}$. By standard metric identification one can make this into a bona fide metric that constitutes an α-preservation of G. We will now determine this metric's doubling dimension. Let $K := \max_i i_k$ be the maximum number of neighborhood parts for any P_i with $i \in [m]$, then:

- Any ball of radius $R < 1 - \epsilon$ contains points only from one part P_i for some $i \in [m]$. By construction it can be covered by $2^{\log_2(3)\lceil \log(K)/\log\lceil \frac{1-\epsilon}{\alpha-1+\epsilon}\rceil\rceil}$ balls of radius $R/2$.
- Any ball of radius $R \geq 1 - \epsilon$ can be covered by $|P(G)|$ (open) balls of radius $1 - \epsilon$, each containing points from exactly one part. Again by construction, each such ball can be covered by $2^{\log_2(3)\lceil \log(K)/\log\lceil \frac{1-\epsilon}{\alpha-1+\epsilon}\rceil\rceil}$ balls of radius $(1-\epsilon)/2$. Hence we can cover the R radius ball by $|P(G)| \cdot 2^{\log_2(3)\lceil \log(K)/\log\lceil \frac{1-\epsilon}{\alpha-1+\epsilon}\rceil\rceil}$ balls of radius $R/2$.

Since $\epsilon > 0$ is chosen small enough such that $\lceil \frac{1-\epsilon}{\alpha-1+\epsilon}\rceil = \lceil \frac{1}{\alpha-1}\rceil$, any ball in our constructed metric space can always be covered by at most $|P(G)| \cdot 2^{\log_2(3)\lceil \log(K)/\log\lceil \frac{1}{\alpha-1}\rceil\rceil}$ balls of half the radius. Thus, our construction has doubling dimension $\lceil \log_2 |P(G)| + \log_2(3)\lceil \log(K)/\log\lceil \frac{1}{\alpha-1}\rceil\rceil\rceil$.

Case $\alpha = 1$. For any integer $m > 0$, by the construction for $\alpha > 1$ case above, we know that for $\alpha = 1 + \frac{1}{m}$ we can α-preserve in a metric space $\mathcal{X}_m$ such that any ball is covered by at most $|P(G)| \cdot 2^{\log_2(3)\lceil \log(K)/\log m\rceil}$ balls of half the radius. Taking $m \to \infty$ we have $\dim_{(\alpha=1)}(\mathcal{X}_\infty) \leq \lceil \log_2(3|P(G)|)\rceil$. □

Proof of Corollary 16. Let $\Delta(G) < C$ for some constant C. Combining Lemmas 11 and 13 tells us

$$\begin{aligned}
\dim_\alpha(G) &\geq (1/2)\Big(\max_U \frac{\log|P(G|_U)|}{\log(4\,\Delta(G|_U)/\alpha)} + \mathbb{1}[\alpha > 1] \cdot \max_U \frac{\log\big|\,\mathrm{C}(G|_U)\big|}{\log(\frac{4\,\Delta(G|_U)}{\alpha-1})}\Big) \\
&\geq (1/2)\Big(\max_U \frac{\log|P(G|_U)|}{\log(4C/\alpha)} + \mathbb{1}[\alpha > 1] \cdot \max_U \frac{\log\big|\,\mathrm{C}(G|_U)\big|}{\log(\frac{4C}{\alpha-1})}\Big) \\
&= (1/2)\Big(\frac{\log|P(G)|}{\log(4C/\alpha)} + \mathbb{1}[\alpha > 1] \cdot \frac{\log\big|\,\mathrm{C}(G)\big|}{\log(\frac{4C}{\alpha-1})}\Big),
\end{aligned}$$

where the last line follows from monotonicity of clique and neighborhood partitions; for $U \subseteq W$, $|P(G|_U)| \leq |P(G|_W)|$ and $|C(G|_U)| \leq |C(G|_W)|$. Meanwhile, Proposition 15 tells us

$$\dim_\alpha(G) = O\Big(\frac{\log|P(G)|}{\log(4/\alpha)} + \mathbb{1}[\alpha > 1] \cdot \frac{\max_{S\in P(G)} \log|C(G|_S)|}{\log(\frac{8}{\alpha-1})}\Big).$$

Monotonicity tells us $\max_{S\in P(G)} \log|C(G|_S)| \leq \log|C(G)|$, completing the proof.

