Spectral Clustering in Directed Networks

Kevin Miller

March 19, 2016

Motivation

Graphs (a.k.a. Networks) are collections of nodes (vertices) and edges (connections) that can be used to model relationships between objects.

- Social Networks (Facebook, Twitter, LinkedIn, ego-nets)
- Protein-protein interaction networks
- Computer Cluster Networks
- Telecommunication Networks
- Buying/Selling Networks (Amazon, Ebay, etc.)

Motivation

Problems:

- Community Detection, Clustering, Partitioning
- Centrality Measures
- Graph Drawing/Visualization
- Diffusion Patterns

We look at Clustering and Community Detection today, via Spectral Clustering

Graph Basics

Graph

Graph $G(V, E)$ is a set of n nodes (vertices) $V = \{1, 2, \ldots, n\}$, with pairs of nodes connected by edges (links) in the set E.

Adjacency Matrix

$$A_{i,j} = \begin{cases} 1 & \textit{if there exists an edge in } E \textit{ from node } i \textit{ to node } j, \\ 0 & \textit{otherwise.} \end{cases}$$

Note we can replace 1 by weight w_e for the weight of edge $e = (i, j)$.

Graph Basics (cont.)

Degree Matrix

$$D = diag(d_1, d_2, \dots, d_n)$$

where $d_i = deg(i)$

Graph Laplacian

Given the adjacency matrix A *and degree matrix* D,

$$L = D - A$$

Other Graph Laplacians:

- $L_{rw} = I - D^{-1}A$
- $L_{sym} = I - D^{-1/2}AD^{-1/2}$

Directed vs Undirected

The edges in a graph represent relationships or flow of information. Not all relationships in the world are mutual, or bidirectional.

Undirected Graph

A graph G is undirected if all of the connections are bidirectional. This is equivalent to A and L being symmetric.

Otherwise, the graph is directed (digraph), with A and L not symmetric.

Undirected Example

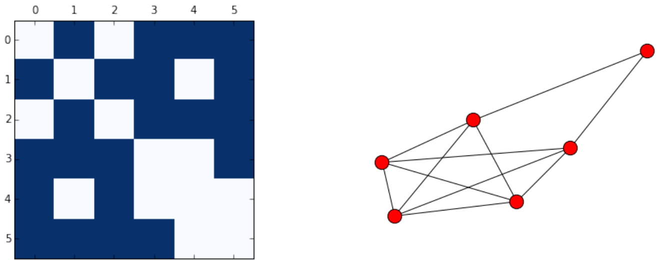
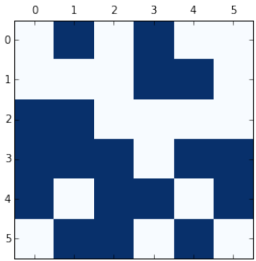


Figure: Undirected Adjacency Matrix and Corresponding Graph

Directed Example



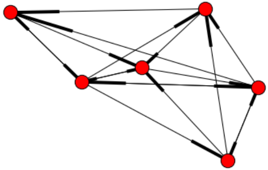


Figure: Directed Adjacency Matrix and Corresponding Graph

Spectral Graph Theory

Spectral Graph Theory : The study of graphs through the lens of the graph's spectral properties (i.e., eigenvalues and eigenvectors of L).

Focus in the field are *undirected* graphs because of nice spectral properties:

- L has n non-negative, real-valued eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$.
- L is symmetric, positive definite
- Smallest eigenvalue of L is 0 with corresponding e-vector $\mathbb{1}$

With digraphs, we get complex eigenvalues and eigenvectors!

Spectral Clustering Overview

- (Main idea) Separate nodes into different groups according to their edge structure
- (Reformulated) We desire to partition the graph such that the weight of edges between **different** groups is "minimized" and that the weight of edges **within** groups is "maximized"

$k = 2$ Clusters Visual

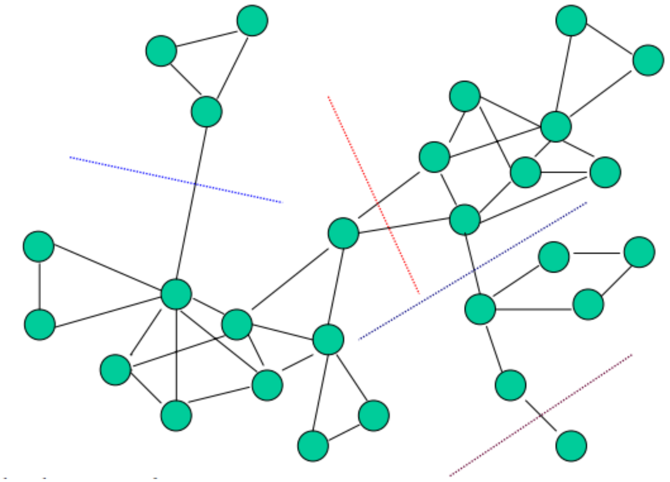


Figure: What Partition to Choose?

Visual from Chis Ding, *A Tutorial on Spectral Clustering* (2007).

