Testing Report on Clustering Algorithms

Lieyu Shi

July 1, 2018

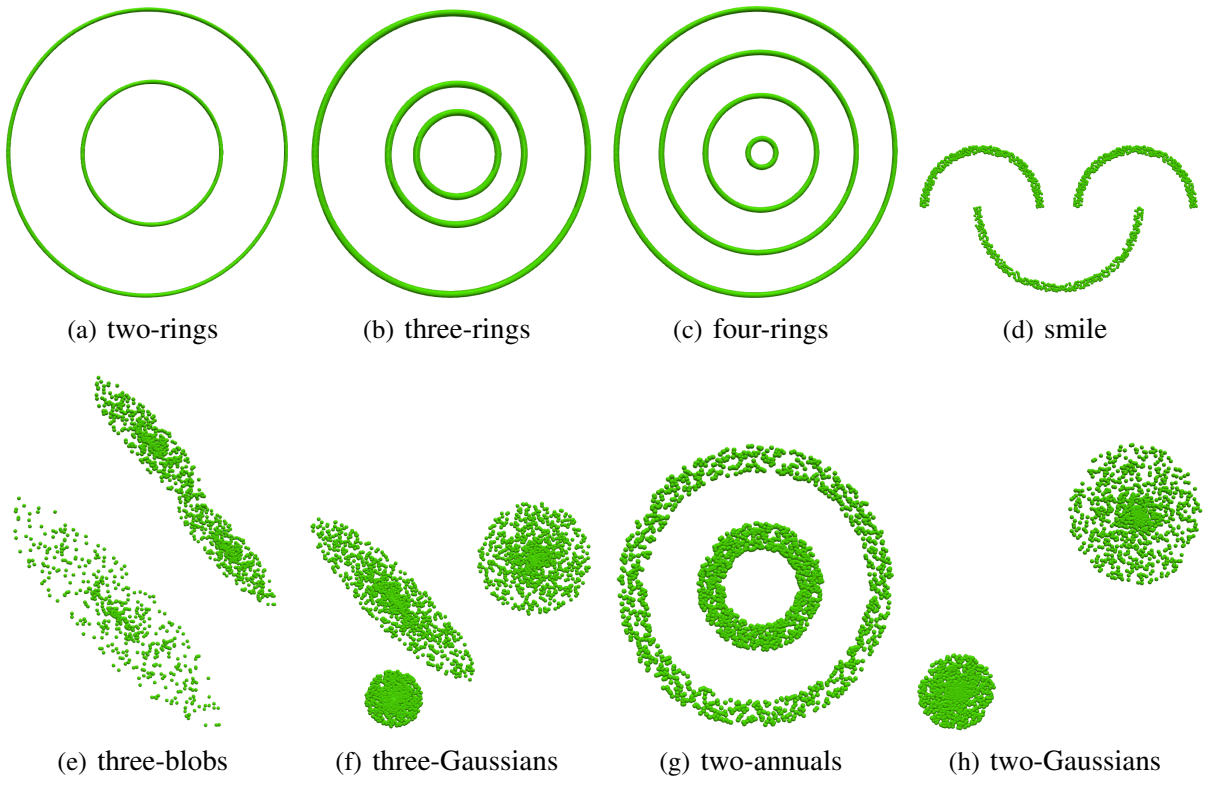
1 Introduction

In this report, we are testing effectiveness and validity of clustering algorithms from [github implementation], which is my C++ implementation for all the prevailing clustering methods in flow visualization. Respectively, it is constituted by, **k-means**, **k-medoids**, **BIRCH**, **agglomerative hierarchical clustering** (AHC), **DBSCAN**, **OPTICS**, **spectral clustering** (SC) and **affinity propagation** (AP).

Besides, some similar clustering algorithm testing results can be seen at [scikit-learn].