□

B.2 Proofs from Preservation in Normed Spaces

Theorem 17. *We have the following.*

(i) ***(General normed spaces)*** *Let $\mathbb{L}$ be the collection of all normed spaces. For all $\alpha \in (1, 2)$ and $n \geq 82$, we have that for at least $1 - 2^{-n/6}$ fraction of $G \in \mathcal{G}_n$:*

$$\dim_\alpha(G, \mathbb{L}) \geq \frac{n}{3\log_2(\frac{16}{\alpha-1})} = \Omega\left(\frac{n}{\log(\frac{16}{\alpha-1})}\right).$$

(ii) ***(Euclidean spaces)*** *For $\alpha = 1$ and $n \geq 0$, we have that for at least $1 - 2^{-n}$ fraction of $G \in \mathcal{G}_n$,*

$$\dim_{(\alpha=1)}(G, \ell_2) \geq \frac{n}{15} - \frac{1}{4}.$$

Proof of Theorem 17(i). Consider the set $S \subset \mathcal{G}_n$ of all graphs with diameter at most 2. If $n \geq 82$, we have (see Corollary 26) $|S| \geq (1 - n^2 e^{-(n-1)/4})|\mathcal{G}_n| \geq (1 - 2^{-n/5})|\mathcal{G}_n| \geq 2^{\binom{n}{2}}/2$. A straightforward application of Lemma 34 on S yields that there exists at least one graph that must require at least $\frac{(\binom{n}{2}-1)\log(2)}{n\log(16/(\alpha-1))} = \Omega(n/\log(\frac{16}{\alpha-1}))$ (doubling) dimensions to $(\alpha > 1)$-preserve any normed space. But we can do better.

Define the set of "low-embedding" graphs:

$$T := \Big\{ G \in \mathcal{G}_n \;\Big|\; \dim_\alpha(G, \mathbb{L}) \leq \frac{(n/3)\log(2)}{\log(16/(\alpha-1))} =: d \Big\}.$$

Assume towards contradiction that $|T| \geq 2^{-n/6}|\mathcal{G}_n|$. Then observe that $|T \cap S| = |T \setminus S^c| \geq |T| - |S^c| \geq (2^{-n/6} - 2^{-n/5})|\mathcal{G}_n| \geq 2^{-n/5}|\mathcal{G}_n| = 2^{\binom{n}{2}-(n/5)}$ (when $n \geq 82$). Thus $\dim_\alpha(T \cap S, \mathbb{L}) \leq \dim_\alpha(T, \mathbb{L}) \leq d$ (see Proposition 7), but by Lemma 34 $\dim_\alpha(T \cap S, \mathbb{L}) \geq \frac{(\binom{n}{2}-\frac{n}{5})\log(2)}{n\log(16/(\alpha-1))} > d$ (for $n \geq 5$). This implies there exists a graph $G \in T \cap S \subseteq T$ where $\dim_\alpha(G) > d$, arriving at a contradiction.

Therefore[10] (when $n \geq 82$), at least $(1 - 2^{-n/6})$ fraction of graphs in $\mathcal{G}_n$ require at least $d = \frac{(n/3)\log(2)}{\log(16/(\alpha-1))}$ (doubling) dimensions to be $(\alpha > 1)$-preserved in any normed space. □

Observation 33. *Let $S \subseteq \mathcal{G}_n$. Suppose $f : S \to \mathcal{X}^n$ α-preserves S in a normed space $\mathcal{X}$. Then there exists $f' : S \to \mathcal{X}^n$ which α-preserves S such that for all $G \in S$*

