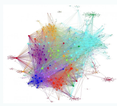
Compressibility Barriers to Neighborhood-Preserving Data Visualizations

Noah Bergam (Columbia University, CS Dept.)
Joint work w. Szymon Snoeck and Nakul Verma

arxiv:2508.07119

COLUMBIA UNIVERSITY

PROBLEM



Many scientists want to visualize big (say, 100k points, 10k dimensions) datasets in 2D or 3D plots, and often resort to (unreliable) tools like t-SNE and UMAP.

Is it even possible to accurately <u>visualize neighborhoods</u> in large, structured datasets?

FINDINGS

Most graphs are maximally difficult.
For most* n-vertex G,

$$\dim_\alpha(G) \gtrsim \frac{\log n}{\log\left(\frac{8}{\alpha}\right)}$$

(Note: $\dim_\alpha(G) \le \log(n)$ for all G)

Sparsity hardly helps in general metrics.
For most n-vertex, k-regular G, $k = O(1)$,

$$\dim_\alpha(G) \gtrsim \frac{\log n}{\log\left(\frac{\log n}{\alpha}\right)}$$

*most = for all but $2^{-\Omega(n)}$ fraction

Sparsity helps in Euclidean space.
For most n-vertex graphs:

$$\dim_{\alpha=1}(G, \ell_2) = \Theta(n)$$

Meanwhile, for *any* degree-k graph:

$$\dim_{\alpha \le 1+1/\sqrt{k}}(G, \ell_2) \lesssim k^2 \log n$$

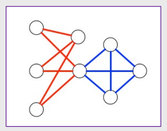
(If α exceeds this threshold, ℓ_2 preservation can become impossible!)

Cluster structure rarely helps.
For G sampled from $(p > q)$ planted partition model on n nodes and k components, w. h. p. ,

If $p = 1$ and $\alpha \le 1$, $\dim_\alpha(G) = \Theta(\log k)$,
otherwise… $\dim_\alpha(G) = \Theta(\log n)$.

$(\alpha \le 1)$- versus $(\alpha > 1)$-preservation.
An interesting phase-change occurs.

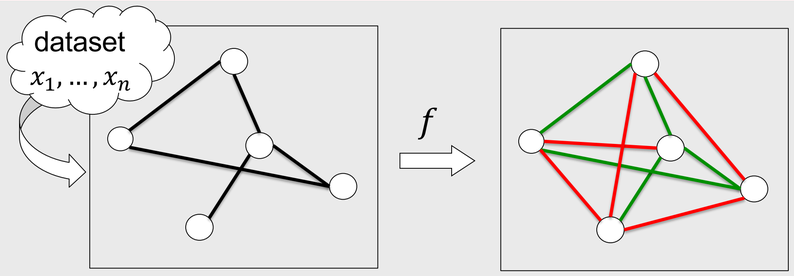
$\dim_{\alpha \le 1}(G)$ depends on the **minimal clique partition of G** (NP-hard to compute).



a size-2 clique partition

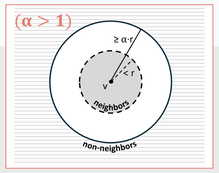
$\dim_{\alpha>1}(G)$ depends on the **# of distinct neighborhoods in G** (often much larger, but easy to compute).

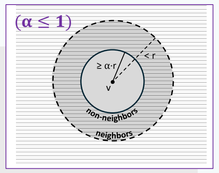
A simple model: α-preservation



ground-truth graph $G = (V, E)$
(representing *neighborhoods*)

embedding in metric space (X, ρ)





Definition: A map $f: V \to X$ is an **α-preservation** of $G = (V, E)$ if

$$(u, v) \in E \implies d\big(f(u), f(v)\big) < 1$$
$$(u, v) \notin E \implies d\big(f(u), f(v)\big) \ge \alpha$$

As α increases, the problem gets harder.
$(\alpha > 1)$ is significantly more desirable than $(\alpha \le 1)$.

Let $\dim(X)$ denote the **doubling dimension** of a metric space.

Definition: **α-preservation dimension** of graph G in collection of metric spaces $\mathcal{X}$

$$\dim_\alpha(G, \mathcal{X}) := \min\{d > 0 : \exists\, X \in \mathcal{X} \text{ and } f: V \to X \text{ such that } \dim(X) = d\}$$

$$\dim_\alpha(G) := \dim_\alpha(G, \{\text{all metric spaces}\})$$

<u>GOAL</u>: characterize preservation dimension for "realistic" graphs (sparse, clustered, etc.).
<u>IDEA</u>: constant preservation dimension suggests that 2D/3D visualization is possible!

DISCUSSION

→ Special cases of α-preservation dimension have been studied before, e.g. $\dim_{\alpha=1}(G, \ell_2)$ = "sphericity".
→ $(\alpha \le 1)$-preservation is a strict generalization of $(1/\alpha)$-distortion embedding. Low-distortion embeddings are well-studied in TCS, but too strict a notion for visualization.
→ Future work: **algorithms and complexity** of α-preservation; and **approximate** α-preservation: how many edges do need to "sacrifice" to preserve in 2D or 3D?