MACHINE LEARNING

ASSIGNMENT 3: CLUSTERING

SARTHAK CHAKRABORTY (16CS30044)

Part 1 (Hierarchical Clustering)

On the given dataset of the accepted papers at the AAAI conference, we performed Hierarchical Clustering, where the clusters are formed according to the most similar topics on which the papers have been written. The output was shown for the clustering that detected **9 clusters**. Two types of hierarchical clustering were performed, where the clusters were formed according to MAX similarity (**Single Linkage**) and MIN similarity (**Complete Linkage**). The similarity was based on the Jaccard Coefficient of two documents.

$$JC_{HS} = JC_{SH} = \frac{|H \cap S|}{|H \cup S|}$$

a) SINGLE LINKAGE

The clusters were formed by maximizing the **MAX similarity** among the clusters, i.e., the ‘distance’ between the two farthest points of clusters.

The final learned clusters with the number of points in them are:

Cluster 0: [52]
Number of Points: 1

Cluster 1: [75]
Number of Points: 1

Cluster 2: [77]
Number of Points: 1

Cluster 3: [107]
Number of Points: 1

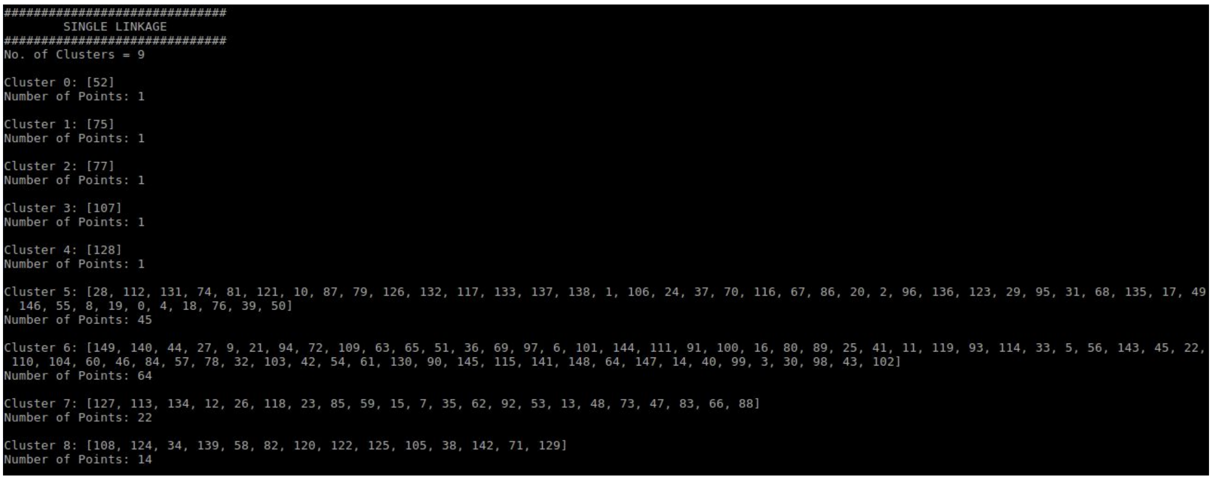
Cluster 4: [128]
Number of Points: 1

Cluster 5: [28, 112, 131, 74, 81, 121, 10, 87, 79, 126, 132, 117, 133, 137, 138, 1, 106, 24, 37, 70, 116, 67, 86, 20, 2, 96, 136, 123, 29, 95, 31, 68, 135, 17, 49, 146, 55, 8, 19, 0, 4, 18, 76, 39, 50]
Number of Points: 45

Cluster 6: [149, 140, 44, 27, 9, 21, 94, 72, 109, 63, 65, 51, 36, 69, 97, 6, 101, 144, 111, 91, 100, 16, 80, 89, 25, 41, 11, 119, 93, 114, 33, 5, 56, 143, 45, 22, 110, 104, 60, 46, 84, 57, 78, 32, 103, 42, 54, 61, 130, 90, 145, 115, 141, 148, 64, 147, 14, 40, 99, 3, 30, 98, 43, 102]
Number of Points: 64

