# t-SNE EXAGGERATES CLUSTERS, PROVABLY

**Noah Bergam**[†], **Szymon Snoeck**[*], **& Nakul Verma**[†]
[†]Computer Science Department; [*]Applied Mathematics Department
Columbia University
{njb2154,sgs2179}@columbia.edu, verma@cs.columbia.edu

## ABSTRACT

Central to the widespread use of t-distributed stochastic neighbor embedding (t-SNE) is the conviction that it produces visualizations whose structure roughly matches that of the input. To the contrary, we prove that (1) the strength of the input clustering, and (2) the extremity of outlier points, *cannot* be reliably inferred from the t-SNE output. We demonstrate the prevalence of these failure modes in practice as well.

## 1 INTRODUCTION

t-SNE and related data visualization methods have become staples in modern exploratory data analysis. They just seem to work: practitioners find that these techniques effortlessly tease out interesting cluster structures in datasets. Consequently they are now used ubiquitously in a wide array of fields, ranging from single-cell genomics to language model interpretability (Kobak & Berens, 2019; Petukhova et al., 2025). The practical success of these techniques has naturally piqued some interest in the theoretical computer science community as well.

Existing analysis of t-SNE has established that, given high-dimensional data with spherical, well-separated cluster structure, t-SNE outputs a visualization which preserves that cluster structure (Arora et al., 2018; Linderman & Steinerberger, 2019). In other words, t-SNE is provably good at generating *true positives* in its visualization of clusters. Curiously, t-SNE's susceptibility to generate *false positives*, i.e. fabricated clusters in the output visualization, has remained largely unstudied. One should note that this is not a purely academic curiosity, since the interpretation of t-SNE outputs have important consequences downstream in the sciences, influencing hypothesis generation, experimental design, and scientific conclusions.

As an illustration of the danger of false positives, consider the 2D t-SNE visualization of a 100-point dataset residing in $\mathbb{R}^{100}$ (depicted on the right).

Based on this plot, it is tempting to conclude that the input dataset obviously contains two distinct clusters. In this case, one would likely design their subsequent data analysis workflow guided by these two salient clusters. However a closer examination of the original (high-dimensional) dataset reveals that the situation perhaps may not be as clear-cut. By standard cluster saliency metrics, for instance, the input dataset appears quite poorly clustered according to the partition that t-SNE so strongly suggests, see Table 1.

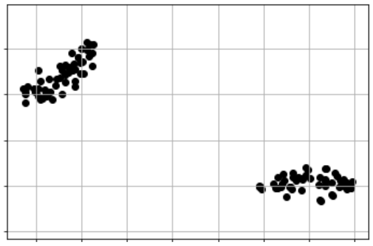
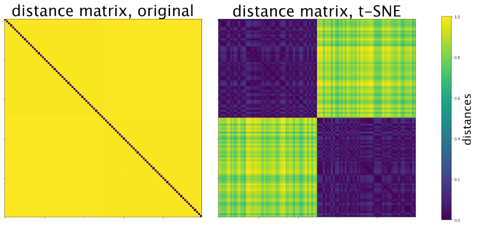


Table 1: Clustering scores (with respect to $k$-means) according to various popular cluster saliency metrics. The range in the first column specifies the possible values that can be attained. A higher value indicates data being highly clustered.

| Cluster Score (range) | t-SNE (2D) | Original Data (100D) |
|---|---|---|
| Silhouette $[-1, 1]$ | .918 | .006 |
| Calinski-Harabasz $[0, \infty]$ | 5590 | 1.61 |
| Dunn Index $[0, \infty]$ | 3.65 | .998 |

The interpoint distance matrix plots on the right further elucidate this discrepancy. t-SNE's two-dimensional visualization (right) features a sizable separation between small intra-cluster and large inter-cluster distances. This separation is not present in the original input data (left), where interpoint distances are near-uniform.



Our work formalizes this phenomenon and other "cluster-happy" behaviors exhibited by t-SNE. Our theoretical analysis, suffused with experiments, shows that one should take positively clustered outputs with a grain of salt. Our contributions are as follows:

- **Misrepresentation of clusters:** We prove that both highly-clustered and arbitrarily *un*-clustered datasets can produce the same maximally clustered visualization, see Theorem 3 and Corollary 4. Moreover, we prove that arbitrarily close inputs can have vastly distinct visualizations, see Theorem 5. We identify the peculiar property of t-SNE that explains these behaviors. We use this understanding to design a targeted adversarial attack that disrupts cluster structure in the output, see Figure 3.
- **Misrepresentation of outliers**: We prove that, regardless of input structure, the resulting t-SNE output is incapable of depicting extreme outliers, in the sense of depicting one point as substantially far away from all the others, see Theorem 8. In practice, on both synthetic and real datasets, we observe a more concerning phenomenon that faraway outliers are often subsumed into the cluster structure of the bulk of points, see Figures 4 and 5.

While there has been some work investigating the shortcomings of t-SNE in various practical settings (see Section 2.2 for a detailed discussion of the relevant literature), to the best of our knowledge this is the first work which theoretically analyzes some of the key limitations of t-SNE.

## 2 Related Work

Confidence in the data visualizations produced by t-SNE and related methods is a somewhat contentious subject in data science (Marx, 2024). Some works argue that these methods have merit in terms of preserving cluster structure, while others warn us about the fundamental issues with them and the broader goal of data visualization.

### 2.1 Performance Guarantees and Analysis of t-SNE

Shaham & Steinerberger (2017) were among the first to provide a guarantee on the visualization produced by optimal SNE embeddings of well-clustered data. Works by Linderman & Steinerberger (2019) and Arora et al. (2018) refined and extended this analysis, showing that t-SNE outputs produced using gradient descent yield well-clustered visualizations so long as the input is sufficiently well-clustered. The latter work established this guarantee in considerable generality, including cases where the input is sampled from a mixture of well-separated log-concave distributions.

Along with these algorithmic performance guarantee results, there is a line of work that seeks to establish a more fundamental understanding of t-SNE as an optimization problem. Cai & Ma (2022), for instance, characterized the distinct phases of gradient-based optimization of t-SNE, and proved an asymptotic equivalence between the early exaggeration phase and spectral clustering. Auffinger & Fletcher (2023) proved a consistency result for a continuous analogue of t-SNE, viewing the optimization problem as producing a map between distributions rather than just point sets. Jeong & Wu (2024) and Weinkove (2024) studied the gradient flow of t-SNE. The former showed mild assumptions under which optima exist, and the latter showed that, even in cases where the gradient flow diverges the relative interpoint distances stabilize in the limit.

### 2.2 Weaknesses and Criticisms

Bunte et al. (2012) were among the first to investigate the potential shortcomings of using KL-divergence in a t-SNE visualization and proposed a generalization to other divergences that may be better suited for specific datasets and user needs. Building upon the precision-recall framework of Venna et al. (2010), Im et al. (2018) extended this result and explored specific intrinsic structures within data that may be less suited for t-SNE. They concluded that while t-SNE is more attuned to reveal intrinsic cluster structure, it usually fails to reveal intrinsic manifold structure.

In terms of analyzing cluster structure specifically, Yang et al. (2021) provided empirical evidence that t-SNE visualizations are prone to *false negatives*. They presented a selection of well-clustered real-world datasets which t-SNE embeddings, even with reasonable parameter-tuning, do not seem to faithfully represent. They also showed that these practical datasets do not abide by the theoretical cluster separation conditions that are required by Arora et al. (2018) analysis. Chari & Pachter (2023) argued that t-SNE and UMAP are unreliable tools for exploratory data analysis. Taking single-cell genomic data as an important real-world example, they provided systematic empirical evidence that these embeddings suffer high distortion, and often misrepresent neighborhood and cluster structure. Curiously, to the best of our knowledge, there is no systematic theoretical study investigating false positive behavior of t-SNE.

More recently, Snoeck et al. (2025) provided theoretical evidence that, not just t-SNE, but any embedding technique that attempts to visualize data in constant dimensions is bound to misrepresent the neighborhood structure in most datasets. This work focuses exclusively on how misrepresentations induced by t-SNE visualizations can lead to false conclusions in terms of data analysis.

## 3 Preliminaries

Given an input dataset[1] $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^D$, the goal of t-SNE is to find an embedding $Y = \{y_1, \ldots, y_n\} \subset \mathbb{R}^d$ (where $d \ll D$, typically $d = 2$) that approximately maintains the neighborhood structure in $X$. t-SNE accomplishes this by assigning affinities to input data points which encode how likely an input point is to be a neighbor to a given point. The goal then is to find a configuration of the embedded points $Y$ that induces a similar neighborhood affinity. Specifically, for $n > 2$, let $P = P(X) \in \mathbb{R}_+^{n \times n}$ and $Q = Q(Y) \in \mathbb{R}_+^{n \times n}$ be the input and embedded *affinity matrices* describing the pairwise neighborhood similarities in the input and the output, respectively. t-SNE constructs $P$ by first computing the affinities for each point $i$ defined as (for any $j \neq i$)[2]

$$P_{j|i}(X; \sigma_i) := \frac{\exp(-\|x_i - x_j\|^2/(2\sigma_i^2))}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2/(2\sigma_i^2))} \qquad P_{i|i} := 0, \tag{1}$$

where $\sigma_i \geq 0$ encodes the (point-dependent) neighborhood scalings[3]. It is worth noting that $P_{\cdot|i}$ is a valid probability distribution over $[n]$. The matrix $P$ is then constructed based on a crucial parameter called the *perplexity*, which is denoted by $\rho \in (1, n-1)$ and can be viewed as a proxy for effective number of neighbors, as follows.

(1) For each $i \in [n]$, select the (unique, see Lemma 13) neighborhood scale $\sigma_i^* \geq 0$ that minimizes the gap between the entropy of $P_{\cdot|j}(X; \sigma_i^*)$ and $\log_2 \rho$.

(2) Define $P = [P_{ij}]_{i,j \in [n]}$ where $P_{ij} := \frac{1}{2n}(P_{i|j}(\sigma_j^*) + P_{j|i}(\sigma_i^*))$ if $i \neq j$ and zero otherwise.

