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

1. Take a glimpse at some graph theory models in biology.

Objectives

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

Contents

1. Preliminaries: Graph Theory

Contents

1. Preliminaries: Graph Theory
2. Graph Theory Models in Biological Networks

Contents

1. Preliminaries: Graph Theory
2. Graph Theory Models in Biological Networks
3. Spectral Graph Theory

Contents

1. Preliminaries: Graph Theory
2. Graph Theory Models in Biological Networks
3. Spectral Graph Theory
4. Two Biological Networks
    - Tissue-specific protein-function associations
    - Drug-target interaction network

Preliminaries: Graph Theory

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.

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.

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

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.

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.

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.

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).

Preliminaries: Graph Theory

Types of Graphs

1. Undirected graph
2. Directed graph (digraph)
3. Weighted graph

Preliminaries: Graph Theory

Distance

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

Preliminaries: Graph Theory

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.

Preliminaries: Graph Theory

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.

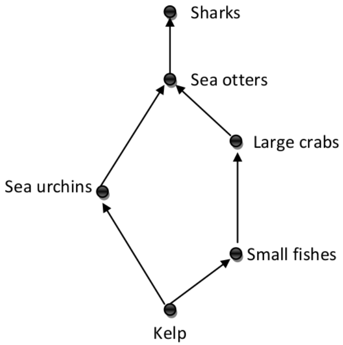
Graph Theory Models in Biological Networks

Food Web

Directed Graph

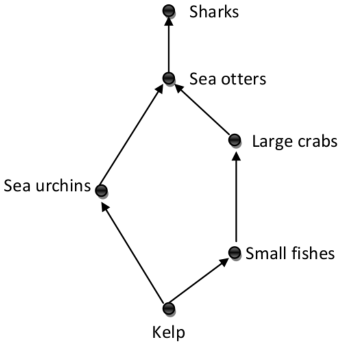
Food Web

Directed Graph



Food Web

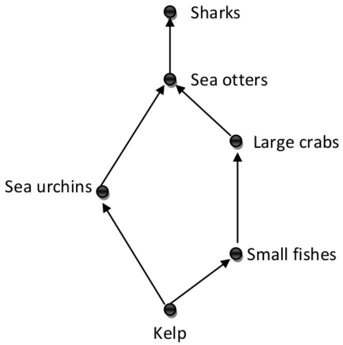
Directed Graph



1. Arrows represent energy flow

Food Web

Directed Graph



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

Food Web

Directed Graph

We can ask:

1. What are the trophic levels?

Food Web

Directed Graph

We can ask:

1. What are the trophic levels?
2. What is a dominant species in a food web?

Food Web

Directed Graph

We can ask:

1. What are the trophic levels?
2. What is a dominant species in a food web?
3. Can we assign weights to the edges?

Food Web

Trophic status

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

Food Web

Trophic status

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

Food Web

Trophic status

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

$$T(u) = \sum_k kn_k,$$

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

Food Web

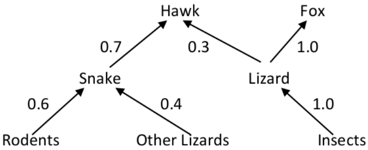
Dominant Species and Weighted Food Webs

1. 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.

Food Web

Dominant Species and Weighted Food Webs

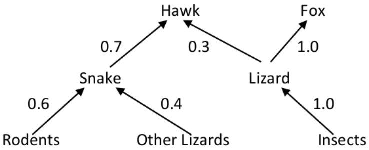
1. A species is *dominant* in a food web if its *trophic status* is greater than the number of species in the food web above level 0.



Food Web

Dominant Species and Weighted Food Webs

1. 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.



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

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

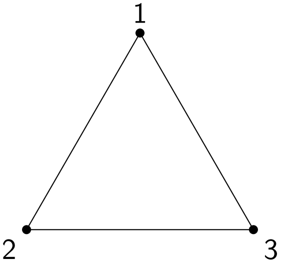
Other models

1. Protein-protein interactions (PPI)
2. Gene-protein interactions
3. Gene regulatory networks (GRN)
4. Signal transduction networks
5. Metabolic and biochemical networks

Spectral Graph Theory

From Graphs to Matrices

Adjacency Matrix



$$\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

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.

