

Comparative Analysis of two Traditional Graph Clustering algorithms

Indranil Bhattacharya

indranil.bhattacharya@csa.iisc.ernet.in

November 28, 2015



Motivation

Problem Definition & Challenges

- Problem Definition

- Challenges

Traditional Graph Clustering Algorithms

- Spectral Clustering using Laplacian

- Hierarchical Clustering using Agglomeration

Experimental Results

- Results on Graph 1

- Results on Graph 2

- Zackary's Karate Club Graph

- Planted ℓ -partition model Graph

Conclusion



Motivation

Problem Definition & Challenges

- Problem Definition

- Challenges

Traditional Graph Clustering Algorithms

- Spectral Clustering using Laplacian

- Hierarchical Clustering using Agglomeration

Experimental Results

- Results on Graph 1

- Results on Graph 2

- Zackary's Karate Club Graph

- Planted ℓ -partition model Graph

Conclusion

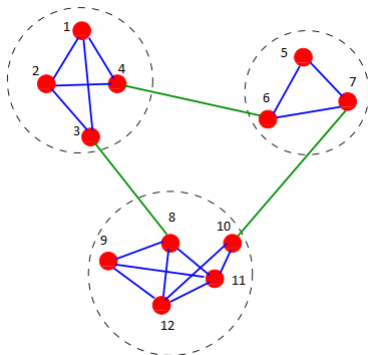


Figure: A Simple Graph with 3 communities

Motivation II

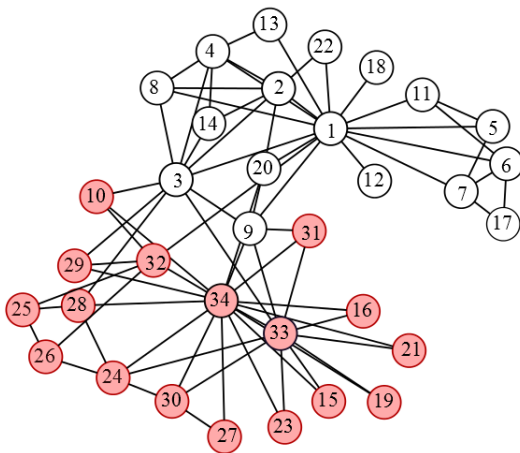


Figure: Zackary's Karate Club after split into 2 communities

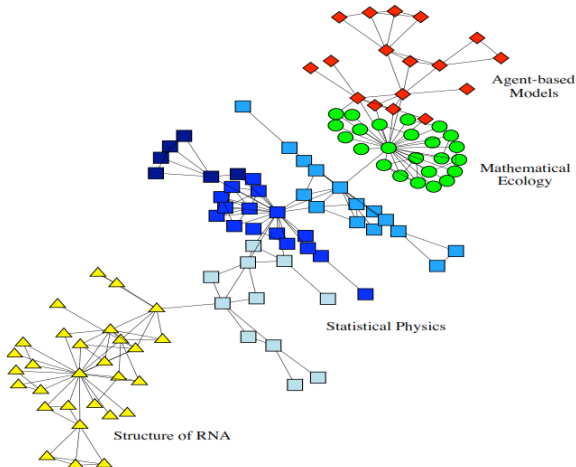


Figure: Collaboration of 118 scientists at Santa Fe Institute



Motivation

Problem Definition & Challenges

- Problem Definition

- Challenges

Traditional Graph Clustering Algorithms

- Spectral Clustering using Laplacian

- Hierarchical Clustering using Agglomeration

Experimental Results

- Results on Graph 1

- Results on Graph 2

- Zackary's Karate Club Graph

- Planted ℓ -partition model Graph

Conclusion

Problem Definition & Challenges

Problem Definition



- ▶ Given a Graph $G = (V, E)$ (V = Set of Nodes, E = Set of Edges), partition G into components (subset of nodes), such that each subset is strongly intra-connected with comparatively *fewer* edges across subsets.
- ▶ Each of these subsets is called a *Community*.
- ▶ The problem is easy if the graph is a disjoint collection of several connected components. (e.g. Use DFS/BFS to find each component)

Problem Definition & Challenges

Challenges



- ▶ The Graph G could be directed.
e.g World Wide Web graph
- ▶ Nodes could be part of multiple Communities, in which case the term “Graph Partition” is ill-defined.
e.g Word Association graph
- ▶ The Graph itself could be multi-partite.
e.g Southern Women Event Participation graph (SWEP)
- ▶ Incorporating both structural and non-structural information in the search of clusters, which will make clustering more consistent, is difficult.
- ▶ Domain specific clustering - identify the peculiar features of classes of graphs, which are bound to become crucial ingredients in the design of suitable algorithms, is not straightforward.