- *(equal scaling) The neighborhood threshold for $f'(G)$ is exactly 1 (c.f. Definition 1).*
- *(centeredness) $(1/n)\sum_{v \in V} f'_G(v) = \vec{0}$.*

Lemma 34. *Let $\mathbb{L}$ be the collection of all normed spaces. For any set $S \subseteq \mathcal{G}_n$ of graphs on n vertices with diameter at most R, and any $\alpha \in (1, 2)$,*

$$\dim_\alpha(S, \mathbb{L}) \geq \frac{\log|S|}{n\log(8R/(\alpha-1))}.$$

Proof. WLOG assume $R < \infty$, and $f : S \to \mathcal{X}^n$ be an $(\alpha > 1)$-preserving map into a normed space $(\mathcal{X}, \|\cdot\|)$ such that for all $G \in S$, $f(G)$ produces an embedded centered at the origin and has the neighborhood threshold of 1 (c.f. Observation 33). Let $B_R(\vec{0}) \subset \mathcal{X}$ be the open ball of radius R about the origin. Observe that $\{f_G(v) : G \in S, v \in V\} \subset B_R(\vec{0})$, since for all $G \in S$ and $v \in V$,

$$\|f_G(v)\| = \left\| f_G(v) - \frac{1}{n}\sum_{u \in V} f_G(u) \right\| \leq (1/n)\sum_{u \in V} \|f_G(v) - f_G(u)\| \leq \operatorname{diam}(f(G)) < 1 \cdot \Delta(G) \leq R.$$

Fix an $((\alpha-1)/4)$-cover N of $B_R(\vec{0})$. Per Observation 3, we may assume $|N| \leq (8R/(\alpha-1))^{\dim(\mathcal{X})}$. Define the projection map $\Phi : \mathcal{X}^n \to N^n$ by $\Phi((x_i)_{i\in[n]}) = (\operatorname{argmin}_{p \in N} \|p - x_i\|)_{i\in[n]}$, where ties are broken in an arbitrary but consistent way.

Consider the map from graphs to their projections onto the net, $\Phi \circ f : S \to N^n$. We claim this map is injective. Suppose, for sake of contradiction, it was not injective. Then there exists distinct $G, G' \in S$ such that $\Phi(f(G)) = \Phi(f(G'))$. Let (u, v) be a pair on which G and G' disagree on an edge. Without loss of generality, we may assume $u \nsim_G v$ and $u \sim_{G'} v$, thus

$$\|f_G(u) - f_G(v)\| \geq \alpha \qquad \text{and} \qquad \|f_{G'}(u) - f_{G'}(v)\| < 1.$$

For convenience, let $\Phi(p) \in N$ denote the projection of a point p onto the net. Then by assumption we have $\Phi(f_G(u)) = \Phi(f_{G'}(u)) =: N_u$ and $\Phi(f_G(v)) = \Phi(f_{G'}(v)) =: N_v$. Then we have

$$\begin{aligned}\|f_G(u) - f_G(v)\| &\leq \|f_G(u) - N_u\| + \|N_u - f_{G'}(u)\| + \|f_{G'}(u) - f_{G'}(v)\| + \|f_{G'}(v) - N_v\| + \|N_v - f_G(v)\| \\ &< \frac{\alpha-1}{4} + \frac{\alpha-1}{4} + 1 + \frac{\alpha-1}{4} + \frac{\alpha-1}{4} = \alpha.\end{aligned}$$

This contradicts the fact that $\|f_G(u) - f_G(v)\| \geq \alpha$ and thereby establishes injectivity. Therefore we have

$$|S| \leq |N^n| = |N|^n = \Big(\frac{8R}{\alpha-1}\Big)^{n\dim(\mathcal{X})}.$$

[10] The constants in the proof can be improved.