Spectral Clustering Algorithm

Spectral Clustering Generalized

Inputs: k (desired # clusters), A (adjacency/similarity matrix of given graph)
Procedure:

- *Compute the desired Laplacian, L.*
- *Compute the first k eigenvectors $x_1, \ldots, x_k$ of L.*
- *Let $X = [x_1 x_2 \ldots x_k]$*
- *For $i = 1, \ldots, n$, let $y_i \in \mathbb{R}^k$ with the k-means clustering algorithm into clusters $C_1, \ldots, C_k$.*

Output: Clusters $C_1, \ldots, C_k$

Success of Spectral Clustering

Unsupervised Machine Learning

- Use Spectral Clustering to cluster data points
- Reveals important relationships in data

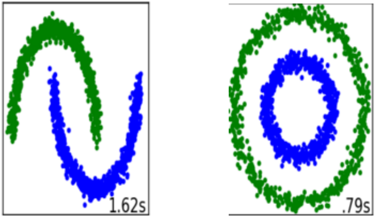


Figure: Half Moon Data Set and Concentric Circles

Visuals from Python Sklearn Website (Accessed March, 2016).

Problems in Spectral Clustering

No prior knowledge of community structure...
How to choose k communities with which to cluster??

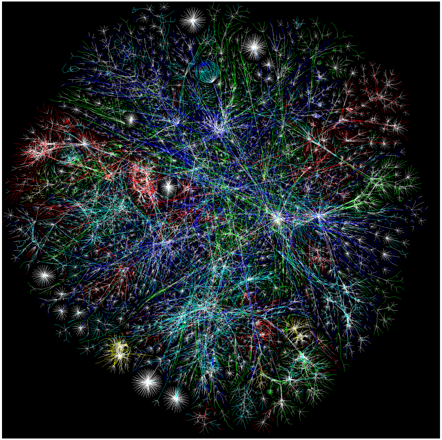


Figure: Map of the Internet

Spectral Clustering on Digraphs

With directed graphs, eigenspaces in complex subspaces make clustering difficult to justify rigorously. Hence the focus on undirected graphs in the field.

Difficulties:

- Lose simple ordering of eigenvalues
- Eigenspace coupling

But... why not try it?

Example - Undirected 3-Community

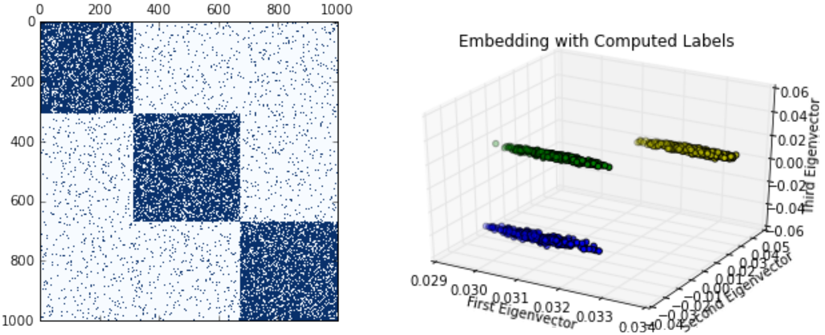


Figure: Adjacency Matrix and Spectral Embedding

Example - Directed 3-Community

The first 3 eigenvalues (and corresponding eigenvectors) were real!

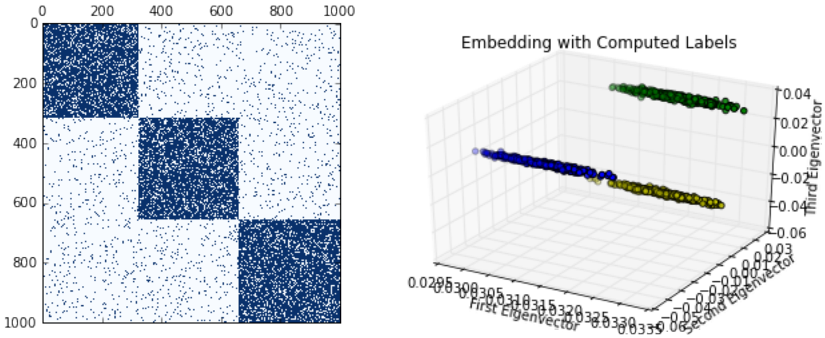


Figure: Adjacency Matrix and Spectral Embedding

Empirical Findings of Spectral Clustering in Digraphs

Main Finding

The number of smallest real eigenvalues of a digraph's Laplacian matrix is an indicator of the number of latent communities. That is, with eigenvalues ordered by magnitude:

$$0 = |\lambda_1| \leq |\lambda_2| \leq \ldots \leq |\lambda_n|$$

the largest value k for which $\lambda_1, \lambda_2, \ldots, \lambda_k \in \mathbb{R}$ indicates the latent number of communities in the digraph.

Implications:

- Eigenvectors corresponding to real eigenvalues are still good for clustering
- Can determine number of clusters to look for BEFORE spectral embedding is done!

Bibliography

- Ulrike von Luxburg. *A Tutorial on Spectral Clustering*. Appears in Statistics and Computing, 17(4), 2007. Accessed online
- Chris Ding. *A Tutorial on Spectral Clustering*. ICML, 2004. Accessed via http://ranger.uta.edu/ chqding/Spectral/spectralA.pdf
- Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. *On spectral clustering: analysis and an algorithm*. Advances in Neural Information Processing Systems, 14(2):849856, 2002.
- Jianbo Shi, Jitendra Malik. *Normalized cuts and image segmentation*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8):888905, 2000.