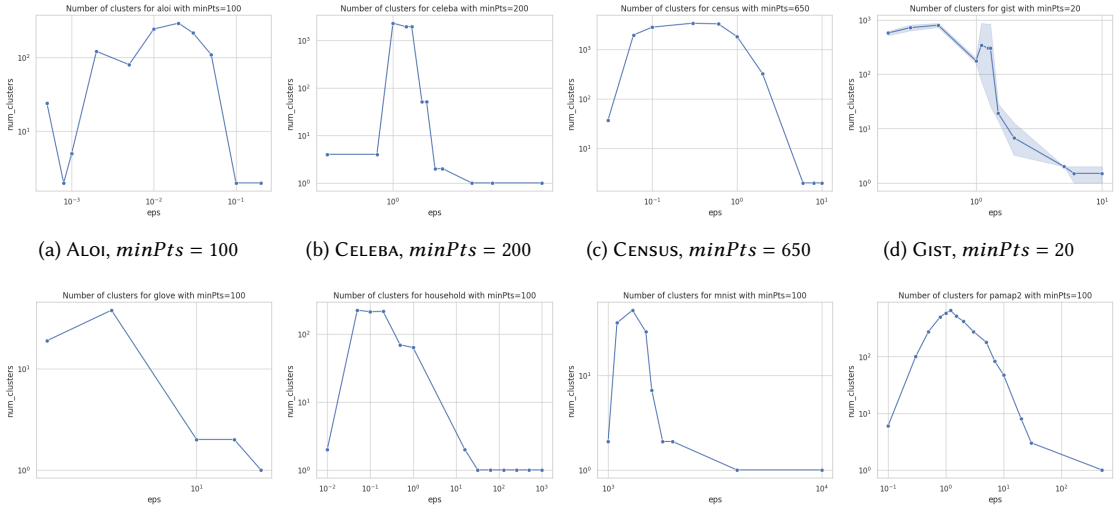
Supplementary material: High-dimensional density-based clustering using locality-sensitive hashing

ANONYMIZED AUTHOR(S)

ACM Reference Format:
Anonymized Author(s). 2024. Supplementary material: High-dimensional density-based clustering using locality-sensitive hashing. 1, 1 (May 2024), 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 CLUSTERING STRUCTURE OF THE DATASETS

Figure 1 provides an overview over the clustering structure found inside the different datasets. For Gist (Fig 1(d)), no baseline results are available. We report on the clustering structure that was found by our algorithm for different values of δ.



(a) Aloi, $minPts = 100$ (b) Celeba, $minPts = 200$ (c) Census, $minPts = 650$ (d) Gist, $minPts = 20$

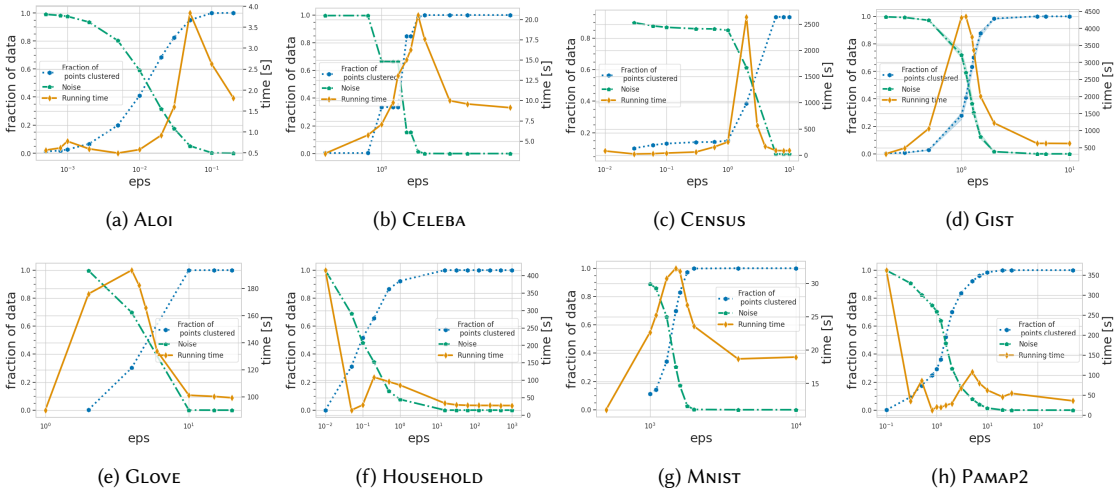
(e) Glove, $minPts = 100$ (f) Household, $minPts = 100$ (g) Mnist, $minPts = 100$ (h) Pamap2, $minPts = 100$

Fig. 1. The number of clusters for different values of epsilon for various datasets.

2 RUNNING TIME VS. CLUSTERING STRUCTURE

Figure 2 relates the running time of our implementation to the clustering structure found in the dataset.

Author's address: Anonymized Author(s).



(a) Aloi (b) Celeba (c) Census (d) Gist

(e) Glove (f) Household (g) Mnist (h) Pamap2

Fig. 2. The fraction of points clustered, the fraction of the data constituted by the largest component and the fraction of noise.