Rearranging the expression, we have that, for any f which $(\alpha > 1)$-preserves S in a normed space $\mathcal{X}$, the doubling dimension of $\mathcal{X}$ is at least $\frac{\log |S|}{n \log(8R/(\alpha-1))}$. □

Proof of Theorem 17(ii). A theorem in Section 3.1 of [Reiterman et al., 1989] says that for $n \geq 38$, $(1 - \frac{1}{n})$-fraction of $G \in \mathcal{G}_n$ have sphericity at least $n/15 - 1$. This lower bound carries over for for $(\alpha = 1)$-preservation in ℓ_2, since $\dim(\ell_2^d) \leq d$. We strengthen this result by showing the lower bound applies to $(1 - \frac{1}{2^n})$-fraction of $\mathcal{G}_n$. We use a similar argument which relies on upper bounds for consistent sign assignments of bounded-degree polynomials.

We specialize Lemma 35 to the $p = 2$ case. Fix any $S \subseteq \mathcal{G}_n$ with $S \geq 2^{-n}|\mathcal{G}_n|$ such that some $f : \mathcal{G}_n \to \ell_2^d$ $(\alpha = 1)$-preserves S. Then by Lemma 35:

$$
\begin{aligned}
& & |S| &\leq \left(\frac{4e(n-1)}{d}\right)^{nd} \\
\implies & & 2^{\binom{n}{2}-n} = 2^{-n}|\mathcal{G}_n| &\leq \left(\frac{4e(n-1)}{d}\right)^{nd} \\
\iff & & \left(\frac{n-3}{2}\right)\log 2 &\leq d \log\left(\frac{4e(n-1)}{d}\right) = d\left(\log\left(\frac{e(n-1)\log 2}{10d}\right) + \log\left(\frac{40}{\log 2}\right)\right) \\
\implies & & \left(\frac{n-3}{2}\right)\log 2 &\leq \left(\frac{n-1}{10}\right)\log 2 + d\log\left(\frac{40}{\log 2}\right) \\
\iff & & d &\geq \frac{2n-7}{5\log_2\left(\frac{40}{\log 2}\right)} \geq \frac{n}{15} - \frac{1}{4}.
\end{aligned}
$$

Thus no more than 2^{-n} fraction of $\mathcal{G}_n$ can be $(\alpha = 1)$-preserved in ℓ_2^d for $d < \frac{n}{15} - \frac{1}{4}$. To finish the proof, it is noted that the doubling dimension of ℓ_2^d is at least d. □

Lemma 35. *For $p \geq 2$ even integer and $S \subseteq \mathcal{G}_n$, if there exists $f : \mathcal{G}_n \to \ell_p^d$ that $(\alpha = 1)$-preserves S, then:*

$$|S| \leq \left(\frac{2ep(n-1)}{d}\right)^{nd}.$$

Proof. For any $G = (V, E) \in S$. Since f $(\alpha = 1)$-preserves G, there exists $r > 0$ and $f_G : V \to \mathbb{R}^d$ such that:

$$
\begin{aligned}
(u, v) \in E &\implies \|f_G(u) - f_G(v)\|_p < r \\
(u, v) \notin E &\implies \|f_G(u) - f_G(v)\|_p \geq r.
\end{aligned}
$$

Pick $\epsilon_n > 0$ such that for all $G = (V, E) \in S$ and $(u, v) \in E$, $\|f_G(u) - f_G(v)\|_q < r - \epsilon_n$ which exists since $|S|$ and $n < \infty$ are finite. Then,

$$\{P_{uv}(x, x') = \|x - x'\|_p^p - r + \epsilon_n \mid x, x' \in \mathbb{R}^d,\ u, v \in V, u \neq v\}$$

is a set of $\binom{n}{2}$ polynomials (since p is even) over nd variables with the property that for all $u, v \in V, u \neq v$:

(i) If $(u, v) \in E$, then $P_{uv}(f_G(u), f_G(v)) < 0$.

(ii) If $(u, v) \notin E$, then $P_{uv}(f_G(u), f_G(v)) > 0$.