Motivation

Problem Definition & Challenges

Problem Definition

Challenges

Traditional Graph Clustering Algorithms

Spectral Clustering using Laplacian

Hierarchical Clustering using Agglomeration

Experimental Results

Results on Graph 1

Results on Graph 2

Zackary's Karate Club Graph

Planted ℓ -partition model Graph

Conclusion

Traditional Graph Clustering Algorithms

Spectral Clustering using Laplacian



Input : Adjacency Matrix A and the number of desired components k of the graph $G = (V, E)$

Algorithm :

- ▶ Compute the normalized Laplacian, $L_{norm} = I - D^{-1}A$,
 D = Degree Matrix of G , I = Identity Matrix
- ▶ Compute k smallest eigen values and corresponding eigen vectors of L_{norm}
- ▶ $X = [v_1 v_2 \dots v_k]$ be the set of k smallest eigen vectors,
 X is an $n * k$ matrix, $n = |V|$
- ▶ Each row of X corresponds to a vertex in G
- ▶ Cluster the n data points in \mathbb{R}^k using any standard “Data Clustering” algorithm (e.g k-means)

Output : Clusters in \mathbb{R}^k correspond to Communities in G



- Squared Euclidean distance

$$\sum_{i=1}^k \sum_{j \in \mathbb{C}_i} \|x_j - c_i\|_2^2$$

where c_i is the centroid of cluster \mathbb{C}_i

- Manhattan distance

$$\sum_{i=1}^k \sum_{j \in \mathbb{C}_i} \|x_j - c_i\|_1$$

- Cosine Similarity measure

$$\sum_{i=1}^k \sum_{j \in \mathbb{C}_i} 1 - \frac{x_j \cdot c_i}{\sqrt{(x_j \cdot x_j)(c_i \cdot c_i)}}$$

Traditional Graph Clustering Algorithms

Hierarchical Clustering using Agglomeration on the eigen matrix X



Input : Adjacency Matrix A and the number of desired components k of the graph $G = (V, E)$

Algorithm :

- ▶ Compute X as is done for the Spectral Clustering algorithm
- ▶ Consider each point in \mathbb{R}^k to be a single cluster
- ▶ Merge 2 clusters which have high “similarity”
- ▶ Repeat the previous step until all the nodes have been merged into a single giant cluster
- ▶ Prune the Dendrogram(cluster hierarchy tree) down from root level-wise depending on the number of clusters desired

Output : Clusters in \mathbb{R}^k correspond to Communities in G

Traditional Graph Clustering Algorithms

Standard Similarity Measures



- ▶ Single Linkage (a.k.a nearest neighbor), uses the smallest distance between objects in two clusters

$$d(r, s) = \min(\text{dist}(x_{ri}, x_{sj})), i \in \{1, 2, \dots, n_r\} j \in \{1, 2, \dots, n_s\}$$

- ▶ Complete Linkage (a.k.a farthest neighbor), uses the largest distance between objects in two clusters

$$d(r, s) = \max(\text{dist}(x_{ri}, x_{sj})), i \in \{1, 2, \dots, n_r\} j \in \{1, 2, \dots, n_s\}$$

- ▶ Average Linkage uses the average distance between all pairs of objects in any two clusters

n_r = number of objects in cluster r ,

x_{ri} = i^{th} object in cluster r , &

“dist” is any or all of the distance measures mentioned before.

Traditional Graph Clustering Algorithms

Hierarchical Clustering using Agglomeration on the Graph G



Input : Adjacency Matrix A and the number of desired components k of the graph $G = (V, E)$

Algorithm :

- ▶ Consider each node in G to be a single cluster
- ▶ Compute the distance between every pair of clusters using some standard “Structural Equivalence Measures defined on G ”
- ▶ Merge 2 clusters having high “similarity”
- ▶ Repeat the previous step until all the nodes have been merged into a single giant cluster
- ▶ Prune the Dendrogram(cluster hierarchy tree) down from root level-wise depending on the number of clusters desired

Output : Communities in G are computed directly without any embedding

Traditional Graph Clustering Algorithms

Standard Measures of Similarity defined on G itself



15

- ▶ Structural Equivalence based on adjacency relationships between vertices

$$d_{ij} = \sqrt{\sum_{k \neq i, j} (A_{ik} - A_{jk})^2}$$

where A is the adjacency matrix of G

- ▶ Structural Equivalence based on Neighborhood Overlap (a.k.a Jaccard Similarity Index)

$$d_{ij} = \frac{|N(i) \cap N(j)|}{|N(i) \cup N(j)|}$$

where $N(i)$ = set of vertices adjacent to vertex i in G

- ▶ Structural equivalence based on Pearson correlation between columns (or rows) of A

$$d_{ij} = \frac{\sum_k (A_{ik} - \mu_i)(A_{jk} - \mu_j)}{n\sigma_i\sigma_j}, \text{ where } \mu_i = \frac{\sum_j A_{ij}}{n}, \sigma_i = \sqrt{\frac{\sum_j (A_{ij} - \mu_i)^2}{n}}$$