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.

Definition (Degree matrix)

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

From Graphs to Matrices

Normalized adjacency matrix

Definition (Normalized adjacency matrix)

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

From Graphs to Matrices

Normalized adjacency matrix

Definition (Normalized adjacency matrix)

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

1. $\overline{A}$ is not necessarily symmetric. But it is column-stochastic, that is, each column has nonnegative entries that sum to 1.

From Graphs to Matrices

Normalized adjacency matrix

Definition (Normalized adjacency matrix)

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

1. $\overline{A}$ 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!

From Graphs to Matrices

Laplacian matrix

Definition (Laplacian matrix)

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

From Graphs to Matrices

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

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

From Graphs to Matrices

Laplacian matrix

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

From Graphs to Matrices

Laplacian matrix

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

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

From Graphs to Matrices

Normalized Laplacian matrix

Proposition

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

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

From Graphs to Matrices

Normalized Laplacian matrix

Proposition

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

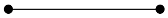
$$x^T \overline{L} x = \sum_{\{i,j\} \in E} \left(\frac{x_i}{\sqrt{d_i}} - \frac{x_j}{\sqrt{d_j}} \right)^2 .$$

If G is d-regular, then this simplifies to

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

Eigenvalues and Eigenvectors of the Laplacian

A single edge

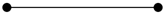


The adjacency matrix and degree matrix are

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

Eigenvalues and Eigenvectors of the Laplacian

A single edge



The adjacency matrix and degree matrix are

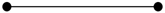
$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \text{ and } D = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

The Laplacians are

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

Eigenvalues and Eigenvectors of the Laplacian

A single edge



The adjacency matrix and degree matrix are

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

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

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

Eigenvalues and Eigenvectors of the Laplacian

A triangle graph

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

Eigenvalues and Eigenvectors of the Laplacian

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

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

Eigenvalues and Eigenvectors of the Laplacian

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

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

Eigenvalues and Eigenvectors of the Laplacian

The path of length 2

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

Eigenvalues and Eigenvectors of the Laplacian

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

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

Eigenvalues and Eigenvectors of the Laplacian

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 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 λ_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 λ_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 $\lambda_2(\overline{L}) > 0$ if and only if G is connected.

The Laplacian

Properties

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

$$0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n :$$

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

The Laplacian

Properties

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

$$0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n :$$

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

The Laplacian

Properties

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

$$0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n :$$

- $\overline{L}$ 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 Laplacian

Properties

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

$$0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n :$$

- $\overline{L}$ 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.

Graph Connectivity

Using Eigenvalues of the Laplacian

The second smallest eigenvalue λ_2 tells us about the connectivity of the graph.

Graph Connectivity

Using Eigenvalues of the Laplacian

The second smallest eigenvalue λ_2 tells us about the connectivity of the graph.

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

Graph Connectivity

Using Eigenvalues of the Laplacian

The second smallest eigenvalue λ_2 tells us about the connectivity of the graph.

1. If G has two disconnected components, $\lambda_2 = 0$.
2. If λ_2 is small, this suggests the graph is nearly disconnected.

Graph Connectivity

Using Eigenvalues of the Laplacian

The second smallest eigenvalue λ_2 tells us about the connectivity of the graph.

1. If G has two disconnected components, $\lambda_2 = 0$.
2. If λ_2 is small, this suggests the graph is nearly disconnected.
3. In short, λ_2 gives us a sort of continuous measure of how well a graph is connected.

Two Biological Systems

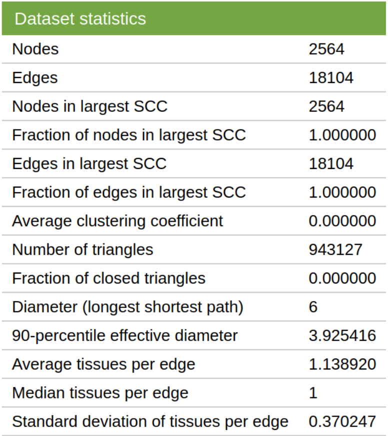
Protein-function associations

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.