Cluster 7: [127, 113, 134, 12, 26, 118, 23, 85, 59, 15, 7, 35, 62, 92, 53, 13, 48, 73, 47, 83, 66, 88]
Number of Points: 22

Cluster 8: [108, 124, 34, 139, 58, 82, 120, 122, 125, 105, 38, 142, 71, 129]
Number of Points: 14



```
##############################
        SINGLE LINKAGE
##############################
No. of Clusters = 9

Cluster 0: [52]
Number of Points: 1

Cluster 1: [75]
Number of Points: 1

Cluster 2: [77]
Number of Points: 1

Cluster 3: [107]
Number of Points: 1

Cluster 4: [128]
Number of Points: 1

Cluster 5: [28, 112, 131, 74, 81, 121, 10, 87, 79, 126, 132, 117, 133, 137, 138, 1, 106, 24, 37, 70, 116, 67, 86, 20, 2, 96, 136, 123, 29, 95, 31, 68, 135, 17, 49
, 146, 55, 8, 19, 0, 4, 18, 76, 39, 50]
Number of Points: 45

Cluster 6: [149, 140, 44, 27, 9, 21, 94, 72, 109, 63, 65, 51, 36, 69, 97, 6, 101, 144, 111, 91, 100, 16, 80, 89, 25, 41, 11, 119, 93, 114, 33, 5, 56, 143, 45, 22,
 110, 104, 60, 46, 84, 57, 78, 32, 103, 42, 54, 61, 130, 90, 145, 115, 141, 148, 64, 147, 14, 40, 99, 3, 30, 98, 43, 102]
Number of Points: 64

Cluster 7: [127, 113, 134, 12, 26, 118, 23, 85, 59, 15, 7, 35, 62, 92, 53, 13, 48, 73, 47, 83, 66, 88]
Number of Points: 22

Cluster 8: [108, 124, 34, 139, 58, 82, 120, 122, 125, 105, 38, 142, 71, 129]
Number of Points: 14
```

[**Note: Here the numbers within each cluster refer to the document number in the given dataset (0 indexing)**]

b) COMPLETE LINKAGE

The clusters were formed by maximizing the **MIN similarity** among the clusters, i.e., the ‘distance’ between the two nearest points of clusters.

The final learned clusters with the number of points in them are:

Cluster 0: [57, 78, 62, 92, 130, 90, 145, 46, 84, 60, 104]
Number of Points: 11

Cluster 1: [11, 94, 32, 143, 59, 15, 7, 35, 97, 6, 101]
Number of Points: 11

Cluster 2: [111, 89, 25, 41, 121, 24, 37, 144, 91, 100, 70, 116, 39, 50]
Number of Points: 14

Cluster 3: [9, 44, 10, 140, 27, 0, 4, 18, 76, 53, 47, 83]
Number of Points: 12

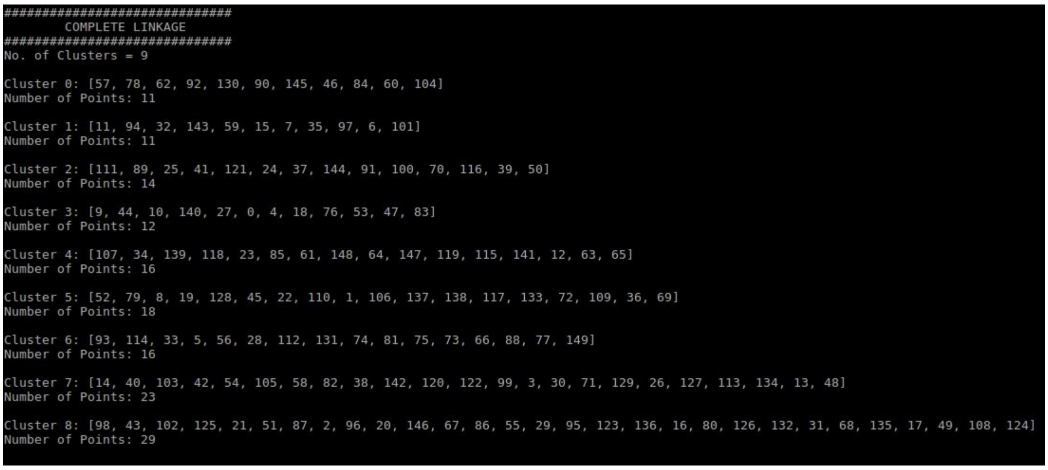
Cluster 4: [107, 34, 139, 118, 23, 85, 61, 148, 64, 147, 119, 115, 141, 12, 63, 65]
Number of Points: 16

