Data Analytics

Assignment -3

Arman Gupta (19220) - Mtech CSA

Implementation Details-

Assuming the graph G(V,E) is given whose adjacency matrix is given by A and Degree Matrix is given by D.

1 . Spectral Clustering Algorithm

Algorithm for 1 iteration :

I. Calculate the laplacian Matrix L by subtracting Adjacency Matrix from Degree Matrix.
II. Find the eigen value and eigen vectors of L and sort the eigen values and eigen vectors in ascending order based on eigen value.
III. Select eigenvector that corresponds to the first non-zero eigen value as the fielder vector, F where F $\in$ $R^{|V| * 1}$.
IV. Separate the nodes into two clusters where one cluster will contain the nodes {i} for which F[i] <= 0 while other cluster will contain the nodes {j} for which F[j] >0 where 0<= i,j < |V|.

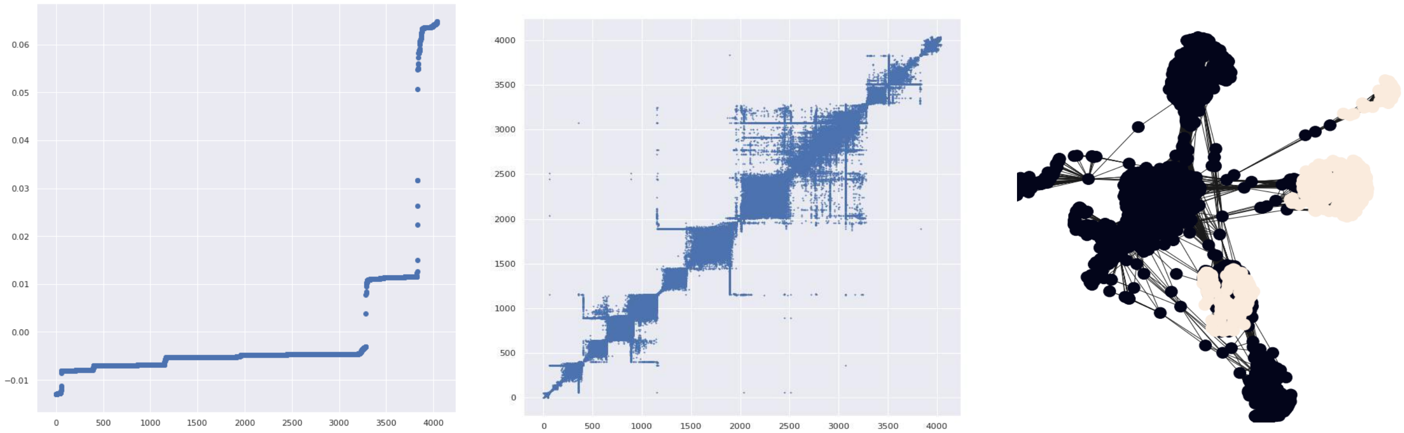
In order to create more than 2 clusters, one of the approach we can use is to keep dividing each cluster of nodes into two more clusters till some condition satisfies. One naive approach is to use the size of the cluster as the stopping condition. Other approach could be calculating the modularity after and before forming clusters and if the after-modularity is greater than before-modularity, we are forming clusters.

2. Louvain Algorithm

I have implemented phase 1 of the algorithm using the slide 11 from lecture 3. In order to know where to stop the algorithm, I am using the patience variable. If the total modularity of the Graph with the assigned cluster does not increase for fixed patience number of the steps , the algorithm stops and give the communities at best modularity.

Question 1

Dataset : Facebook



(a) (b) (c)

Fig 1. (a) Represent the Fielder vector where x axis is node number sorted using Fiedler vector values. (b) Represent the Associated Adjacency matrix after running 1 iteration of Spectral Clustering. (c) represents the graph partition after 1 iteration with 3285 in cluster 1 and 754 in cluster 2.

