## SPECTRAL GRAPH ANALYSIS

With Applications of Graph Theory in Biological Networks

Ronald & Aditya

**IISER** Mohali

November 21, 2022

### **Abstract**

Spectral graph theory can be used to analyze the topological properties (e.g., connectivity) of graphs. Each graph has a Laplacian matrix whose eigenvalues and eigenvectors reveal many properties of the graph. We look at discrete mathematical (graph theoretical) models for biological networks, then study some mathematics of spectral graph theory and in particular, properties of the Laplacian.

# **Objectives**

Take a glimpse at some graph theory models in biology.

# **Objectives**

- Take a glimpse at some graph theory models in biology.
- ② Discuss some spectral graph theory and apply it to two biological networks.

1 Preliminaries: Graph Theory

- 1 Preliminaries: Graph Theory
- Graph Theory Models in Biological Networks

- Preliminaries: Graph Theory
- @ Graph Theory Models in Biological Networks
- Spectral Graph Theory

- 1 Preliminaries: Graph Theory
- @ Graph Theory Models in Biological Networks
- Spectral Graph Theory
- Two Biological Networks
  - ► Tissue-specific protein-function associations
  - Drug-target interaction network

PRELIMINARIES: GRAPH THEORY

Degree of a Vertex, a Simple Graph

A Graph G = (V, E) has set V of vertices and set E of edges connecting the vertices.

Degree of a Vertex, a Simple Graph

A Graph G = (V, E) has set V of vertices and set E of edges connecting the vertices.

If two vertices  $i, j \in V$  are connected by an edge, we write  $\{i, j\} \in E$ .

Degree of a Vertex, a Simple Graph

A Graph G = (V, E) has set V of vertices and set E of edges connecting the vertices.

If two vertices  $i, j \in V$  are connected by an edge, we write  $\{i, j\} \in E$ .

The degree of a vertex is the number of edges connected to it.

Degree of a Vertex, a Simple Graph

A Graph G = (V, E) has set V of vertices and set E of edges connecting the vertices.

If two vertices  $i, j \in V$  are connected by an edge, we write  $\{i, j\} \in E$ .

The degree of a vertex is the number of edges connected to it.

A *simple graph* is one where there is at most an edge between any two pair of vertices and there are no self-loops (edges that join the same vertex).

Types of Graphs

- Undirected graph
- ② Directed graph (digraph)
- Weighted graph

Distance

The distance  $\delta(i,j)$  from i to j is the length of the shortest path from i to j in G.

Distance

The distance  $\delta(i,j)$  from i to j is the length of the shortest path from i to j in G.

The diameter of G is the maximum value of  $\delta(i,j)$  taken over all distinct pairs of vertices.

Distance

The distance  $\delta(i,j)$  from i to j is the length of the shortest path from i to j in G.

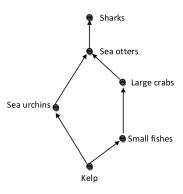
The diameter of G is the maximum value of  $\delta(i,j)$  taken over all distinct pairs of vertices.

The most common algorithms for calculating the shortest paths are *Dijkstra's* greedy algorithm and *Floyd's* dynamic algorithm.

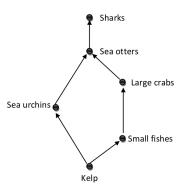
GRAPH THEORY MODELS IN BIOLOGICAL NETWORKS

Directed Graph

#### Directed Graph

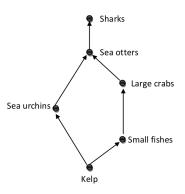


#### Directed Graph



Arrows represent energy flow

#### Directed Graph



- Arrows represent energy flow
- ② Directed graphs represent predator-prey interactions

Directed Graph

We can ask:

What are the trophic levels?

Directed Graph

#### We can ask:

- What are the trophic levels?
- What is a dominant species in a food web?

#### Directed Graph

#### We can ask:

- What are the trophic levels?
- What is a dominant species in a food web?
- On we assign weights to the edges?

#### Trophic status

- Two ways to assigning trophic levels:
  - Shortest path
  - Longest path

#### Trophic status

- Two ways to assigning trophic levels:
  - Shortest path
  - Longest path
- 2 Both of the above methods have shortcomings.

#### Trophic status

- Two ways to assigning trophic levels:
  - Shortest path
  - Longest path
- Observe Both of the above methods have shortcomings.
- We can define trophic status of a species u by:

$$T(u) = \sum_{k} k n_{k},$$

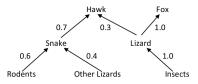
where  $n_k$  is the number of species whose longest path to u has length k.

