UNIVERSITÀ DI PISA

University of Pisa

Laurea Magistrale (MSc) in Artificial Intelligence and Data Engineering

Project

Data Mining and Machine Learning

FEDERATED DBSCAN BASED ON GRID

Alessio Serra, Valerio Giannini

https://github.com/ValeGian/DMML_FederatedDBSCAN

Academic year 2020-2021

INTRODUCTION TO THE PROBLEM

Federated Learning (FL): Can we train a model, in a "collaborative" way, without transferring the data to a central processing server?

Local Data Owner:

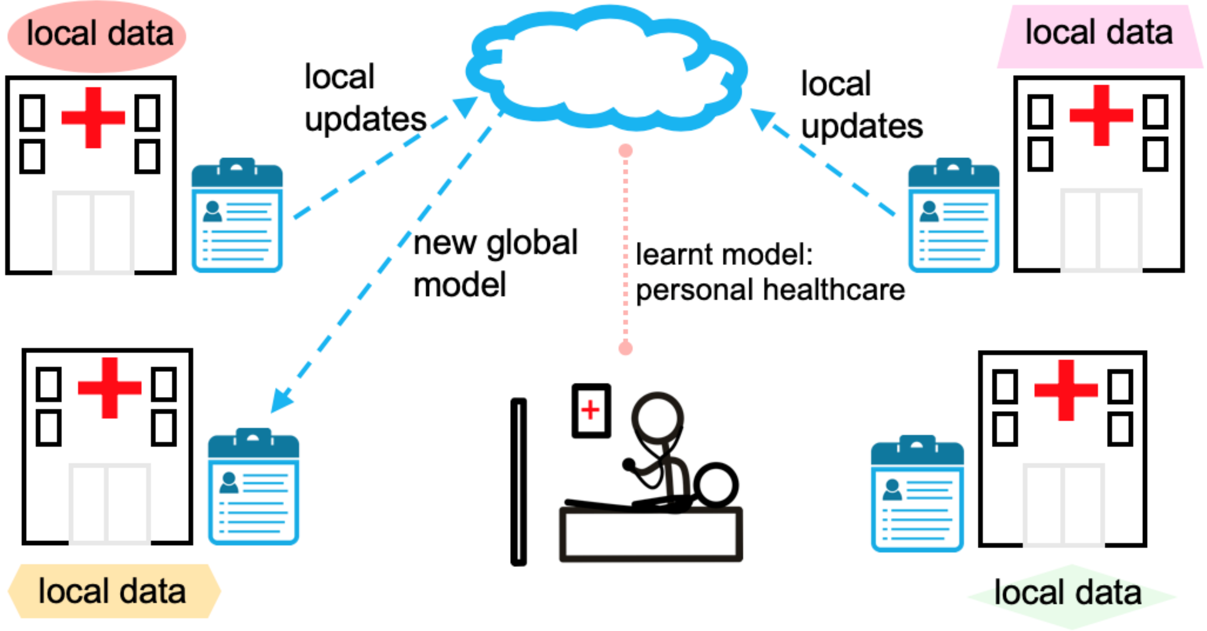
- Contains several samples described with same attributes.
- Can perform local processing.
- Can assign points to clusters depending on the result they receive from the server.

Central Server:

- Aggregate local models and consolidate the global model.
- Send results to individual Data Owners.

Pratical example:

- **Healthcare domain** in which patient data cannot be transmitted.



STEPS OF THE ALGORITHM

LOCAL	SERVER
Partition the space with a granularity fixed (L), assuming the same range of features for all nodes.	
Evaluates the number of points in each cell and transmits information about non-empty cells to the server.	
	For each cell, add the contributions of all owner of the data.
	Define dense cell the cell with at least MinPts.
	Evaluate clustering by expanding a cluster along adjacent dense cells.
	Return to each local information on cluster membership of each cell.
Assign all the points relating to the cells dense to its cluster.	
Assign the remaining points to the cluster of dense adjacent cell closest to the point. Otherwise, the point is considered an	

PARAMETERS TO SET

Real parameter of federated algorithm:

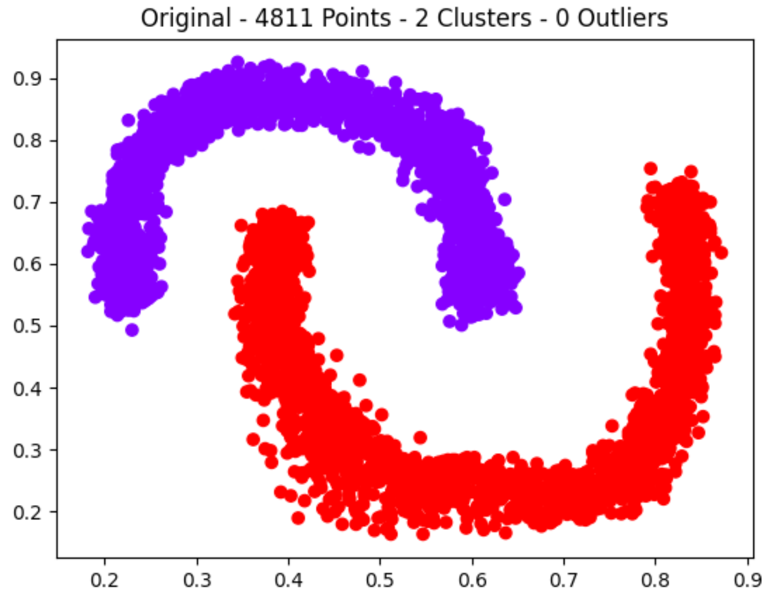
- L = fix the granularity of the cell.
- MinPts = determine the minimum number of points in a dense cell.

Parameters to simulate locally a distributed execution:

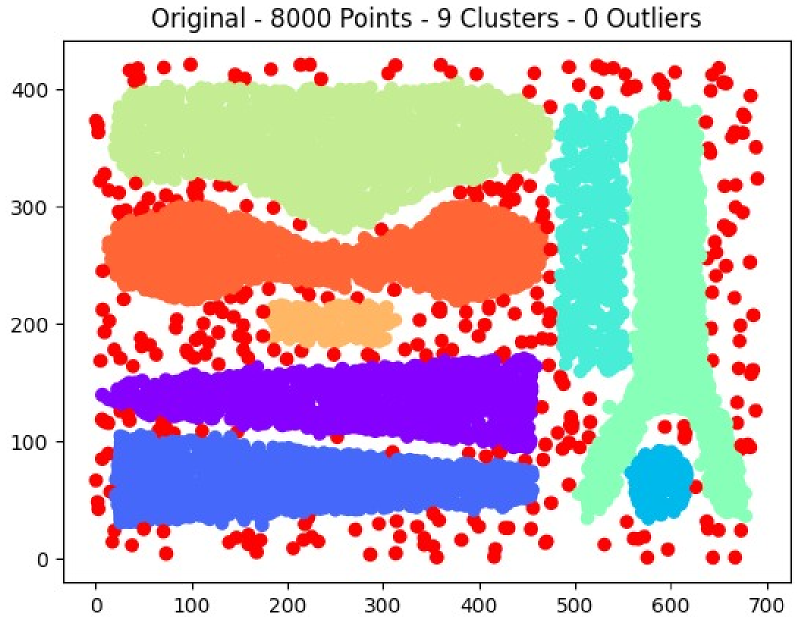
- M = Number of nodes.
- Partitioning methods.

DATASET ANALIZED

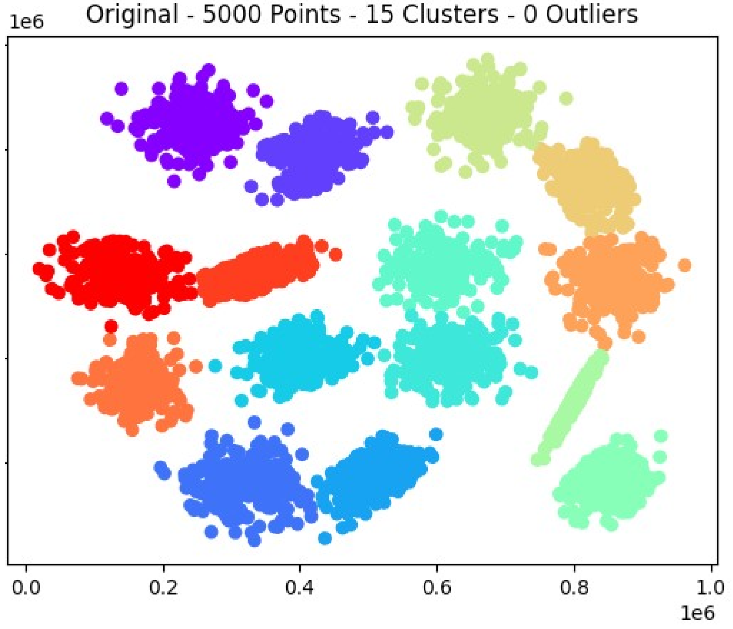
BANANA (4800)


Original - 4811 Points - 2 Clusters - 0 Outliers

CLUTO-T8.8K (8000)

