Graph Analysis: Community Detection

Giri Iyengar

Cornell University

gi43@cornell.edu

March 19, 2018

Overview

Overview

Graph Partitioning

Partitioning a Graph

In mathematics, the graph partition problem is defined on data represented in the form of a graph G = (V,E), with V vertices and E edges, such that it is possible to partition G into smaller components with specific properties. For instance, a k-way partition divides the vertex set into k smaller components. A good partition is defined as one in which the number of edges running between separated components is small. Uniform graph partition is a type of graph partitioning problem that consists of dividing a graph into components, such that the components are of about the same size and there are few connections between the components.

Source: Wikipedia

Graph Partitioning

Partitioning a Graph

Important applications of graph partitioning include scientific computing, partitioning various stages of a VLSI design circuit and task scheduling in multi-processor systems. Recently, the graph partition problem has gained importance due to its application for clustering and detection of cliques in social, pathological and biological networks. For a survey on recent trends in computational methods and applications see Buluc et al. (2013).

Source: Wikipedia

Graph Partitioning

- Natural clusters have nodes with many edges between them
- And few edges across clusters
- A random walk on the graph will **infrequently** go from one cluster to another
- Adjacency matrix will have **block-like** structure
- Eigenvalues of Adjacency / Laplacian matrix will be revealing

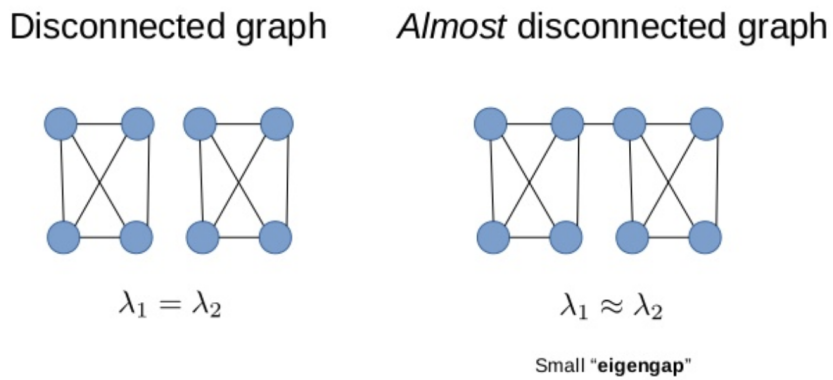

Figure: Carlos Castillo, Sapienza University

Overview

Spectral Clustering

Spectral Clustering

In multivariate statistics and the clustering of data, spectral clustering techniques make use of the spectrum (eigenvalues) of the similarity matrix of the data to perform dimensionality reduction before clustering in fewer dimensions. The similarity matrix is provided as an input and consists of a quantitative assessment of the relative similarity of each pair of points in the dataset.

Source: Wikipedia

Spectral Clustering

- Initially used for Image Segmentation

Spectral Clustering

- Initially used for Image Segmentation
- Generally recognized as a technique to cluster data

Spectral Clustering

- Initially used for Image Segmentation
- Generally recognized as a technique to cluster data
- An alternative to K-Means

Spectral Clustering

- Initially used for Image Segmentation
- Generally recognized as a technique to cluster data
- An alternative to K-Means
- Widely used in Graph Partitioning / Community Detection

Spectral Clustering: Definitions

- Graph: $G(V, E)$
- Weighted Graph. That is between v_i and v_j, there is a weight w_{ij}
- Undirected. $w_{ij} == w_{ji}$
- Define degree of vertex $d_i = \sum_{j=1}^{n} w_{ij}$

Spectral Clustering: Definitions

- W(A,B) $= \sum_{i \in A, j \in B} w_{ij}$. $A, B \subset V$
- $|A|$: Number of vertices in A
- $vol(A) = \sum_{i \in A} d_i$
- $A \cap \bar{A} = \emptyset$

Spectral Clustering: Properties of Laplacian

- $f^T L f = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2$
- $f^T L f = f^T D f - f^T A f = \sum_{i=1}^{n} d_i f_i^2 - \sum_{i,j=1}^{n} w_{ij} f_i f_j$
- $f^T L f = \frac{1}{2}(\sum_{i=1}^{n} d_i f_i^2 - 2\sum_{i,j=1}^{n} w_{ij} f_i f_j + \sum_{j=1}^{n} d_j f_j^2)$
- L is symmetric (follows from symmetry of D and A)
- L is positive semi-definite
- Smallest eigenvalue of L is 0 and has $\mathbb{1}^n$ as the eigenvector

Spectral Clustering: Properties of Laplacian

Number of connected components

Number of connected components in the Graph is given by multiplicity of eigenvalue 0. Trivial to see from previous slide.