To avoid the so-called *the crowding problem* (see Van der Maaten & Hinton (2008) for details), the output affinity matrix $Q$ is computed based on a t-distribution. Specifically, for $i \neq j$

$$Q_{ij}(Y) := \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k,l;k \neq l}(1 + \|y_k - y_l\|^2)^{-1}} \qquad Q_{ii} := 0. \tag{2}$$

As indicated before, the objective then is to minimize the gap between the input and the output affinities. This is accomplished by penalizing the relative entropy (KL-divergence) from $P$ to $Q$, where these affinity matrices are viewed as probability distributions.

$$\text{minimize}_Y \ \mathcal{L}_X(Y) := \mathrm{KL}(P(X) \| Q(Y)) = \sum_{\substack{i,j \\ i \neq j}} P_{ij}(X) \log\Big(\frac{P_{ij}(X)}{Q_{ij}(Y)}\Big).$$

[1] Without loss of generality, we shall assume that the input dimension $D = n - 1$.

[2] When $X$ and $\sigma_i^*$ are clear from context, we will often drop it from the notation.

[3] We define $P_{j|i}(X; 0) := \lim_{\sigma_i \to 0} P_{j|i}(X; \sigma_i)$.

This highly non-convex objective is usually optimized by initializing at a good starting point via an *early exaggeration phase*, followed by performing standard gradient descent methods and returning an embedding $Y$ that corresponds to a local minimum of the objective. Our central task is to study the nature of the these (local minimum) embeddings returned by t-SNE and their relation to the space of input datasets.

**Definition 1.** *For an $(n > 2)$-point dataset $X \subset \mathbb{R}^{n-1}$ and perplexity parameter $\rho \in (1, n-1)$, define*

$$\text{t-SNE}_\rho(X) := \{Y \subset \mathbb{R}^d : \nabla_Y \mathcal{L}_X(Y) = 0\}$$

*as the set of outputs $Y \subset \mathbb{R}^d$ that are stationary to the t-SNE objective on a given input $X$.*

*Furthermore, for a set of $n$-point datasets $\mathcal{X}_n$, we define $\text{t-SNE}_\rho(\mathcal{X}_n) = \bigcup_{X \in \mathcal{X}_n} \text{t-SNE}_\rho(X)$. If $\mathcal{X}_n$ is the set of* all *$n$-point datasets, we denote $\text{t-SNE}_\rho(\mathcal{X}_n)$ as $\mathsf{Im}(\text{t-SNE}_{\rho,\text{n}})$.*

All the supporting proofs for our formal statements can be found in the Appendix, and the code related to our empirical demonstrations is available on Github at `https://github.com/njbergam/tsne-exaggerates-clusters`.

## 4 Misrepresentation of Cluster Structure

Previous works by Linderman & Steinerberger (2019) and Arora et al. (2018) have identified that clustered inputs induce clustered t-SNE visualizations in a suitable sense. A key question for practitioners left unanswered by these analyses is: when does a clustered output imply a clustered input? More generally, what information can be deduced about the input given a visualization? We answer this question by providing theoretical and practical evidence that the strength of cluster structure in the input cannot be reliably inferred from the low-dimensional visualization. In particular, we prove that (i) similarly clustered t-SNE visualizations do not imply similarly clustered inputs, and (ii) distinctly clustered visualizations do not imply distinctly different inputs.

To quantify the strength of the cluster structure in a dataset, we employ well-known cluster indices such as the average silhouette score (Rousseeuw, 1987), the Calinski-Harabasz index (Caliński & Harabasz, 1974), and the Dunn index (Dunn, 1974). For sake of readability, we focus on presenting our results with respect to the average silhouette score. Our results hold identically for the other indices as well (see Appendix A).

**Definition 2.** *Given a partition $C_1 \sqcup C_2 \sqcup \cdots \sqcup C_k = [n]$ of $n$ points $\{x_1, \ldots, x_n\} = X$, the **silhouette score** of a point $x_i$ (w.r.t. the partition), denoted $\mathcal{S}(i)$, is the normalized difference between the average within- and the closest across-cluster distances from $x_i$:*

$$\mathcal{S}(i) := \frac{b(i) - a(i)}{\max\{b(i), a(i)\}} \qquad a(i) := \sum_{j \in C^{(i)}} \frac{\|x_i - x_j\|}{|C^{(i)}| - 1} \qquad b(i) := \min_{\substack{m \in [k] \\ C_m \neq C^{(i)}}} \sum_{j \in C_m} \frac{\|x_i - x_j\|}{|C_m|},$$

*where $C^{(i)}$ is the cluster to which $i$ belongs. Note that if $|C^{(i)}| = 1$, then $\mathcal{S}(i)$ is defined to be zero. The **average silhouette score** then is simply the average across all points in $X$:*

$$\bar{\mathcal{S}}(X; C_{m \in [k]}) := \frac{1}{n} \sum_{i \in [n]} \mathcal{S}(i).$$

Observe that the (average) silhouette score ranges from $-1$ to $1$ with a score of $1$ being assigned to maximally clustered data, $-1$ to incorrectly clustered data, and $0$ to unclustered data.

Defining the strength of a clustering with respect to this cluster index, we show that any stationary t-SNE output can be produced by an arbitrarily unclustered input:

**Theorem 3.** *Fix any $n > k > 1$, and $n$-point dataset $X \subset \mathbb{R}^{n-1}$ with partition $C_1 \sqcup \cdots \sqcup C_k = [n]$ such that $|C_{m \in [k]}| > 1$ and $\bar{\mathcal{S}}(X; C_{m \in [k]})$ is well defined. For all $0 < \epsilon \leq 1$, there exists $n$-point dataset $X_\epsilon \subset \mathbb{R}^{n-1}$ such that*

$$\bar{\mathcal{S}}(X_\epsilon; C_{m \in [k]}) = \epsilon \cdot \bar{\mathcal{S}}(X; C_{m \in [k]}),$$

*yet, for any $\rho \in (1, n-1)$:*

$$\text{t-SNE}_\rho(X) = \text{t-SNE}_\rho(X_\epsilon).$$

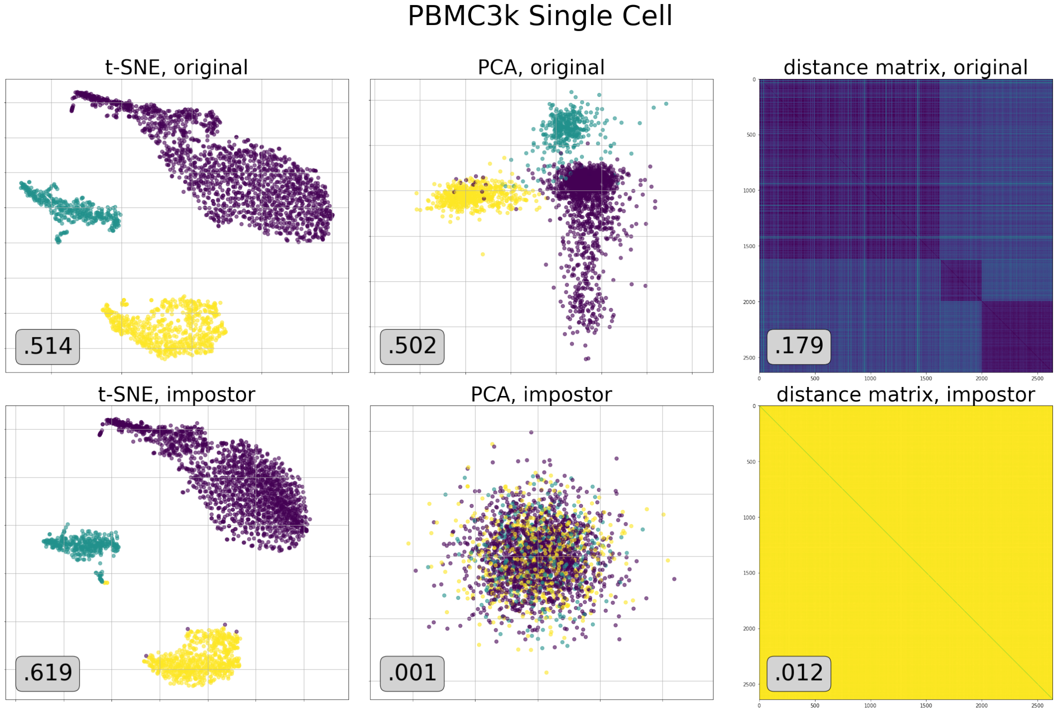


Figure 1: Visualizations of single-cell data (top row) versus an arbitrarily unclustered impostor dataset (bottom row). Based on the 2D t-SNE visualization (left column), it is difficult do distinguish which dataset (real or impostor) may have produced the plot. Plotting the high-dimensional interpoint distances (right column) confirms that the imposter dataset is unclustered in some sense. As a reference we also plot the 2D PCA visualization (center column) to indicate that this issue does not occur with other methods. The numbers on the bottom left of each figure shows the cluster salience in terms of the average silhouette score for the 2D t-SNE plot (left), 2D PCA plot (center), and high-dimensional input (right) for the real dataset (top) and the impostor dataset (bottom). Note that the color coding in all of the scatter plots corresponds to a DBSCAN clustering (Ester et al., 1996) of the top left t-SNE plot.

It is important to understand the implications of this result. For any high-dimensional dataset $X$ (regardless of how clustered it is), we can find an arbitrarily unclustered impostor dataset $X_\epsilon$ such that *all* t-SNE stationary points (local as well as global) of $X$ and $X_\epsilon$ match perfectly! In other words it is *impossible* to distinguish between $X$ and $X_\epsilon$ based on the low-dimensional t-SNE visualization.

As a consequence, the same maximally clustered visualization can be produced by a sequence of impostor datasets ranging from maximally clustered to arbitrarily unclustered,

**Corollary 4.** *For all $n \geq 4$ even, and partition $C_1 \sqcup C_2 = [n]$ such that $|C_1| = |C_2| = \frac{n}{2}$. There exist a sequence of $n$-point datasets in $\mathbb{R}^{n-1}$, $\{X_\epsilon\}_{0<\epsilon\leq 1}$, with*

$$\bar{S}(X_\epsilon; C_1, C_2) = \epsilon$$

*such that for any $\rho \in (1, n-1)$, we have $Y \in \bigcap_{0<\epsilon\leq 1} \text{t-SNE}_\rho(X_\epsilon)$ with*

$$\bar{S}(Y; C_1, C_2) = 1.$$

The above shows that $Y$, a perfectly clustered visualization, is a local (and global, see the proof in Appendix) minimizer for any member of a set of inputs that range from being maximally clustered to being arbitrarily unclustered. Thus, even from a *perfectly* clustered visualization, the strength of the input's cluster structure cannot be inferred.