Dataset : bitcoin

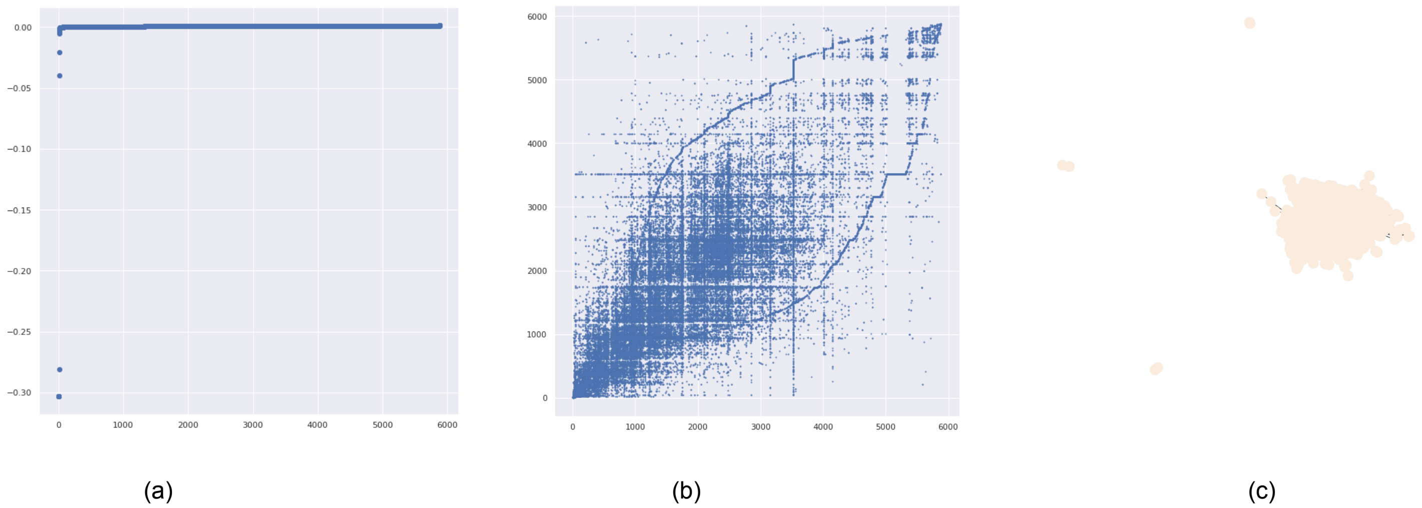


Fig 2. (a) Represent the Fielder vector where x axis is node number sorted using Fiedler vector values. (b) Represent the Associated Adjacency matrix after running 1 iteration of Spectral Clustering. (c) represents the graph partition after 1 iteration.

Question2

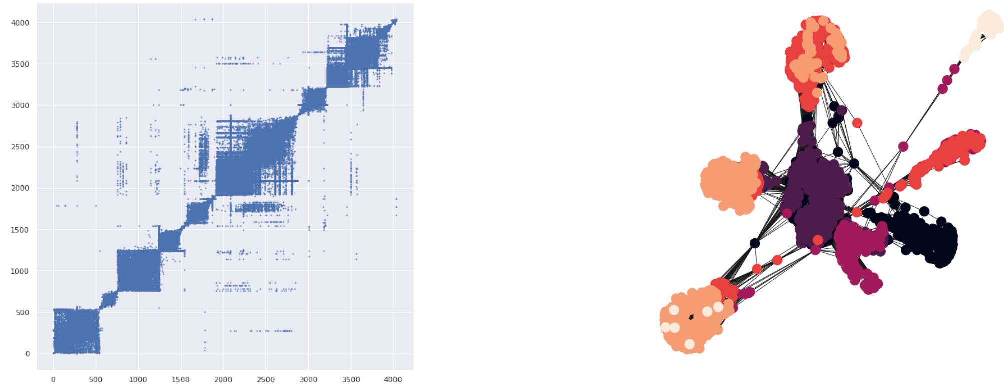
As I have mentioned in Spectral Clustering Implementation ,I am calculating the modularity after and before forming clusters and if the after-modularity is greater than before-modularity, we are forming clusters. Using this stopping criterion , I have got 8 communities for Facebook dataset and 1 community for bitcoin dataset.
Though, we are getting few communities in facebook dataset but we are not getting more communities in bitcoin dataset and it makes sense because modularity suffers a resolution limit and, therefore, it is unable to detect small communities which is quite obvious from above graph partition and even the sorted feidler vector plot shows the small communities.
[Ref : https://en.wikipedia.org/wiki/Modularity_(networks)]