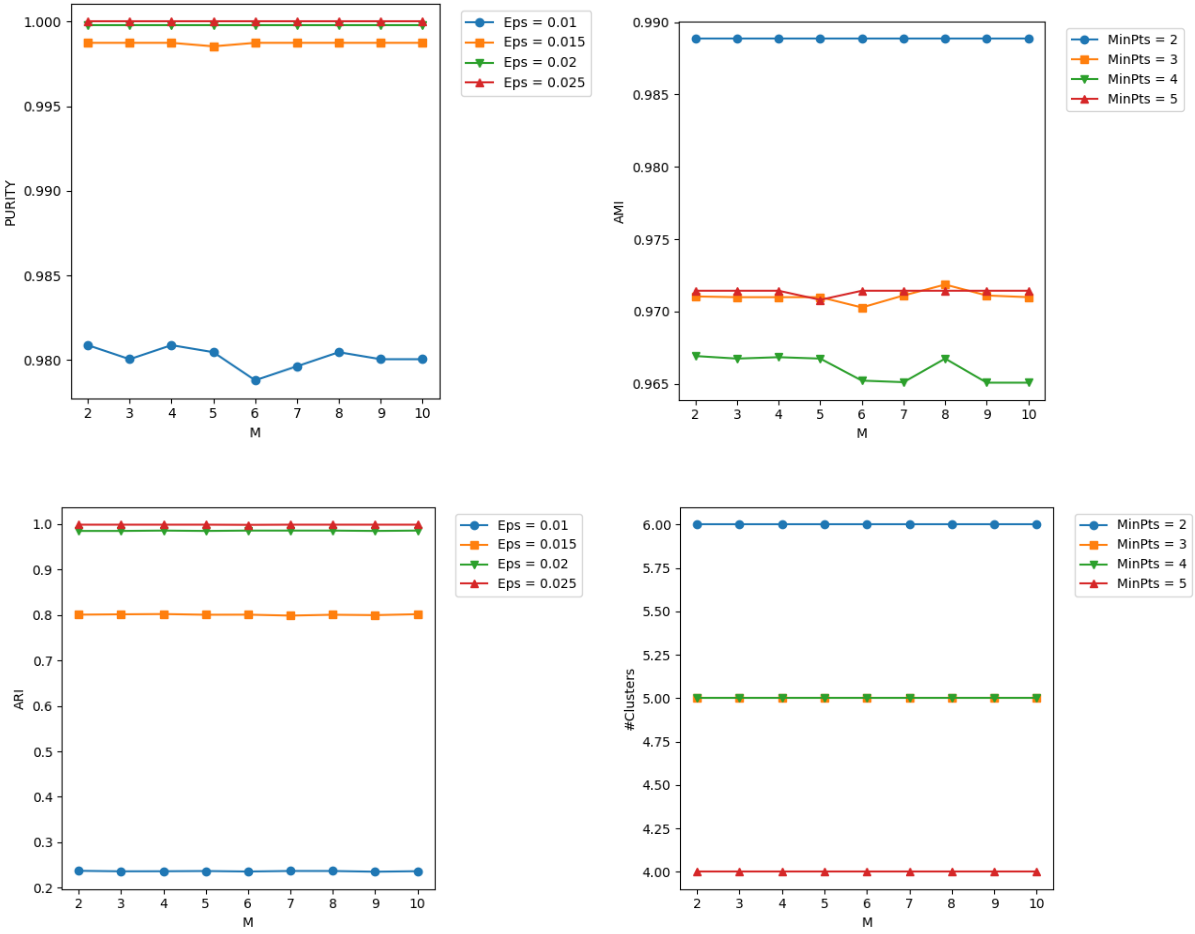

Original - 8000 Points - 9 Clusters - 0 Outliers

S-SET-1 (5000)

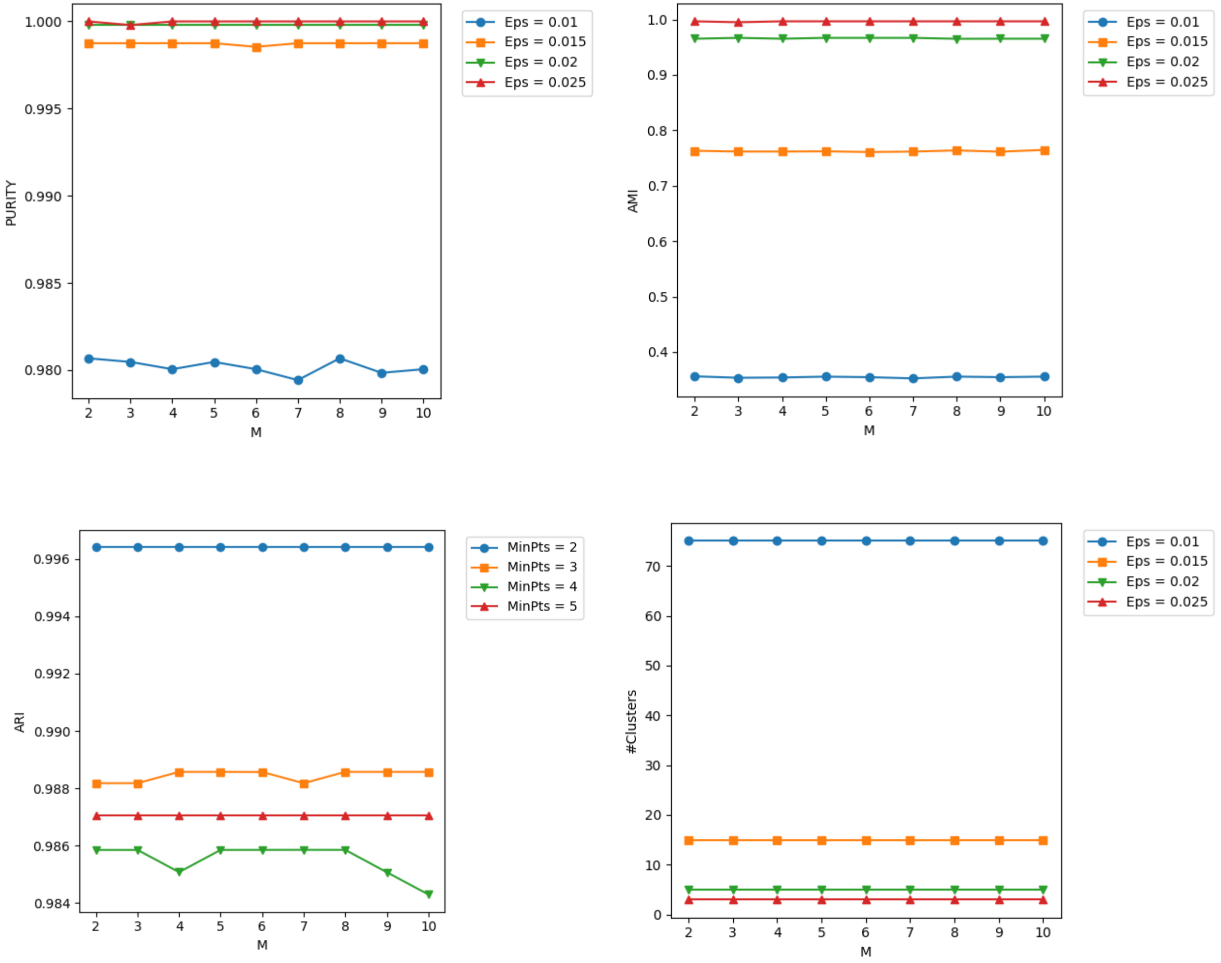

Original - 5000 Points - 15 Clusters - 0 Outliers

PARAMETERS REDUCTION

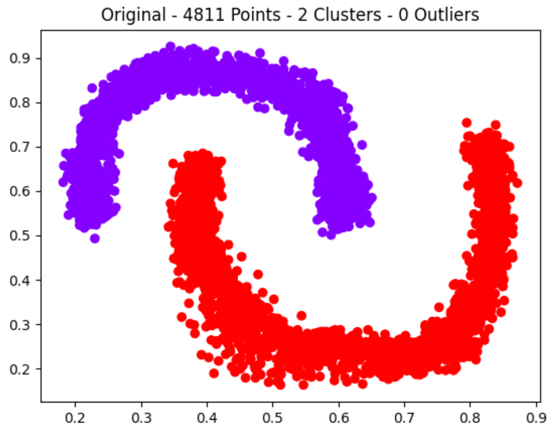
Stratified Partitioning

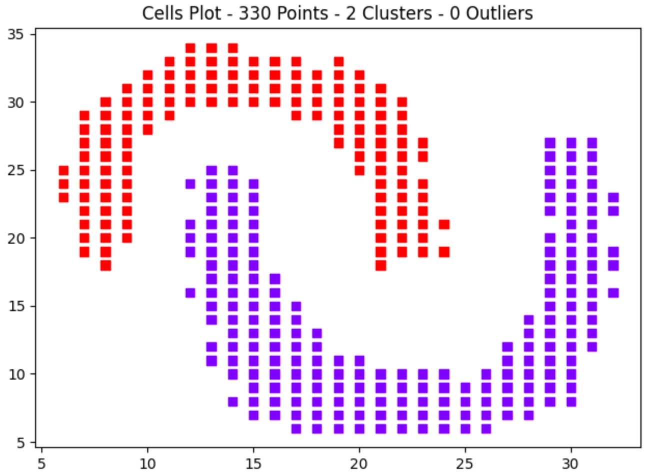


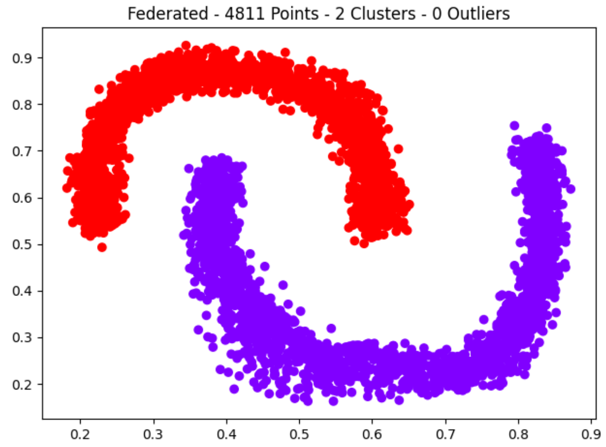
Separated Partitioning



RESULT WITH BANANA







L = 0.03   MinPts = 2

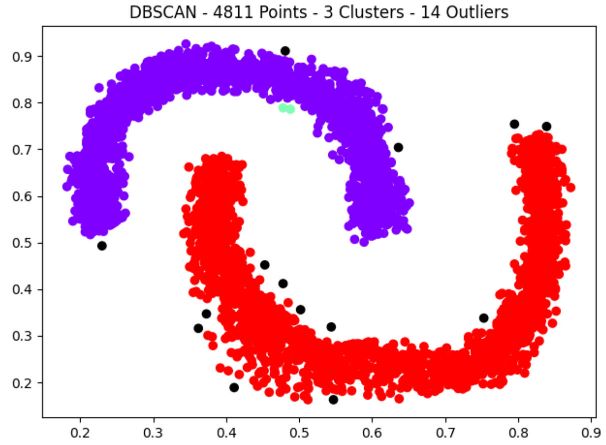
PURITY: 1.0
ARI: 1.0
AMI: 1.0
PRECISION-BCUBED: 1.0
RECALL-BCUBED: 1.0