Therefore, each $G \in S$ will produce a unique non-zero sign assignment of $\{P_{uv} | u, v \in V, u \neq v\}$, so by Lemma 36:

$$|S| \leq \left(\frac{4ep\binom{n}{2}}{nd}\right)^{nd} = \left(\frac{2ep(n-1)}{d}\right)^{nd}.$$

□

Lemma 36 (Warren [1968]). *If $\{p_1, \ldots, p_m\}$ is a set of polynomials of degree at most $D \geq 1$ in N variables with $m \geq N$, then the number of consistent non-zero sign assignments to the p_i is at most $(4eDm/N)^N$, that is, $\left|\left\{\big(\text{sign}(p_1(x)), \ldots, \text{sign}(p_m(x))\big) \ : \ x \in \mathbb{R}^N \text{ s.t. } p_i(x) \neq 0 \text{ for } i \in [m]\right\}\right| \leq \big(4eDm/N\big)^N$.*

Proof of Proposition 20. Fix any $G \in \mathcal{G}_n$, and let $P(G) = \{P_1, \ldots, P_m\}$ be the m parts of its clique partition. WLOG, assume each vertex in G has degree at least 1 (since isolated vertices can be embedded separately).

Case $\alpha \in (0, \frac{1}{\sqrt{2}})$. A simple α-packing of m points in a unit ℓ_2 ball achieves the desired α-preserving embedding. Specifically, by Observation 29(ii) (take $n = m$, $r = 1$ and $\epsilon = \alpha$) we know that m points can be α-packed in an (open) unit ball in ℓ_2^d with $d = \lceil \frac{4\log(m+1)}{2-4\alpha^2} \rceil$. Thus, the mapping where all vertices of the partition P_i of the input graph G is mapped to the i-th point of the packing in unit ball yields an α-preserving embedding in ℓ_2, with (doubling) dimension at most $\left\lceil \log_2(5) \left\lceil \frac{4\log(|P(G)|+1)}{2-4\alpha^2} \right\rceil \right\rceil$.

Case $\alpha \in (\frac{1}{\sqrt{3}}, 1)$. A low-distortion embedding of a regular simplex of m points in ℓ_2 can be used to get a good $(\alpha < 1)$-preservation. In particular, consider a regular unit simplex of m points in ℓ_2. By Lemma 28 we know that for any $0 < \epsilon \leq \frac{1}{2}$, a $\sqrt{\frac{1+\epsilon}{1-\epsilon}}$-distortion embedding of these m points exists in ℓ_2^d with $d = \lceil \frac{12}{\epsilon^2} \log m \rceil$. Hence by picking $\epsilon = \frac{1-\alpha^2}{1+\alpha^2}$, we get an α-distortion and therefore an α-preservation in ℓ_2^d for $\alpha \in (\frac{1}{\sqrt{3}}, 1)$, cf. Observation 32.

Case $\alpha \in \left[1, \frac{1}{\sqrt{1-\min(1,(1/4\lambda_G))}}\right)$. Fix an arbitrary $G \in \mathcal{G}_n$. Let $A = A(G)$ be the adjacency matrix of G with maximum eigenvalue λ_G, and let A^c be the adjacency matrix of the complement graph. Define $D := A^c + (1 - \frac{1}{\lambda_G})A$ as a squared interpoint distance matrix over the n vertices. Using a theorem of Schoenberg [Bavaud, 2011], we can verify that the (non-squared) distances of D are ℓ_2^n isometrically embeddable: for $u \in \mathbb{R}^n$ with $\|u\|_2 = 1$ and $\mathbf{1}^T u = 0$, we have

$$u^T D u = u^T \Big(\mathbf{1}\mathbf{1}^T - I_n - \frac{1}{\lambda_G} A\Big) u = -\|u\|_2^2 - \frac{1}{\lambda_G} u^T A u \leq -1 + \frac{1}{\lambda_G} |u^T A u| \leq 0.$$