Another Methods that I have tried:

- I have used the size of the cluster as the hyperparameter for the stopping condition ie. If the size of some cluster is less than 200 , my algorithm stops finding more clusters from that cluster. Using the above method, I have got 7 communities for Facebook dataset and 27 clusters for bitcoin dataset.
- I have tried another method where I am finding the variance of the feidler vector corresponding to somecluster and if the variance is below certain threshold, I am not forming the sub clusters. But using this method,I was getting very low clusters and this cluster formation varies a lot with slight change in the threshold so I avoid using this method.
- I have also tried forming clusters using Kmean instead of using sign and there I was avoid forming subclusters if the mean of the two clusters are quite close to each other. I didn't get the enough clusters using this method as well.
- I have also considered using the conductance , ratio cut and normalised cut to come up with the stopping criteria but the main issue that I have face is that these metrics are defined only when we have formed the clusters. We cannot calculate these metrics for a single set of nodes. Hence, we don't have a way to find these metrics before forming cluster.

Question3

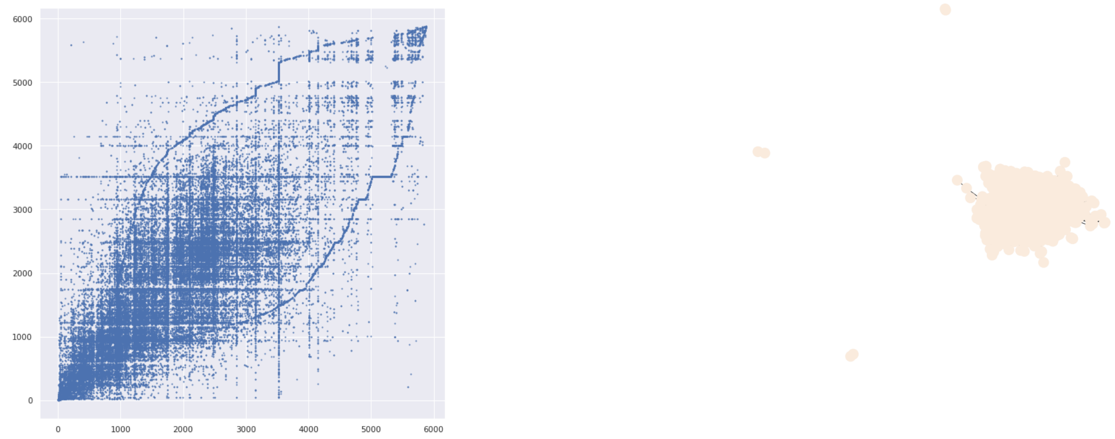
Dataset : Facebook



(a) (b)

Fig 3 (a) Associated adjacency matrix sorted by associated sorted sub graph Fiedler vectors. (b) Graphs representing the partitions.

Dataset : Bitcoin



(a) (b)

Fig 4 (a) Associated adjacency matrix sorted by associated sorted sub graph Fiedler vectors. (b) Graphs representing the partitions.

Note in Bitcoin dataset, I am getting only 1 cluster using the modularity as the stopping criterion so is indicated here also which is why Fig 2 and Fig 4 are indicating the same thing. For seeing the results with other mentioned stopping criterion, see the last section.

A Very Important Observation About the NetworkX Graph Plotting Library :
The graph plotting using networkX does not seem to be good for the larger graphs. I have tried plotting the graph with the disjoint clusters but those graphs overlapped too somehow using this library. In Fig 5 ,we can see the overlapping between disjoint clusters present in the graph. Hence we cannot completely trust the networkX library for seeing the graph partitions for large graph.