Outliers Federated

MinPts \ L	0.002	0.004	0.006	0.008	0.01	0.012	0.014	0.016	0.018	0.02	0.022	0.024	0.026	0.028	0.03	0.032	0.034	0.036	0.038
2.0	3902.0	1622.0	549.0	203.0	89.0	48.0	23.0	12.0	7.0	3.0	3.0	1.0	2.0	1.0	0.0	1.0	0.0	0.0	0.0
3.0	4725.0	3379.0	1460.0	585.0	213.0	100.0	46.0	31.0	15.0	9.0	4.0	1.0	2.0	1.0	0.0	1.0	0.0	0.0	0.0
4.0	4793.0	4460.0	2639.0	1089.0	442.0	232.0	83.0	38.0	24.0	21.0	5.0	1.0	3.0	2.0	0.0	1.0	0.0	0.0	0.0
5.0	4811.0	4787.0	3516.0	1752.0	873.0	362.0	193.0	80.0	51.0	28.0	12.0	3.0	3.0	2.0	1.0	1.0	0.0	0.0	0.0
6.0	4811.0	4811.0	4229.0	2605.0	1362.0	523.0	297.0	130.0	59.0	40.0	25.0	19.0	5.0	2.0	0.0	1.0	0.0	0.0	0.0
7.0	4811.0	4811.0	4649.0	3271.0	1745.0	816.0	479.0	200.0	85.0	49.0	27.0	19.0	21.0	6.0	3.0	1.0	0.0	0.0	0.0

Outliers DBSCAN

MinPts \ Eps	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009	0.01	0.011	0.012	0.013	0.014	0.015	0.016	0.017	0.018	0.019
2.0	4639.0	3375.0	2021.0	1180.0	757.0	433.0	298.0	204.0	139.0	104.0	70.0	56.0	45.0	34.0	27.0	22.0	19.0	17.0	15.0
3.0	4811.0	4357.0	3035.0	1902.0	1245.0	681.0	470.0	298.0	213.0	152.0	98.0	74.0	59.0	42.0	37.0	32.0	23.0	21.0	17.0
4.0	4811.0	4747.0	3864.0	2694.0	1805.0	1037.0	692.0	458.0	292.0	208.0	141.0	102.0	78.0	56.0	43.0	33.0	27.0	25.0	20.0
5.0	4811.0	4806.0	4429.0	3468.0	2432.0	1417.0	938.0	654.0	398.0	284.0	196.0	151.0	112.0	79.0	56.0	44.0	31.0	27.0	21.0
6.0	4811.0	4811.0	4708.0	4056.0	3010.0	1840.0	1275.0	878.0	602.0	373.0	259.0	193.0	139.0	102.0	74.0	54.0	45.0	31.0	25.0
7.0	4811.0	4811.0	4795.0	4411.0	3596.0	2314.0	1586.0	1108.0	749.0	499.0	344.0	244.0	178.0	136.0	101.0	70.0	52.0	39.0	31.0



ε = 0.019   MinPts = 2

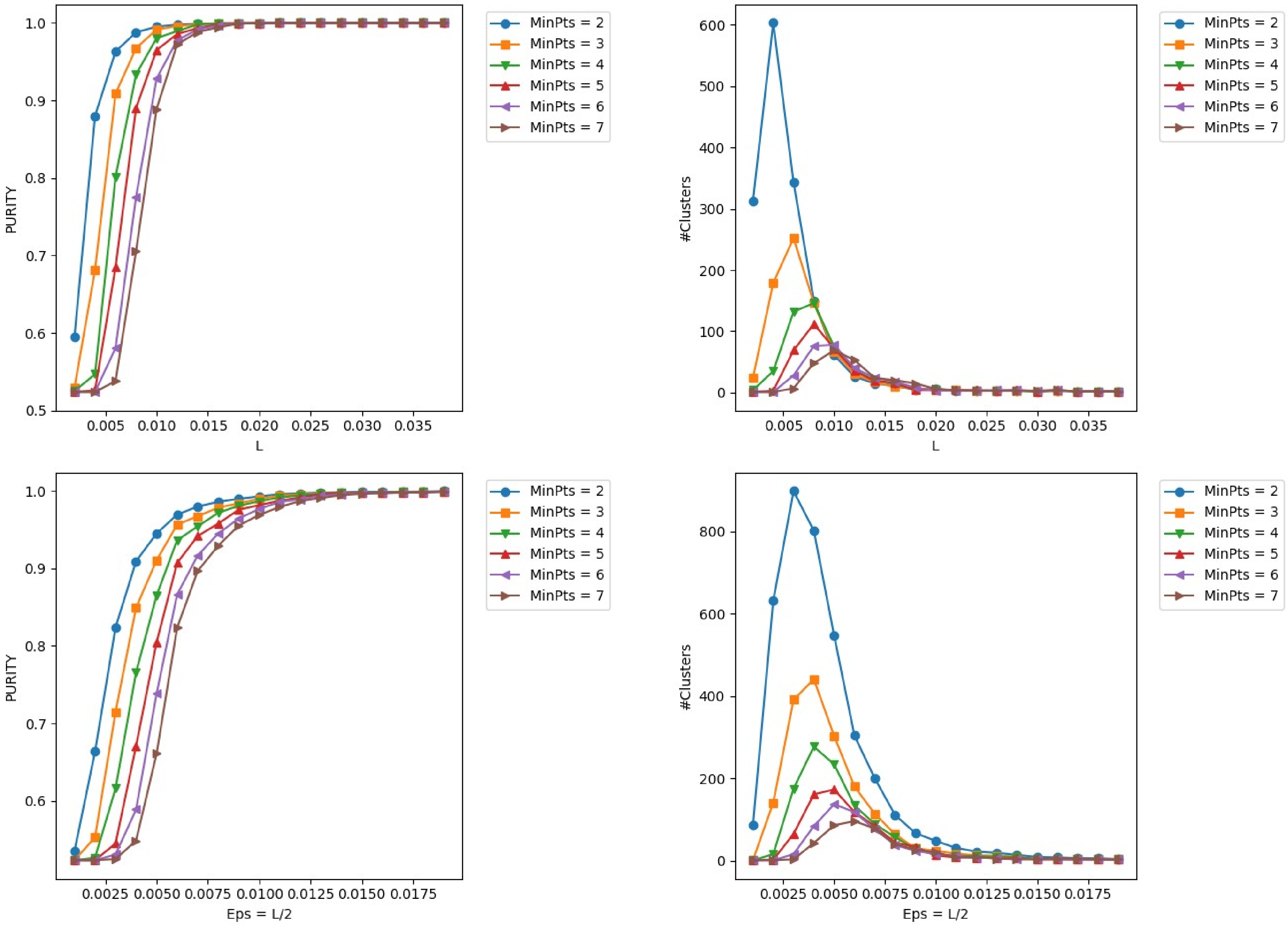
PURITY: 0.9994
ARI: 0.9935
AMI: 0.9828
PRECISION-BCUBED:
0.9990

