Exploring Applications of Graphs Theory in Biology

Kangsheng Qi[1, 2]

[1] *Department of Computer Science, The University of Texas at Austin*
[2] *Department of Molecular Biosciences, The University of Texas at Austin*

January 4, 2025

Abstract

Biological networks capture a wide range of interactions and relationships between biological entities, and have proven valuable for gaining insights and addressing key challenges. As the size and complexity of these networks continue to grow, understanding their potential applications becomes increasingly important. In this manuscript, we explore the integration of graph theory into computational biology and demonstrate several applications of this approach. We aim to highlight the usefulness of graph-based methods in biological research and motivate further advancements in this field.

1 Introduction

Mathematical graphs were originally developed to describe and visualize relationships between variables or individual objects, and numerous methods and frameworks have been designed with that goal in mind. In recent years, graph theory concepts and techniques have been applied across a range of computer science-related fields, such as the analysis of social networks, internet networks, electrical circuits, and many other domains [1, 2, 3].

Advances in next-generation sequencing technology have enabled the mapping of large and complex biological networks. These networks represent molecular and protein interactions, gene expression correlations, regulatory relationships, metabolic pathways, protein structures, and other biological systems. The growing availability and complexity of biological network data have prompted computational biologists to tackle problems through the framework of discrete biological networks [4].

This has raised important questions and discussions regarding how biological networks and their analysis can contribute to answering key biological questions. In this manuscript, we provide an overview of fundamental graph theory as it is applied in biological contexts. We also explore several prominent applications of graph theory in biology.

We provide a brief outline for the rest of the manuscript. We remark that this may get continuously updated as the manuscript get refined and extended.

- Section 2 provides a foundation in graph theory, including key notations and commonly used graph models in biological analysis.
- Section 3 focuses on the application of graph theory to protein and gene function prediction.
- Section 4 focuses on the application of graph theory to evolution-related studies.
- Section 5 focuses on the application of graph theory to metabolic networks.
- Section 6 presents a discussion of the current state of the field and potential directions for future research.
- Section 7 concludes the manuscript.

2 Graph Theory Foundation

In this section, we discuss key graph theory metrics and tools commonly used in biological analysis.

- Section 2.1 introduces the standard graph notation.
- Section 2.2 outlines several commonly used metrics for analysis.
- Section 2.3 outlines various popular graph models employed in analysis.

2.1 Graph Notation

In this section, we briefly introduce the graph theory notation and fundamental concepts used throughout the manuscript, most of which follow standard practices in the literature. A graph is *undirected* if its edges do not have directions, and *directed* if they do. Similarly, a graph is *weighted* if its edges are assigned different numerical values, and *unweighted* if all edges share the same value.

We denote a graph $G := (V, E, W)$, where V represents the set of vertices, E represents the set of edges, and W represents the set of weights assigned to the edges. We let $n := |V|$ denote the number of vertices, and $m := |E| = |W|$ denote the number of edges. In the case of an unweighted graph, all elements of W would equal 1, and the notation can be simplified to $G := (V, E)$. We say a graph $G' = (V', E', W')$ is a subgraph of the graph G if $V' \subseteq V$, $E' \subseteq E$, and all edges in E' inherit their weights from E. Further, G' is an *induced subgraph* if for all edges $(u, v) \in V' \times V'$: $(u, v) \in E' \iff (u, v) \in E$.

Two vertices u and v of G are said to be *reachable* if there exists a path between them. In an undirected graph, any sequence of edges connecting u to v suffices. In a directed graph, however, the edges must follow a consistent direction from u to v. A undirected graph is *connected* if there exists a path between every pair of vertices. A directed graph is *connected* if for all vertices $u, v \in V$, u can reach v and vice versa. Given this, we can partition G into subsets of *connected components*, where each connected component is a connected induced subgraph of G.

A graph G can be represented using various methods and data structures. One common representation is through the **adjacency matrix**, $\mathbf{A} \in \mathbb{R}^{m \times m}$, which is defined as

$$\mathbf{A}_{u,v} := \begin{cases} w_{(u,v)} & \text{if } (u,v) \in E \\ 0 & \text{else} \end{cases}$$

for $u, v \in [m]$. For undirected graphs, $\mathbf{A}$ is symmetric, and only the upper or lower diagonal half needs to be stored to save space.