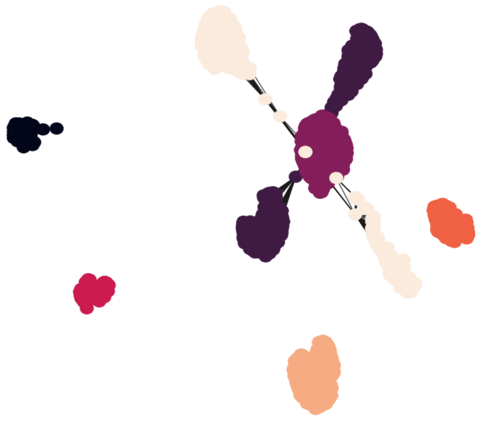


Fig 5 : Plotting the large graphs with disjoint clusters

Question 4
I have implemented the Louvain Algorithm as described above. Following are the plots for each dataset.

Dataset : Facebook
Running the phase1 of Louvain Algorithm on the Facebook Dataset, I have got 25 clusters at the maximum modularity which was 0.3598 . Fig 6 indicates the communities.

Dataset: Bitcoin
Running the phase1 of Louvain Algorithm on the Bitcoin Dataset, I have got 402 clusters at the maximum modularity of 0.0118 (patience = 20) . Fig 7 indicates the communities

PTO.