Spectral Clustering Algorithm

- Compute the un-normalized Laplacian L
- Compute the first k eigenvectors of L. These are the smallest k eigenvectors
- U matrix of the first k eigenvectors is an $n \times k$ matrix
- Take the $\mathbb{R}^k$ row-vectors of U
- Perform k-way K-Means on these vectors

Variations of Spectral Clustering

- Instead of the un-normalized Laplacian, start with normalized Laplacian matrices
- Instead of solving the eigenvalue problem, solve the generalized eigenvalue problem instead

Spectral Clustering for Image Segmentation

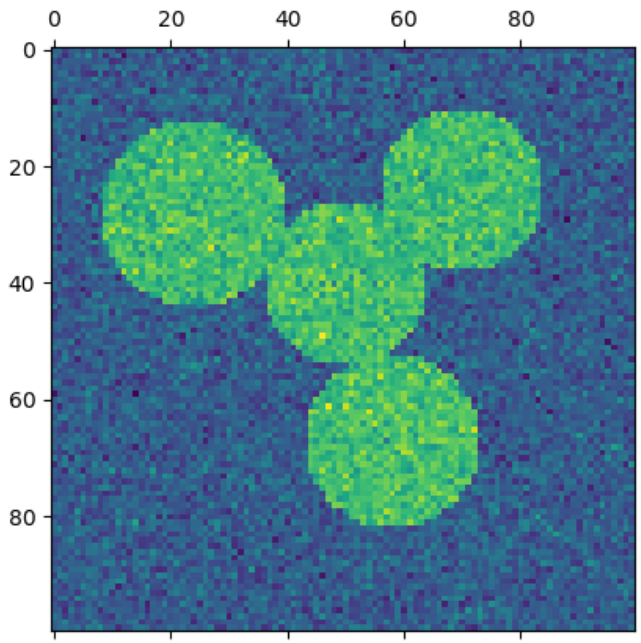

Figure: Source - Scikit-learn

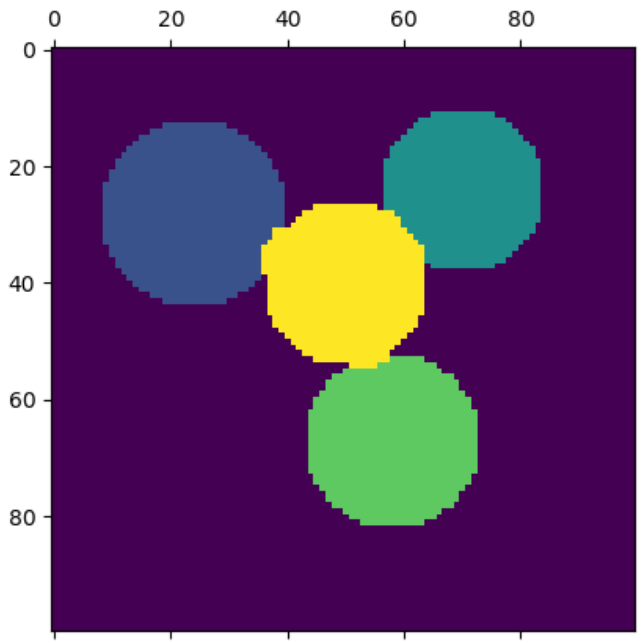

Figure: Source - Scikit-learn

Overview

Markov Cluster Algorithm

MCL Simulation

MCL Algorithm

MCL

The MCL algorithm is short for the Markov Cluster Algorithm, a fast and scalable unsupervised cluster algorithm for graphs (also known as networks) based on simulation of (stochastic) flow in graphs.

