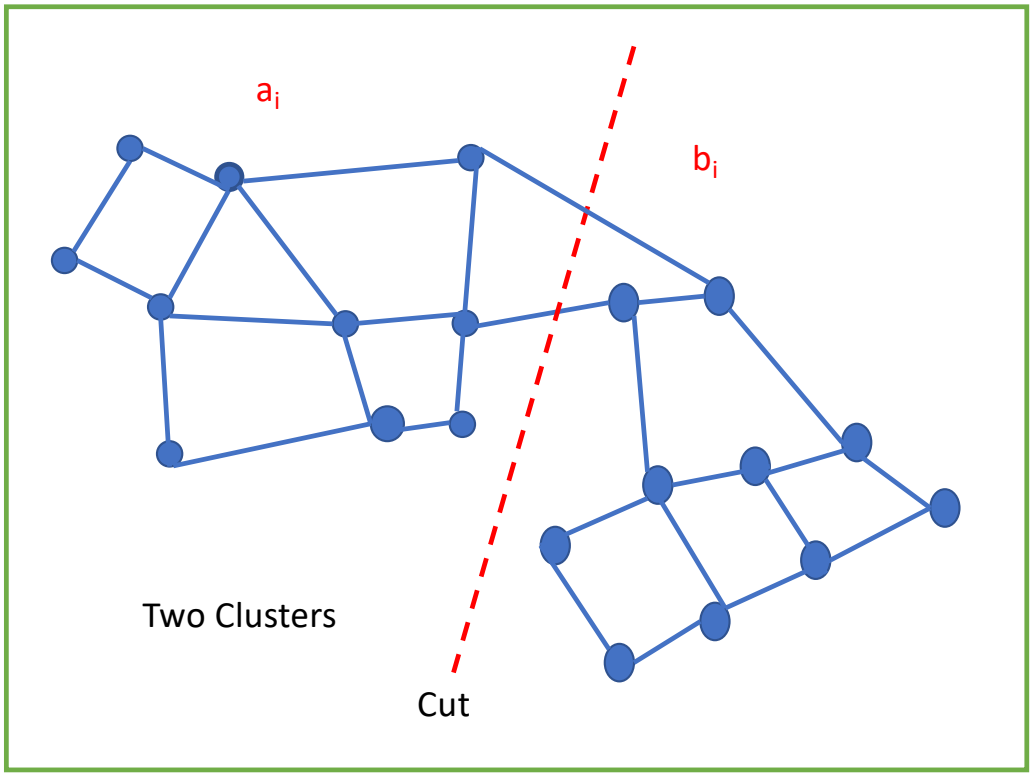
Lecture 8: Finding Clusters in Graphs

Agenda

- Clustering for graphs

Source: Sections III.3 and IV.6 and VII.1 in Linear Algebra and Learning from Data (2019) by **Gilbert Strang**

Clusters in Graphs



Clustering for graphs. Take a large graph. Let us divide it into two clusters.

Cut between two reasonably equal parts of the graph reasonably same size

Question is, how do you find such a cut by an algorithm? What's an algorithm that would find that cut? So that's a problem

Problem: Find position x,y to minimize distance
Minimize $\Sigma \|a_i - x\|^2 + \| b_i - y\|^2$
(a's ∪ b's=all nodes and a's ∩ b's=empty)

That would be not a satisfactory clustering. We're looking for clusters that are good sized clusters. So, minimize that

Clusters in Graphs

Lot of different algorithms to do clustering. Some are more directly attacking this problem
Others use matrices that we associate with the graph. Let us talk about two of those algorithms

What would be the best choice of the x once we've decided on the a's? And what would be the best choice of the y once we've decided on the b's

Given a's= $a_1, a_2, \ldots a_k$, What is the best x?