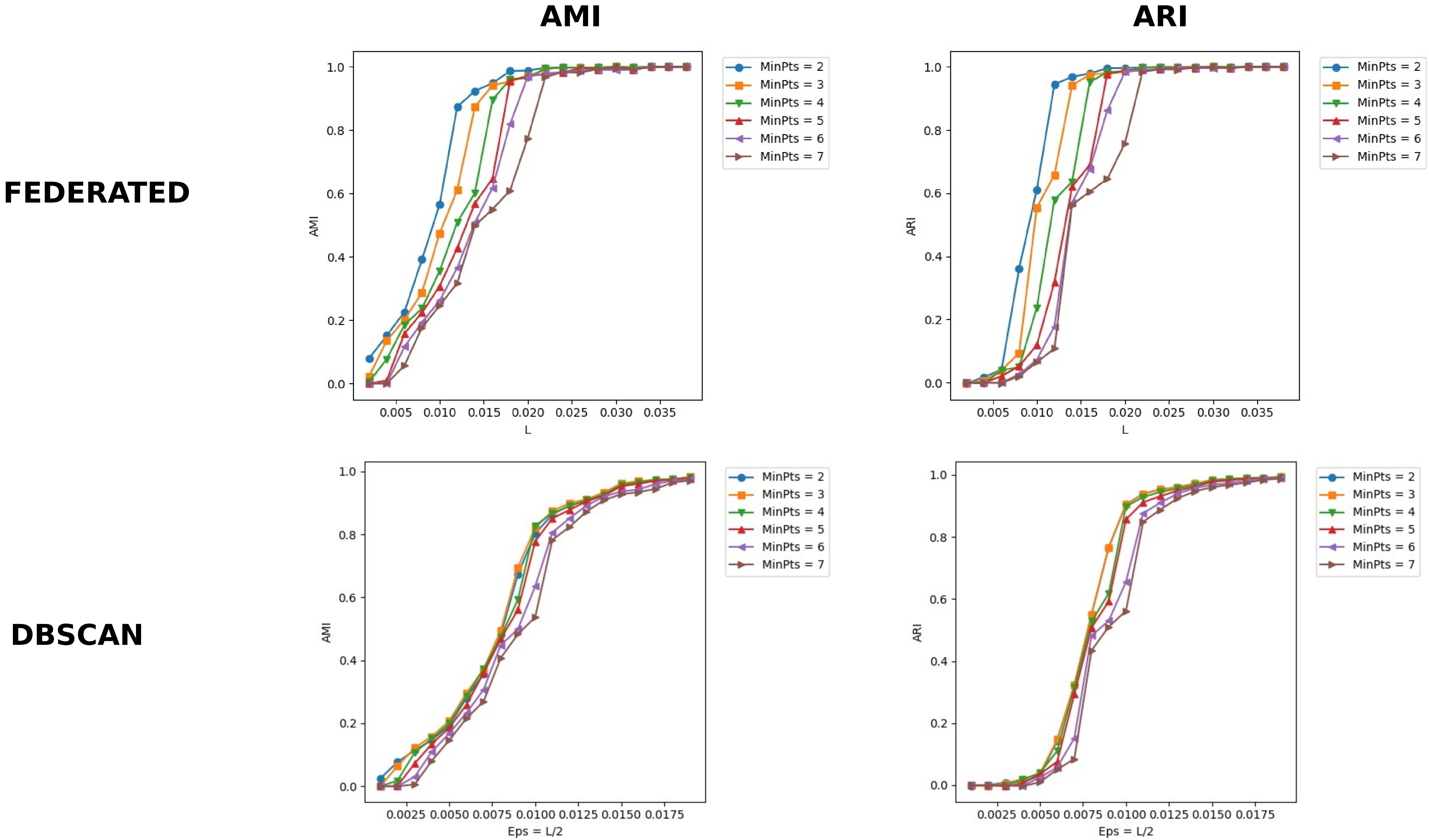
PURITY

CLUSTERS

FEDERATED

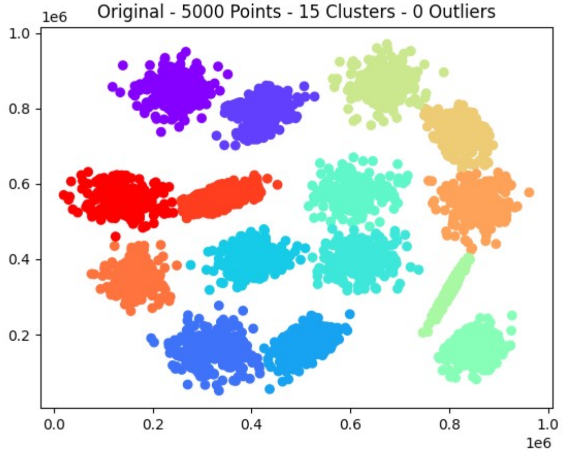
DBSCAN



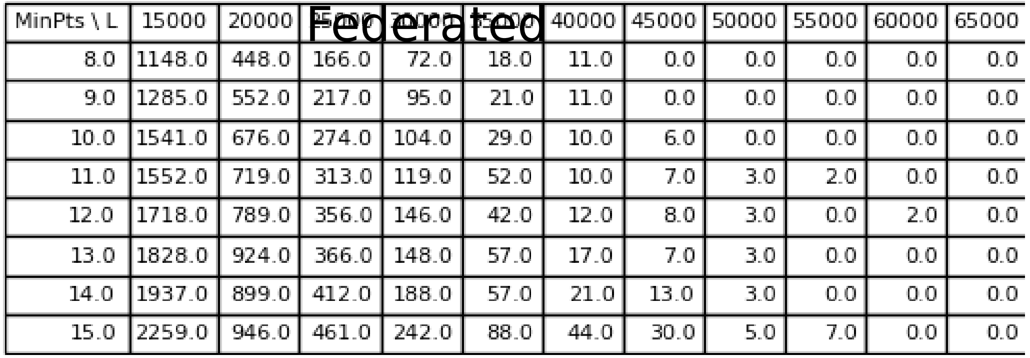


AMI
ARI
FEDERATED
DBSCAN
AMI
ARI
L
Eps = L/2
0.0
0.2
0.4
0.6
0.8
1.0
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0.0025
0.0050
0.0075
0.0100
0.0125
0.0150
0.0175
MinPts = 2
MinPts = 3
MinPts = 4
MinPts = 5
MinPts = 6
MinPts = 7

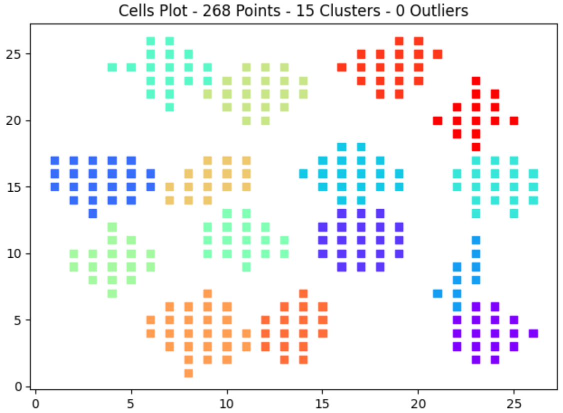
RESULT WITH S-SET-1

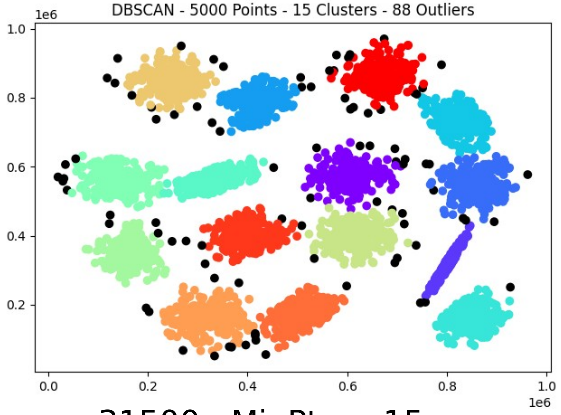


Outliers
Federated



MinPts \ L	15000	20000	25000	30000	35000	40000	45000	50000	55000	60000	65000
8.0	1148.0	448.0	166.0	72.0	18.0	11.0	0.0	0.0	0.0	0.0	0.0
9.0	1285.0	552.0	217.0	95.0	21.0	11.0	0.0	0.0	0.0	0.0	0.0
10.0	1541.0	676.0	274.0	104.0	29.0	10.0	6.0	0.0	0.0	0.0	0.0
11.0	1552.0	719.0	313.0	119.0	52.0	10.0	7.0	3.0	2.0	0.0	0.0
12.0	1718.0	789.0	356.0	146.0	42.0	12.0	8.0	3.0	0.0	2.0	0.0
13.0	1828.0	924.0	366.0	148.0	57.0	17.0	7.0	3.0	0.0	0.0	0.0
14.0	1937.0	899.0	412.0	188.0	57.0	21.0	13.0	3.0	0.0	0.0	0.0
15.0	2259.0	946.0	461.0	242.0	88.0	44.0	30.0	5.0	7.0	0.0	0.0

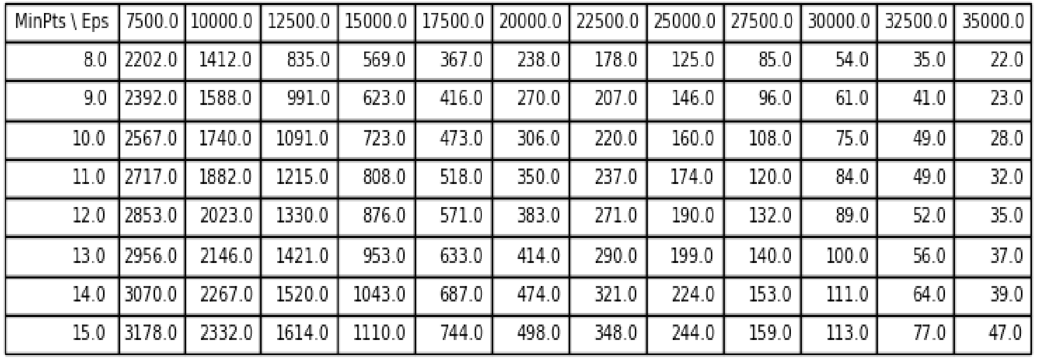




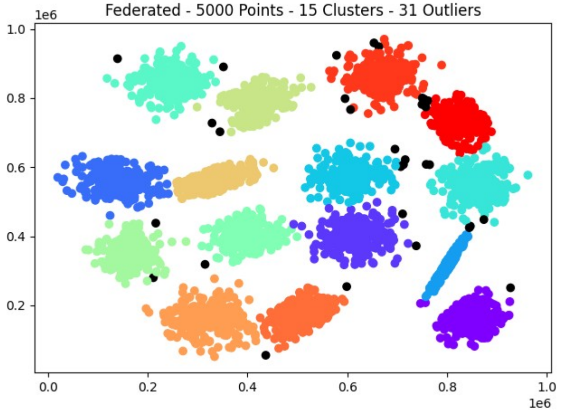
ε = 31500 MinPts = 15

PURITY: 0.9842
ARI: 0.9763
AMI: 0.9741
PRECISION-BCUBED:
0.9818
RECALL-BCUBED:

Outliers



MinPts \ Eps	7500.0	10000.0	12500.0	15000.0	17500.0	20000.0	22500.0	25000.0	27500.0	30000.0	32500.0	35000.0
8.0	2202.0	1412.0	835.0	569.0	367.0	238.0	178.0	125.0	85.0	54.0	35.0	22.0
9.0	2392.0	1588.0	991.0	623.0	416.0	270.0	207.0	146.0	96.0	61.0	41.0	23.0
10.0	2567.0	1740.0	1091.0	723.0	473.0	306.0	220.0	160.0	108.0	75.0	49.0	28.0
11.0	2717.0	1882.0	1215.0	808.0	518.0	350.0	237.0	174.0	120.0	84.0	49.0	32.0
12.0	2853.0	2023.0	1330.0	876.0	571.0	383.0	271.0	190.0	132.0	89.0	52.0	35.0
13.0	2956.0	2146.0	1421.0	953.0	633.0	414.0	290.0	199.0	140.0	100.0	56.0	37.0
14.0	3070.0	2267.0	1520.0	1043.0	687.0	474.0	321.0	224.0	153.0	111.0	64.0	39.0
15.0	3178.0	2332.0	1614.0	1110.0	744.0	498.0	348.0	244.0	159.0	113.0	77.0	47.0

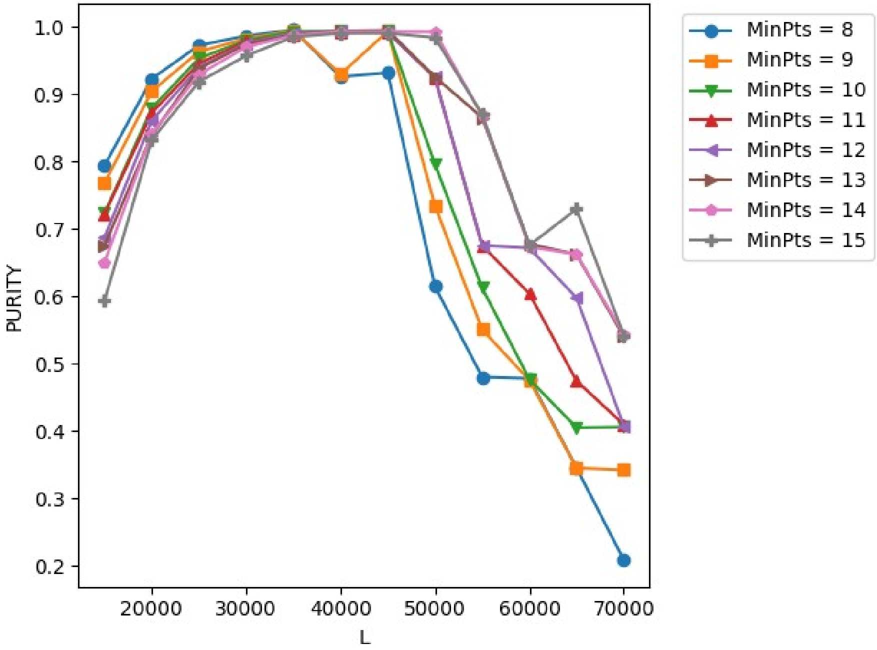


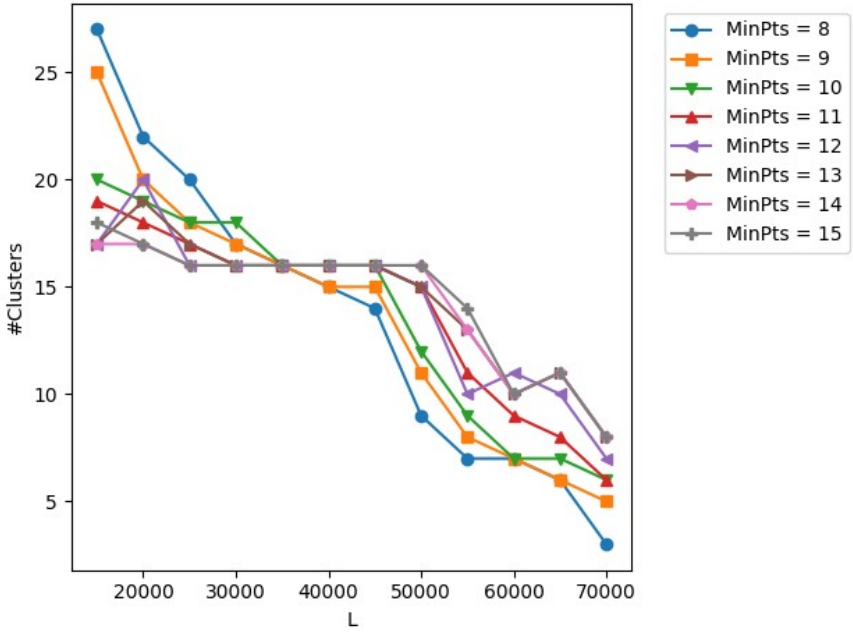
L = 34750 MinPts = 9

PURITY: 0.9962
ARI: 0.9885
AMI: 0.9860
PRECISION-BCUBED:
0.9902
RECALL-BCUBED:

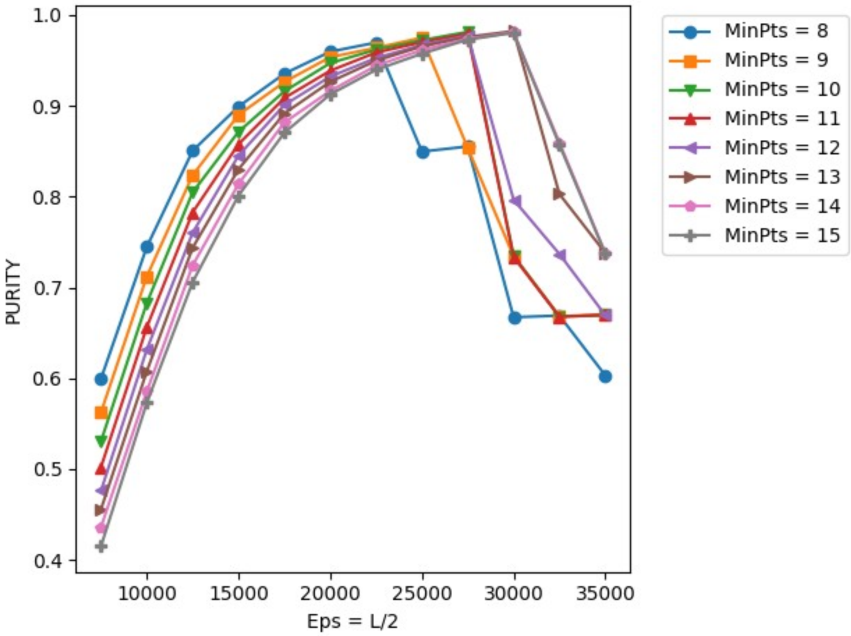
PURITY **# CLUSTERS**

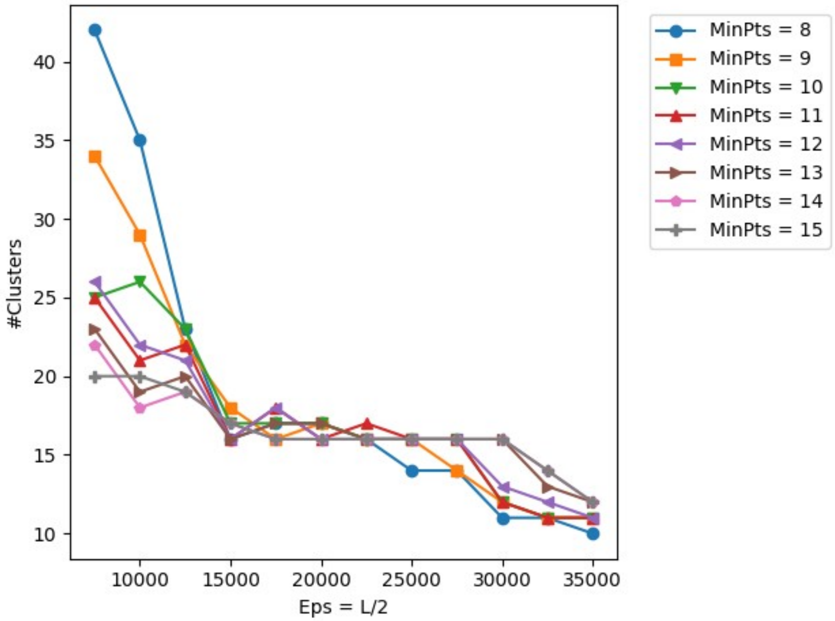
FEDERATED





DBSCAN

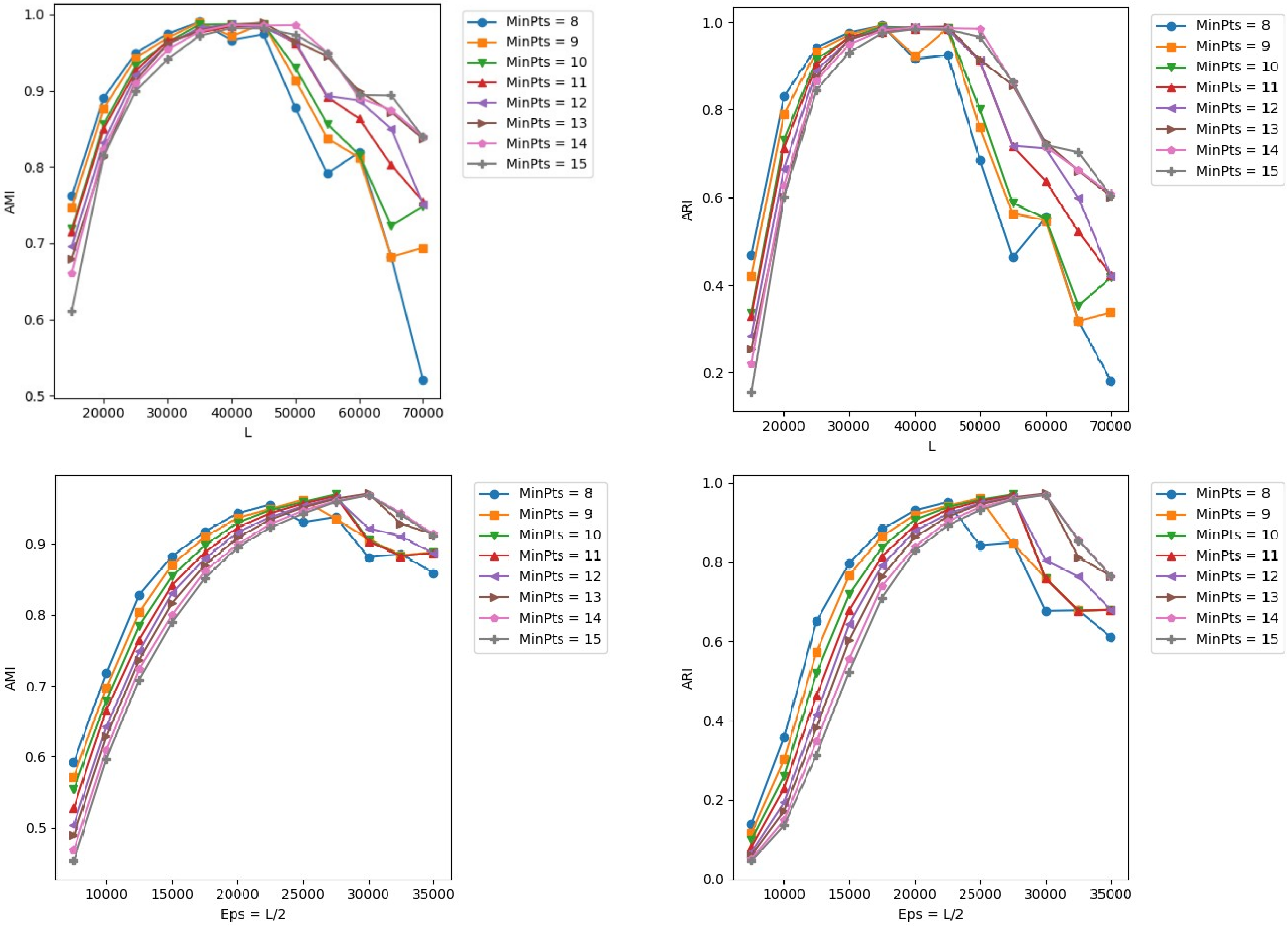




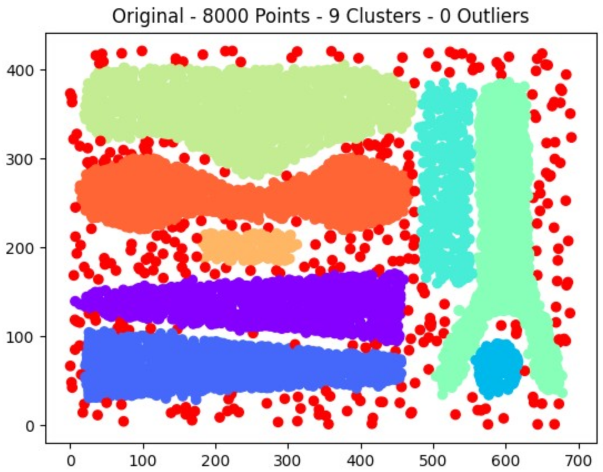
AMI ARI

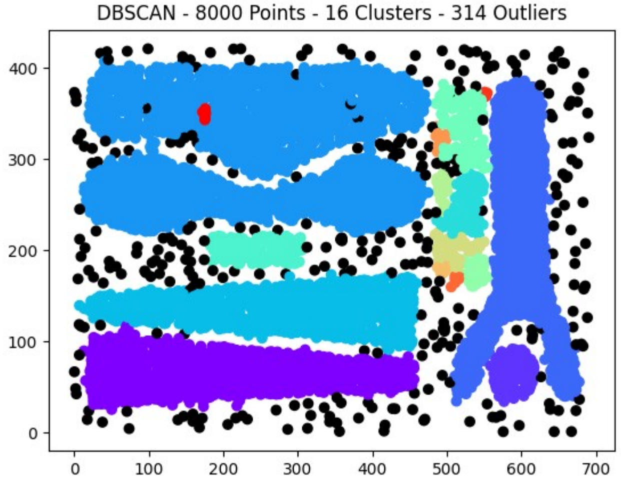
FEDERATED

DBSCAN



RESULTS WITH CLUTO

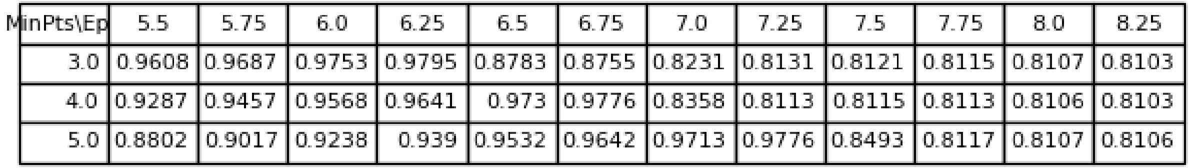

Original - 8000 Points - 9 Clusters - 0 Outliers


DBSCAN - 8000 Points - 16 Clusters - 314 Outliers

ε = 7.25   MinPts = 5

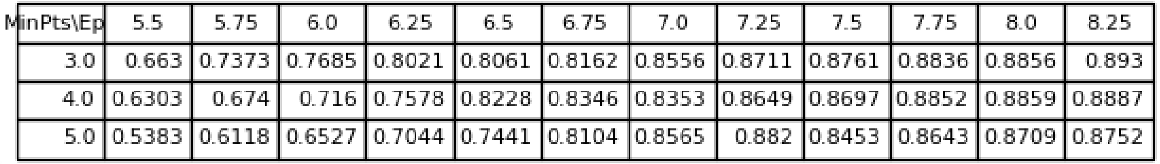
PURITY: 0.9776
ARI: 0.8305
AMI: 0.882
PREC-BCUBED: 0.9427
REC-BCUBED: 0.5400
F-SCORE: 0.6867

PURITY_DBSCAN



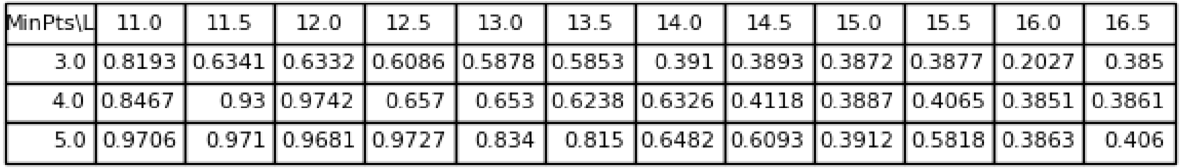
MinPts\Ep	5.5	5.75	6.0	6.25	6.5	6.75	7.0	7.25	7.5	7.75	8.0	8.25
3.0	0.9608	0.9687	0.9753	0.9795	0.8783	0.8755	0.8231	0.8131	0.8121	0.8115	0.8107	0.8103
4.0	0.9287	0.9457	0.9568	0.9641	0.973	0.9776	0.8358	0.8113	0.8115	0.8113	0.8106	0.8103
5.0	0.8802	0.9017	0.9238	0.939	0.9532	0.9642	0.9713	0.9776	0.8493	0.8117	0.8107	0.8106

AMI_DBSCAN



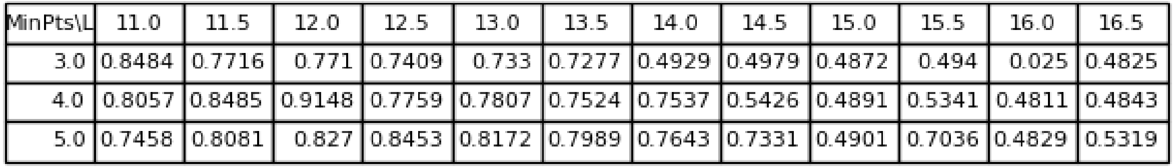
MinPts\Ep	5.5	5.75	6.0	6.25	6.5	6.75	7.0	7.25	7.5	7.75	8.0	8.25
3.0	0.663	0.7373	0.7685	0.8021	0.8061	0.8162	0.8556	0.8711	0.8761	0.8836	0.8856	0.893
4.0	0.6303	0.674	0.716	0.7578	0.8228	0.8346	0.8353	0.8649	0.8697	0.8852	0.8859	0.8887
5.0	0.5383	0.6118	0.6527	0.7044	0.7441	0.8104	0.8565	0.882	0.8453	0.8643	0.8709	0.8752