Protein-function associations

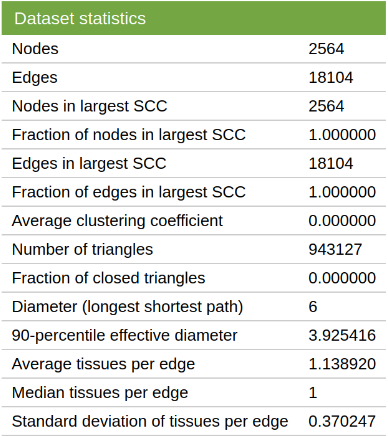
Tissue-specific protein-function associations



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

Protein-function associations

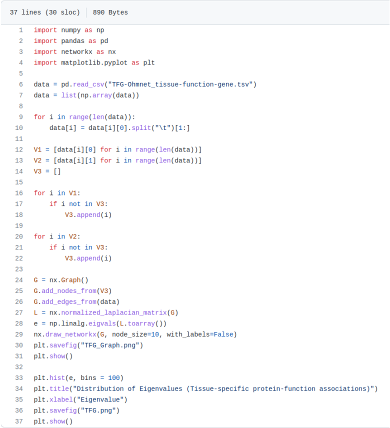
Tissue-specific protein-function associations



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

Protein-function associations

Code on GitHub



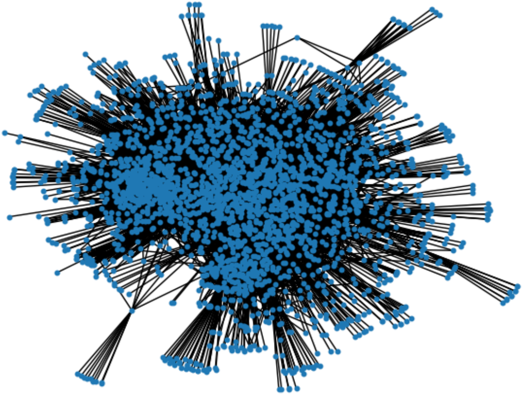
37 lines (30 sloc) | 890 Bytes

```python
import numpy as np
import pandas as pd
import networkx as nx
import matplotlib.pyplot as plt

data = pd.read_csv("TFG-Ohmnet_tissue-function-gene.tsv")
data = list(np.array(data))

for i in range(len(data)):
    data[i] = data[i][0].split("\t")[1:]

V1 = [data[i][0] for i in range(len(data))]
V2 = [data[i][1] for i in range(len(data))]
V3 = []

for i in V1:
    if i not in V3:
        V3.append(i)

for i in V2:
    if i not in V3:
        V3.append(i)

G = nx.Graph()
G.add_nodes_from(V3)
G.add_edges_from(data)
L = nx.normalized_laplacian_matrix(G)
e = np.linalg.eigvals(L.toarray())
nx.draw_networkx(G, node_size=10, with_labels=False)
plt.savefig("TFG_Graph.png")
plt.show()

plt.hist(e, bins = 100)
plt.title("Distribution of Eigenvalues (Tissue-specific protein-function associations)")
plt.xlabel("Eigenvalue")
plt.savefig("TFG.png")
plt.show()
```

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

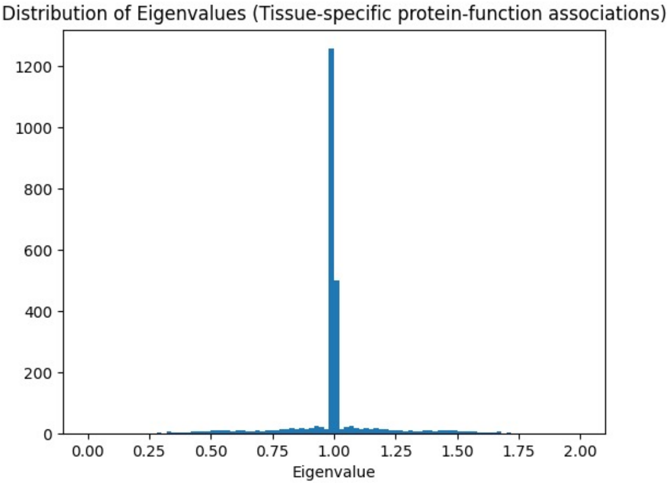
Protein-function associations

Plotting the Graph



Protein-function associations

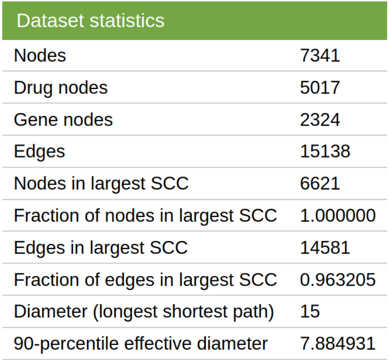
Plotting the Histogram of the Eigenvalues



Drug-target interaction network

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.