Note that the existence of an impostor $X_\epsilon$ is not just theoretical; it can be constructed practically as well (see Appendix A.5 for an explicit construction). Hence this phenomenon can be demonstrated in real-world scenarios, see Figure 1. In this case, we select a preprocessed version of the well-known PBMC3k single-cell genomics dataset (2638 points, 50 dimensions; 10x Genomics (2019))

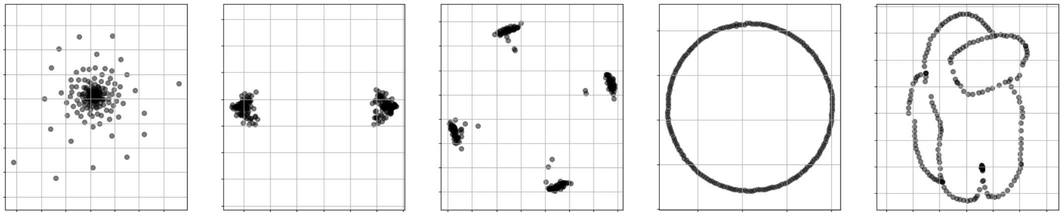


Figure 2: Myriad 2D t-SNE visualizations, all produced by small perturbations of the same 200-point input dataset. Each perturbation satisfies the conditions of Theorem 5 for $\epsilon = 0.01$.

as $X$. We show that there is an arbitrary unclustered impostor dataset $X_\epsilon$ that is essentially indistinguishable from the real dataset in terms of its 2D t-SNE visualization. In short, similarity in t-SNE visualization does not necessarily imply similarity in the input space.

Symmetrically, similarity in the input space does not guarantee similarity in the t-SNE visualizations. In fact, any two drastically different visualizations can be produced by arbitrarily close inputs:

**Theorem 5.** *Fix any $n > 2$ and $\rho \in (1, n-1)$. For all $\epsilon > 0$ and all $Y, Y' \in \mathsf{Im}(\text{t-SNE}_{\rho,\text{n}})$, there exists $n$-point datasets $X = \{x_1, \dots, x_n\}$ and $X' = \{x'_1, \dots, x'_n\} \subset \mathbb{R}^{n-1}$ such that $\forall i \neq j$*

$$1 - \epsilon \leq \frac{\|x_i - x_j\|^2}{\|x'_i - x'_j\|^2} \leq 1 + \epsilon,$$

*yet $Y \in \text{t-SNE}_\rho(X)$ and $Y' \in \text{t-SNE}_\rho(X')$.*

Thus even minor perturbations of the input dataset can develop into massive changes in the visualization. Figure 2 demonstrates this phenomenon quite clearly. We start with a dataset $X$ that is a regular unit simplex (all pairwise distances are unit length). By systematically perturbing the input $X$ ever so slightly ($\epsilon \leq 0.01$), t-SNE produces strikingly different outputs.

The key observation behind our main Theorems 3 and 5 is the simple yet counter-intuitive fact[4] that t-SNE is not only invariant under multiplicative scaling of the input squared distances, but also additive scaling of these distances. Specifically given a dataset $X = \{x_1, \dots, x_n\}$, for any dataset $X' = \{x'_1, \dots, x'_n\}$ and $C \in \mathbb{R}$ such that, $\|x'_i - x'_j\|^2 = \|x_i - x_j\|^2 + C \geq 0$ for $i \neq j$, we have $\text{t-SNE}_\rho(X) = \text{t-SNE}_\rho(X')$ (see Lemma 16 for a formal statement). As a consequence, for any input dataset, we can simply pump up the interpoint distances and construct an impostor dataset which has the same visualization profile but is arbitrarily close to a regular simplex (and hence is arbitrarily unclustered)[5]. This observation also leads to the following seemingly bizarre fact.

**Lemma 6.** *Fix any $n > 2$ and $\rho \in (1, n-1)$. For any $\epsilon > 0$, define the set of $\epsilon$-perturbations of a unit simplex as $\Delta_\epsilon := \{X = \{x_1, \dots, x_n\} \subset \mathbb{R}^{n-1} : \forall i \neq j, \|x_i - x_j\|^2 \in [1 - \epsilon, 1 + \epsilon]\}$. Then, for all $\epsilon > 0$*

$$\mathsf{Im}(\text{t-SNE}_{\rho,\text{n}}) = \text{t-SNE}_\rho(\Delta_\epsilon).$$

In other words, there is a set of datasets $\Delta_\epsilon$ arbitrarily close to a regular unit simplex that generates *all* possible stationary t-SNE outputs! The instability of t-SNE on such datasets (c.f. Figure 2) has real-world consequences since many high-dimensional datasets fall into this regime (Beyer et al., 1999; Aggarwal et al., 2001) due to the concentration of measure phenomenon (Ledoux, 2001). In particular, such datasets are susceptible to single-point adversarial attacks. Consider a dataset $X$ sampled from a mixture of two high-dimensional Gaussians. t-SNE, as expected, reveals the two underlying clusters (c.f. Figure 3, first panel). However, we can add just a single "poison" point to $X$ and destroy the clustered visualization (see Figure 3 second panel). This failure mode of t-SNE is also observed on a real high-dimensional datasets (see Figure 5 left vs. center).

The success of the poison point attack can be attributed to additive invariance. Given an input dataset in $\Delta_\epsilon$ from a clustered, high-dimensional distribution, the set of interpoint distances occupy a tight band between $1 - \epsilon$ and $1 + \epsilon$. Since t-SNE is invariant under additive scaling, the dataset appears identically as if all the distances are in the range $[0, 2\epsilon]$. Thus, from t-SNE's perspective,

[4] To the best of our knowledge, no theory or practical work on t-SNE has studied this observation formally.

[5] See Algorithm 1 for a formalization of this process.

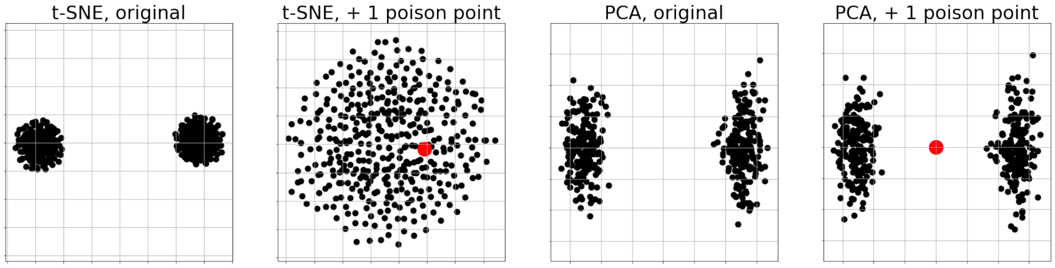


Figure 3: t-SNE versus PCA plots in response to the injection of a single "poison" point in the input dataset. The original dataset, visualized in panels 1 and 3, consists of 400 points sampled from a mixture of two well-separated Gaussians in $\mathbb{R}^{2000}$. The poison point is then placed at the mean of the previously sampled points; the resulting 401-point dataset is visualized in panels 2 and 4.

the variation between inter-cluster distance ($\approx 2\epsilon$) and intra-cluster ($\approx 0$) is large. However, when the poison point is added at the mean, the minimum distance from any point to the rest of the set is approximately halved. As a result, almost all distances remain in the range $[1-\epsilon, 1+\epsilon]$, but, as t-SNE sees it, the effective inter-cluster ($\approx (1+\epsilon) - \frac{1}{2}(1-\epsilon) = \frac{1}{2} + \frac{3}{2}\epsilon$) and intra-cluster ($\approx (1-\epsilon) - \frac{1}{2}(1-\epsilon) = \frac{1}{2} - \frac{1}{2}\epsilon$) gap has been reduced, causing the cluster structure to go unrecognized in some cases.

In the next section, we explore this phenomenon on a real-world dataset (Figure 5), where we contrast it with t-SNE's strikingly indifferent response to the injection of outlier points.

## 5 Misrepresentation of Outliers

Most analysis on t-SNE, including the previous section, is concerned with whether it faithfully depicts global structure, specifically cluster structure. In this section, we consider how t-SNE represents points that drastically deviate from the global structure: namely, *outliers*. It is natural to hope that data visualization methods can enable the identification of outliers. Unfortunately, we find that t-SNE cannot fulfill this desideratum, as it arbitrarily suppresses the severity of outliers in its depiction of certain datasets.

An intuitive explanation of this phenomenon can be made based on the asymmetry of the input and output affinity matrices of t-SNE. Roughly speaking, the input affinity behaves like a normalized, symmetrized nearest neighbor graph, where the log of the perplexity roughly corresponds to the number of neighbors. Meanwhile, the output affinity behaves more like a radius neighborhood graph, at least in the sense that each point's neighborhood scale is the same. This means the output affinity is optimized to represent the outlier point in close proximity with at least some points, even if it was extremely far from those points in the input.

To begin to formalize this observation, we provide a geometric definition of an outlier.

**Definition 7.** *Fix $X \subset \mathbb{R}^D$, $x_0 \in \mathbb{R}^D$, and $\alpha \in \mathbb{R}_+$. We say $X$ is an $(\alpha, x_0)$-**outlier configuration** if there exists a hyperplane separating $x_0$ and $X \setminus \{x_0\}$ with margin width at least*

$$\alpha \cdot \max\{1, \operatorname{diam}(X \setminus \{x_0\})\},$$

*Define the **outlier number** of a dataset, denoted $\alpha(X)$, as the largest $\alpha$ for which there exists $x_0 \in X$ such that $X$ is an $(\alpha, x_0)$-outlier configuration.*

This definition can be generalized to accommodate more than one outlier, but for the purposes of theoretical analysis we consider just one. Note that the outlier extremity $\alpha$ is defined relative to the diameter of the rest of the points, unless that diameter is below 1. The choice of a threshold here is important and intuitive: it allows us to have a suitable notion of outlier in extreme cases such as when $\operatorname{diam}(X \setminus \{x_0\}) = 0$.

Our main theorem establishes that any stationary t-SNE output, *regardless of its input*, is incapable of depicting extreme outliers.