2 Testbed Data Sets

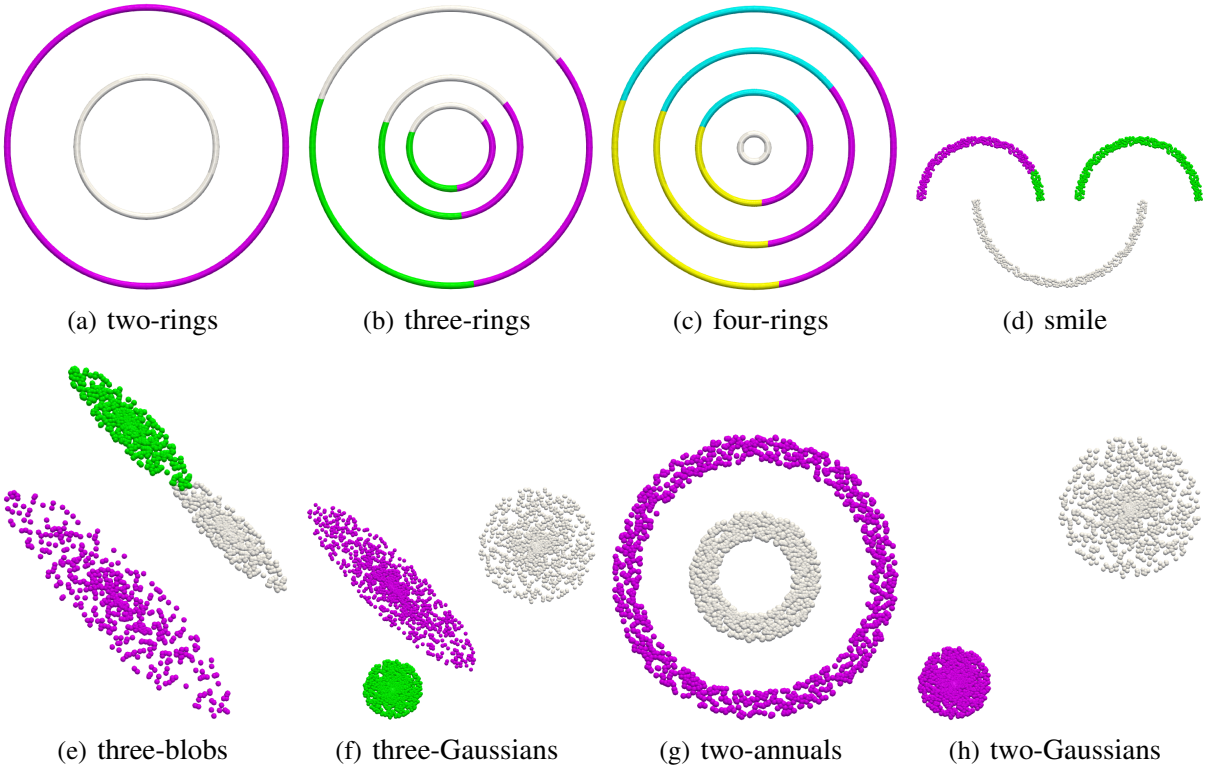
We are using 8 point cloud data sets which are either regular or irregular shapes to testify the effectiveness of the clustering algorithms. The intrinsic cluster numbers are a priori known from natural intuition. The data sets are illustrated in Fig. 1.



(a) two-rings (b) three-rings (c) four-rings (d) smile

(e) three-blobs (f) three-Gaussians (g) two-annuals (h) two-Gaussians

Figure 1: 8 groups of point cloud data sets which has explicitly predefined number of clusters.



(a) two-rings (b) three-rings (c) four-rings (d) smile

(e) three-blobs (f) three-Gaussians (g) two-annuals (h) two-Gaussians

Figure 2: SC with k-means clustering results on point cloud data sets.