Cluster 5:[52, 79, 8, 19, 128, 45, 22, 110, 1, 106, 137, 138, 117, 133, 72, 109, 36, 69]
Number of Points: 18

Cluster 6: [93, 114, 33, 5, 56, 28, 112, 131, 74, 81, 75, 73, 66, 88, 77, 149]
Number of Points: 16

Cluster 7: [14, 40, 103, 42, 54, 105, 58, 82, 38, 142, 120, 122, 99, 3, 30, 71, 129, 26, 127, 113, 134, 13, 48]
Number of Points: 23

Cluster 8: [98, 43, 102, 125, 21, 51, 87, 2, 96, 20, 146, 67, 86, 55, 29, 95, 123, 136, 16, 80, 126, 132, 31, 68, 135, 17, 49, 108, 124]
Number of Points: 29



```
##############################
        COMPLETE LINKAGE
##############################
No. of Clusters = 9

Cluster 0: [57, 78, 62, 92, 130, 90, 145, 46, 84, 60, 104]
Number of Points: 11

Cluster 1: [11, 94, 32, 143, 59, 15, 7, 35, 97, 6, 101]
Number of Points: 11

Cluster 2: [111, 89, 25, 41, 121, 24, 37, 144, 91, 100, 70, 116, 39, 50]
Number of Points: 14

Cluster 3: [9, 44, 10, 140, 27, 0, 4, 18, 76, 53, 47, 83]
Number of Points: 12

Cluster 4: [107, 34, 139, 118, 23, 85, 61, 148, 64, 147, 119, 115, 141, 12, 63, 65]
Number of Points: 16

Cluster 5: [52, 79, 8, 19, 128, 45, 22, 110, 1, 106, 137, 138, 117, 133, 72, 109, 36, 69]
Number of Points: 18

Cluster 6: [93, 114, 33, 5, 56, 28, 112, 131, 74, 81, 75, 73, 66, 88, 77, 149]
Number of Points: 16

Cluster 7: [14, 40, 103, 42, 54, 105, 58, 82, 38, 142, 120, 122, 99, 3, 30, 71, 129, 26, 127, 113, 134, 13, 48]
Number of Points: 23

Cluster 8: [98, 43, 102, 125, 21, 51, 87, 2, 96, 20, 146, 67, 86, 55, 29, 95, 123, 136, 16, 80, 126, 132, 31, 68, 135, 17, 49, 108, 124]
Number of Points: 29
```

[Note: Here the numbers within each cluster refer to the document number in the given dataset (0 indexing)]

Part 2 (Girvan Newman Clustering)

On the given dataset of the accepted papers at the AAAI conference, we performed Girvan Newman Graph Clustering, where the clusters are formed according to the most similar topics on which the papers have been written. The output was shown for the clustering that detected **9 clusters**.

Initial input graph G is built where nodes are corresponding to the articles and edges between two nodes is added depending on their similarity (Jaccard Coefficient) exceeding some use- defined threshold. The **Threshold** that I chose is **0.18**, which is used for the results shown below.