#### Dominant Species and Weighted Food Webs

A species is dominant in a food web if its trophic status is greater than the number of species in the food web above level 0.

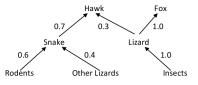
#### Dominant Species and Weighted Food Webs

A species is dominant in a food web if its trophic status is greater than the number of species in the food web above level 0.



#### Dominant Species and Weighted Food Webs

A species is dominant in a food web if its trophic status is greater than the number of species in the food web above level 0.



We can also define flow-based trophic level (TL) in a weighted food web:

$$\mathsf{TL}(i) = 1 + \sum_{i} w_{ij} + \mathsf{TL}(\mathsf{food}\;\mathsf{source}\;j)$$

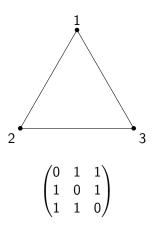
### Other models

- Protein-protein interactions (PPI)
- ② Gene-protein interactions
- Gene regulatory networks (GRN)
- Signal transduction networks
- Metabolic and biochemical networks

SPECTRAL GRAPH THEORY

## From Graphs to Matrices

#### Adjacency Matrix



The graph and the matrix has the same information.

## From Graphs to Matrices

#### Adjacency Matrix

## Definition (Adjacency matrix)

The adjacency matrix  $A \in \{0,1\}^{n \times n}$  is defined by

$$A_{ij} = \begin{cases} 1 & \text{if } \{i,j\} \in E, \\ 0 & \text{otherwise.} \end{cases}$$

## From Graphs to Matrices

#### Adjacency Matrix

## Definition (Adjacency matrix)

The adjacency matrix  $A \in \{0,1\}^{n \times n}$  is defined by

$$A_{ij} = \begin{cases} 1 & \text{if } \{i,j\} \in E, \\ 0 & \text{otherwise.} \end{cases}$$

- **1** A is symmetric for an undirected graph.
- ② A has diagonal elements zero if there are no self-loops.

#### Adjacency Matrix

### Definition (Adjacency matrix)

The adjacency matrix  $A \in \{0,1\}^{n \times n}$  is defined by

$$A_{ij} = \begin{cases} 1 & \text{if } \{i,j\} \in E, \\ 0 & \text{otherwise.} \end{cases}$$

- **1** A is symmetric for an undirected graph.
- ② A has diagonal elements zero if there are no self-loops.

### Definition (Degree matrix)

The degree matrix  $D \in \mathbb{R}^{n \times n}$  is defined as the diagonal matrix with diagonal entries  $(d_1, \ldots, d_n)$ .

Normalized adjacency matrix

### Definition (Normalized adjacency matrix)

The normalized adjacency matrix is defined by  $\overline{A} = AD^{-1}$ .

Normalized adjacency matrix

### Definition (Normalized adjacency matrix)

The normalized adjacency matrix is defined by  $\overline{A} = AD^{-1}$ .

• A is not necessarily symmetric. But it is column-stochastic, that is, each column has nonnegative entries that sum to 1.

Normalized adjacency matrix

### Definition (Normalized adjacency matrix)

The normalized adjacency matrix is defined by  $\overline{A} = AD^{-1}$ .

- f A is not necessarily symmetric. But it is column-stochastic, that is, each column has nonnegative entries that sum to 1.
- ② Suppose  $\rho \in \mathbb{R}^n$  is a probability distribution,  $\overline{A}\rho$  is another probability distribution. Therefore, we have a *random walk* on the graph!

Laplacian matrix

### Definition (Laplacian matrix)

The Laplacian matrix is defined as L = D - A.

Laplacian matrix

### Definition (Laplacian matrix)

The Laplacian matrix is defined as L = D - A.

### Definition (Normalized Laplacian matrix)

The normalized Laplacian matrix is defined as

$$\overline{L} = D^{-1/2}LD^{-1/2} = \mathbb{I} - D^{-1/2}AD^{-1/2}.$$

Laplacian matrix

**1** L and  $\overline{L}$  are always symmetric.

#### Laplacian matrix

- ullet L and  $\overline{L}$  are always symmetric.
- ② They are best thought of as quadratic forms: for any  $x \in \mathbb{R}^n$ ,

$$x^T L x = \sum_i d_i x_i^2 - \sum_{(i,j) \in E} x_i x_j = \sum_{\{i,j\} \in E} (x_i - x_j)^2.$$

Normalized Laplacian matrix

### Proposition

 $\forall x \in \mathbb{R}^n$ , we have

$$x^T \overline{L} x = \sum_{\{i,j\} \in E} \left( \frac{x_i}{\sqrt{d_i}} - \frac{x_j}{\sqrt{d_j}} \right)^2.$$