For an undirected graph, the degree of a vertex v is defined as $\deg(v) := \sum_{(v,u) \in E} w_{(v,u)}$. For an *unweighted* graph, the degree of a vertex v is the number of vertices adjacent to v. We define the diagonal matrix $\mathbf{D} \in \mathbb{R}^{m \times m}$ as the **degree matrix** of G such that

$$\mathbf{D}_{u,v} := \begin{cases} \deg(v) & \text{if } u = v \\ 0 & \text{if } u \neq v \end{cases}$$

For directed graph, we define $\deg_{\text{out}}(v)$ as the sum of the weights of the edges leaving v, and $\deg_{\text{in}}(v)$ as the sum of the weights for the edges directed towards v. Similarly, we define two diagonal matrices, $\mathbf{D}_{\text{in}}$ and $\mathbf{D}_{\text{out}}$, for the in-degrees and out-degrees, respectively.

For an undirected graph, the **Laplacian matrix** $\mathbf{L} \in \mathbb{R}^{m \times m}$ follows naturally and is defined as $\mathbf{L} := \mathbf{D} - \mathbf{A}$. For directed graphs, we define analogous Laplacians, $\mathbf{L}_{\text{in}}$ and $\mathbf{L}_{\text{out}}$, based on in-degrees and out-degrees. The Laplacian matrix has the important property of being positive semi-definite, meaning that all of its eigenvalues are non-negative . The **normalized Laplacian matrix** is defined as $\widehat{\mathbf{L}} := \mathbf{D}^{-\frac{1}{2}}(\mathbf{D} - \mathbf{A})\mathbf{D}^{-\frac{1}{2}}$.

In addition to storing graphs as matrices, a graph G can also be represented using an **adjacency list**. For an undirected graph, the adjacency list consists of a list L_V of length n where each element corresponds to a vertex u in G and contains a list of adjacent vertices v. For a directed graph, two adjacency lists, L_{in} and L_{out} is needed to record the incoming edges and outgoing edges for each vertex. The adjacency list representation offers advantages in terms of space efficiency for sparse graphs and easier implementation for graph traversal algorithms like depth first search (DFS) and breadth first search (BFS).

For unweighted graphs, we say graph G and graph $H := (U, F)$ are isomorphic if and only if $|V| = |U|$ and there exists a bijective function $\pi : V \to U$ such that $(v, v') \in E \iff (\pi(v), \pi(v')) \in F$. In other words, G and H are isomorphic if there is a one-to-one correspondence between their vertices and edges. For weighted graphs, the weights of the graphs adds additional complexity in order to claim isomorphic. We claim G and H have a *node correspondence* if there exists a bijective function $f : V \to U$ that maps each $v \in V$ to a vertex $v' \in U$.

2.2 Graph Theory Metrics

To provide a quantitative analysis of a graph, numerous metrics have been developed to lend insights into the structural organization, topology, and characteristics of the query graph. These metrics can largely be separated into focusing on the global features of the graph or focusing on the local features.

Global Focused

A key tool for global analysis of a graph is **average degree**, which is calculated as $\frac{\sum_{v \in V} \deg(v)}{m}$, and the **degree distribution**, which describes how the degrees of vertices in the graph are distributed. The degree distribution can be transformed into a probability distribution and fitted to various known distributions. A well-known example of this type of analysis is the study of **scale-free networks**, whose degree distribution follows a power law. This will be discussed further in Section 2.3.

Another widely used measure is the **density** of the graph. It is computed as $\text{density}(G) = \frac{2m}{n(n-1)}$. Graph density provides insight into evolutionary studies, which will be explored in Section 4.

Many network metrics can be defined both at the global level and for local subgraphs. For example, the **global clustering coefficient** measures the tendency of the entire graph to form clusters, which are densely connected subsets of vertices. The first attempt to calculate this measure was made by R. Duncan Luce and Albert Perry [5]. Their work focuses on the concept of triplets, which can be either open triplets (three vertices connected by two edges) or closed triplets (three vertices connected by three edges). The clustering coefficient is calculated as the number of closed triplets divided by the total number of triplets. This metric, however, loses nuance for weighted graphs, where edge weights—important for understanding network structure—are ignored. For weighted graphs, the global clustering coefficient can instead be calculated as the sum of the values of all closed triplets divided by the sum of the values of all triplets [6].

Other instances include **average shortest path** and **average diameter** (longest shortest path), both of which can be calculated via Dijkstra's Algorithm [7] or Bellman-Ford Algorithm [8]. Both of these metrics help to ascertain the connectedness and speed of information prorogation of the network.