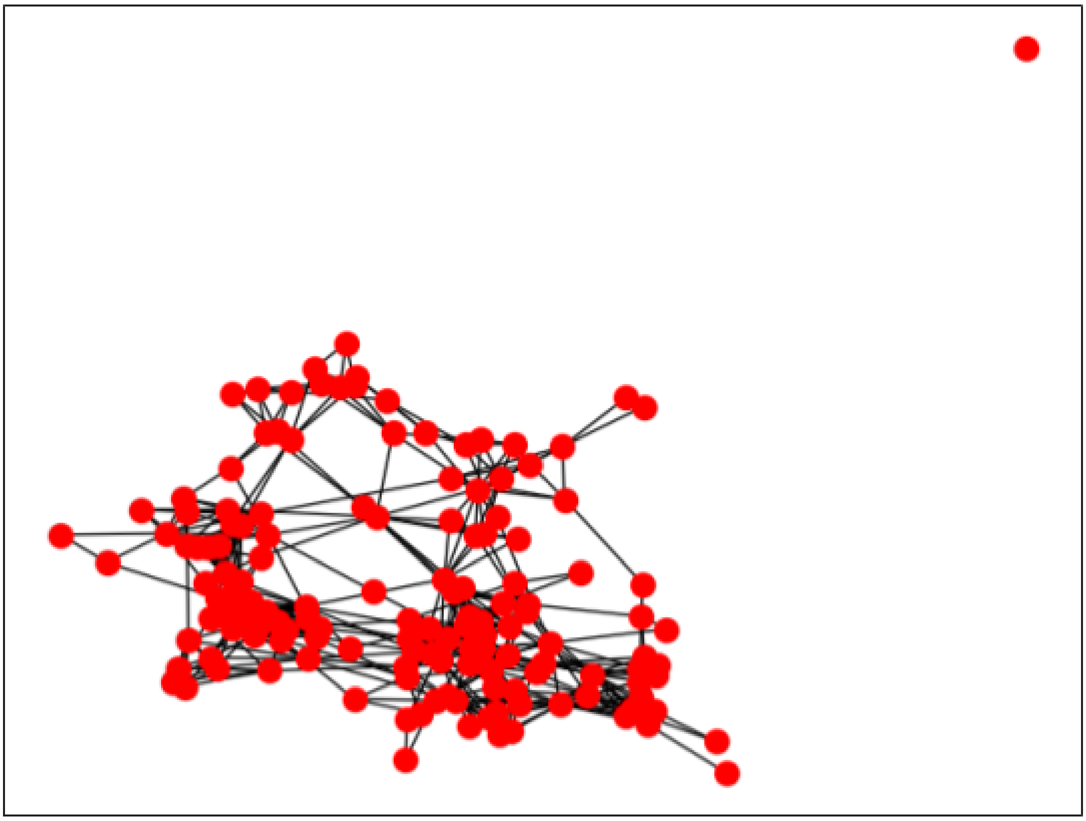


Fig. 1: Input Graph G

The final learned clusters and the number of nodes in them are:

Cluster 0: [0, 1, 2, 4, 133, 135, 8, 137, 138, 17, 146, 19, 20, 132, 27, 29, 31, 49, 136, 50, 52, 55, 67, 68, 70, 76, 79, 86, 87, 95, 96, 98, 106, 39, 18, 116, 117, 123, 124, 126]
Number of Points: 40

Cluster 1: [129, 66, 99, 134, 71, 73, 75, 13, 47, 113, 3, 53, 88, 83, 30, 127]
Number of Points: 16

Cluster 2: [130, 5, 11, 141, 14, 145, 147, 148, 149, 22, 40, 42, 45, 46, 51, 54, 56, 57, 60, 61, 64, 78, 84, 90, 93, 103, 104, 110, 114, 115, 119, 125, 21]
Number of Points: 33

Cluster 3: [97, 101, 6, 33, 10, 140, 24, 121, 37]
Number of Points: 9

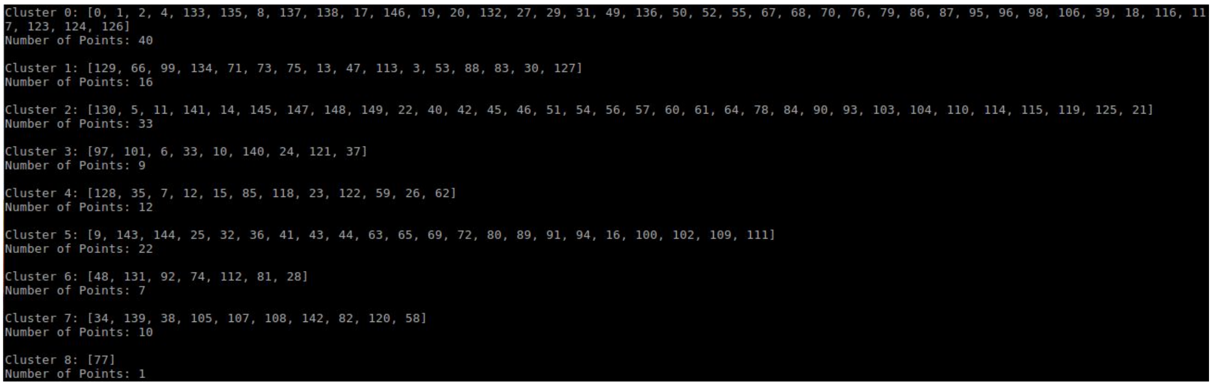
Cluster 4: [128, 35, 7, 12, 15, 85, 118, 23, 122, 59, 26, 62]
Number of Points: 12

