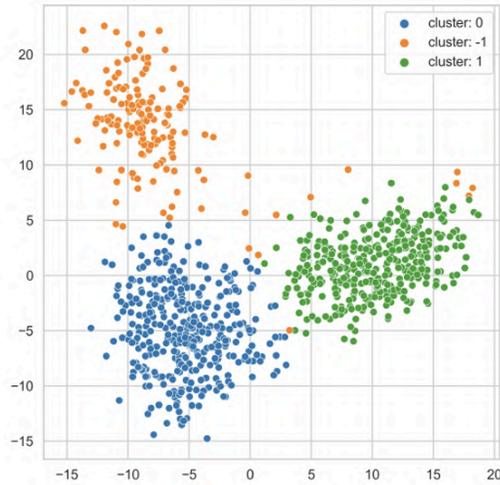
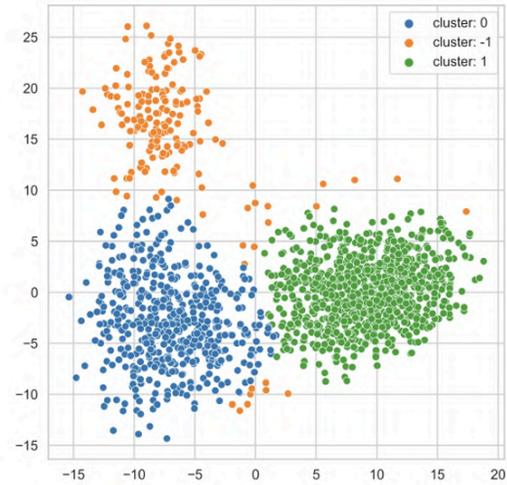


4. Why PCA + t-SNE and not UMAP [2]? UMAP is a dimension reduction algorithm commonly used with clustering - would SS-DBSCAN work with UMAP as well as it does with t-SNE? If not, this would be a critical limitation that potential users would want to know about.