PURITY_FEDERATED

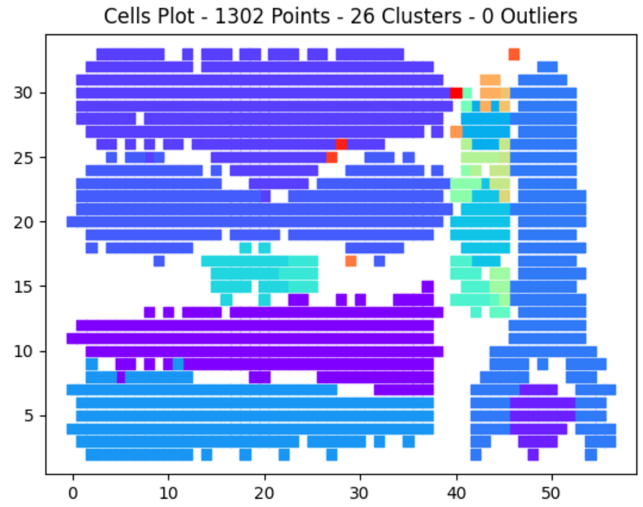


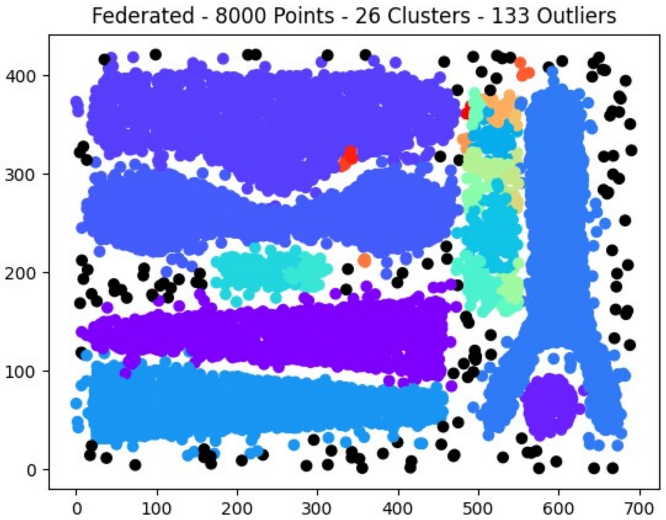
MinPts\L	11.0	11.5	12.0	12.5	13.0	13.5	14.0	14.5	15.0	15.5	16.0	16.5
3.0	0.8193	0.6341	0.6332	0.6086	0.5878	0.5853	0.391	0.3893	0.3872	0.3877	0.2027	0.385
4.0	0.8467	0.93	0.9742	0.657	0.653	0.6238	0.6326	0.4118	0.3887	0.4065	0.3851	0.3861
5.0	0.9706	0.971	0.9681	0.9727	0.834	0.815	0.6482	0.6093	0.3912	0.5818	0.3863	0.406

AMI_FEDERATED



MinPts\L	11.0	11.5	12.0	12.5	13.0	13.5	14.0	14.5	15.0	15.5	16.0	16.5
3.0	0.8484	0.7716	0.771	0.7409	0.733	0.7277	0.4929	0.4979	0.4872	0.494	0.025	0.4825
4.0	0.8057	0.8485	0.9148	0.7759	0.7807	0.7524	0.7537	0.5426	0.4891	0.5341	0.4811	0.4843
5.0	0.7458	0.8081	0.827	0.8453	0.8172	0.7989	0.7643	0.7331	0.4901	0.7036	0.4829	0.5319


Cells Plot - 1302 Points - 26 Clusters - 0 Outliers


Federated - 8000 Points - 26 Clusters - 133 Outliers

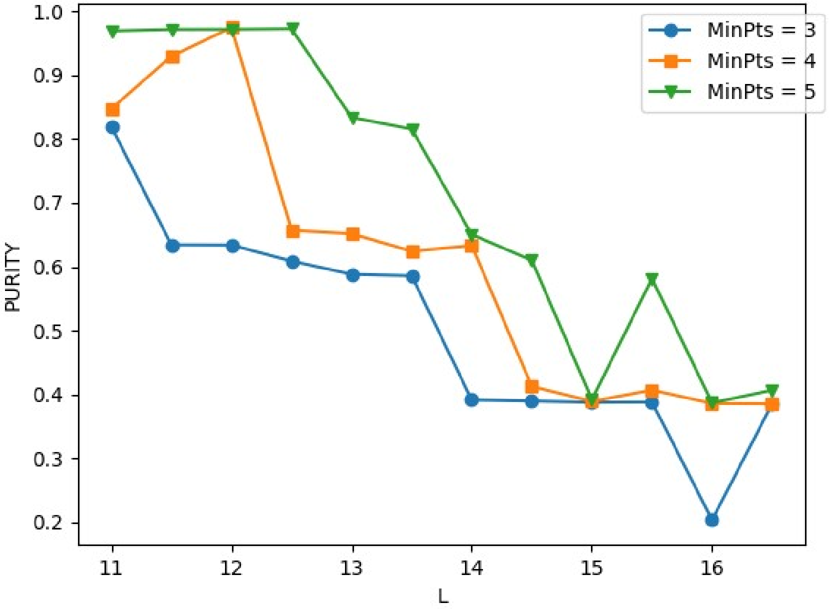
L = 12   MinPts = 4

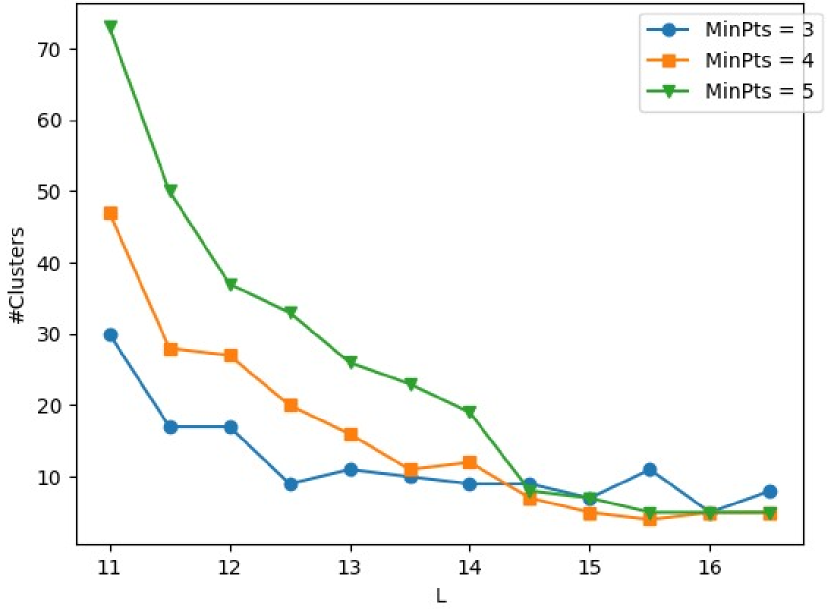
PURITY: 0.9748
ARI: 0.9536
AMI: 0.9148
PREC-BCUBED:
0.9504
REC-BCUBED: 0.9148