Modularity Q [9] is a measure that evaluates how effectively a network's partition separates its vertices into modules, where the vertices within each module are densely connected. Essentially, modularity compares the distribution of edges in the given network to that of a randomly generated network. Several variations of modularity calculations exist, including a method that uses eigenvectors [10]. A commonly used formula [11] is $Q := \frac{1}{2m} \sum_{u,v \in V} (\mathbf{A}_{u,v} - \frac{\deg(u) \cdot \deg(v)}{2m}) \cdot \mathbb{I}_{u,v}$, where $\mathbb{I}_{u,v}$ is an indicator function that returns 1 if u, v are in the same partition and 0 otherwise. This concept can be extended to subgraphs or connected components of a network, making it a locally focused measure. The relationship between modular networks and the concept of modularity will be further explored in Section 2.3.

Centrality is another important metric, used to assess the significance of vertices within the network. **Centralization** quantifies how close the graph's structure is to a star topology. It is calculated as $\frac{n}{n-2}(\frac{\max \deg(v \in V)}{n-1} - \text{density}(G))$ [12]. Higher centralization indicates that a few vertices are highly connected, acting as hubs within the network.

Cluster analysis is another common approach for analyzing the entire network. It involves applying clustering algorithms to group similar vertices and partition the network into distinct clusters [12, 13]. The choice of the clustering algorithm is critical to the quality of the results.

Spectral analysis is a major approach employed to analyze graphs, which explores the graph through the eigenvalues and eigenvectors of typically the Laplacian matrix and the adjacency matrix. These analyses are mathematically rigorous, and are informative of the structure, connectedness, and similarity between graphs. Interested readers can refer to these works [14, 15] for more information.

Local Focused

Local analysis focuses on a specific subgraph, connected component, or even an individual vertex, providing more detailed insights into the structure and function of the network. A key area of local analysis is centrality, which can be defined in various ways. **Degree centrality** measures the significance of the vertex in terms of interactions with other vertices, and it is simply calculated as $C_{\text{degree}}(u) := \deg(u)$. Vertices with high degree centrality are often referred to as hubs. **Closeness centrality** [16] quantifies how efficiently a vertex can spread information to other vertices within the network. It is calculated as $C_{\text{close}}(u) := \frac{1}{\sum_{v \in V} \text{dist}(u,v)}$, where $\text{dist}(u, v)$ represents the shortest path distance between vertices u and v. Closeness centrality is particularly useful in evolutionary studies and in metabolic networks, where it helps identify core metabolites [17]. **Eccentricity centrality** measures the accessibility of a vertex by considering its farthest distance from any other vertex in the network. It is calculated as $C_{\text{ecc}}(u) := \frac{1}{\max_{v \in V} \text{dist}(u,v)}$, which is the reciprocal of the largest shortest path. This metric provides insight into the relative importance of a vertex in terms of its reachability.

Betweenness centrality [18] also relies on the concept of shortest paths. It measures the number of shortest paths between pairs of vertices that pass through the query vertex. It is calculated as $C_{\text{betw}}(u) := \sum_{(u \neq v \neq w) \in V} \frac{\sigma_{v,w}(u)}{\sigma_{v,w}}$, where $\sigma_{v,w}$ is the total number of shortest paths between v and w, and $\sigma_{v,w}(u)$ is the number of those paths that pass through u. Vertices with high betweenness centrality often act as "bottlenecks" in the network, as they are crucial for information propagation between many pairs of vertices [12]. This centrality is useful in metabolic network analysis [19] and protein-protein interaction network analysis [20]. **Eigenvector centrality** [21], first introduced by Edmund Landau, at the age of 18, in 1895 to rank chess tournament players, measures the influence of a vertex based on the centrality of its neighbors [22]. For an undirected graph, eigenvector centrality C_{eigv} is computed by first finding the largest eigenvalue $|\lambda_{\max}|$ and its associated eigenvector C_{eigv} such that $\mathbf{A}C_{\text{eigv}} = \lambda_{\max} C_{\text{eigv}}$. The eigenvector centrality for each vertex is the corresponding value in C_{eigv}. Eigenvector centrality is utilized in the Google Page Rank Algorithm [22], and within biology, this metric is ideal for identifying hubs within protein interaction networks [23]. Finally, **subgraph centrality** [24] is another measure of vertex importance that considers the number of closed walks originating from the vertex, which corresponds to the vertex's participation in subgraphs. It is computed as $C_{\text{sub}}(u) := \sum_{k=0}^{\infty} \frac{(\mathbf{A}^k)_{u,u}}{k!}$. Subgraph centrality is particularly useful for understanding the molecular structure of proteins and other biomolecules.