Table 1: Clustering Measurements for k-means

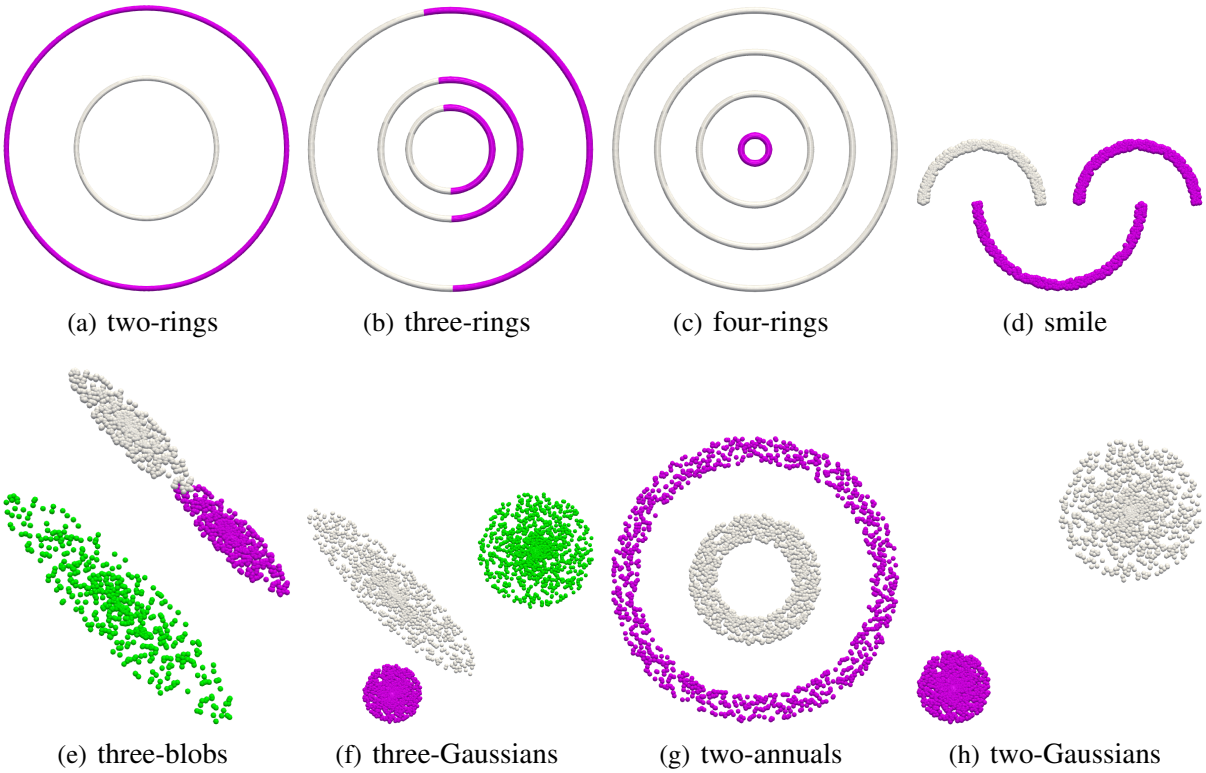
Analysis	Rings2	Rings3	Rings4	ThreeBlobs	TwoGaussians	ThreeGaussians	Smile	twoAnnuals
Silhouette	0.35036	0.32969	0.34062	0.62953	0.84261	0.43559	0.51457	0.31943
DB-Index	1.18475	0.92186	0.99275	0.5473	0.21670	1.04981	0.68930	1.23915
Gamma	0.47702	0.46599	0.51877	0.78544	0.96881	0.69176	0.70273	0.41861

3 Spectral Clustering

Spectral clustering (SC) has two versions available, i.e., one with k-means as post-processing [1] and the other with eigenvector rotation minimization as post-processing [2]. The SC-kmeans suffers stability problem which might be able to detect meaningful shapes due to drawback of k-means, while SC-eigenrotation suffers from being unable to find reasonably optimal number of clusters as indicated in [self-tuning SC github]. K-means++ [3] can be applied to enhance careful seeding for initial samples of k-means of SC spectral clustering, which solves stability and local minimum problem for SC with k-means, but SC with eigenrotation however, the final clustering results is very sensitive to scaling factor setting which is hardly able to be designed.

The SC with k-means is illustrated in Fig. 2 and we can observe that SC with k-means can not work well in Fig. 2(b), 2(c) or 2(d) while working well for all other data sets.

The SC clustering with eigenvector rotation minimization is illustrated in Fig. 3, and we observe that SC with eigenrotation also fails to detect natural clusters for Fig. 3(b), 3(c) and 3(d).



(a) two-rings (b) three-rings (c) four-rings (d) smile

(e) three-blobs (f) three-Gaussians (g) two-annuals (h) two-Gaussians

Figure 3: SC clustering with eigenvector rotation minimization results on point cloud data sets.