Cluster 5: [9, 143, 144, 25, 32, 36, 41, 43, 44, 63, 65, 69, 72, 80, 89, 91, 94, 16, 100, 102, 109, 111]
Number of Points: 22

Cluster 6: [48, 131, 92, 74, 112, 81, 28]
Number of Points: 7

Cluster 7: [34, 139, 38, 105, 107, 108, 142, 82, 120, 58]
Number of Points: 10

Cluster 8: [77]
Number of Points: 1



```
Cluster 0: [0, 1, 2, 4, 133, 135, 8, 137, 138, 17, 146, 19, 20, 132, 27, 29, 31, 49, 136, 50, 52, 55, 67, 68, 70, 76, 79, 86, 87, 95, 96, 98, 106, 39, 18, 116, 11
7, 123, 124, 126]
Number of Points: 40

Cluster 1: [129, 66, 99, 134, 71, 73, 75, 13, 47, 113, 3, 53, 88, 83, 30, 127]
Number of Points: 16

Cluster 2: [130, 5, 11, 141, 14, 145, 147, 148, 149, 22, 40, 42, 45, 46, 51, 54, 56, 57, 60, 61, 64, 78, 84, 90, 93, 103, 104, 110, 114, 115, 119, 125, 21]
Number of Points: 33

Cluster 3: [97, 101, 6, 33, 10, 140, 24, 121, 37]
Number of Points: 9

Cluster 4: [128, 35, 7, 12, 15, 85, 118, 23, 122, 59, 26, 62]
Number of Points: 12

Cluster 5: [9, 143, 144, 25, 32, 36, 41, 43, 44, 63, 65, 69, 72, 80, 89, 91, 94, 16, 100, 102, 109, 111]
Number of Points: 22

Cluster 6: [48, 131, 92, 74, 112, 81, 28]
Number of Points: 7

Cluster 7: [34, 139, 38, 105, 107, 108, 142, 82, 120, 58]
Number of Points: 10

Cluster 8: [77]
Number of Points: 1
```

The output graph with 9 clusters is shown below.