Hence the same ℓ_2^n embedding is also an $\big(\alpha' = (1 - 1/\lambda_G)^{-1/2}\big)$-preservation of G: if $i \not\sim j$, then $D_{ij} = 1$, whereas if $i \sim j$, then $D_{ij} = (1 - 1/\lambda_G)^{-1/2}$. Given this realization of points in $\mathbb{R}^n$, we can apply Lemma 28 (with $\epsilon = \frac{1}{4\lambda_G} \leq \frac{1}{4}$; note that λ_G is at least the average vertex degree, which is at least 1 in this graph by assumption) to conclude neighborhood preservation of G is possible in $\mathbb{R}^d$ with $d = \lceil 12 \cdot 16\lambda_G^2 \log n \rceil$ with a slight degradation in the α-parameter. In particular, α-preservation is possible in $\mathbb{R}^d$ for $\alpha \in \left(1, \big(1 - \frac{1}{4\lambda_G}\big)^{-1/2}\right) \subseteq \left(1, \alpha' \sqrt{\frac{1-\epsilon}{1+\epsilon}}\right)$, cf. Observation 32. The extra factor of $\log_2(5)$ is an artifact of switching from ℓ_2 dimension to doubling dimension. The proposition statement follows by replacing G with $G/C(G)$ (see proof of Proposition 19 for the definition), therefore n with $|C(G)|$, in the argument.

Case $\alpha > (1 - \frac{1}{\lfloor n/2 \rfloor})^{-1/2}$. For simplicity of our discussion assume n is even, and consider a complete bipartite graph $G \in \mathcal{G}_n$ with parts S_0 and S_1 each containing $n/2$ vertices. For convenience label the vertices $1, \ldots, n$ (in any order).

We shall show that if an α-preservation of G exists in ℓ_2 then necessarily $\alpha \leq (1 - \frac{1}{\lfloor n/2 \rfloor})^{-1/2}$. Let $\{x_1, \ldots, x_n\}$ be an α-preservation of this graph in ℓ_2 with neighborhood threshold r (see Definition 1). WLOG we can assume the embedding resides in $\mathbb{R}^n$. Let $u \in \mathbb{R}^n$ be such that $u_i = \mathbf{1}[i \in S_0] - \mathbf{1}[i \in S_1]$. Clearly $u^T \mathbf{1} = 0$. Then by a theorem of Schoenberg [Bavaud, 2011], $u^T D u \leq 0$, where $D \in \mathbb{R}^{n \times n}$ is the squared interpoint distance matrix: $D_{ij} = \|x_i - x_j\|_2^2$. By definition of u, this implies $\sum_{i \not\sim j} D_{ij} \leq \sum_{i \sim j} D_{ij}$. Applying the definition of α-preservation (i.e. embedded pairwise distance of any edge-connected pair of vertices is at most r, otherwise is at least αr), and counting the number of edges and non-edges we have

$$\alpha^2 r^2 (n^2/2 - n) \leq \sum_{i \not\sim j} D_{ij} \leq \sum_{i \sim j} D_{ij} \leq r^2 n^2 / 2.$$

The requirement on α for ℓ_2 preservation follows. When $n \geq 4$ is odd, a similar analysis holds: let G be a complete bipartite graph on $n - 1$ (even) vertices and let the remaining n-th vertex have no edges. Define u the same as before with $u_n = 0$. The same analysis yields the stated restriction on α. □

B.3 Proofs from Preservation of Clustered Data

Theorem 23. *Fix $n \geq k \geq 1$ and let $\{S_1, \ldots, S_k\}$ be a partition of V (of size n). Let $c \geq 1$ be such that $\max_{i\in[k]} |S_i| \leq \frac{cn}{k}$, and $0 < q \leq p \leq 1$ with $q < 1$. Then for all $\alpha \in (0, 2)$ with probability at least $1 - \exp(-\Omega(n^{\min(2(p+q-pq),1)}))$ over $G \sim \mathtt{PP}_{p,q}(S_1, \ldots, S_k)$:*

$$\dim_\alpha(G) \geq \frac{1}{\log(8/\alpha)}\Big((1-\xi_{p,q})\log n + \xi_{p,q}\log(k/2c)\Big),$$

where $\xi_{p,q} := p - q + pq$ encodes the cluster saliency *of the planted partition model.*