4 K-means

Table 2: Clustering Measurements for k-medoids

Analysis	Rings2	Rings3	Rings4	ThreeBlobs	TwoGaussians	ThreeGaussians	Smile	twoAnnuals
Silhouette	0.35041	0.32964	0.33035	0.62990	0.84262	0.6613	0.39252	0.31958
DB-Index	1.18475	0.92184	1.02663	0.54828	0.21670	0.49241	0.94201	1.23882
Gamma	0.47679	0.46511	0.52877	0.78550	0.96881	0.79967	0.60784	0.41863

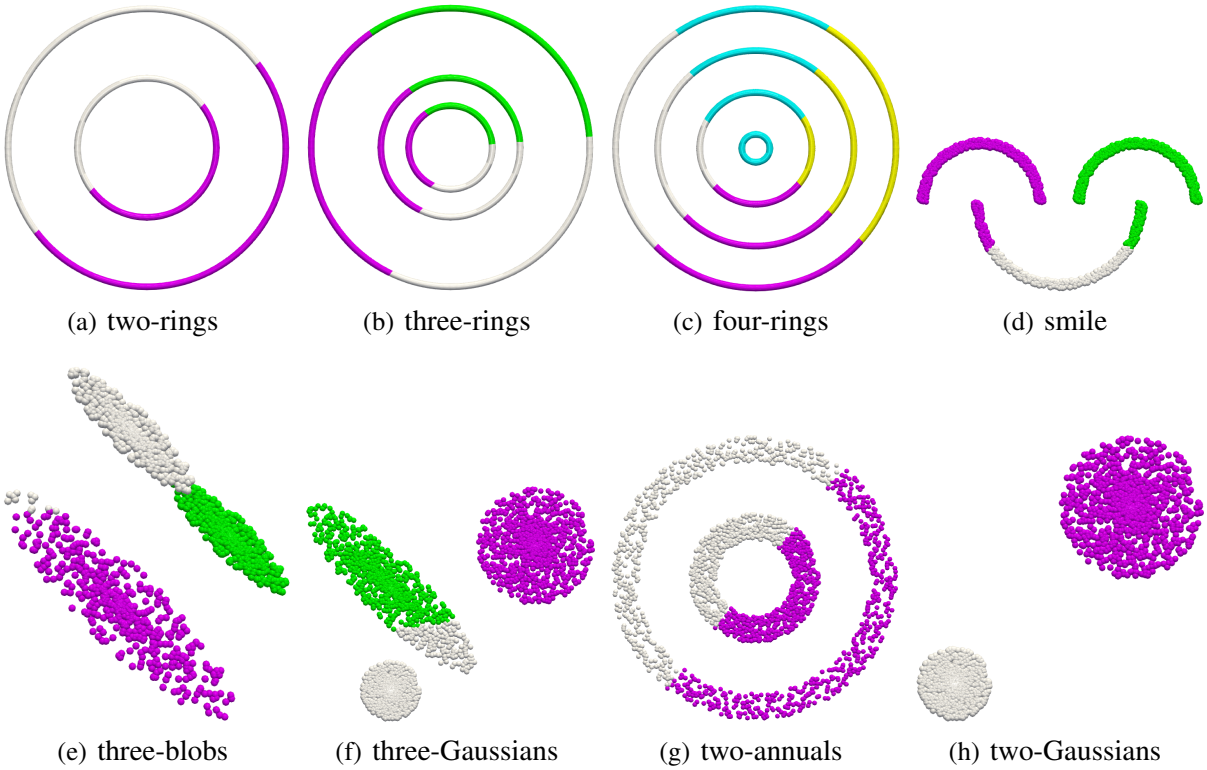
We implemented C++ as initial sampling method and tested the clustering effect illustrated in Fig. 4 and in Table. 1. Obviously that k-means could only detect convex and spherical shape of clusters.

5 K-medoids

Similar to k-means in Sec. 4, k-medoids also only detects convex (not strictly spherical) shape of clusters as illustrated in Fig. 5 and in Table. 2. There is only slight difference between the results of k-means and k-medoids, while computational overhead is highly increased due to medoid computing (could be median from samples or by iteration as in [geometric median], and the results show that with former method).

6 AHC

As is well-known, AHC has three linkages, i.e., single, complete and average. The results of single, complete and average linkage are respectively shown in Fig. 6, 7 and 8. Only single-linkage AHC could authentically detect all



(a) two-rings (b) three-rings (c) four-rings (d) smile

(e) three-blobs (f) three-Gaussians (g) two-annuals (h) two-Gaussians

Figure 4: K-means clustering results on point cloud data sets. Notice that only two spherical clusters in Fig. 4(h) can be authentically detected.

complicated shapes of clusters, while complete-linkage and average-linkage could not, this is due to that single-linkage AHC is beneficial for detecting chaining effects and shapes.