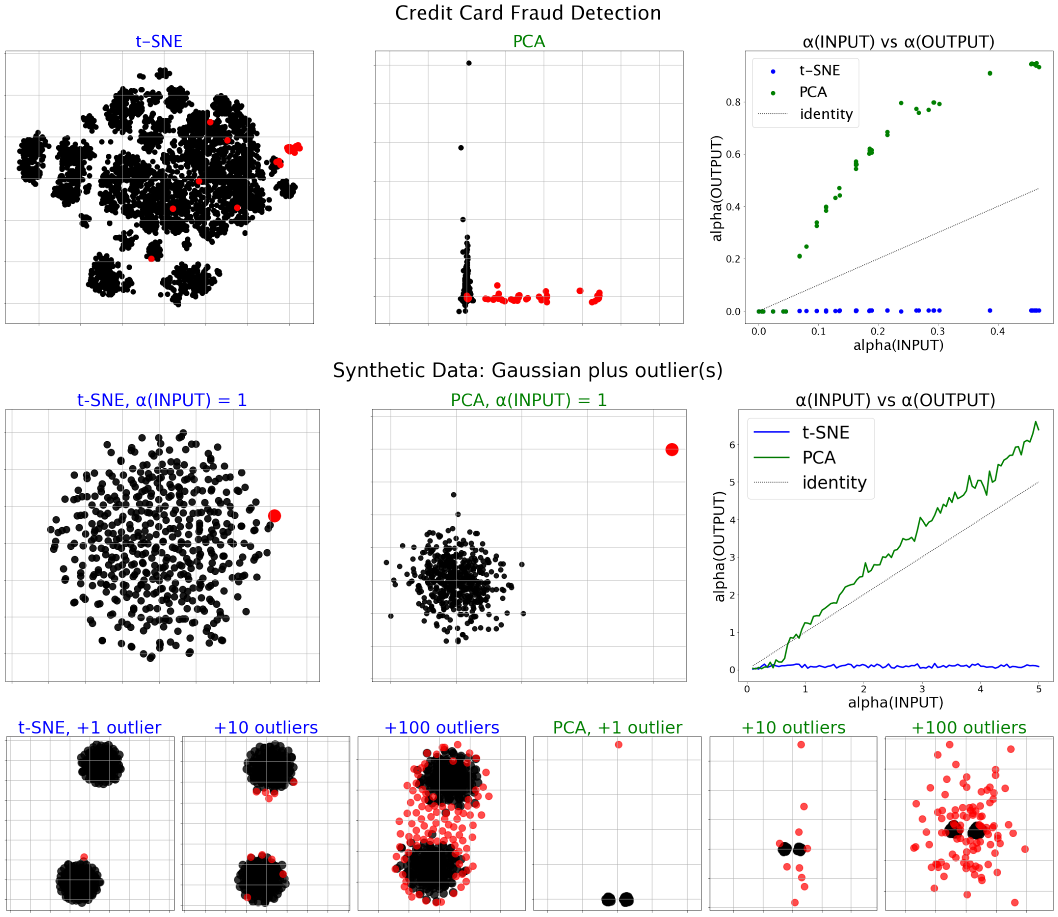


Figure 4: t-SNE's versus PCA's response to $\alpha$-outliers. Top row: on a dataset that tracks financial activity, around $1\%$ of which is fraudulent, t-SNE fails while PCA largely succeeds at separating fraudulent (red) from non-fraudulent (black) points. Note that each of the fraudulent data points is an $(\alpha > 0)$-outlier with respect to the non-fraudulent group; the top right figure shows how t-SNE and PCA register those $\alpha$-values in their output. Middle row: a similar analysis on a synthetic dataset comprised of a Gaussian sample plus a single $\alpha$-outlier, with varying values of $\alpha$. Bottom row: mixture of two Gaussians plus 1, 10, and 100 $\alpha$-outliers. Despite a large gap ($\alpha > 1$) between the outliers and the two clusters, t-SNE is unable to separate them.

**Theorem 8.** *Fix $n > 2$ and $\rho \in (1, n-1)$. Let $Y = \{y_0, y_1, \ldots, y_{n-1}\} \in \mathsf{Im}(\text{t-SNE}_{\rho,\text{n}})$ be a stationary t-SNE embedding. Without loss of generality let $y_0$ be the outlier point. Then we have:*

$$\alpha(Y) = \alpha(Y, y_0) \leq \sqrt{1 + \left(1 + \frac{2}{n-2}\right)\left(\frac{8}{1 + \sum_{i=1}^{n-1} P_{0|i}(X)}\right)} = 3 + o(1)$$

*for all $X = \{x_0, x_1, \ldots, x_{n-1}\}$ such that $Y \in \text{t-SNE}_\rho(X)$.*

The result is proven via analysis of the t-SNE gradient: we argue that if the outlier is too far away, its gradient is nonzero, thus violating stationarity. Key to this analysis is a comparison between the aggregate behavior of the outlier point's affinities in the input versus the output; in other words, the comparison between $\sum_{i=1}^n P_{i0}$ and $\sum_{i=1}^n Q_{i0}$. This is where the fundamental asymmetry of t-SNE comes in. While the latter is dependent on the position of the outlier point $y_0$, per Lemma 19, the former has a lower bound of $1/(2n)$ due to the normalization of the conditional affinity probabilities.

The input-agnostic nature of this result is striking: even if the input is an extreme outlier configuration, a t-SNE output cannot depict its extremity past roughly $\alpha = 3$. This behavior stands in stark contrast to that of principal component analysis (PCA), as shown in Figure 4 on both real and synthetic data models. PCA tends to preserve the $\alpha$ outlier number, while t-SNE seldom depicts outliers past $\alpha > 0.2$ in practice, and sometimes even depicts them as within the convex hull of the

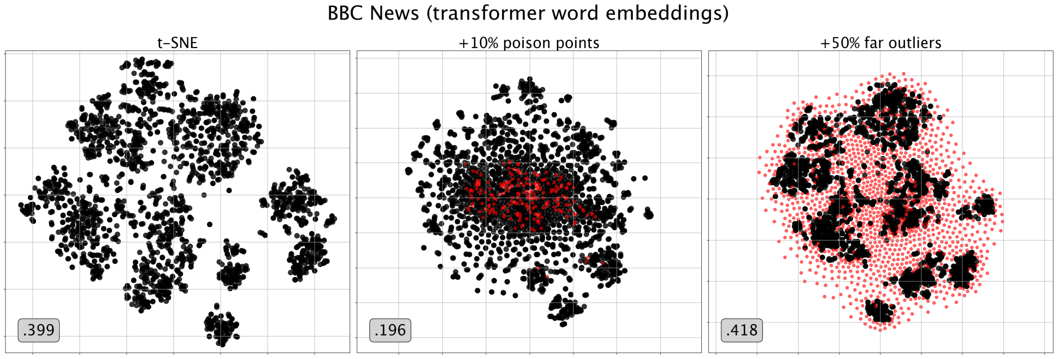


Figure 5: t-SNE's response to the injection of poison points (middle) and $\alpha$-outliers (right) on the BBC News Article dataset. Middle: injecting poison points (red) to the original dataset (black) significantly disrupts the underlying cluster structure. Right: while injecting ($\alpha > 1$)-outliers (red) does not disrupt the underlying cluster structure (black), the extreme outliers themselves are not well separated. The bottom left label in each plot denotes silhouette score of the t-SNE projected original points (without the injected points) with respect to the true labels (business, entertainment, politics, sport, tech).

rest of the points (hence $\alpha = 0$). Furthermore, when faced with multiple outliers, (Figure 4, bottom) t-SNE gracefully accommodates them into the global structure of the bulk of the data.

Our result suggests that t-SNE is an inappropriate tool to use in situations involving outlier detection. Consider, for instance, a dataset of financial transactions where the goal is to detect fraudulent user, studied by Pozzolo et al. (2015). In this dataset, only $0.172\%$ percent of the points ($492$ out of $284,807$) are fraudulent and by many standard statistical metrics register as outliers. Comparing the t-SNE and PCA plots on a random representative subset of this data ($5050$ points, of which $50$ are fraudulent), we see that t-SNE mixes the frauds with the bulk of the points while PCA keeps them separated for the most part, see Figure 4, top row.

Finally, note the distinction between t-SNE's muted response to outliers and its dramatic sensitivity to poison points. We illustrate this distinction on a dataset of BBC news articles (Greene & Cunningham, 2006), see Figure 5. Given RoBERTa (Liu et al., 2019) sentence embeddings of these articles ($n = 2225, D = 1024$), we find that injecting $220$ poison points (see Appendix B.1 for the explicit construction) can halve the silhouette score of the t-SNE embedding with respect to the ground-truth labelling, whereas injecting $1100$ large-$\alpha$-outliers slightly improves the silhouette score.

## 6 Discussion

Our study of t-SNE has established in considerable generality that one cannot infer the *degree* of cluster structure or the *extremity* of outliers from a t-SNE plot, see Theorems 3, 5, and 8. The proofs and intuitions behind these statements guided us to the surprising empirical observation that one cannot even infer the *existence* of clusters or outliers. In particular, the injection of a small subset of adversarially chosen points can largely mask the cluster structure, while sizable injections of outlier points are masked within the cluster structure, see Figures 3, 4, 5, and 7.

We have identified two properties of t-SNE that give rise to these idiosyncratic behaviors: (1) additive invariance with respect to the squared interpoint distances, and (2) the asymmetry between the input and output affinity matrices. While we have uncovered significant false positive failure modes that arise from these properties, we cannot completely rule out their utility. Additive invariance, while brittle under certain adversarial perturbations, may be robust to certain random perturbations. Indeed, adding random noise to a dataset is approximately equivalent to adding a constant to the interpoint distances due to concentration of measure. Additive invariance effectively allows t-SNE to ignore such noise, see Figure 6. This phenomenon is worthy of further study.

t-SNE belongs to a wide selection of data visualization techniques that are yet to be understood fully (McInnes et al., 2018; Jacomy et al., 2014; Tang et al., 2016; Amid & Warmuth, 2019). Our hope is that this work inspires the reader to explore this fascinating landscape further and pursue the essential question: what can be provably deduced from a visualization?

ACKNOWLEDGMENTS

NB was supported by the Irving Institute for Cancer Research and ONR N00014-24-1-2700.

## REFERENCES

10x Genomics. PBMCs from C57BL/6 mice (v1, 150x150), Single Cell Immune Profiling Dataset by Cell Ranger v3.1.0. `https://www.10xgenomics.com/`, 2019.

Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional space. *International Conference on Database Theory (ICDT)*, pp. 420–434, 2001.

Ehsan Amid and Manfred K Warmuth. TriMap: Large-scale dimensionality reduction using triplets. *Computing Research Repository (CoRR)*, abs/1910.00204, 2019.

Sanjeev Arora, Wei Hu, and Pravesh K Kothari. An analysis of the t-SNE algorithm for data visualization. *Conference on Learning Theory (COLT)*, pp. 1455–1462, 2018.

Antonio Auffinger and Daniel Fletcher. Equilibrium distributions for t-distributed stochastic neighbour embedding. *Computing Research Repository (CoRR)*, abs/2304.03727, 2023.

Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? *International Conference on Database Theory (ICDT)*, pp. 217–235, 1999.