Proof of Theorem 23. The proof of the theorem follows from showing (for any $G = (V, E) \sim \mathtt{PP}_{p,q}(S_1, \ldots, S_k)$): (i) $\Delta(G) \leq 2$ with probability at least $1 - \exp(-\Omega(n))$, and (ii) $|\kappa(G)| \leq \left(\frac{2cn}{k}\right)^{\xi_{p,q}}$ with probability at least $1 - \exp(-\Omega(n^{2p+2q-2pq}))$. Thus by applying Lemma 11 and Proposition 10 we get that for $G \sim \mathtt{PP}_{p,q}(S_1, \ldots, S_k)$:

$$\mathbb{P}\left[\dim_\alpha(G) < \frac{1}{\log(8/\alpha)}\log\left(\frac{n}{(2cn/k)^{\xi_{p,q}}}\right)\right] \leq \mathbb{P}\left[G \text{ has } |\kappa(G)| > \left(\frac{2cn}{k}\right)^{\xi_{p,q}} \text{ or } \Delta(G) > 2\right]$$
$$\leq \exp(-\Omega(n^{2p+2q-2pq})) + \exp(-\Omega(n)) \leq \exp(-\Omega(n^{\min(2p+2q-2pq,1)})).$$

Bounding the diameter. The diameter bound follows directly from Lemma 25: (since $0 \leq q \leq p$) with probability at least $1 - n^2 e^{-q^2(n-1)}$, the diameter of G is at most 2.

Bounding the clique number. We bound the probability that $G \sim \mathtt{PP}_{p,q}(S_1, \ldots, S_k)$ contains a clique of size greater than $m := \lceil \left(\frac{2cn}{k}\right)^{p+q-pq} \rceil$. For any $S \subseteq V$ and $G = (V, E) \sim \mathtt{PP}_{p,q}(S_1, \ldots, S_k)$, define the random variables $Y_S := \mathbb{1}[G|_S \text{ is a clique}]$. Then $C_m := \sum_{S\subseteq V:|S|=m} Y_S$ is the number of m-sized cliques in G. Therefore:

$$\begin{aligned}
\mathbb{P}_{G\sim\mathtt{PP}_{p,q}(S_1,\ldots,S_k)}\Big[G \text{ contains an } m \text{ sized clique}\Big] &= \mathbb{P}[C_m \geq 1] \leq \mathbb{E}[C_m] = \sum_{S\subseteq V:|S|=m} \mathbb{E}[Y_S] \\
&= \sum_{S\subseteq V:|S|=m} \mathbb{P}(G|_S \text{ is a clique}) \\
&= \sum_{S\subseteq V:|S|=m} p^{\binom{|S\cap S_1|}{2}+\cdots+\binom{|S\cap S_k|}{2}} q^{\sum_{1\leq i<j\leq k}|S\cap S_i|\cdot|S\cap S_j|} \\
&= q^{\binom{m}{2}} \sum_{S\subseteq V:|S|=m} (p/q)^{\binom{|S\cap S_1|}{2}+\cdots+\binom{|S\cap S_k|}{2}},
\end{aligned}$$

where $\binom{\sum_i x_i}{2} = \sum_i \binom{x_i}{2} + \sum_{i<j} x_i x_j$ holds for all $x_1, \ldots, x_n$ non-negative integers. Define:

$$A := \{S \subseteq V : |S| = m, \exists i \in [k] \text{ s.t. } |S \cap S_i| > m/2\},$$
$$B := \{S \subseteq V : |S| = m, \forall i \in [k] \text{ s.t. } |S \cap S_i| \leq m/2\}.$$