Motivation

Problem Definition & Challenges

Problem Definition

Challenges

Traditional Graph Clustering Algorithms

Spectral Clustering using Laplacian

Hierarchical Clustering using Agglomeration

Experimental Results

Results on Graph 1

Results on Graph 2

Zackary's Karate Club Graph

Planted ℓ -partition model Graph

Conclusion

Results on Graph 1 I

Graph 1 : Figure 1 from Fortunato's Review Paper [1]

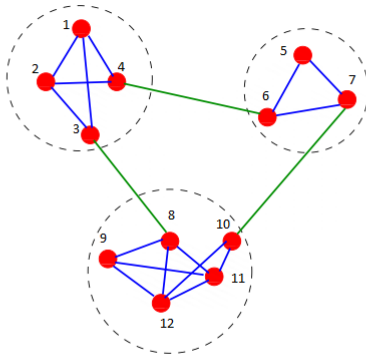


Figure 1: A simple graph with three communities, highlighted by the dashed circles.



Spectral Clustering Using Laplacian :

- ▶ All the 3 different distance measures viz. Squared Euclidean, Manhattan, Cosine Similarity based distance extracted the exact set of clusters.

Hierarchical Agglomerative Clustering on X :

- ▶ All the 9 possible combinations of different distance measures and different types of linkages viz. Single, Complete & Average linkage retrieved the exact set of clusters.

Hierarchical Agglomerative Clustering on A(G) :

- ▶ Using all the 3 Structural Equivalence based methods, only Single linkage criterion failed to achieve the desired set of clusters.

Results on Graph 2 I

Graph 2 : Figure 9 from Fortunato's Review Paper [1]

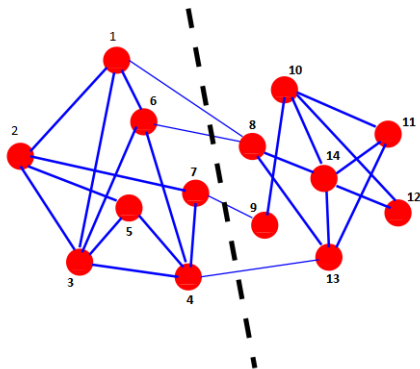


FIG. 9 Graph partitioning. The dashed line shows the solution of the minimum bisection problem for the graph illustrated, i. e. the partition in two groups of equal size with minimal number of edges running between the groups.



Spectral Clustering Using Laplacian :

- ▶ Only with Cosine Similarity based distance measure, the set of clusters benchmarked by Newmann et. al. were obtained. The other two distance measures mis-clustered vertices 8 and 9 into two different clusters.

Hierarchical Agglomerative Clustering on X :

- ▶ Both Complete and Average linkage criterion were able to extract the desired set of clusters with almost all the distance metrics.
- ▶ Complete linkage performed better than the Average one using Cosine similarity distance metric.



Hierarchical Agglomerative Clustering on A(G) :

- ▶ In the first Structural Equivalence based method, only Single linkage criterion failed to achieve the desired set of clusters.
- ▶ In the Jaccard Similarity based methods, only Average linkage criterion failed to achieve the desired set of clusters.
- ▶ But using Pearson Correlation based measure, all the 3 linkage methods obtained the benchmark results.

Zackary's Karate Club Graph I

Graph 3 : Zackary's Karate Club Graph [3]

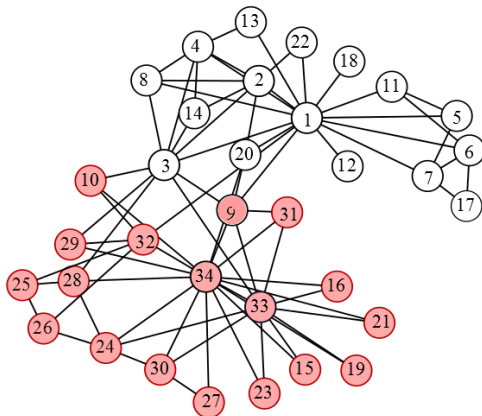


Figure: Zackary's Karate Club after split into 2 communities : A standard benchmark in Community detection



Spectral Clustering Using Laplacian :

- ▶ Only with Manhattan distance measure, the benchmarked results were obtained. The other two distance measures mis-clustered vertices 3 and 9 into two different clusters.
- ▶ Most of known Graph Clustering algorithms happen to make this same mistake with vertices 3 and 9.

Hierarchical Agglomerative Clustering on X :

- ▶ Only with Cosine Similarity metric, Complete and Average linkage criterion were able to extract the desired set of clusters.
- ▶ All other distance measure based methods failed miserably.