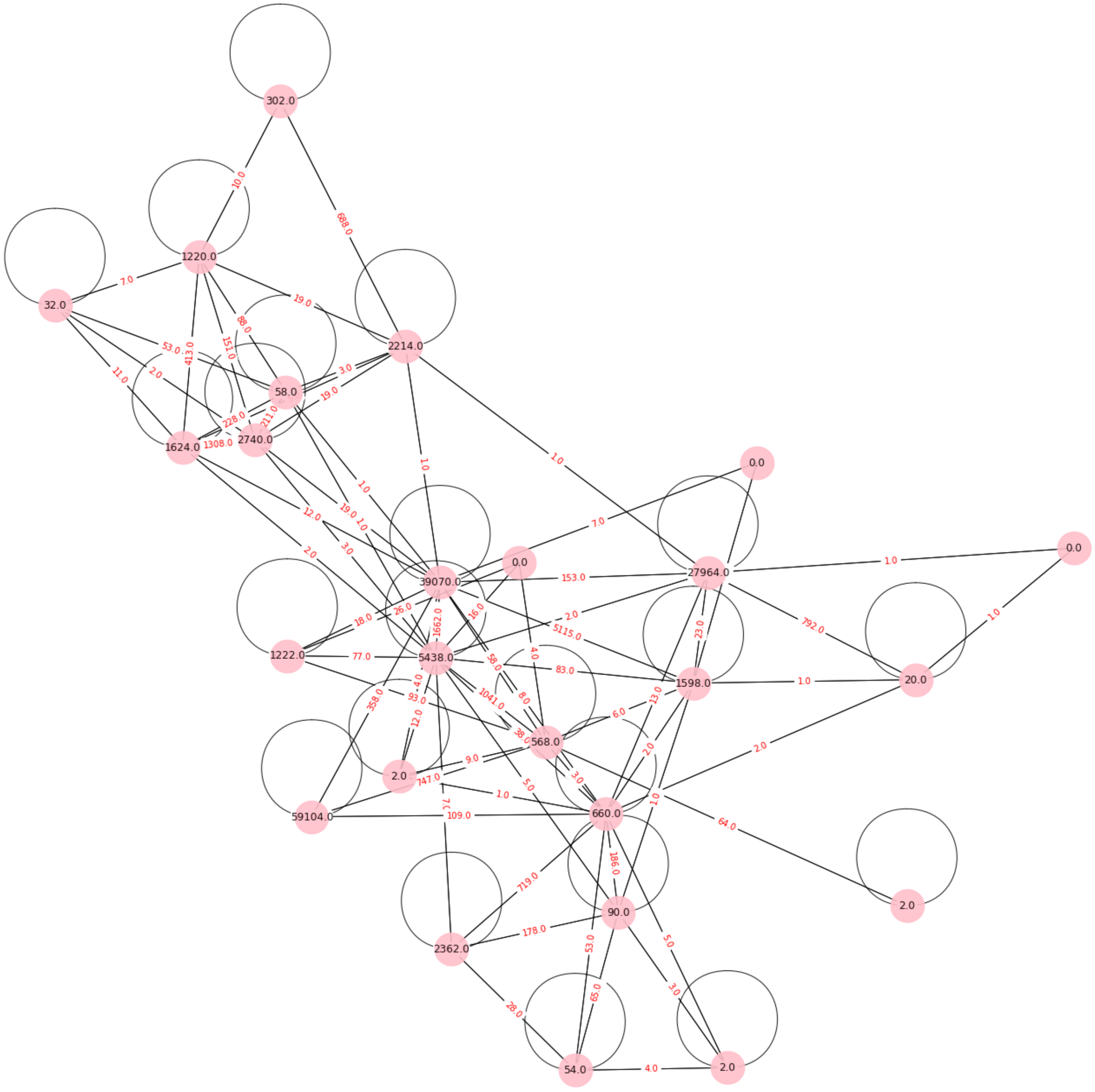


Fig 6. Facebook : Communities are shown in pink color and edge labels indicates the number of edges from one community to other community. Node labels indicates number of edges within the community.

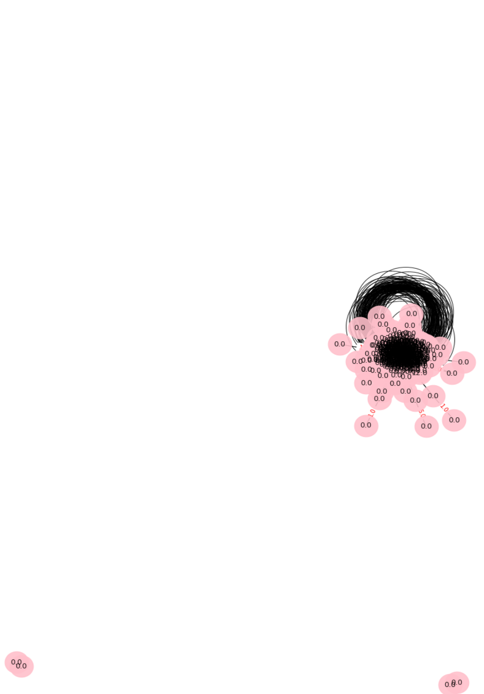


Fig 7. Bitcoin : Communities are shown in pink color and edge labels indicates the number of edges from one community to other community. Node labels indicates number of edges within the community.

Question 5

As I have mentioned earlier in Louvain Implementation section , in order to know where to stop the algorithm and pick the best decomposition of nodes into communities, I am using the patience variable. If the total modularity of the Graph with the assigned cluster does not increase for fixed patience number of the steps , the algorithm stops and give the communities at best modularity.

Question 6

Run Time for Both Algorithms on both dataset:

- Spectral Decomposition Runtime For Facebook : 0.61 Minutes
- Louvain Decomposition Runtime For Facebook : 26.59 Minutes
- Spectral Decomposition Runtime For Bitcoin : 0.62 Minutes
- Louvain Decomposition Runtime For Bitcoin : 22.86 Minutes