Kerstin Bunte, Sven Haase, Michael Biehl, and Thomas Villmann. Stochastic neighbor embedding (SNE) for dimension reduction and visualization using arbitrary divergences. *Neurocomputing*, 90:23–45, 2012.

T Tony Cai and Rong Ma. Theoretical foundations of t-sne for visualizing high-dimensional clustered data. *Journal of Machine Learning Research (JMLR)*, 23(1):13581–13634, 2022.

T. Caliński and J Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3 (1):1–27, 1974.

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

J. C Dunn. Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4(1): 95–104, 1974.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. *Knowledge Data Discovery (KDD)*, 96 (34):226–231, 1996.

D. Greene and P. Cunningham. Practical solutions to the problem of diagonal dominance in kernel document clustering. *International Conference in Machine Learning (ICML)*, 2006.

Daniel Jiwoong Im, Nakul Verma, and Kristin Branson. Stochastic neighbor embedding under $f$-divergences. *Computing Research Repository (CoRR)*, abs/1811.01247, 2018.

Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS One*, 9(6):e98679, 2014.

Seonghyeon Jeong and Hau-Tieng Wu. Convergence analysis of t-SNE as a gradient flow for point cloud on a manifold. *Computing Research Repository (CoRR)*, abs/2401.17675, 2024.

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

Michel Ledoux. *The concentration of measure phenomenon*. Number 89. American Mathematical Society, 2001.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Computing Research Repository (CoRR)*, abs/1907.11692, 2019.

Vivien Marx. Seeing data as t-SNE and UMAP do. *Nature Methods*, 21(6):930–933, 2024.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *Computer Research Repository (CoRR)*, abs/1802.03426, 2018.

Alina Petukhova, Joao P Matos-Carvalho, and Nuno Fachada. Text clustering with large language model embeddings. *International Journal of Cognitive Computing in Engineering*, 6:100–108, 2025.

Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson, and Gianluca Bontempi. Calibrating probability with undersampling for unbalanced classification. *IEEE Symposium Series on Computational Intelligence*, pp. 159–166, 2015.

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

Isaac J Schoenberg. Metric spaces and completely monotone functions. *Annals of Mathematics*, 39 (4):811–841, 1938.

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

Szymon Snoeck, Noah Bergam, and Nakul Verma. Compressibility barriers to neighborhood-preserving data visualizations. *Computing Research Repository (CoRR)*, abs/2508.07119, 2025.

Jian Tang, Jingzhou Liu, Ming Zhang, and Qiaozhu Mei. Visualizing large-scale and high-dimensional data. *International Conference on World Wide Web (WWW)*, pp. 287–297, 2016.

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

Jarkko Venna, Jaakko Peltonen, Kristian Nybo, Helena Aidos, and Samuel Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research (JMLR)*, 11(2), 2010.

Ben Weinkove. Stochastic neighborhood embedding and the gradient flow of relative entropy. *Computing Research Repository (CoRR)*, abs/2409.16963, 2024.

Zhirong Yang, Yuwei Chen, and Jukka Corander. t-SNE is not optimized to reveal clusters in data. *Computing Research Repository (CoRR)*, abs/2110.02573, 2021.

# A Appendix: Misrepresentation of Cluster Structure

## A.1 Additional Experiments

In Figure 6, we plot a sample from a mixture of two Gaussians in 250, 500, 1000, 2000, and 4000 dimensions. Notice that as the dimension of the Gaussian increases, the interpoint distance matrix of the input points (bottom) approaches a simplex but the t-SNE corresponding visualization (top) remains qualitatively unchanged.

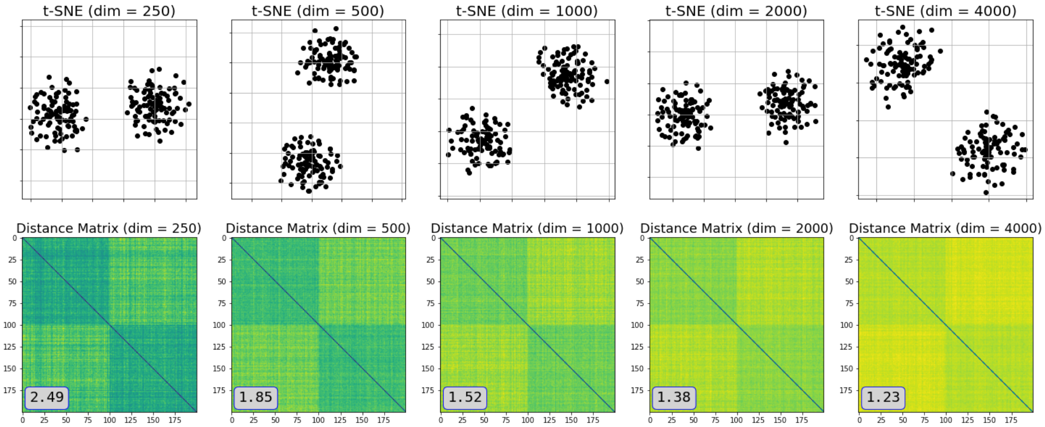


Figure 6: t-SNE's interplay with Gaussian concentration of measure.

## A.2 Calinski-Harabasz Index

For an $n$-point dataset $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^{n-1}$ and a partition of the dataset into clusters $C_1 \sqcup C_2 \sqcup \cdots \sqcup C_k = [n]$ with $n > k > 1$, the Calinski-Harabasz Index is defined as the ratio of the distance between cluster centers to the internal distance to a cluster's center. Let $E$ be the function sending $S \subseteq [n]$ to $\mathbb{R}^{n-1}$ such that:

$$E(S) = \frac{1}{|S|} \sum_{i \in S} x_i.$$

Then the Calinski-Harabasz Index is defined as[6]:

$$\mathrm{CH}(X; C_{m\in[k]}) = \frac{\frac{1}{k-1} \sum_{m\in[k]} |C_m| \cdot \|E(C_m) - E([n])\|^2}{\frac{1}{n-k} \sum_{m\in[k]} \sum_{i \in C_m} \|x_i - E(C_m)\|^2}.$$

It ranges from 0 to $\infty$ with a score of $\infty$ being assigned to perfectly clustered data, 1 to unclustered data and 0 to incorrectly clustered data.

Now we provide an analogue to Theorem 3 with respect to the Calinski-Harabasz Index:

**Theorem 9.** *Fix any $n > k > 1$, and $n$-point dataset $X \subset \mathbb{R}^{n-1}$ with partition $C_1 \sqcup \cdots \sqcup C_k = [n]$ such that $\mathrm{CH}(X; C_{m\in[k]}) > 1$. For all $1 < \epsilon \leq \mathrm{CH}(X; C_{m\in[k]})$, there exists $n$-point dataset $X_\epsilon \subset \mathbb{R}^{n-1}$ such that*

$$\mathrm{CH}(X_\epsilon; C_{m\in[k]}) = \epsilon,$$

*yet, for any $\rho \in (1, n-1)$:*

$$\text{t-SNE}_\rho(X) = \text{t-SNE}_\rho(X_\epsilon).$$

[6] If the denominator and numerator are 0, then $\mathrm{CH}(X; C_{m\in[k]}) := 1$. If only the denominator is 0, then $\mathrm{CH}(X; C_{m\in[k]}) := \infty$.

**Corollary 10.** *For all $n \geq 4$ even, and partition $C_1 \sqcup C_2 = [n]$ such that $|C_1| = |C_2| = \frac{n}{2}$. There exist a sequence of $n$-point datasets in $\mathbb{R}^{n-1}$, $\{X_\epsilon\}_{1<\epsilon\leq\infty}$, with*

$$\mathrm{CH}(X_\epsilon; C_1, C_2) = \epsilon$$

*such that for any $\rho \in (1, n-1)$, $\bigcap_{1<\epsilon\leq\infty}$ t-SNE$_\rho(X_\epsilon)$ contains $n$-point dataset $Y \subseteq \mathbb{R}^2$ with*

$$\mathrm{CH}(Y; C_1, C_2) = \infty.$$

***Proof of Theorem 9.*** First, let us assume that $\mathrm{CH}(X; C_{m\in[k]}) < \infty$. Let $g$ be the function from Corollary 18, and let $f(C) = \mathrm{CH}(g(C); C_{m\in[k]})$. Note that $f$ is continuous whenever the denominator of $\mathrm{CH}(\,\cdot\,; C_{m\in[k]})$ is non-zero which is always the case for $C \in [0,1]$. Therefore, the image of $f$ on $[0,1)$ contains the interval $(f(1), f(0)] = (1, \mathrm{CH}(X; C_{m\in[k]})]$. Thus, for all $\epsilon \in (1, \mathrm{CH}(X; C_{m\in[k]})]$, there exists $C \in [0,1)$ such that $X_\epsilon = g(C)$ satisfies the hypothesis.

If $\mathrm{CH}(X; C_{m\in[k]}) = \infty$, then $f$ is continuous on $(0,1)$ only. Thus for all $\epsilon \in (1, \mathrm{CH}(X; C_{m\in[k]}))$, there exists $C \in (0,1)$ such that $X_\epsilon = g(C)$ satisfies the hypothesis and for $\epsilon = \mathrm{CH}(X; C_{m\in[k]}))$, $X_\epsilon = X$ satisfies the hypothesis. □

For proof of Corollary 10 see Appendix A.4.

### A.3 Dunn Index

For an $n$-point dataset $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^{n-1}$ and a partition of the dataset into clusters $C_1 \sqcup C_2 \sqcup \cdots \sqcup C_k = [n]$ with $|C_{m\in[k]}| > 1$, the Dunn index measures the ratio between the minimum inter-cluster distance and maximum intra-cluster distance. Specifically, the Dunn index is given by the expression[7]

$$\mathrm{DI}(X; C_{m\in[k]}) = \frac{\min_{m,l\in[k], m\neq l, i\in C_m, j\in C_l} \|x_i - x_j\|}{\max_{m\in[k], i,j\in C_m} \|x_i - x_j\|}.$$

It ranges from 0 to $\infty$ with a score of 0 being assigned to incorrectly clustered data, 1 to unclustered data, and $\infty$ to perfectly clustered data.