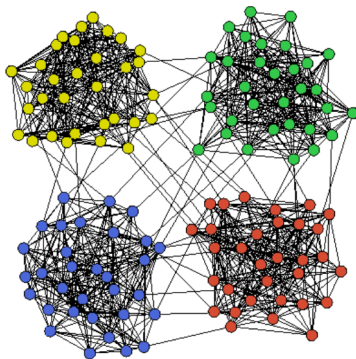


Hierarchical Agglomerative Clustering on $A(G)$:

- ▶ The first Structural Equivalence based method did complete blunder by putting vertices 1 & 34 itself in same cluster.
- ▶ Neighborhood overlap method too did the same mistake as above.
- ▶ But using Pearson Correlation based measure, again, both complete and average linkage methods obtained the benchmark results.

Planted ℓ -partition model Graph I

Graph 4 : Planted ℓ -partition model [4] : Benchmark of Girvan & Newman



Planted I-partition Benchmark By Girvann and Newmann

Figure: Planted ℓ -partition Graph ($l=4$, $g=32$, $p_{in}=15/31$, $p_{out}=1/96$): Girvan & Newman benchmark in Community detection



Spectral Clustering Using Laplacian :

- ▶ All the 3 different distance measures viz. Squared Euclidean, Manhattan, Cosine Similarity based distance extracted the exact set of clusters.

Hierarchical Agglomerative Clustering on X :

- ▶ All the 3 linkage criterion were able to extract the desired set of clusters with all the distance metrics.

Hierarchical Agglomerative Clustering on A(G) :

- ▶ The first Structural Equivalence based method was not able to find the proper clusters.
- ▶ Neighborhood overlap method performed good in this case as in most of the cases it found out the “best” clusters possible.
- ▶ As expected, Pearson Correlation based measure, again obtained the benchmark results.



Motivation

Problem Definition & Challenges

Problem Definition

Challenges

Traditional Graph Clustering Algorithms

Spectral Clustering using Laplacian

Hierarchical Clustering using Agglomeration

Experimental Results

Results on Graph 1

Results on Graph 2

Zackary's Karate Club Graph

Planted ℓ -partition model Graph

Conclusion

Conclusion

Observations from these Results :



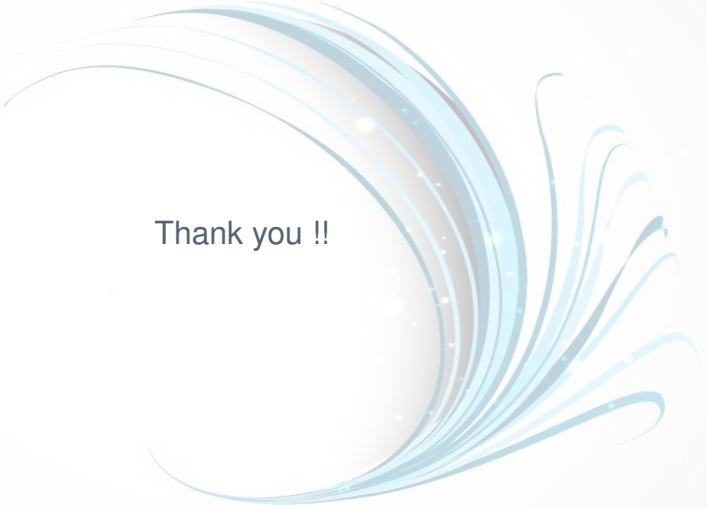
- ▶ All the 3 distance metrics can be used interchangeably for clustering on $X \in \mathbb{R}^{n \times k}$.
- ▶ In Hierarchical clustering, using complete linkage criterion is better than the rest.
- ▶ If Structural Equivalence on Graph adjacency metric is chosen, then Pearson Correlation is empirically proven to be the best choice for clustering.



If one tries to look for a very general method, that should give good results on any type of graphs, one is inevitably forced to make very general assumptions on the structure of the graph and on the properties of communities. In this way one neglects a lot of specific features of the system, that may lead to a more accurate detection of the clusters. Informing a method with features characterizing some types of graphs makes it far more reliable to detect the community structure of those graphs than a general method, even if its applicability may be limited. - Santo Fortunato. [1]



- [1] Fortunato, Santo. Community detection in graphs. *Physics Reports* 486.3 (2010): 75-174.
- [2] Fortunato, Santo. Castellano Claudio. Community Structure in Graphs. *Physics Reports* arXiv:0712.2716.
- [3] W. W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research* 33, 452-473 (1977).
- [4] Condon, A. karp, R.M. Algorithms for graph partitioning on the planted partition model. *Random Struct. Algor.* 18 (2001) 116-140.



Thank you !!