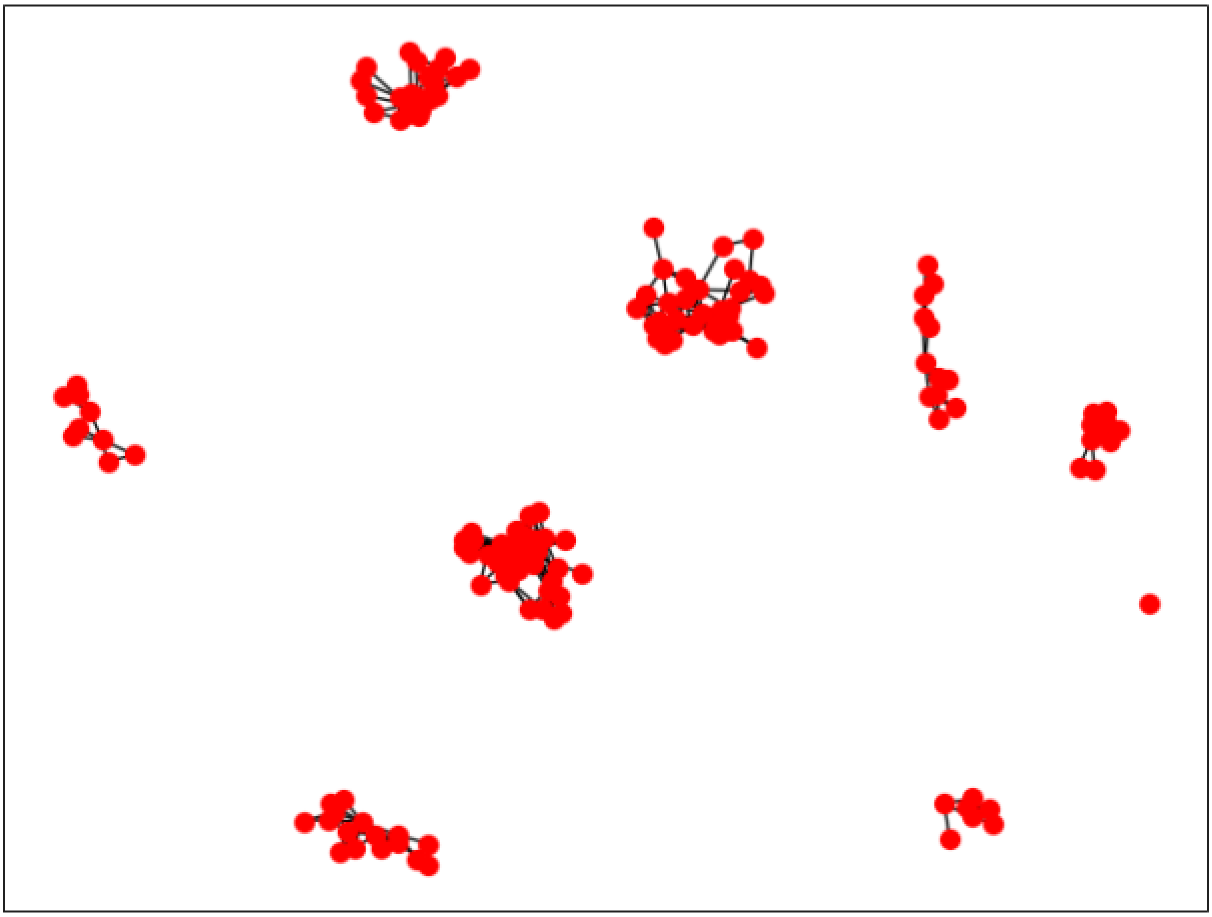


Fig. 2: Output Graph showing 9 clusters

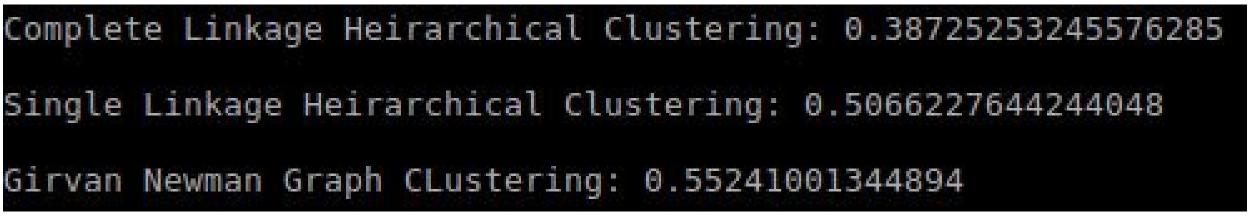
Part 3 (NMI Calculation)

For the clusters that we obtained in Part 1 and Part 2, we calculate NMI values. We take the gold-standard label from the dataset where the accepted papers are clustered according to the High-Level Domains. There are a total of 9 High-Level Domains in the dataset.

$$NMI(Y,C) = \frac{2 \times I(Y;C)}{[H(Y) + H(C)]}$$

where, Y = Class Labels
C = Cluster Labels
H(.) = Entropy
I(Y;C) = Mutual Information between Y and C

The NMI values for the above-obtained clusters are:

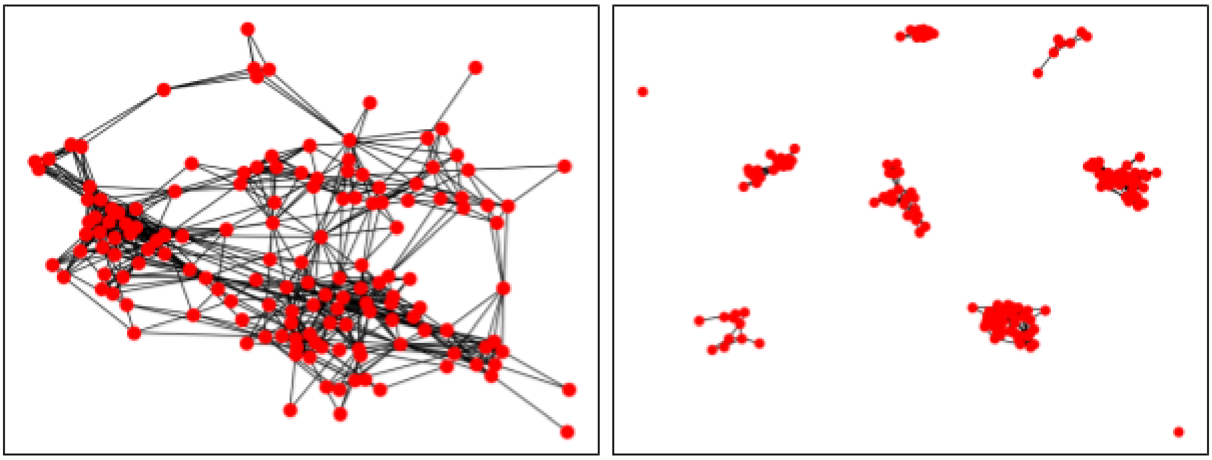


```
Complete Linkage Heirarchical Clustering: 0.38725253245576285

Single Linkage Heirarchical Clustering: 0.5066227644244048

Girvan Newman Graph CLustering: 0.55241001344894
```

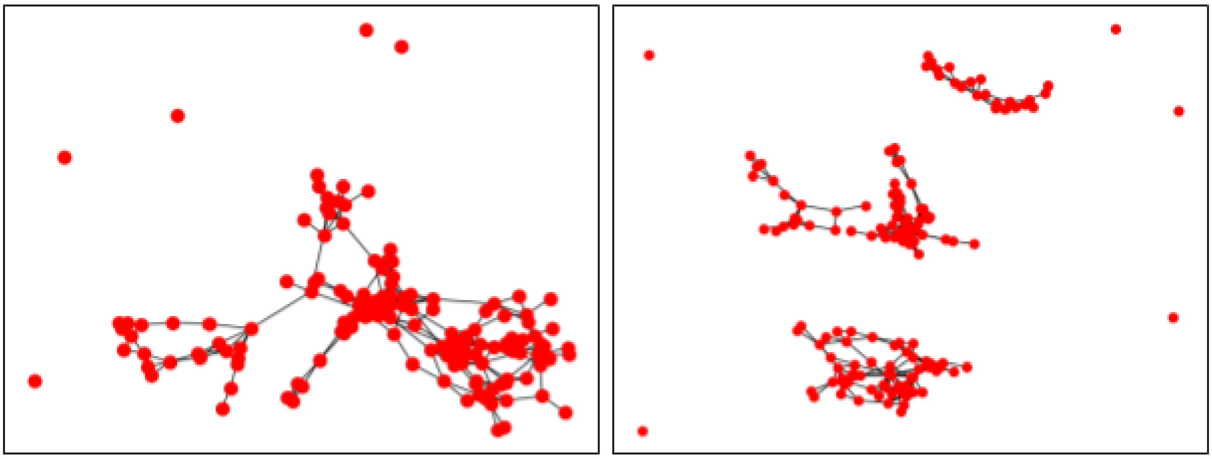
Thus, as we see from the NMI value that Girvan Newman clustering algorithm obtains a higher NMI value than the other two clustering algorithms. Thus Girvan Newman Algorithm performs better in this particular dataset when we have to obtain 9 clusters.

Below, we show the input and output graphs in Part 2 for various thresholds and their NMI values:

NMI values for Girvan Newman clustering algorithm for different threshold are as follows:



a) Threshold = 0.15, NMI = 0.62746



b) Threshold = 0.21, NMI = 0.51298

Thus, NMI value depends on the threshold as well since the initial edges will be different and hence the clusters formed will be different.