Now we provide an analogue to Theorem 3 with respect to the Dunn Index:

**Theorem 11.** *Fix any $n > k > 1$, and $n$-point dataset $X \subset \mathbb{R}^{n-1}$ with partition $C_1 \sqcup \cdots \sqcup C_k = [n]$ such that $|C_{m\in[k]}| > 1$ and $\mathrm{DI}(X; C_{m\in[k]}) > 1$. For all $1 < \epsilon \leq \mathrm{DI}(X_\epsilon; C_{m\in[k]})$, there exists $n$-point dataset $X_\epsilon \subset \mathbb{R}^{n-1}$ such that*

$$\mathrm{DI}(X_\epsilon; C_{m\in[k]}) = \epsilon,$$

*yet, for any $\rho \in (1, n-1)$:*

$$\text{t-SNE}_\rho(X) = \text{t-SNE}_\rho(X_\epsilon).$$

**Corollary 12.** *For all $n \geq 4$ even, and partition $C_1 \sqcup C_2 = [n]$ such that $|C_1| = |C_2| = \frac{n}{2}$. There exist a sequence of $n$-point datasets in $\mathbb{R}^{n-1}$, $\{X_\epsilon\}_{1<\epsilon\leq\infty}$, with*

$$\mathrm{DI}(X_\epsilon; C_1, C_2) = \epsilon$$

*such that for any $\rho \in (1, n-1)$, $\bigcap_{1<\epsilon\leq\infty}$ t-SNE$_\rho(X_\epsilon)$ contains $n$-point dataset $Y \subseteq \mathbb{R}^2$ with*

$$\mathrm{DI}(Y; C_1, C_2) = \infty.$$

***Proof of Theorem 11.*** First, let us assume that $\mathrm{DI}(X; C_{m\in[k]}) < \infty$. Let $g$ be the function from Corollary 18, and $f(C) = \mathrm{DI}(g(C); C_{m\in[k]})$. Fix $i, j \in [n]$ such that:

$$\min_{m,l\in[k], m\neq l, i'\in C_m, j'\in C_l} \|x_{i'} - x_{j'}\| = \|x_i - x_j\|,$$

[7] If the denominator and numerator are 0, then $\mathrm{DI}(X; C_{m\in[k]}) := 1$. If only the denominator is 0, then $\mathrm{DI}(X; C_{m\in[k]}) := \infty$.

and $t, r \in [n]$ such that:

$$\max_{m\in[k],i',j'\in C_m} \|x_{i'} - x_{j'}\| = \|x_r - x_t\|,$$

Then:

$$f(C) = \frac{\sqrt{(1-C)\cdot\|x_i - x_j\| + C}}{\sqrt{(1-C)\cdot\|x_r - x_t\| + C}},$$

since $g$ preserves the ordering of the distances. Thus, $f$ is continuous on $[0, 1)$ and the image of $f$ on $[0, 1)$ is $(f(0), f(1)] = (1, \mathrm{DI}(X; C_{m\in[k]})]$. Therefore, for all $\epsilon \in (1, \mathrm{DI}(X; C_{m\in[k]})]$, there exists $C \in [0, 1)$ such that $X_\epsilon = g(C)$ satisfies the hypothesis.

If $\mathrm{DI}(X; C_{m\in[k]}) = \infty$, then $f$ is continuous on $(0, 1)$ only. Thus for all $\epsilon \in (1, \mathrm{DI}(X; C_{m\in[k]}))$, there exist $C \in (0, 1)$ such that $X_\epsilon = g(C)$ satisfies the hypothesis and for $\epsilon = \mathrm{DI}(X; C_{m\in[k]})$, $X_\epsilon = X$ satisfies the hypothesis. □

For proof of Corollary 12 see Appendix A.4.

### A.4 Proofs

The main effort of this section will be to prove Lemma 6 which gives us Theorems 3 and 5. We first introduce a number of technical lemmas that collectively show that t-SNE is invariant under additive and multiplicative scaling of the input.

**Lemma 13.** *Let $H(\cdot)$ denote the entropy function. For any $n > 2$, $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^{n-1}$ and $\rho \in (1, n-1)$, there is a unique $\sigma_i \geq 0$ that minimizes*

$$\Big|H(P_{\cdot|i}(X; \sigma_i)) - \log_2 \rho\Big|.$$

*Proof.* This follows easily from the fact that $H(P_{\cdot|i}(X; \sigma))$ is a continuous, strictly increasing function of $\sigma$ (see e.g. Lemma 4.2 of Jeong & Wu (2024)), where $\lim_{\sigma\to\infty} H(P_{\cdot|i}(X; \sigma)) = \log_2(n-1)$ and $H(P_{\cdot|i}(X; 0)) \in (0, \log_2(n-1))$. □

**Definition 14.** *For any $n \geq 1$, dataset $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^{n-1}$, and $C \geq 0$, define $X_{+C} = \{x'_1, \ldots, x'_n\} \subset \mathbb{R}^{n-1}$ such that for all $i \neq j$*

$$\|x'_i - x'_j\|^2 = \|x_i - x_j\|^2 + C.$$

**Lemma 15.** *Fix any $n \geq 1$. For all $n$-point datasets $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^{n-1}$ and $C \geq 0$, there exists $X_{+C} = \{x'_1, \ldots, x'_n\} \subset \mathbb{R}^{n-1}$ such that for all $i \neq j$, $\|x'_i - x'_j\|^2 = \|x_i - x_j\|^2 + C$.*

*Proof.* Let $D$ be the squared inter-point distance matrix of $X$. Thus, the inter-point squared distance matrix of $X_{+C}$ is $D_{+C} = D + C \cdot (11^T - I_n)$. By a famous theorem by Schoenberg (1938), $X_{+C}$ is isometrically embeddable in $\mathbb{R}^{n-1}$ with respect to $\ell_2$ metric if and only if $\forall u \in \mathbb{R}^n$ with $u^T\vec{1} = 0$, $u^T D_{+C} u \leq 0$ holds. Indeed,

$$u^T D_{+C} u = u^T D u + C \cdot (u^T\vec{1})(\vec{1}^T u) - C \cdot u^T u = u^T D u - C \cdot \|u\|^2 \leq 0,$$

where the final inequality uses the fact that $D$ is embeddable. □

**Lemma 16.** *Fix any $n > 2$. For all $n$-point datasets $X \subset \mathbb{R}^{n-1}$, $\rho \in (1, n-1)$, and $C \geq 0$:*

$$\text{t-SNE}_\rho(X) = \text{t-SNE}_\rho(X_{+C}).$$

*Proof.* It is sufficient to show that the input affinity matrices for $X$ and $X_{+C}$ are identical. Indeed, for all $i, j \in [n], i \neq j$ and for all $\sigma_i > 0$

$$\begin{aligned} P_{j|i}(X; \sigma_i) &= \frac{\exp\left(-\|x_i - x_j\|_2^2/(2\sigma_i^2)\right)}{\sum_{k\neq i}\exp\left(-\|x_i - x_k\|_2^2/(2\sigma_i^2)\right)} \\ &= \frac{\exp\left(-(\|x_i - x_j\|_2^2 + C)/(2\sigma_i^2)\right)}{\sum_{k\neq i}\exp\left(-(\|x_i - x_k\|_2^2 + C)/(2\sigma_i^2)\right)} = P_{j|i}(X_{+C}; \sigma_i). \end{aligned}$$

If $\sigma_i = 0$, then $P_{j|i}(X)$ is purely a function of the ordering of the squared interpoint distances, which is unaffected by adding a constant. □

**Lemma 17.** *Fix any $n > 2$. For all $n$-point datasets $X \subset \mathbb{R}^{n-1}$, $\rho \in (1, n-1)$, and $C > 0$:*

$$\text{t-SNE}_\rho(X) = \text{t-SNE}_\rho(C \cdot X).$$

*Proof.* First note that for any dataset $X$ and its scaling $C \cdot X$, and all $\sigma_i \geq 0$, we have the following:

$$P_{j|i}(C \cdot X; C \cdot \sigma_i) = \frac{\exp(-C^2 \cdot \|x_i - x_j\|^2/(2C^2 \cdot \sigma_i^2))}{\sum_{k=1,k\neq j}^n \exp(-C^2 \cdot \|x_i - x_k\|^2/(2C^2 \cdot \sigma_i^2))} = P_{j|i}(X; \sigma_i).$$

Let $H(\cdot)$ denote the entropy function. By the above, $H(P_{\cdot|i}(X;\sigma_i)) = H(P_{\cdot|i}(C \cdot X; C \cdot \sigma_i))$. Let $\sigma_i^*$ and correspondingly $\gamma_i^*$ be the (unique, per Lemma 13) neighborhood scalings that satisfy the perplexity condition for $X$ and $C \cdot X$ respectively (see Section 3). Then $\gamma_i^* = C \cdot \sigma_i^*$.

Therefore $P_{\cdot|i}(C \cdot X; \gamma_i^*) = P_{\cdot|i}(C \cdot X; C \cdot \sigma_i^*) = P_{\cdot|i}(X; \sigma_i^*)$, yielding the result. □

Using the additive and multiplicative invariance of t-SNE, we now prove Lemma 6:

***Proof of Lemma 6.*** Fix any $\epsilon > 0$. It suffices to show that $\mathsf{Im}(\text{t-SNE}_{\rho,\text{n}}) \subseteq \text{t-SNE}_\rho(\Delta_\epsilon)$. Fix any $Y \in \mathsf{Im}(\text{t-SNE}_{\rho,\text{n}})$, there exists a $n$-point dataset $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^{n-1}$ such that:

$$Y \in \text{t-SNE}_\rho(X).$$

Using additive and multiplicative invariance, we will manipulate $X$ such that it is in $\Delta_\epsilon$ which by Lemma 16 and Lemma 17 will not change the output. Let $D_{\max} = \max_{i\neq j}\|x_i - x_j\|^2$ and $D_{\min} = \min_{i,j\in[n],i\neq j}\|x_i - x_j\|^2$. WLOG, assume that $D_{\max} \neq 0$ otherwise $X_{+1} \in \Delta_\epsilon$. Set $A = \frac{1}{2\epsilon}\big|(1-\epsilon)D_{\max} - (1+\epsilon)D_{\min}\big|$ and $B = \frac{1+\epsilon}{D_{\max}+A}$. Note that since $A \geq 0$ and $D_{\max} > 0$, $B$ is well defined and strictly greater than 0. Then the dataset $B \cdot (X_{+A}) = \{x_1', \ldots, x_n'\}$ exists by Lemma 15 and is such that:

$$\text{t-SNE}_\rho(X) = \text{t-SNE}_\rho(B \cdot (X_{+A}))$$

by Lemma 16 and Lemma 17. Moreover, for all $i \neq j$:

$$\|x_i' - x_j'\|^2 = \frac{1+\epsilon}{D_{\max}+A} \cdot (\|x_i - x_j\|^2 + A) \leq 1 + \epsilon,$$

and

$$\begin{aligned}\|x_i' - x_j'\|^2 &\geq (1+\epsilon)\frac{D_{\min}+A}{D_{\max}+A} \\ &\geq (1+\epsilon)\frac{D_{\min} + \frac{1}{2\epsilon}\big((1-\epsilon)D_{\max} - (1+\epsilon)D_{\min}\big)}{D_{\max} + \frac{1}{2\epsilon}\big((1-\epsilon)D_{\max} - (1+\epsilon)D_{\min}\big)} \\ &\geq (1+\epsilon)\frac{(1-\epsilon)(D_{\max} - D_{\min})}{(1+\epsilon)(D_{\max} - D_{\min})} = 1 - \epsilon,\end{aligned}$$

where the second inequality follows because $A \geq \frac{1}{2\epsilon}\big((1-\epsilon)D_{\max} - (1+\epsilon)D_{\min}\big) > -D_{\max}$. □

The above lemmas give us the following useful corollary that will allow us to prove Theorem 3, Theorem 9, and Theorem 11.