Geometrically, what x should be here? X is a bunch of points, and we're looking for the middle of those points. It'll make the sum of the distances squared-- a minimum. What is x?

x=Centroids of a's=(sum of a's)/(number of a's)

Talking about k-means algorithm. Here k=2, i.e., 2-means separations or partitions

Clusters in Graphs

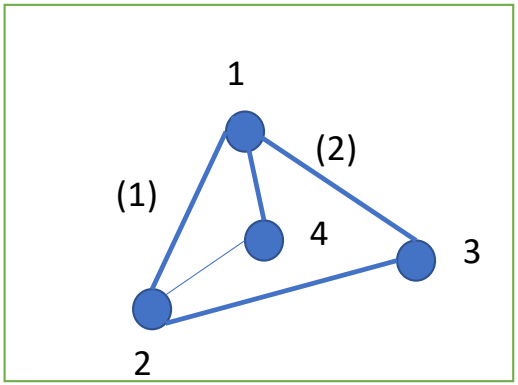
- K-means (2-means here)
- (1) Given a's and b's find centroids x and y; and (2) Given x, and y form best clusters
- Each node goes with the closer of x,y
- So points are closer to x-- put them in the cluster around x
- Come back to step one. Changed the clusters. They'll have different centroids
- So we repeat step one-- find the centroids for the two new clusters. Then we come to step two
- Find the ones that should go with the two centroids, and back and forth
- Popular algorithm, k-means. k would be to have k clusters

Backup Slides

Clusters in Graphs

- Second Solution Method is ”Spectral Clustering”. Spectral graph theory, spectral clustering.
- What's that word spectral about? What does that mean? It is the eigenvalues
- So spectral theory, spectral clustering is using the eigenvalues of some matrix
- Start with Laplacian Matrix. That's the key connection of linear algebra to graph theory, is the properties of this graph
- Laplacian matrix $L = A^TA$, where A is the incidence matrix of the graph
- $\Rightarrow L = A^TA = D - B$ where D is degree matrix (diagonal) and B is adjacency matrix
- Incidence matrix A, that's m by n-- edges and nodes. It's rectangular. but we’re forming A^TA here. So we’re forming a symmetric, positive, semi-definite matrix. So L is symmetric, positive semi-definite.

Clusters in Graphs



A^T=(nxm) and A=(mxn). So A^TA=(nxn)

A=

1	-1	1	0	0
2	-1	0	1	0
3	So on			
4				
5				

And let me take edge 2, going from 1 to node 3, so it would have a minus 1 and a 1 there, and so on. So that's the incidence matrix A.

$D_{(4x4)}$=

3			
	3		
		2	
			2

Degree matrix will be 4 by 4. So three edges going in, node 2, three edges going in, node 3 has just two edges. And node 4 has just two edges

$B_{(4x4)}$=

0	1	1	1
1	0	1	1
1	1	0	0
1	1	0	0

B is 4x4. It tells us which node is connected to which node. Edges that connect a node to itself, so 0's on diagonal. All of 2 and 4 and 3 are connected to 1. So we have 1's there. Node 2-- all three nodes are connected to node 2. So we'll have—the second column and row will have all three 1's

Clusters in Graphs

$$L=D-B=\begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 2 & 0 \\ -1 & -1 & 0 & 2 \end{pmatrix}$$

Is it a positive definite matrix? Is it singular or is it not singular? Is there a vector in its null space, or is there not a vector in its null space? Can you solve Lx equals all 0's? We see that vector of all 1's will be a solution to L

$$Lx=\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad \text{for} \quad x=\begin{pmatrix} C \\ C \\ C \\ C \end{pmatrix}$$

Lx = 0 as for a whole line of one-dimensional null space of L has dimension 1 (See slide 10-Appendix I).
It's got 1 basis vector, 1, 1, 1, 1. So that's a first fact, that this positive, semi-definite matrix, L, has λ_1 equals 0. And the eigenvector is constant-- C, C, C, C-- the 1-dimensional eigenspace. Or 1, 1, 1, 1 is the typical eigenvector.

dim of Null space N(L)=1

L=D-B has $\lambda_1=0$ $\quad x_1=\begin{pmatrix} C \\ C \\ C \\ C \end{pmatrix}$

Back to graph clustering. The idea of graph clustering is to look at the Fiedler eigenvector. This is called the x_2 and is the eigenvector for the smallest positive eigenvalue for a λ_{min} excluding 0-- so the smallest eigenvalue of L-- this is called the Fiedler vector (Appendix II)

1st evalue=0=λ_1 and 2nd evalue= λ_{min} (no zeros)

Clusters in Graphs

How do you decide the clusters from L? How does L propose two clusters, the a's and b's? And here's the answer

Positive components of evectors and negative components of evectors

$x_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$

X_2=evectors of λ_{min} of L

Eigenvector got some +ve and some -ve components.
The components with +ve numbers of this eigenvector-the +ve components of x . And there are -ve components of this eigenvector.
And those are the two clusters
So two clusters are decided by the eigenvector by +ve or –ve signs of the components.
The +ve signs go in one and the -ve signs go in another.