PURITY

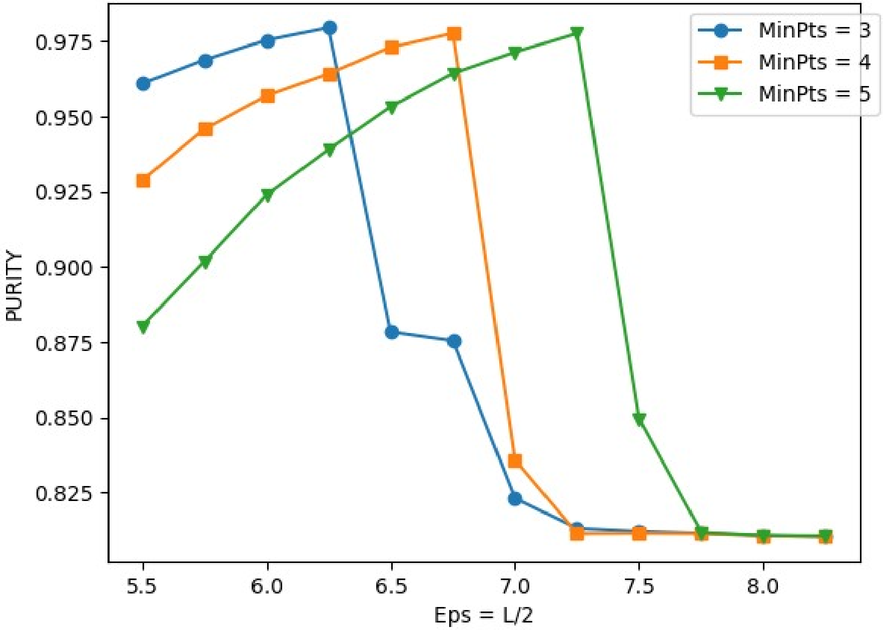
# CLUSTERS

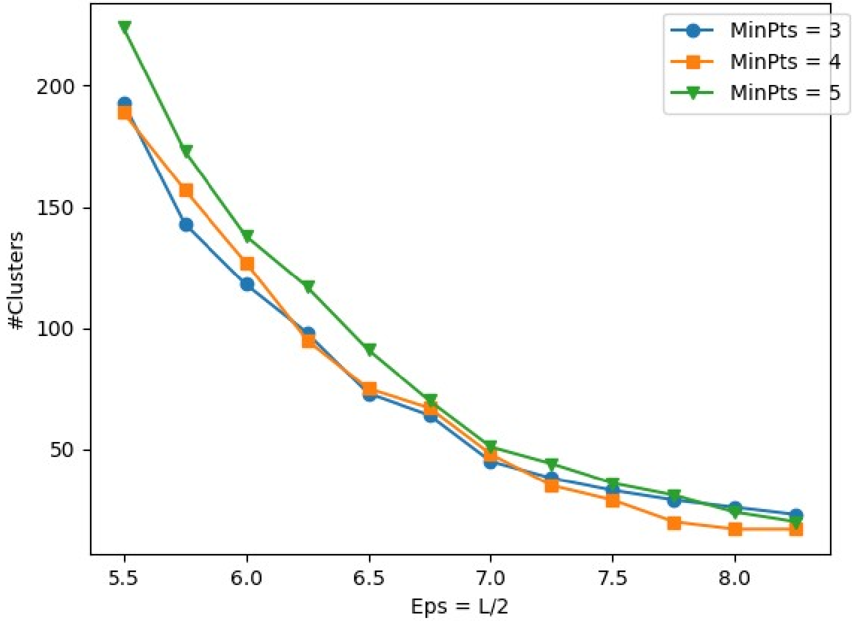
FEDERATED

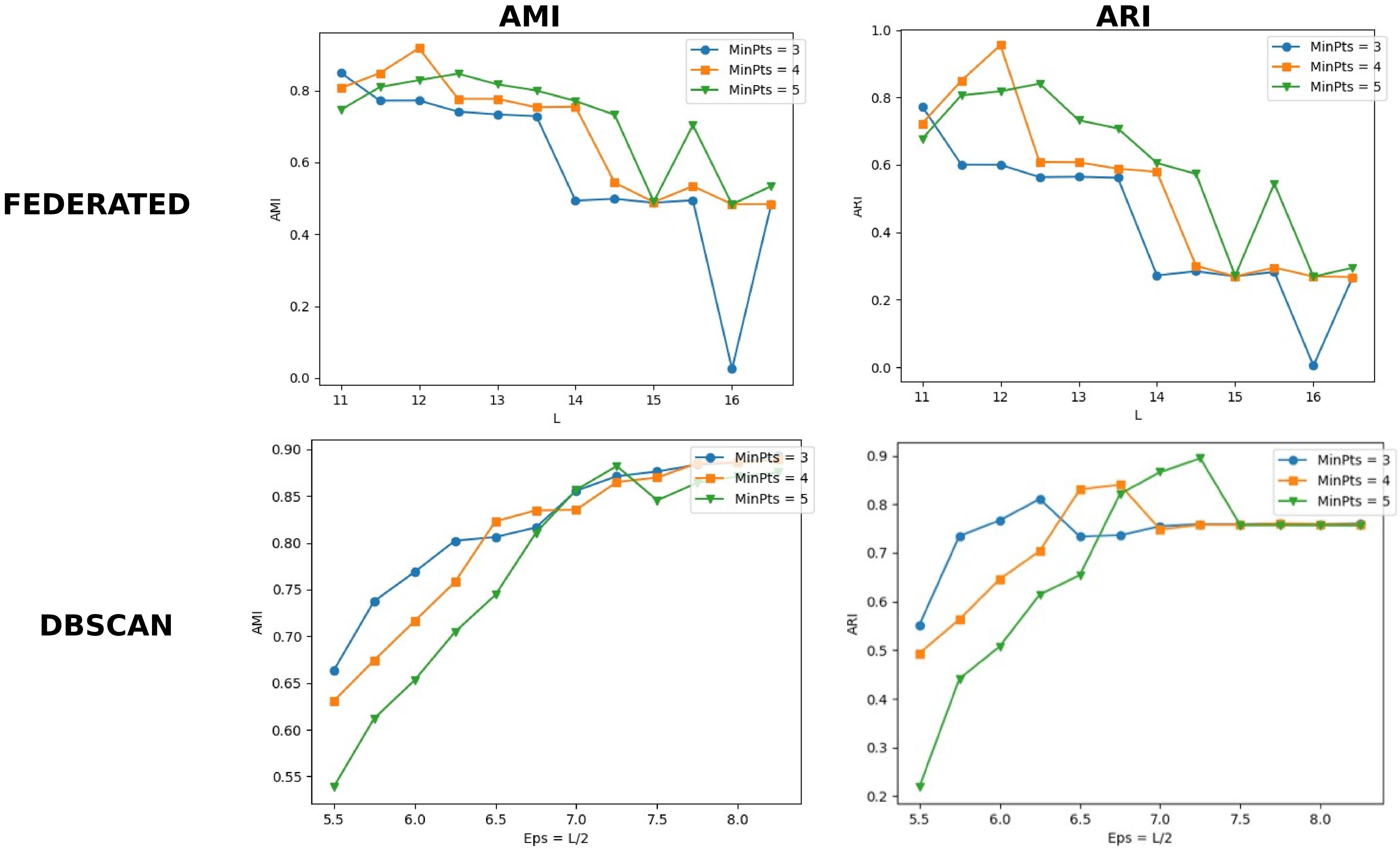




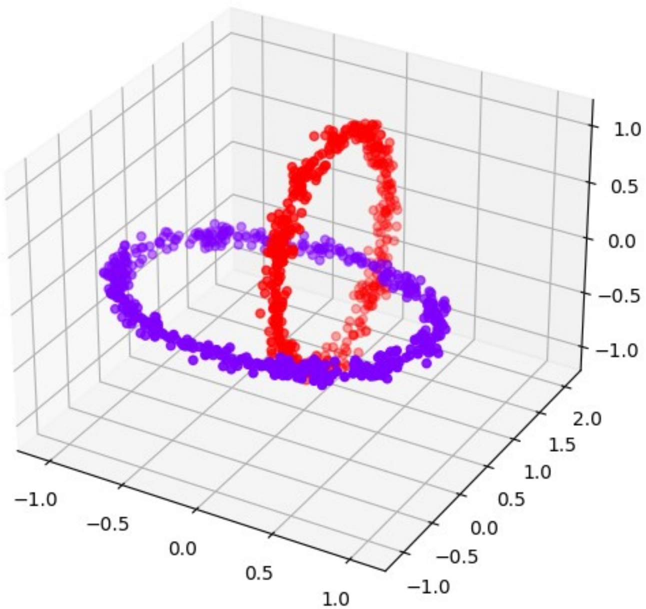
DBSCAN

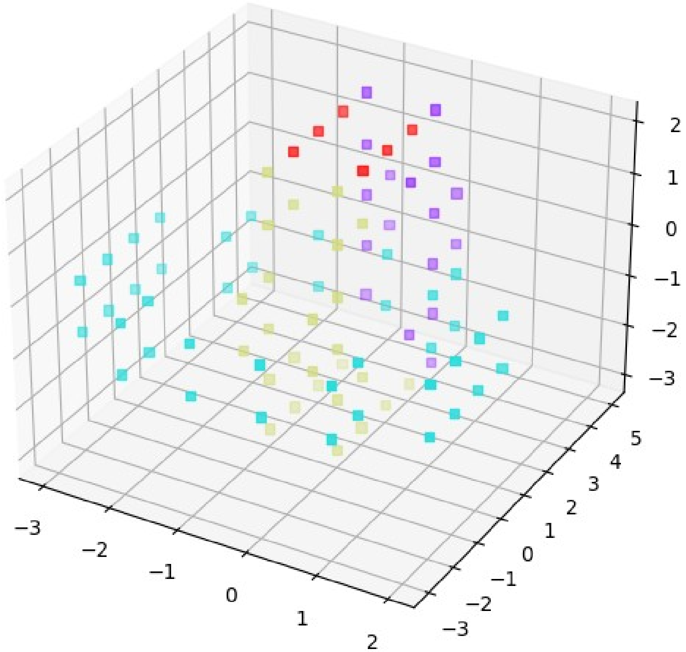


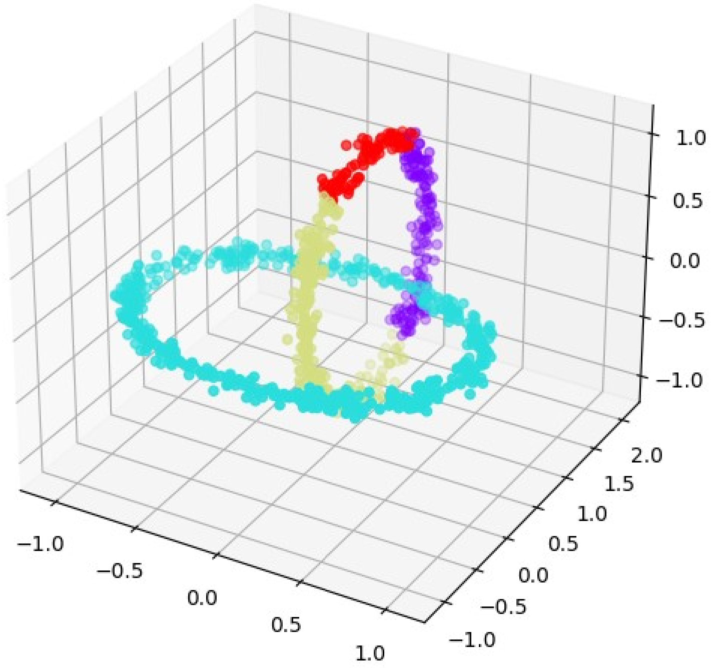



AMI
ARI
FEDERATED
DBSCAN
MinPts = 3
MinPts = 4
MinPts = 5
AMI
L
11
12
13
14
15
16
0.0
0.2
0.4
0.6
0.8
ARI
1.0
Eps = L/2
5.5
6.0
6.5
7.0
7.5
8.0
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

RESULTS CHAINLINK


1.0
0.5
0.0
−0.5
−1.0
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0


2
1
0
−1
−2
−3
5
4
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2


1.0
0.5
0.0
−0.5
−1.0
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0