**Corollary 18.** *Fix any $n > 2$, and $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^{n-1}$. There exists a well-defined, continuous function $g : [0,1] \to \mathbb{R}^{n \times n-1}$ such that:*

$$C \mapsto ((1-C) \cdot X)_{+C},$$

*and for all $\rho \in (1, n-1)$ and $C \in [0,1)$ :*

$$\text{t-SNE}_\rho(X) = \text{t-SNE}_\rho(g(C)).$$

*Proof.* $g$ is well defined by Lemma 15 and WLOG continuous since it continuously transforms the distances in $X$:

$$\forall i, j \in [n], i \neq j, \qquad \|g(C)_i - g(C)_j\| = \sqrt{(1-C) \cdot \|x_i - x_j\|^2 + C}.$$

Moreover, by Lemmas 16 and 17, for all $C \in [0,1)$ :

$$\text{t-SNE}_\rho(X) = \text{t-SNE}_\rho(g(C)).$$

□

Using the above lemmas, Theorem 3, Corollary 4, and Theorem 5 are proven.

**Theorem 3.** *Fix any $n > k > 1$, and $n$-point dataset $X \subset \mathbb{R}^{n-1}$ with partition $C_1 \sqcup \cdots \sqcup C_k = [n]$ such that $|C_{m\in[k]}| > 1$ and $\bar{\mathcal{S}}(X; C_{m\in[k]})$ is well defined. For all $0 < \epsilon \leq 1$, there exists $n$-point dataset $X_\epsilon \subset \mathbb{R}^{n-1}$ such that*

$$\bar{\mathcal{S}}(X_\epsilon; C_{m\in[k]}) = \epsilon \cdot \bar{\mathcal{S}}(X; C_{m\in[k]}),$$

*yet, for any $\rho \in (1, n-1)$:*

$$\text{t-SNE}_\rho(X) = \text{t-SNE}_\rho(X_\epsilon).$$

***Proof of Theorem 3.*** Let $g$ be the function from Corollary 18, and $f(C) = \bar{\mathcal{S}}(g(C); C_{m\in[k]})$. Note that $f$ is continuous for $C \in [0,1]$ since $g$ is continuous, and $\bar{\mathcal{S}}(\,\cdot\,; C_{m\in[k]})$ is continuous whenever for all $i \in [n]$, $a(i), b(i) \neq 0$ which follows from $\bar{\mathcal{S}}(X; C_{m\in[k]})$ being well-defined and the definition of $g$. Therefore, the image of $f$ on $[0,1)$ contains the interval $(f(1), f(0)] = (0, \bar{\mathcal{S}}(X; C_{m\in[k]})]$ (or if $\bar{\mathcal{S}}(X; C_{m\in[k]}) \leq 0$, $[\bar{\mathcal{S}}(X; C_{m\in[k]}), 0)$). Thus, for all $\epsilon \in (0,1]$, there exists $C \in [0,1)$ such that $X_\epsilon = g(C)$ satisfies the hypothesis. □

Now we can prove Corollary 4, Corollary 10, and Corollary 12 simultaneously:

***Proof of Corollaries 4, 10, and 12.*** The proof proceeds by showing a dataset and its output who have an average silhouette score of 1, Calinski-Harabasz index of $\infty$, and Dunn index of $\infty$, and then applies Theorem 3, Theorem 9, and Theorem 12 respectively. WLOG fix partition $C_1 \sqcup C_2 = [n]$ with $C_1 = [1, n/2]$ and $C_2 = [n/2+1, n]$. Consider the $n$-point dataset, $\mathcal{X} = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^{n-1}$, such that for all $i \in C_1$, $x_i = \vec{0}$, and for all $i \in C_2$, $x_i = \vec{e_1}$.

Routine calculations show that the conditional input affinities are:

$$P_{i|j} = \begin{cases} \frac{1}{\frac{n}{2} - 1 + \frac{n}{2}\exp\left(-\frac{1}{2\sigma_j^2}\right)} & i \in C(j), i \neq j \\ \frac{\exp\left(-\frac{1}{2\sigma_j^2}\right)}{\frac{n}{2} - 1 + \frac{n}{2}\exp\left(-\frac{1}{2\sigma_j^2}\right)} & i \notin C(j) \\ 0 & i = j. \end{cases}$$

By symmetry, $\sigma_j = \sigma_i$ for all $i, j \in [n]$. Hence, let $\sigma$ be the neighborhood size for all $j \in [n]$ which is non-zero and well defined for $\rho \in [1, n-1]$. Thus the symmetrized input affinities are:

$$P_{ij} = \begin{cases} \frac{1}{\frac{n^2}{2} - n + \frac{n^2}{2}\exp\left(-\frac{1}{2\sigma^2}\right)} & i \in C(j), i \neq j \\ \frac{\exp\left(-\frac{1}{2\sigma^2}\right)}{\frac{n^2}{2} - n + \frac{n^2}{2}\exp\left(-\frac{1}{2\sigma^2}\right)} & i \in C_1, j \in C_2 \\ 0 & i = j. \end{cases}$$

Any set $\mathcal{Y} = \{y_1, \ldots, y_n\} \subseteq \mathbb{R}$ is a global minimizer if $P_{ij} = Q_{ij}$ for all $i, j \in [n]$. In this case, this is achieved if $y_{i\in C_1} = 0$ and $y_{i\in C_2} = \sqrt{\exp\left(\frac{1}{2\sigma^2}\right) - 1}$. Furthermore, since $\mathcal{Y}$ can be isometrically embedded in $\mathbb{R}^d$ for all $d \geq 1$, this result holds for t-SNE embeddings of all dimensions.

To finish the proof note that for all $i \in [n]$, $a(i) = 0$ when defined with respect to $\mathcal{Y}$ and partition $C_1 \sqcup C_2$. □

**Theorem 5.** *Fix any $n > 2$ and $\rho \in (1, n-1)$. For all $\epsilon > 0$ and all $Y, Y' \in \mathsf{Im}(\text{t-SNE}_{\rho,\text{n}})$, there exists $n$-point datasets $X = \{x_1, \ldots, x_n\}$ and $X' = \{x'_1, \ldots, x'_n\} \subset \mathbb{R}^{n-1}$ such that $\forall i \neq j$*

$$1 - \epsilon \leq \frac{\|x_i - x_j\|^2}{\|x'_i - x'_j\|^2} \leq 1 + \epsilon,$$

*yet $Y \in \text{t-SNE}_\rho(X)$ and $Y' \in \text{t-SNE}_\rho(X')$.*

***Proof of Theorem 5.*** The proof is immediate by application of Lemma 6. □

### A.5 Impostor dataset construction

The construction of an impostor dataset based on an input dataset is done as follows.

**Algorithm 1** Impostor Dataset Creation

```
Require: Dataset $X = \{x_1, \ldots, x_n\}$ with at least two distinct points, and tolerance $\epsilon > 0$
Construct squared interpoint distance matrix of $X$, denote it by $D$
Form $D' \leftarrow \frac{\epsilon}{\max_{i,j} D_{ij}} \cdot D + (11^\top - I_n)$
Run classical multidimensional scaling on $D'$ to obtain its Euclidean embedding
    $X_\epsilon = \{x'_1, \ldots, x'_n\} \subset \mathbb{R}^{n-1}$.
return $X_\epsilon$
```

## B Appendix: Misrepresentation of Outliers

### B.1 Additional Experiments

We provide a comparison of t-SNE and PCA on the BBC news dataset. For ease of presentation, we take a three-cluster, $(n = 1204)$-size subset (business, sports, tech) and we analyze what happens under injection of 120 poison points versus 120 far outliers.

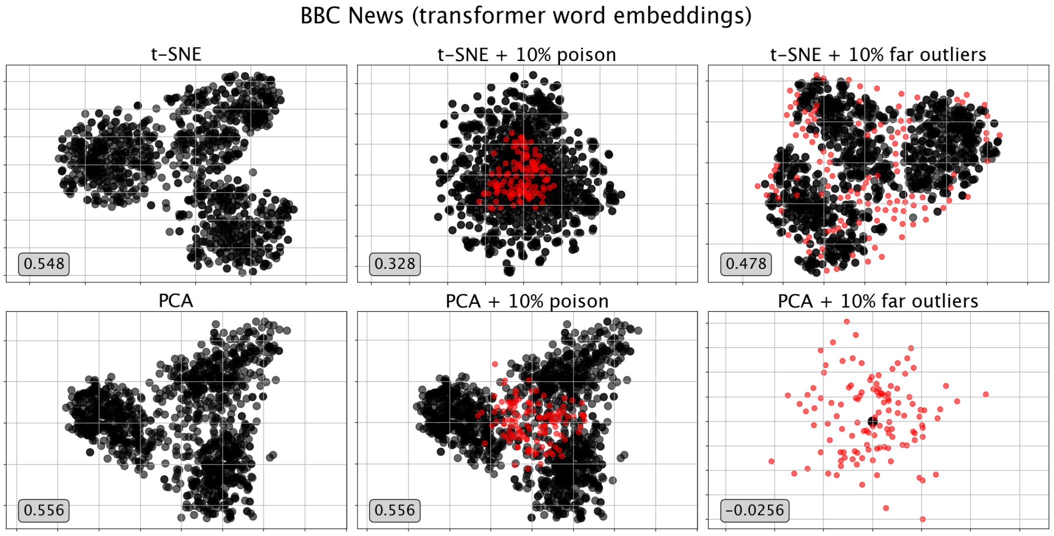


Figure 7: t-SNE vs. PCA on poison points versus outlier points on a three-cluster subset of the BBC news dataset. The label on the bottom left is silhouette score of the plot (sans injected points) with respect to their ground-truth labels.