Note that $A \sqcup B = \{S \subseteq V : |S| = m\}$. Continuing onwards, we have

$$\begin{aligned}
&= q^{\binom{m}{2}}\left(\sum_{S\in A}(p/q)^{\binom{|S\cap S_1|}{2}+\cdots+\binom{|S\cap S_k|}{2}} + \sum_{S\in B}(p/q)^{\binom{|S\cap S_1|}{2}+\cdots+\binom{|S\cap S_k|}{2}}\right) \\
&\leq q^{\binom{m}{2}}\left(|A|(p/q)^{\binom{m}{2}} + |B|(p/q)^{(\frac{m^2}{4}-\frac{m}{2})}\right),
\end{aligned}$$

where for the first exponent, we use $\sum_i \binom{x_i}{2} \leq \binom{\sum_i x_i}{2}$, and for the second exponent, we use Hölder's inequality: $\langle x, y\rangle \leq \|x\|_\infty \|y\|_1$, therefore $\sum_i \binom{|S\cap S_i|}{2} = \frac{1}{2}\sum_i (|S \cap S_i| - 1)|S \cap S_i| \leq \frac{1}{2}\left(\frac{m}{2} - 1\right) m$. We can upper bound $|B|$ by $\binom{n}{m}$ and $|A|$ by $\binom{n}{m} \cdot \mathbf{1}[\frac{m}{2} < \frac{cn}{k}]$ because $\max_i |S \cap S_i| \leq \max_i |S_i| \leq cn/k$. Thus,

$$\begin{aligned}
&\leq q^{\binom{m}{2}}\left(\binom{n}{m}(p/q)^{\binom{m}{2}} \cdot \mathbb{1}\left[m < \frac{2cn}{k}\right] + \binom{n}{m}(p/q)^{(\frac{m^2}{4}-\frac{m}{2})}\right) \\
&= \binom{n}{m} p^{\binom{m}{2}}\left(\mathbb{1}\left[m < \frac{2cn}{k}\right] + (q/p)^{m^2/4}\right).
\end{aligned}$$

From here we can evaluate two cases:

- If $m = \lceil \left(\frac{2cn}{k}\right)^{p+q-pq} \rceil \geq 2cn/k$, then our bound is $\binom{n}{m} q^{m^2/4} p^{m^2/4 - m/2} = \exp(-\Omega(n^{2p+2q-2pq}))$.
- If $m = \lceil \left(\frac{2cn}{k}\right)^{p+q-pq} \rceil < 2cn/k$, then it must be the case that $p < 1$, so our upper bound becomes $\binom{n}{m}\left(p^{\binom{m}{2}} + p^{(\frac{m^2}{4} - \frac{m}{2})} q^{m^2/4}\right) = \exp(-\Omega(n^{2p+2q-2pq}))$.

Therefore, $\mathbb{P}_{G \sim \mathtt{PP}_{p,q}(S_1,\ldots,S_k)}\left[|\kappa(G)| \geq m\right] \leq \exp(-\Omega(n^{2p+2q-2pq}))$. □

Proof of Proposition 24. For any $i \in [k]$ and $u, v \in S_i$, the probability that $N(u) = N(v)$ (see Definition 12) is at most

$$(p^2 + (1-p)^2)^{|S_i|} \cdot (q^2 + (1-q)^2)^{n-|S_i|} \leq (q^2 + (1-q)^2)^{n(1-c/k)} \leq \max(q, 1-q)^{2n(1-c/k)}.$$

For any $u \in S_i$ and $v \in S_j$ with $i \neq j$, the probability that $N(u) = N(v)$ is at most

$$(pq + (1-p)(1-q))^{|S_i|+|S_j|} \cdot (q^2 + (1-q)^2)^{n-|S_i|-|S_j|} \leq \max(q, 1-q)^{2n}.$$

Applying Lemma 25, with probability at least $1 - n^2\left(\max(q, 1-q)^{2n(1-c/k)} + e^{-q^2(n-1)}\right)$, we have $|C(G)| = n$ and $\Delta(G) \leq 2$. The lower bound follows from applying Lemma 13. □