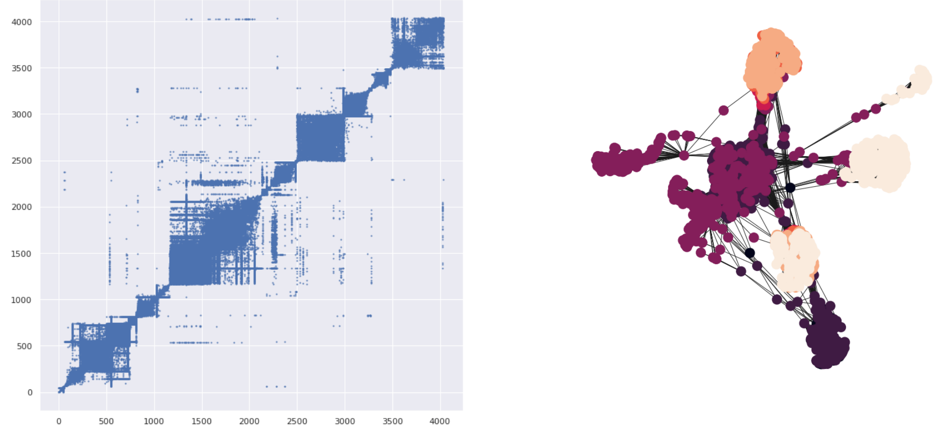
Question 7

Originally, Spectral Clustering algorithm is not designed for automatically detecting the number of clusters in the graph which is the same issue we have faced while implementing the question 2. We came up with various stopping criteria to stop spectral clustering at some point. While in case of Louvain Algorithm, we are able to determined the number of communities automatically. Moreover, using the modularity based stopping criterion in case of Spectral

Clustering, I am able to find only 6 clusters in case of facebook dataset and 1 cluster in bitcoin dataset and not able to find more subclusters within those huge clusters and moreover, we are not able to detect the smaller disjoint communities (reason i have mentioned earlier) in case of spectral clustering with the mentioned stopping criterion. While in case of Louvain, not only we are able to detect the smaller disjoint communities as we can see in bitcoin dataset but also able to detect the subcommunities within the huge clusters and that's why we are getting 25 communities in facebook dataset and 402 communities in bitcoin dataset. Thus, in my opinion phase 1 of Louvain Algorithm is performing better than the spectral clustering with modularity based stopping criterion.

Result using Cluster Size Based Stopping Criterion for Automating the determination of the right set of communities using the spectral decomposition method-

Dataset- Facebook
Using the mentioned stopping criterion , I have got 7 clusters for facebook dataset

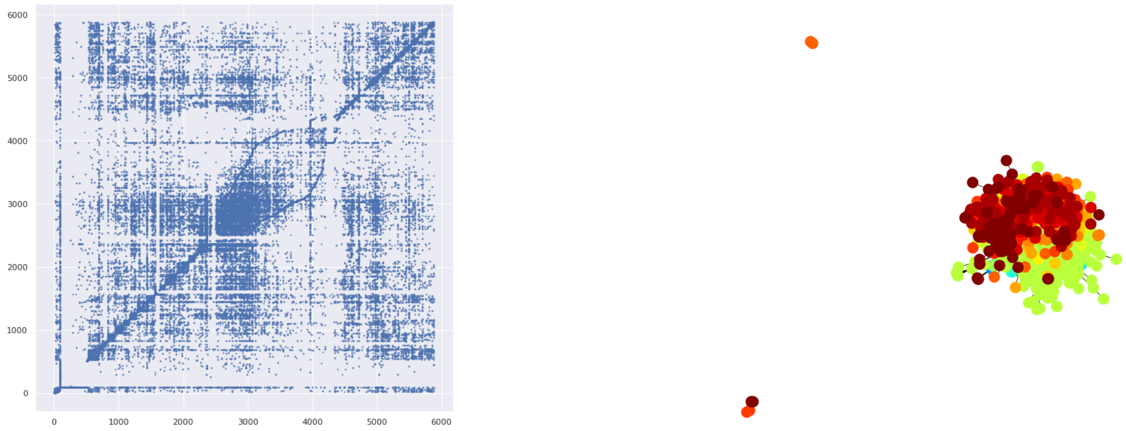


(a)

(b)

Fig 8 .(a) Associated adjacency matrix sorted by associated sorted sub graph Fiedler vectors using cluster size based stopping criterion. (b) Graphs representing the clusters.

Dataset - Bitcoin
Using the mentioned stopping criterion , I have got 27 clusters for bitcoin dataset which I tried representing in Fig 9 (b) but the clusters are not represented very well using NetworkX graph plotting library.



(a) (b)

Fig 9 .(a) Associated adjacency matrix sorted by associated sorted sub graph Fiedler vectors using cluster size based stopping criterion. (b) Graphs representing the clusters.

Note : I have used decimal point precision tolerance of 5e-14 in determining the non zero eigen value (i.e every eigen value below 5e-14 is considered 0) because of the fact that I was getting -1.4e-14 eigen value for the facebook dataset where I should be getting 0.