Heavily used in clustering Proteins based on their Amino acid sequences

MCL Algorithm

- Finds clusters by simulating random walks
- Assumption: Random walks in a graph with tend to spend more time in Natural clusters
- Random walks will infrequently go between clusters
- Within a cluster, much more likely to find longer-length random walk between any pair of nodes

MCL Algorithm

- Start with a Column Stochastic matrix of the Graph
- **Expansion Step**: Power iteration of the Matrix
- **Inflation Step**: Hadamard power followed by normalization
- Repeat Expansion and Inflation till convergence

MCL Algorithm

- M_1: Column Stochastic Matrix
- **Expansion**: $M_2 = M_1 \times M_1$
- **Hadamard**: $M_3 = [m_{2ij}^r]$. Typically $r = 2$
- **Normalization**: $M_1 = [\frac{m_{3ij}}{\sum_j m_{3ij}}]$

Hadamard power operation on Matrix

$$\begin{bmatrix} 0.8 & 0.6 & 0.9 \\ 0.2 & 0.4 & 0.1 \end{bmatrix} \qquad \begin{bmatrix} 0.64 & 0.36 & 0.81 \\ 0.04 & 0.16 & 0.01 \end{bmatrix} \rightarrow \begin{bmatrix} 0.94 & 0.69 & 0.99 \\ 0.06 & 0.31 & 0.01 \end{bmatrix}$$

MCL Algorithm

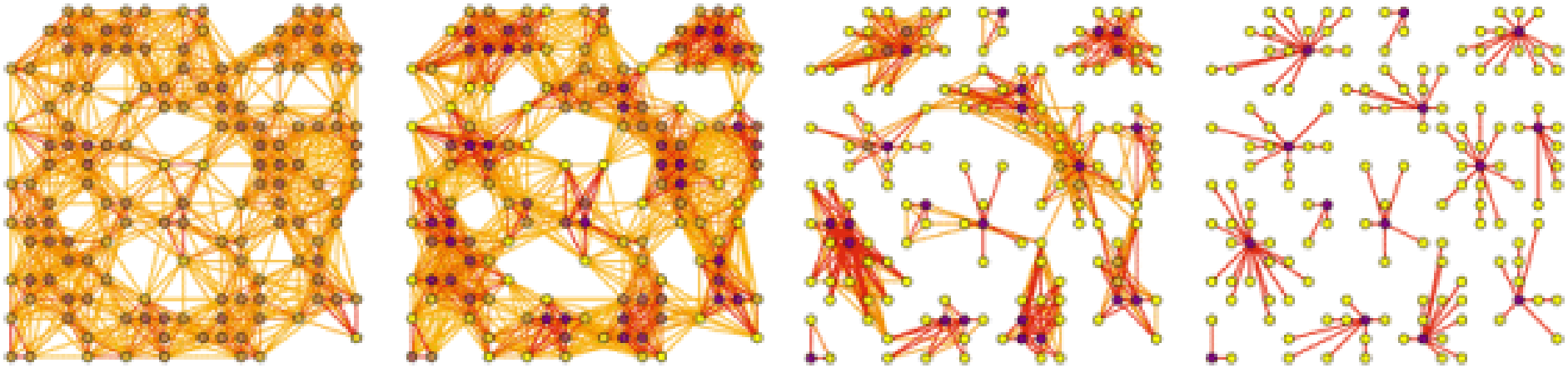

Figure: Source - Stijn van Dongen


thank you
danke
謝謝
ngiyabonga
teşekkür ederim
gracias
merci
obrigado
dziękuję
bedankt
спасибо
sukriya
kop khun krap
go raibh maith agat
mochchakkeram
terima kasih
grazie
arigatō
dakujem
мерси
감사합니다
ありがとう
ευχαριστώ
dank je
tapadh leat
hvala
mauruuru
sagolun
takk
taiku
рахмат
Баярлалаа
blagodaram
misaotra
matondo
paldies
asante
manana
хвала
enkosi
xièxie
shukriya
dhanyavadagalu
diolch
rahmet
tanemirt
najis tuke
kam sah hamnida
mèsi
dèkuji
sobodi
gracie
kiitos
dankie
dhanyavad
nandi
vinaka
mersi
barka
welalin
tack
spas
mahalo
grazzi
obrigada
murakoze
tenki
djiere dieuf
chnorakaloutioun
gratias ago
gràcies
sulpáy
trugarez
mamnun
дякую
faafetai lava
köszönöm
bayarlalaa