(a) Aloi (b) Celeba (c) Census (d) Gist

(e) Glove (f) Household (g) Mnist (h) Pamap2

Fig. 3. Running times for different for various datasets.

3 RUNNING TIME COMPARISON

Figure 3 relates the running time of the different implementations to each other. If a implementation is missing data points, it means that certain ε configurations could not be run. If an implementation is missing at all from a certain dataset, it means that not a single ε-value led to a run that succeeded within the time limit.

Aloi						
Failure probability δ	$\varepsilon = 0.0008$		$\varepsilon = 0.01$		$\varepsilon = 0.1$	
	Time [s]	ARI	Time [s]	ARI	Time [s]	ARI
0.01	0.89	0.9423	1.8	0.9569	2.88	1.0
0.1	0.62	0.9423	0.58	0.9534	2.62	1.0
0.5	0.48	0.9374	0.47	0.9185	1.44	0.9124
0.9	0.37	0.9243	0.44	0.6455	1.04	0.0084

Table 1. ARI for different epsilon and failure probabilities for the aloi dataset.

Celeba						
Failure probability δ	$\varepsilon = 0.8$		$\varepsilon = 1.6$		$\varepsilon = 4.0$	
	Time [s]	ARI	Time [s]	ARI	Time [s]	ARI
0.01	7.98	0.1926	29.04	0.9909	10.75	1.0
0.1	5.75	0.1926	16.29	0.9808	9.58	1.0
0.5	3.47	0.1926	7.71	0.8141	5.85	1.0
0.9	2.58	0.1926	4.49	0.1628	4.78	0.0

Table 2. ARI for different epsilon and failure probabilities for the celeba dataset.

Census						
Failure probability δ	$\varepsilon = 0.03$		$\varepsilon = 1.0$		$\varepsilon = 8.0$	
	Time [s]	ARI	Time [s]	ARI	Time [s]	ARI
0.01	496.56	0.9544	1122.9	0.9494	103.75	0.7305
0.1	23.42	0.9544	253.61	0.95	86.88	0.7305
0.5	22.12	0.9589	53.38	0.9409	62.36	0.7305
0.9	21.12	0.9484	30.96	0.9317	56.69	-0.0032

Table 3. ARI for different epsilon and failure probabilities for the celeba dataset.

Glove						
Failure probability δ	$\varepsilon = 2.0$		$\varepsilon = 4.0$		$\varepsilon = 10.0$	
	Time [s]	ARI	Time [s]	ARI	Time [s]	ARI
0.01	328.48	0.9437	353.1	0.9167	116.13	0.827
0.1	176.14	0.931	193.69	0.8268	101.1	0.732
0.5	84.56	0.8759	89.51	0.6274	93.01	0.4849
0.9	41.92	0.4592	46.64	0.2785	53.89	0.0111

Table 4. ARI for different epsilon and failure probabilities for the glove dataset.

4 INFLUENCE OF FAILURE PROBABILITY δ ON ARI SCORE

Tables 1—7 give an overview of the relation of the failure probability δ that we set as worst-case misclassification probability for a core point in our implementation to the ARI score.

Mnist						
Failure probability δ	$\varepsilon = 1000$		$\varepsilon = 1600$		$\varepsilon = 4000$	
	Time [s]	ARI	Time [s]	ARI	Time [s]	ARI
0.01	36.11	0.9938	43.4	0.9749	25.2	1.0
0.1	22.7	0.9945	31.83	0.9129	18.7	1.0
0.5	14.41	0.9819	17.62	0.5472	14.77	0.0
0.9	10.34	0.6297	13.9	0.0372	13.49	0.0

Table 5. ARI for different epsilon and failure probabilities for the Mnist dataset.

Household						
Failure probability δ	$\varepsilon = 0.05$		$\varepsilon = 1.0$		$\varepsilon = 128$	
	Time [s]	ARI	Time [s]	ARI	Time [s]	ARI
0.01	21.6	0.7333	84.73	0.9815	29.06	1.0
0.1	14.21	0.7327	86.15	0.981	27.8	1.0
0.5	17.76	0.695	49.43	0.9706	24.13	1.0
0.9	9.8	0.6159	35.42	-0.0062	20.54	0.0

Table 6. ARI for different epsilon and failure probabilities for the household dataset.

Pamap2						
Failure probability δ	$\varepsilon = 0.3$		$\varepsilon = 2.0$		$\varepsilon = 30$	
	Time [s]	ARI	Time [s]	ARI	Time [s]	ARI
0.01	27.44	0.7866	47.76	0.6857	57.47	0.9983
0.1	35.6	0.7866	29.7	0.6792	54.81	0.9983
0.5	16.24	0.783	17.35	0.5429	153.09	0.9973
0.9	9.31	0.7325	13.5	0.0432	25.07	0.0096

Table 7. ARI for different epsilon and failure probabilities for the pamap2 dataset.