Normalized Laplacian matrix

### Proposition

 $\forall x \in \mathbb{R}^n$ , we have

$$x^T \overline{L} x = \sum_{\{i,j\} \in E} \left( \frac{x_i}{\sqrt{d_i}} - \frac{x_j}{\sqrt{d_j}} \right)^2.$$

If G is d-regular, then this simplifies to

$$x^T \overline{L} x = \frac{1}{d} \sum_{\{i,j\} \in E} (x_i - x_j)^2.$$

A single edge



The adjacency matrix and degree matrix are

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$
 and  $D = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ .

A single edge



The adjacency matrix and degree matrix are

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$
 and  $D = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ .

The Laplacians are

$$L = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} = \overline{L}.$$

A single edge



The adjacency matrix and degree matrix are

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$
 and  $D = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ .

The Laplacians are

$$L = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} = \overline{L}.$$

The eigenvalues of  $\overline{L}$  are given by  $\lambda_1=0, \lambda_2=2$  whose corresponding eigenvectors are

$$\frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix} \text{ and } \frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\end{pmatrix}.$$

A triangle graph

$$A = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}, \text{ and } D = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

A triangle graph

$$A = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}, \text{ and } D = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

The Laplacian matrix is

$$L = \begin{pmatrix} 1 & -1/2 & -1/2 \\ -1/2 & 1 & -1/2 \\ -1/2 & -1/2 & 1 \end{pmatrix}.$$

A triangle graph

$$A = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}, \text{ and } D = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

The Laplacian matrix is

$$L = \begin{pmatrix} 1 & -1/2 & -1/2 \\ -1/2 & 1 & -1/2 \\ -1/2 & -1/2 & 1 \end{pmatrix}.$$

The eigenvalues of  $\overline{L}$  are  $\{0, 3/2, 3/2\}$  with corresponding eigenvectors

$$\frac{1}{\sqrt{3}} \begin{pmatrix} 1\\1\\1 \end{pmatrix}, \frac{1}{\sqrt{6}} \begin{pmatrix} 2\\-1\\-1 \end{pmatrix}, \frac{1}{\sqrt{2}} \begin{pmatrix} 0\\1\\-1 \end{pmatrix}.$$

The path of length 2

$$A = egin{pmatrix} 0 & 1 & 0 \ 1 & 0 & 1 \ 0 & 1 & 0 \end{pmatrix}, \text{ and } D = egin{pmatrix} 1 & 0 & 0 \ 0 & 2 & 1 \end{pmatrix}.$$

The path of length 2

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \text{ and } D = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 1 \end{pmatrix}.$$

The Laplacian matrix is

$$L = \begin{pmatrix} 1 & -1/\sqrt{2} & 0 \\ -1/\sqrt{2} & 1 & -1/\sqrt{2} \\ 0 & -1/\sqrt{2} & 1 \end{pmatrix}.$$

The path of length 2

$$A = egin{pmatrix} 0 & 1 & 0 \ 1 & 0 & 1 \ 0 & 1 & 0 \end{pmatrix}, \text{ and } D = egin{pmatrix} 1 & 0 & 0 \ 0 & 2 & 1 \end{pmatrix}.$$

The Laplacian matrix is

$$L = \begin{pmatrix} 1 & -1/\sqrt{2} & 0 \\ -1/\sqrt{2} & 1 & -1/\sqrt{2} \\ 0 & -1/\sqrt{2} & 1 \end{pmatrix}.$$

The eigenvalues of  $\overline{L}$  are  $\{0,1,2\}$  with corresponding eigenvectors

$$\frac{1}{2}\begin{pmatrix}1\\\sqrt{2}\\1\end{pmatrix},\frac{1}{\sqrt{2}}\begin{pmatrix}-1\\0\\1\end{pmatrix},\frac{1}{2}\begin{pmatrix}-1\\\sqrt{2}\\-1\end{pmatrix}.$$

# Eigenvalues of the Laplacian

### Proposition

For any graph G with normalized Laplacian  $\overline{L}$ , if  $\lambda_1$  is the smallest eigenvalue of  $\overline{L}$  then  $\lambda_1=0$  with multiplicity equal to the number of connected components of G.

# Eigenvalues of the Laplacian

### Proposition

For any graph G with normalized Laplacian  $\overline{L}$ , if  $\lambda_1$  is the smallest eigenvalue of  $\overline{L}$  then  $\lambda_1=0$  with multiplicity equal to the number of connected components of G.