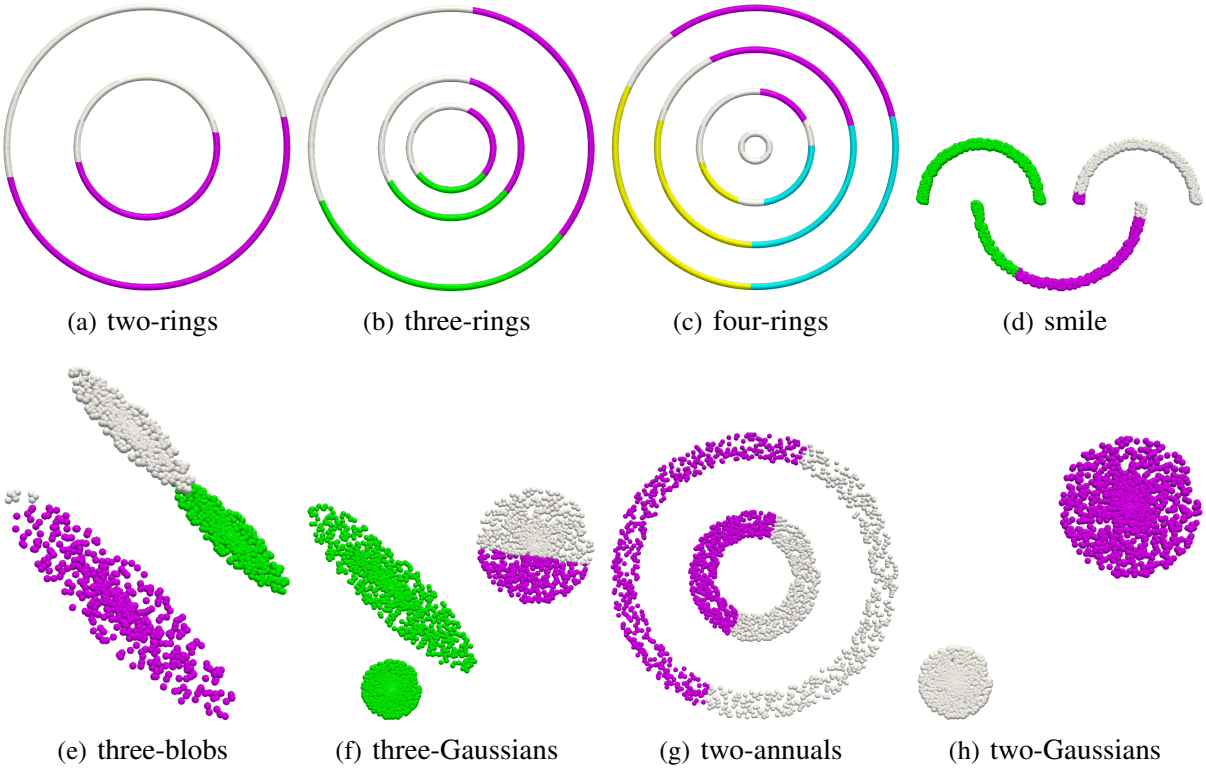
Table 3: Clustering Measurements for single-linkage AHC

Analysis	Rings2	Rings3	Rings4	ThreeBlobs	TwoGaussians	ThreeGaussians	Smile	twoAnnuals
Silhouette	0.11765	0.034080	0.01460	0.63116	0.84262	0.6505	0.45144	0.15816
DB-Index	3.74e7	3.46e7	2.63e8	0.55390	0.21670	0.51799	0.85903	2.46e7
Gamma	0.10727	0.12525	0.15601	0.78461	0.96881	0.78309	0.62838	0.15032

7 BIRCH

8 DBSCAN

Some of clustering results is illustrated in Fig. 9. With careful selections of parameters, DBSCAN is able to detect naturally meaningful shapes of clusters.



(a) two-rings (b) three-rings (c) four-rings (d) smile

(e) three-blobs (f) three-Gaussians (g) two-annuals (h) two-Gaussians

Figure 5: K-medoids clustering results on point cloud data sets. Note that only Fig. 5(h) can be authentically detected.

9 OPTICS

10 AP

11 Validity Measurement Comparisons

Besides, we implemented a new validity measurement as in [4] by minimizing a validity index w.r.t. minimal spanning tree distances among all clusters. From Table 4 we can see this validity measure is small when the clustering is correct.