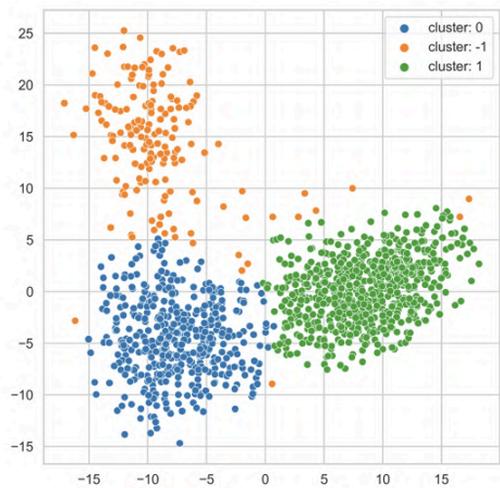
Original Results using PCA+t-SNE



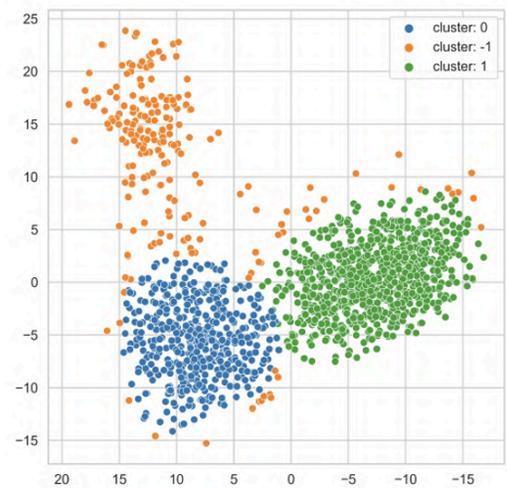
(a) Data size = 1000



(b) Data size = 3000



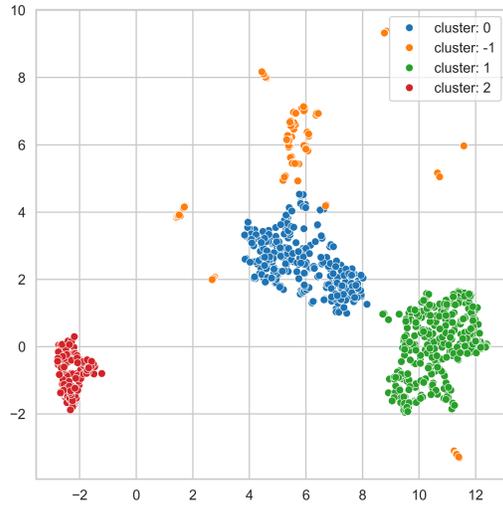
(c) Data size = 4000



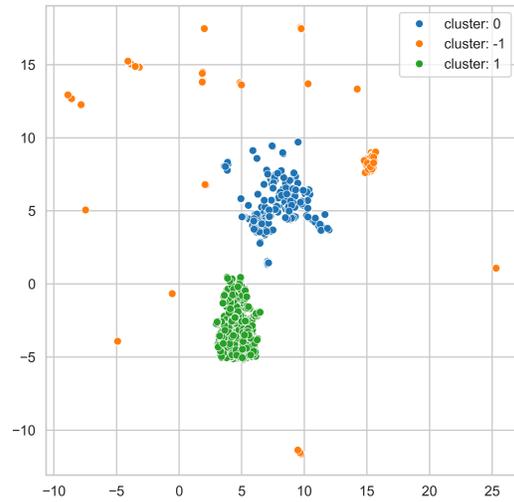
(d) Data size = 5000

Fig. 2. Comparison of SS-DBSCAN results for different data sizes.

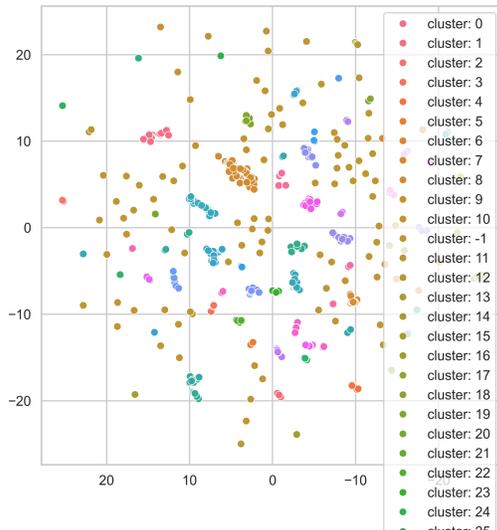
New Results Using UMAP



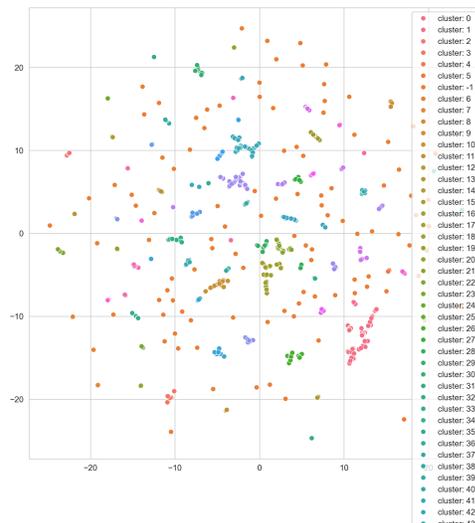
Data size 1000



Data size 2000



Data size 5000



Data size 10000

Observations made

1. From these experiments, I observed that UMAP produced highly compact clusters for smaller dataset sizes. However, as the dataset size increased, UMAP introduced significant noise and formed numerous small clusters, unlike our original approach,

which maintained a consistent number of clusters across all dataset sizes and demonstrated superior noise handling.

2. When combined with SS-DBSCAN, UMAP yielded suboptimal results. Similarly, other clustering algorithms, including DBSCAN, HDBSCAN, and OPTICS, also performed poorly when paired with UMAP.
3. The dataset was expected to contain two clusters. With PCA+t-SNE, SS-DBSCAN successfully identified two clusters across varying dataset sizes, while DBSCAN and HDBSCAN achieved this only for smaller datasets. OPTICS exhibited poor performance throughout. In contrast, UMAP performed well on small datasets but struggled significantly as the size increased.

Why that happens and possible solutions

Issue	Why does it happen?	Solution
UMAP distorts global structure	UMAP emphasizes local manifold learning too much	Increasing n_neighbors may help
UMAP clusters too tightly	min_dist might be too low	Increasing min_dist might help
PCA+t-SNE works better because PCA removes noise	UMAP works directly on raw data	Maybe we should try to run PCA before UMAP
t-SNE's perplexity helps cluster separation	UMAP lacks a direct equivalent	Maybe if we use t-SNE after UMAP instead
UMAP struggles with varying densities	t-SNE adapts dynamically	It might be better to stick with t-SNE if necessary.