### Corollary

For any graph G, the second smallest eigenvalue  $\lambda_2(L) > 0$  if and only if G is connected.

#### **Properties**

For an undirected graph G and its Laplacian matrix  $\overline{L}$  with eigenvalues

$$0 = \lambda_1 \le \lambda_2 \le \cdots \le \lambda_{n-1} :$$

•  $\overline{L}$  is positive-semidefinite (that is,  $\lambda_i \geq 0$ ).

#### **Properties**

For an undirected graph G and its Laplacian matrix  $\overline{L}$  with eigenvalues

$$0 = \lambda_1 \le \lambda_2 \le \cdots \le \lambda_{n-1} :$$

- $\overline{L}$  is positive-semidefinite (that is,  $\lambda_i \geq 0$ ).
- Every row sum and column sum of  $\overline{L}$  is zero.

#### **Properties**

For an undirected graph G and its Laplacian matrix  $\overline{L}$  with eigenvalues

$$0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{n-1}$$
:

- $\overline{L}$  is positive-semidefinite (that is,  $\lambda_i \geq 0$ ).
- Every row sum and column sum of  $\overline{L}$  is zero.
- The second smallest nonzero eigenvalue of  $\overline{L}$  (could be zero) is the algebraic connectivity (or Fiedler value) of G.

#### **Properties**

For an undirected graph G and its Laplacian matrix  $\overline{L}$  with eigenvalues

$$0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{n-1}$$
:

- $\overline{L}$  is positive-semidefinite (that is,  $\lambda_i \geq 0$ ).
- Every row sum and column sum of  $\overline{L}$  is zero.
- The second smallest nonzero eigenvalue of  $\overline{L}$  (could be zero) is the algebraic connectivity (or Fiedler value) of G.
- The trace of the Laplacian matrix  $\overline{L}$  is equal to 2m where m is the number of edges of the considered graph.

Using Eigenvalues of the Laplacian

The second smallest eigenvalue  $\lambda_2$  tells us about the connectivity of the graph.

Using Eigenvalues of the Laplacian

The second smallest eigenvalue  $\lambda_2$  tells us about the connectivity of the graph.

**1** If *G* has two disconnected components,  $\lambda_2 = 0$ .

Using Eigenvalues of the Laplacian

The second smallest eigenvalue  $\lambda_2$  tells us about the connectivity of the graph.

- **1** If *G* has two disconnected components,  $\lambda_2 = 0$ .
- ② If  $\lambda_2$  is small, this suggests the graph is nearly disconnected.

Using Eigenvalues of the Laplacian

The second smallest eigenvalue  $\lambda_2$  tells us about the connectivity of the graph.

- **1** If *G* has two disconnected components,  $\lambda_2 = 0$ .
- ② If  $\lambda_2$  is small, this suggests the graph is nearly disconnected.
- **3** In short,  $\lambda_2$  gives us a sort of continuous measure of how well a graph is connected.

TWO BIOLOGICAL SYSTEMS

Tissue-specific protein-function associations

This is a collection of protein-function association networks for a large number of human tissues. Nodes represent human proteins, tissues, and biological functions. Edges represent associations that indicate different biological roles of proteins in various tissues and cell types.

#### Tissue-specific protein-function associations

Dataset statistics	
Nodes	2564
Edges	18104
Nodes in largest SCC	2564
Fraction of nodes in largest SCC	1.000000
Edges in largest SCC	18104
Fraction of edges in largest SCC	1.000000
Average clustering coefficient	0.000000
Number of triangles	943127
Fraction of closed triangles	0.000000
Diameter (longest shortest path)	6
90-percentile effective diameter	3.925416
Average tissues per edge	1.138920
Median tissues per edge	1
Standard deviation of tissues per edge	0.370247

#### Tissue-specific protein-function associations

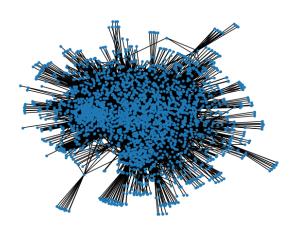
Dataset statistics	
Nodes	2564
Edges	18104
Nodes in largest SCC	2564
Fraction of nodes in largest SCC	1.000000
Edges in largest SCC	18104
Fraction of edges in largest SCC	1.000000
Average clustering coefficient	0.000000
Number of triangles	943127
Fraction of closed triangles	0.000000
Diameter (longest shortest path)	6
90-percentile effective diameter	3.925416
Average tissues per edge	1.138920
Median tissues per edge	1
Standard deviation of tissues per edge	0.370247