In both Figure 5 and Figure 7, poison points are picked as follows: first we run a $k$-means algorithm on the original dataset; then, for each poison point, we pick one of the these means and 10 random points of the dataset, and we average these two quantities (the idea is to connect the points in a way that contradicts the ground-truth three-clustering). We found $k = 2$ worked well. We pick outlier points as normal vectors centered at the mean of the dataset with variance 32 (the diameter of the original dataset is roughly 1.5).

### B.2 Proofs

**Lemma 19.** *Fix $n \geq 2$ and $Y = \{y_0, \ldots, y_{n-1}\} \subset \mathbb{R}^d$. Let $\beta := \operatorname{diam}(Y \setminus \{y_0\})$ and $\gamma := \min_{j \in [n]} \|y_0 - y_j\|$. Then*

$$\sum_{i=1}^{n} Q_{0i} \leq \frac{1}{2 + (n-2) \cdot \frac{1+\gamma^2}{1+\beta^2}}.$$

*Proof.* Let $Z_0 = \sum_{i=1}^{n} \frac{1}{1+\|y_i - y_0\|^2}$ and $Z_{1:n} = \sum_{i,j:i \neq j} \frac{1}{1+\|y_i - y_j\|^2}$. Then

$$\sum_{i=1}^{n} Q_{0i} = \frac{Z_0}{2Z_0 + Z_{1:n}} = \frac{1}{2 + Z_{1:n}/Z_0}.$$

Now observe that

$$\begin{aligned}
\frac{Z_{1:n}}{Z_0} &= \frac{\sum_{i,j:i \neq j}(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{j=1}^{n}(1 + \|y_0 - y_j\|^2)^{-1}} \\
&\geq \frac{(n-1)(n-2)(1 + \max_{i,j \in [n]} \|y_i - y_j\|^2)^{-1}}{(n-1)(1 + \min_{j \in [n]} \|y_0 - y_j\|^2)^{-1}} \\
&= \frac{(n-2)(1+\gamma^2)}{1+\beta^2}.
\end{aligned}$$

Plugging this back into the previous equation gives the statement. □

**Lemma 20.** *Fix $n \geq 2$ and $Y = \{y_0, y_1, ..., y_{n-1}\} \subset \mathbb{R}^d$. If $Y$ is a $(\alpha, y_0)$-outlier configuration such that $\alpha = \alpha(Y)$, then there exists $v \in \mathbb{R}^d$ such that for all $i \in [n]$:*

$$\|y_i - y_0\| \cdot \frac{\alpha}{\sqrt{1+\alpha^2}} \leq (y_i - y_0) \cdot v \leq \|y_i - y_0\|.$$

*Proof.* Fix $i \in [n]$, let $\beta := \operatorname{diam}(Y \setminus \{y_0\})$, and WLOG let $y_0 = 0$. Take $v$ as in Definition 7. Then by Cauchy-Schwarz, $(y_i - y_0) \cdot v \leq \|y_i - y_0\|$. To prove the other side of the inequality, we only need to lower bound the cosine of the angle between $y_i - y_0$ and $v$:

$$(y_i - y_0) \cdot v = \|y_i - y_0\| \cdot \cos(\angle(v, y_i)).$$

Since $v$ is the maximum-margin hyperplane between $y_0 = 0$ and $Y \setminus \{y_0\}$, it holds that $u = v \cdot (\alpha \max\{1, \beta\})$ is in the convex hull of $Y \setminus \{y_0\}$. Indeed, $\|u\| = \inf_{y \in \operatorname{conv}(Y \setminus \{y_0\})} \|y\|$. Thus, we know that the closed ball $\overline{B_\beta(u)}$ contains $\operatorname{conv}(Y \setminus \{y_0\})$. Therefore, there exists $t \in \mathbb{R}^d$ such that $\|t\| \leq \beta$, $u + t = y_i$, and $u \cdot t \geq 0$. Hence

$$\cos(\angle(v, y_i)) = \frac{v \cdot y_i}{\|y_i\|} = \frac{v \cdot (u+t)}{\sqrt{\|u\|^2 + \|t\|^2 - 2u \cdot t}} \geq \frac{\alpha \max(\beta, 1)}{\sqrt{\alpha^2 \max(\beta, 1)^2 + \beta^2}} \geq \frac{\alpha}{\sqrt{1+\alpha^2}},$$

completing the proof. □

**Theorem 8.** *Fix $n > 2$ and $\rho \in (1, n-1)$. Let $Y = \{y_0, y_1, \ldots, y_{n-1}\} \in \mathsf{Im}(\text{t-SNE}_{\rho,\text{n}})$ be a stationary t-SNE embedding. Without loss of generality let $y_0$ be the outlier point. Then we have:*

$$\alpha(Y) = \alpha(Y, y_0) \leq \sqrt{1 + \Big(1 + \frac{2}{n-2}\Big)\Big(\frac{8}{1 + \sum_{i=1}^{n-1} P_{0|i}(X)}\Big)} = 3 + o(1)$$

*for all $X = \{x_0, x_1, \ldots, x_{n-1}\}$ such that $Y \in \text{t-SNE}_\rho(X)$.*

*Proof.* Fix $Y \in \mathsf{Im}(\text{t-SNE}_{\rho,\text{n}})$ and define $\gamma = \min_i \|y_i - y_0\|$. WLOG, let $y_0$ be the outlier point and assume $\gamma > 0$ otherwise the hypothesis goes through trivially. Since $Y$ is stationary, $\frac{\partial L}{\partial y_0} = 0$.

Pick $v$ as in Lemma 20 and observe:

$$\begin{aligned}
0 = \frac{\partial \mathcal{L}}{\partial y_0} \cdot v = \sum_{i=1}^{n-1} \frac{(P_{i0} - Q_{i0})(y_0 - y_i) \cdot v}{1 + \|y_0 - y_i\|^2} \\
&\geq \frac{\alpha}{\sqrt{1+\alpha^2}} \sum_{i=1}^{n-1} P_{i0} \frac{\|y_0 - y_i\|}{1 + \|y_0 - y_i\|^2} - \sum_{i=1}^{n-1} Q_{i0} \frac{\|y_0 - y_i\|}{1 + \|y_0 - y_i\|^2} \\
&\geq \frac{\alpha}{\sqrt{1+\alpha^2}} \sum_{i=1}^{n-1} P_{i0} \frac{\|y_0 - y_i\|}{1 + \|y_0 - y_i\|^2} - \sum_{i=1}^{n-1} Q_{i0} \frac{\|y_0 - y_i\|}{1 + \|y_0 - y_i\|^2} \\
&\geq \frac{\alpha}{\sqrt{\alpha^2+1}} \frac{\gamma}{1 + (\gamma+\beta)^2} \sum_{i=1}^{n-1} P_{i0} - \frac{\gamma+\beta}{1+\gamma^2} \sum_{i=1}^{n-1} Q_{i0} \\
&\geq \frac{\alpha}{\sqrt{1+\alpha^2}} \frac{\gamma}{1 + (\gamma+\beta)^2} \frac{1 + \sum_{i=1}^{n-1} P_{0|i}}{2n} - \frac{\gamma+\beta}{1+\gamma^2} \frac{1}{2 + (n-2)\frac{1+\gamma^2}{1+\beta^2}}
\end{aligned}$$

where, in the third line, we use Lemma 19 and the fact that $\sum_{i=1}^{n} P_{i|0} = 1$. Multiplying by $\frac{1+\gamma^2}{\gamma+\beta} \cdot \frac{2n}{1+\sum_{i=1}^{n-1} P_{0|i}} > 0$ and rearranging, we get that:

$$\begin{aligned}
\frac{\alpha}{\sqrt{1+\alpha^2}} \cdot \frac{1+\gamma^2}{\gamma+\beta} \cdot \frac{\gamma}{1+(\gamma+\beta)^2} &\leq \frac{1}{2 + (n-2) \cdot \frac{1+\gamma^2}{1+\beta^2}} \cdot \frac{2n}{1 + \sum_{i=1}^{n-1} P_{0|i}} \\
&\leq \frac{1+\beta^2}{(n-2)(1+\gamma^2)} \cdot \frac{2n}{1 + \sum_{i=1}^{n-1} P_{0|i}} \\
&= \frac{1+\beta^2}{1+\gamma^2} \cdot \left(1 + \frac{2}{n-2}\right) \cdot \frac{2}{1 + \sum_{i=1}^{n-1} P_{0|i}}.
\end{aligned}$$

Recall, by definition of $\alpha$-outlier configuration, that $\gamma \geq \alpha \cdot \max\{\beta, 1\}$. Rearranging, we have:

$$\begin{aligned}
\left(1 + \frac{2}{n-2}\right) \cdot \frac{2}{1 + \sum_{i=1}^{n-1} P_{0|i}} &\geq \frac{\alpha}{\sqrt{1+\alpha^2}} \cdot \frac{\gamma}{\gamma+\beta} \cdot \frac{1+\gamma^2}{1+(\gamma+\beta)^2} \cdot \frac{1+\gamma^2}{1+\beta^2} \\
&\geq \frac{\alpha}{\sqrt{1+\alpha^2}} \frac{\gamma^3}{(\gamma+\beta)^3} \frac{1+\gamma^2}{1+\beta^2} \\
&\geq \frac{\alpha}{\sqrt{1+\alpha^2}} \frac{\alpha^3 \max\{\beta,1\}^3}{(\alpha \max\{\beta,1\} + \beta)^3} \frac{1 + \alpha^2 \max\{\beta,1\}^2}{1+\beta^2} \\
&\geq \frac{\alpha}{\sqrt{1+\alpha^2}} \frac{\alpha^3}{(\alpha + \frac{\beta}{\max\{\beta,1\}})^3} \frac{1+\alpha^2}{2} \\
&\geq \frac{\alpha}{\sqrt{1+\alpha^2}} \frac{\alpha^3}{(1+\alpha)^3} \frac{1+\alpha^2}{2} \\
&= \frac{\alpha^4 \sqrt{1+\alpha^2}}{2(1+\alpha)^3}.
\end{aligned}$$

Assume $\alpha \geq 3$ (or else the hypothesis holds trivially), then the above is lower-bounded by $(\alpha^2-1)/4$. Solving for $\alpha$, we find $\alpha \leq \sqrt{1 + \left(1 + \frac{2}{n-2}\right) \cdot \left(\frac{8}{1+\sum_{i=1}^{n-1} P_{0|i}}\right)}$.

□