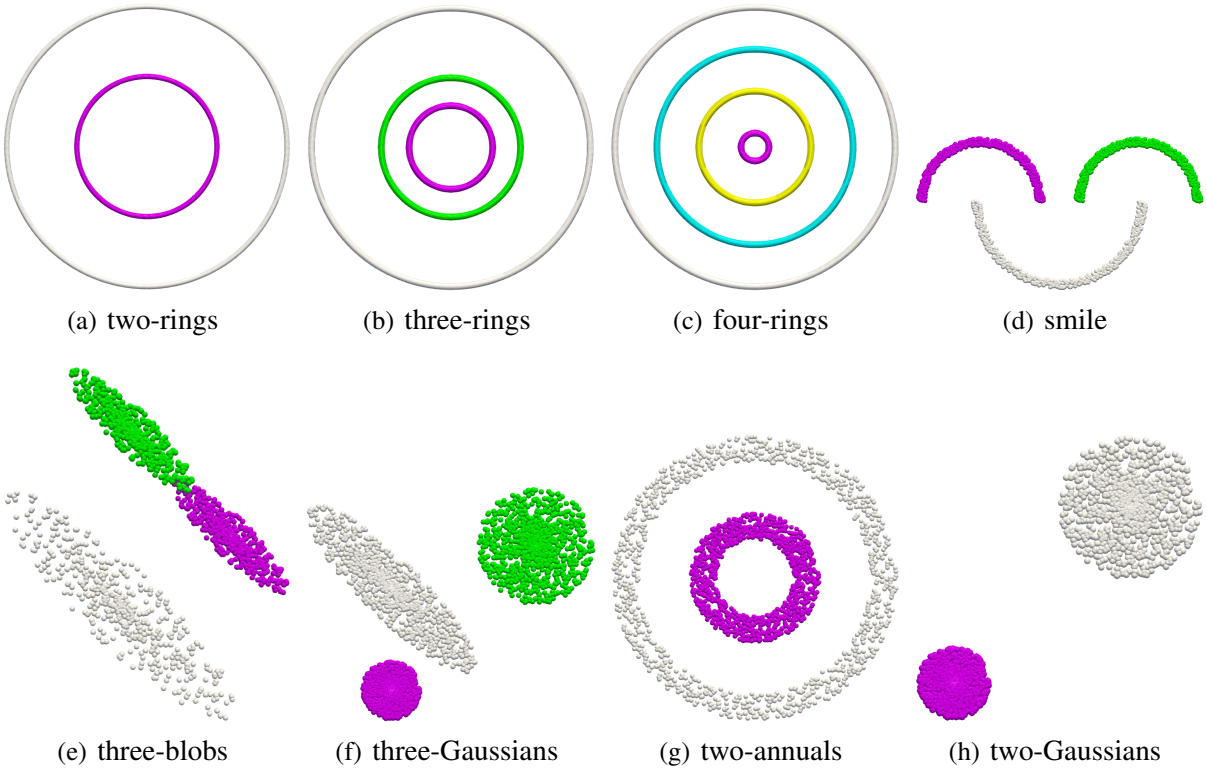
Table 4: Validity Measurement for clustering

Analysis	Rings2	Rings3	Rings4	ThreeBlobs	TwoGaussians	ThreeGaussians	Smile	twoAnnuals
K-means	8.34e-3	1.14e-2	8.25e-3	1.09e-2	1.46e-3	1.89e-3	4.73e-3	8.80e-3
AHC-sing[1]	1.87e-9	1.31e-9	1.03e-9	2.36e-3	1.46e-3	1.61e-3	2.98e-4	2.29e-3
SC-kmeans	1.87e-9	1.15e-2	5.43e-3	2.36e-3	1.46e-3	1.61e-3	1.23e-3	2.29e-3

References

[1] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, pages 849–856. MIT Press, 2001.

[2] Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, NIPS'04, pages 1601–1608, Cambridge, MA, USA, 2004. MIT Press.