#### Code on GitHub

```
37 lines (30 sloc) 890 Bytes
 2 import pandas as pd
  3 import networks as nx
 4 import matplotlib.pyplot as plt
 6 data = pd.read_csv("TFG-Ohmnet_tissue-function-gene.tsv")
 7 data = list(np.array(data))
 9 for i in range(len(data)):
       data[i] = data[i][0].solit("\t")[i:]
12 V1 = [data[i][0] for i in range(len(data))]
13 V2 = [data[i][1] for i in range(len(data))]
14 V3 = []
16 for 1 in V1:
     if i not in V3:
          V3_annend(1)
20 for i in V2:
     if i not in V3:
           V3.append(i)
24 G = nx.Graph()
25 G.add_nodes_from(V3)
26 G.add edges from(data)
27 L = nx.normalized_laplacian_matrix(6)
28 e = np.linalq.eigvals(L.toarray())
29 nx.draw_networkx(G, node_size=10, with_labels=False)
30 plt.savefig("TFG_Graph.png")
31 plt.show()
33 plt.hist(e, bins = 100)
34 plt.title("Distribution of Eigenvalues (Tissue-specific protein-function associations)")
35 plt.xlabel("Eigenvalue")
 36 plt.savefig("TEG.npp")
37 plt.show()
```

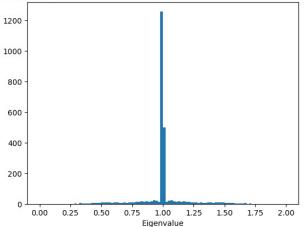
https://github.com/huidr/spectral-graph-analysis/tfg.py

Plotting the Graph



#### Plotting the Histogram of the Eigenvalues

Distribution of Eigenvalues (Tissue-specific protein-function associations)



This is a drug-target interaction network that contains information on which genes (i.e., proteins encoded by genes) are targeted by drugs that are on the U.S. market. Drug targets are molecules that play a critical role in the transport, delivery or activation of the drug.

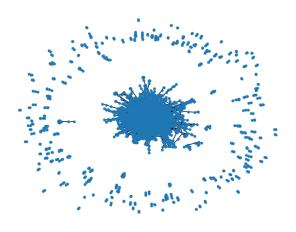
Dataset statistics	
Nodes	7341
Drug nodes	5017
Gene nodes	2324
Edges	15138
Nodes in largest SCC	6621
Fraction of nodes in largest SCC	1.000000
Edges in largest SCC	14581
Fraction of edges in largest SCC	0.963205
Diameter (longest shortest path)	15
90-percentile effective diameter	7.884931

#### Code on GitHub

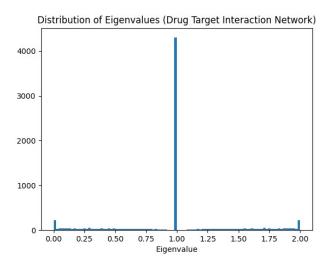
```
37 lines (30 sloc) 868 Bytes
 1 import numpy as no
 2 import pandas as pd
 3 import networks as nx
     import matplotlib.pyplot as plt
     data = pd.read csv("ChG-Miner miner-chem-gene.tsv")
     data = list(np.array(data))
 9 for i in range(len(data)):
       data[i] = data[i][8].split("\t")
12 V1 = [data[i][0] for i in range(len(data))]
13 V2 = [data[i][1] for i in range(len(data))]
14 V3 = []
16 for 1 in V1:
      if i not in V3:
         V3.append(i)
28 for 1 to V21
      if i not in V3:
            V3.append(i)
24 G = nx.Graph()
25 G. add nodes from(V3)
26 G.add_edges_from(data)
27 L = nx.normalized laplacian matrix(G)
28 e = np.linalg.eigvals(L.toarray())
29 nx.draw_networkx(G, node_size=10, with_labels=False)
30 plt.savefig("ChG Graph.png")
31 plt.show()
33 plt.hist(e, bins = 100)
34 plt.title("Distribution of Eigenvalues (Drug Target Interaction Network)")
35 plt.xlabel("Eigenvalue")
36 plt.savefig("ChG.png")
37 plt.show()
```

https://github.com/huidr/spectral-graph-analysis/chg.py

Plotting the Graph



Plotting the Histogram of the Eigenvalues



The dataset of the two biological networks we used are available at:

https://snap.stanford.edu/biodata/

All of our work is on GitHub.

https://github.com/huidr/spectral-graph-analysis

#### References



Introduction to Spectral Graph Theory. http://users.cms.caltech.edu/vidick/notes/CMS139/spectral.pdf