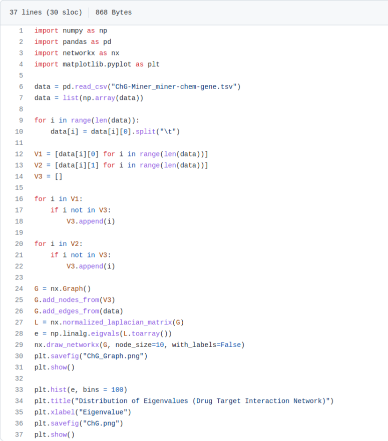
Drug-target interaction network



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

Drug-target interaction network

Code on GitHub



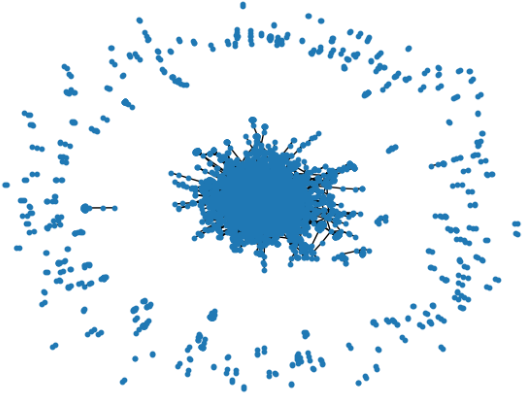
37 lines (30 sloc) | 868 Bytes

```python
import numpy as np
import pandas as pd
import networkx as nx
import matplotlib.pyplot as plt

data = pd.read_csv("ChG-Miner_miner-chem-gene.tsv")
data = list(np.array(data))

for i in range(len(data)):
    data[i] = data[i][0].split("\t")

V1 = [data[i][0] for i in range(len(data))]
V2 = [data[i][1] for i in range(len(data))]
V3 = []

for i in V1:
    if i not in V3:
        V3.append(i)

for i in V2:
    if i not in V3:
        V3.append(i)

G = nx.Graph()
G.add_nodes_from(V3)
G.add_edges_from(data)
L = nx.normalized_laplacian_matrix(G)
e = np.linalg.eigvals(L.toarray())
nx.draw_networkx(G, node_size=10, with_labels=False)
plt.savefig("ChG_Graph.png")
plt.show()

plt.hist(e, bins = 100)
plt.title("Distribution of Eigenvalues (Drug Target Interaction Network)")
plt.xlabel("Eigenvalue")
plt.savefig("ChG.png")
plt.show()
```

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

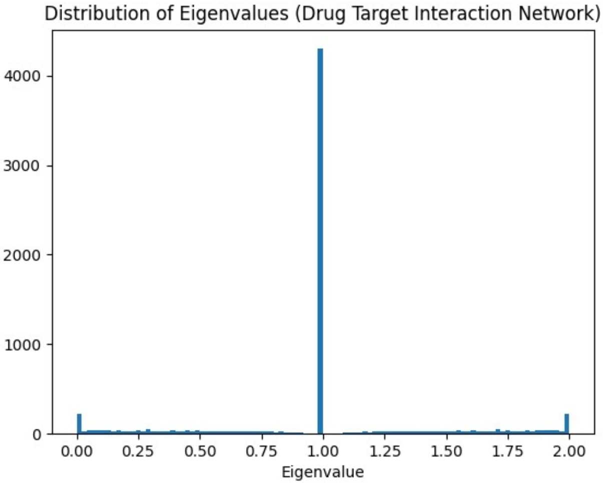
Drug-target interaction network

Plotting the Graph



Drug-target interaction network

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

M. Cozzens. *Food Webs, Competition Graphs, and Habitat Formation*. Math. Model. Nat. Phenom., Vol. 6, No. 6, 2011, pp. 22-38.

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