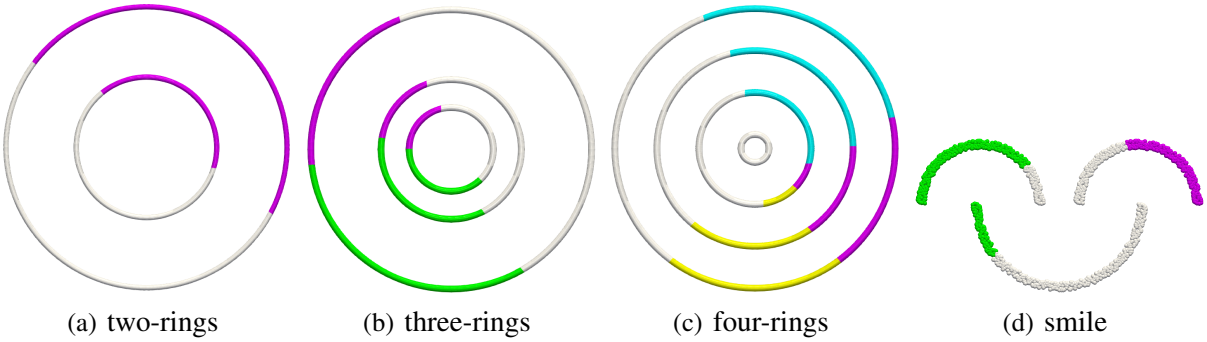
(a) two-rings (b) three-rings (c) four-rings (d) smile

(e) three-blobs (f) three-Gaussians (g) two-annuals (h) two-Gaussians

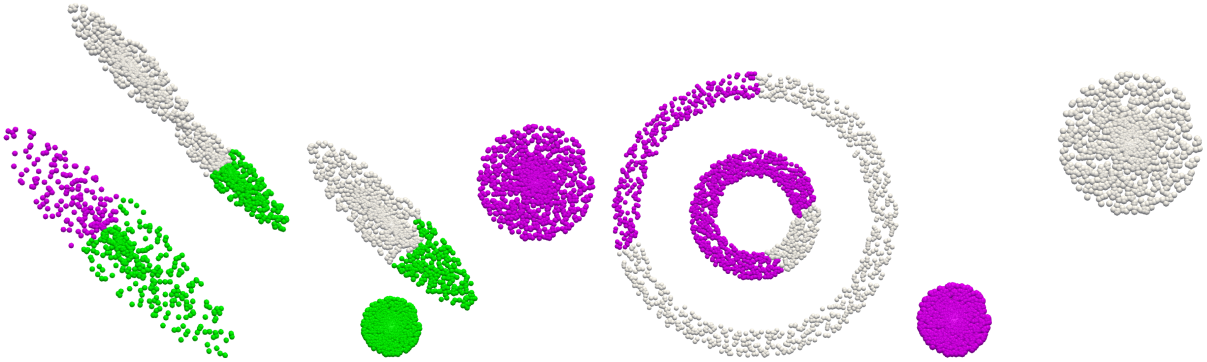
Figure 6: Single-linkage AHC clustering results on point cloud data sets. Note that only single-linkage AHC could precisely detect all shapes of clusters.

[3] David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.

[4] N. A. Yousri, M. S. Kamel, and M. A. Ismail. A novel validity measure for clusters of arbitrary shapes and densities. In *2008 19th International Conference on Pattern Recognition*, pages 1–4, Dec 2008.