Now what's the relation between those two eigenvectors of L? They are Orthogonal.
These are eigenvectors of a symmetric matrix. orthogonal to these components add to 0. That dot product is just, add up the components

So we have a bunch of +ve components and a bunch of -ve components. They have the same sum, because the dot product with that is 0

So, x_1 and x_2 are two clusters

Appendix I

From Slide 8,We take a 2-dimensional L

Lx= $\begin{bmatrix} 3 & -1 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

Lx=0

$\Longrightarrow 3x_1 - x_2 = 0$

$-x_1 + 3x_2 = 0$

$\Longrightarrow x_1 = 3x_2$

$\Longrightarrow 8x_2 = 0$

$\Longrightarrow x_2 = 0$ since $8 \neq 0$

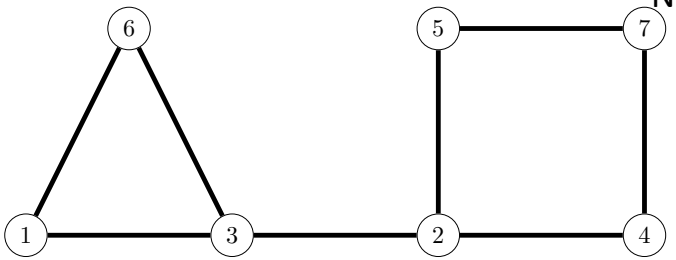
Hence Lx = $\begin{Bmatrix} 0 \\ 0 \end{Bmatrix}$

Appendix II

Basic Fiedler Method
First, Laplacian Matrix L = D − B

If G is a simple connected graph with n vertices and if L is the Laplacian matrix for G then L has n real eigenvalues satisfying $0=\lambda_1 <\lambda_2 \leq\lambda_3 \leq...\leq\lambda_n$
Fiedler Value or the algebraic connectivity of a graph is the second smallest eigenvalue of its Laplacian matrix L

A Fiedler Vector of a graph is an eigenvector corresponding to the Fiedler Value.



Now we have

$$L = \begin{bmatrix} 2 & 0 & -1 & 0 & 0 & -1 & 0 \\ 0 & 3 & -1 & -1 & -1 & 0 & 0 \\ -1 & -1 & 3 & 0 & 0 & -1 & 0 \\ 0 & -1 & 0 & 2 & 0 & 0 & -1 \\ 0 & -1 & 0 & 0 & 2 & 0 & -1 \\ -1 & 0 & -1 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & -1 & -1 & 0 & 2 \end{bmatrix}$$

Source: Graph Theory Justin Wyss-Gallifent July 21, 2021

Appendix II

Eigenvalues in order are: 0, 0.3588, 2.0000, 2.2763, 3.0000, 3.5892, 4.7757

Fiedler Value is 0.3588. A Fiedler Vector is an eigenvector corresponding to this. Any nonzero multiple of the following unit vector will suffice:

$$\bar{v} = \begin{bmatrix} 0.48 \\ -0.15 \\ 0.31 \\ -0.35 \\ -0.35 \\ 0.48 \\ -0.42 \end{bmatrix}$$

we can achieve a “reasonable” partition into two subgraphs by separating the vertices according to the sign of the values in a Fiedler Vector $\bar{v}$ where each entry corresponds to a vertex.

This means we group together the vertices i with v_i = + and we group together the vertices i with v_i = −. In the case that vi = 0 we simply have to make a choice.
So, we have partition P ({1,3,6},{2,4,5,7}).

Source: Graph Theory Justin Wyss-Gallifent July 21, 2021