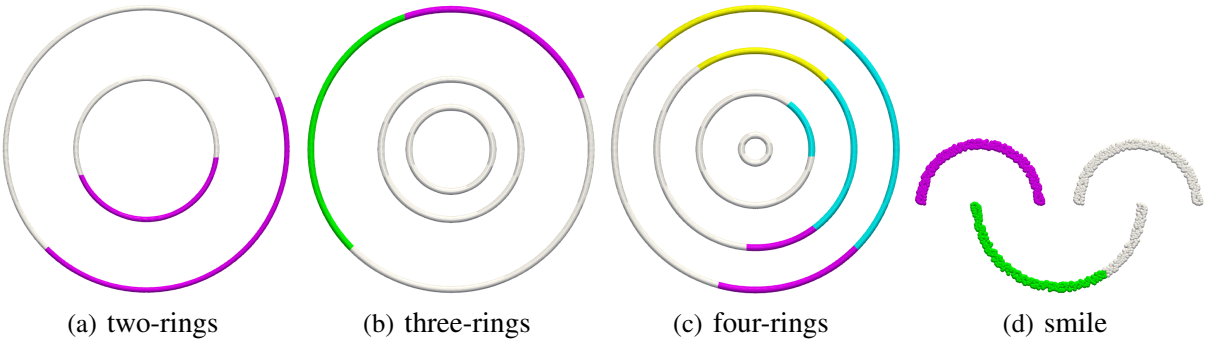


(a) two-rings (b) three-rings (c) four-rings (d) smile

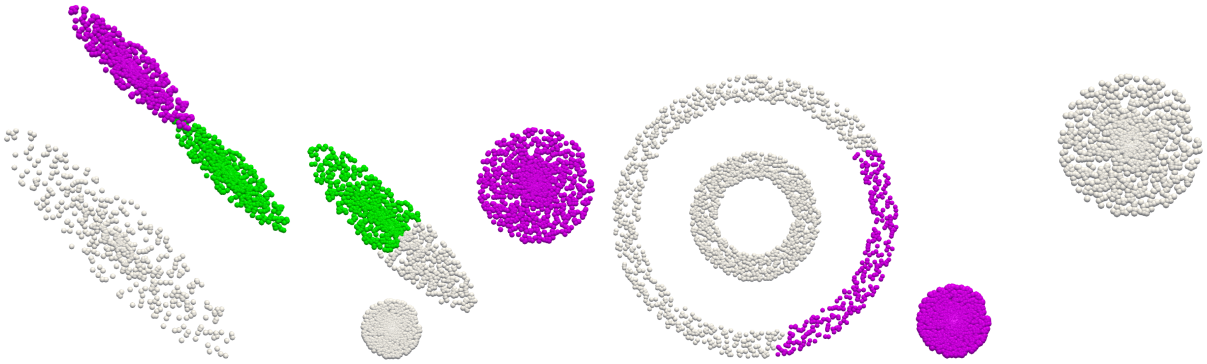


(e) three-blobs (f) three-Gaussians (g) two-annuals (h) two-Gaussians

Figure 7: Complete-linkage AHC clustering results on point cloud data sets.

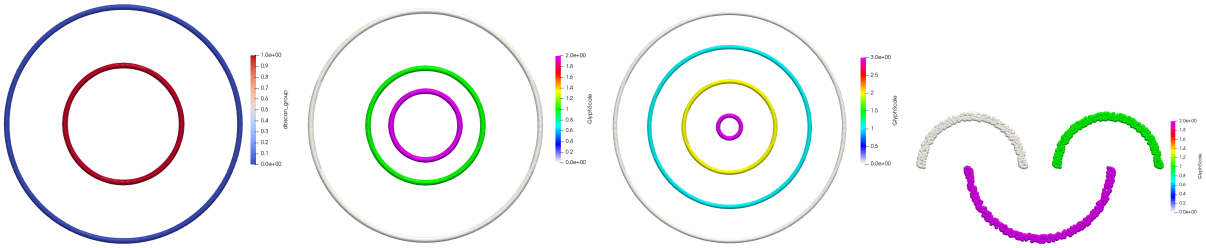


(a) two-rings (b) three-rings (c) four-rings (d) smile

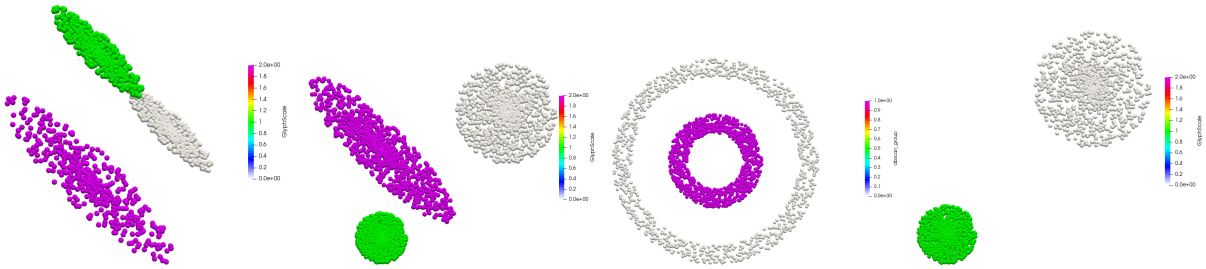


(e) three-blobs (f) three-Gaussians (g) two-annuals (h) two-Gaussians

Figure 8: Average-linkage AHC clustering results on point cloud data sets.



(a) two-rings (b) three-rings (c) four-rings (d) smile



(e) three-blobs (f) three-Gaussians (g) two-annuals (h) two-Gaussians

Figure 9: DBSCAN clustering results on point cloud data sets.