The **vertex matching index** quantifies the similarity between two vertices by comparing their neighbors, which can be extended to compare vertices across multiple graphs, particularly when there is a known node correspondence. This metric can be especially helpful in multi-graph analysis.

2.3 Prominent Graph Models and Engineered Networks

To understand statistical properties and characteristics of real-world networks, mathematicians have developed numerous graph models that captures some significant property which are informative.

We will describe several of these models that are useful in biology applications.

Erdős–Rényi

One of the most well-known graph models is the Erdős-Rényi Model [25]. developed by Paul Erdős and Alfréd Rényi in 1961 to represent the properties of random graphs. In this model, edges between vertices are randomly created with a probability p, independently of one another. The degree distribution of the graph largely follows a binomial distribution, with the exact range depending on p. Given this, the probability of the degree of a vertex is

$$\Pr[\deg(v) = k] = \binom{n-1}{k} p^k (1-p)^{n-k-1} \approx \frac{(np)^k e^{-np}}{k!}$$

The Erdős-Rényi model is critical for testing whether a real-world graph arises from randomness or if some underlying biological factors are driving the observed structure. In this model, the probability p must be chosen carefully, as it significantly affects subsequent analyses. Notably, the product of the number of vertices and p provides a rough estimate of the number of connected components in the graph [26].

Watts-Strogatz

The **small-world** property describes networks with a high clustering coefficient and small average path length between vertices. In 1998, Duncan Watts and Steven Strogatz introduced a model characterized by this property [27]. The Watts–Strogatz model is defined by three parameters: n (number of vertices), k (mean degree of each vertex), and p (rewiring probability). To construct the graph, vertices are first organized into a ring lattice, where each vertex is connected to k neighbors, half on either side, such that $(u, v) \in E \iff 0 < |u - v| \mod (n - \frac{k}{2} - 1) \leq \frac{k}{2}$. Then, for each vertex, half of its rightmost neighbors are rewired to random vertices with probability p. The Watts–Strogatz model exhibits several notable properties: the degree distribution approximates a delta function centered around k, and it maintains the small-world property even as the network grows in size [28].

Networks with small-world properties are common in both biology and real-world systems. Social media networks, for example, exhibit small-world characteristics [29], as do metabolic networks in biological systems [30]. In this model, the parameters p and k are critical for later analyses.

Barabási–Albert

The Barabási–Albert model, developed by Albert-László Barabási and Réka Albert, captures the property of scale-free networks. In this model, a network is initialized with n_0 vertices with $n >> n_0$. At each step, a new vertex is added and forms edges to existing vertices with a probability proportional to their degree. Formally, for a new vertex, the probability of forming an edge to vertex v is $p_v := \frac{\deg(v)}{\sum_{u \in V'} \deg(u)}$ where V' represents the current set of vertices in the graph.

Given this formulation, it is intuitive to see that vertices with more connections, larger degree, are more likely to attract new edges, a process known as preferential attachment. Interestingly, unlike the Watts–Strogatz model, the clustering coefficient in the Barabási–Albert model is not independent of network size and follows a power-law distribution [31].

The Barabási–Albert is frequently used in biology analysis due to the fact that scale-free network is common in real-world network, this will be discussed in details in the next paragraph.

Scale-free

The concept of scale-free networks was first observed by Derek de Solla Price, who noted that citation networks follow a power-law degree distribution [32]. Barabási and Albert later found similar characteristics in sub-networks of the World Wide Web [33], eventually coining the term "scale-free network." A scale-free network is defined by a degree distribution that follows a power law, $P(k) \sim k^{-\gamma}$, where γ typically lies between 2 and 3.

A key concept related to scale-free networks is **preferential attachment**, where new vertices are more likely to form connections with vertices that already have many connections. This idea is also reflected in the Barabási–Albert model, capturing the essence of scale-free networks.

In biological systems, scale-free networks are widespread. Protein–protein interaction networks [34], metabolic networks [35], and other types of networks [12] exhibit scale-free properties. In fact, Barabási and Albert's research even suggest that the scale-free nature of biological networks may provide resilience against random failures.

Modular Network

Modular networks are characterized by high modularity, meaning that the network can be partitioned into distinct modules or communities. In biology, modularity is important as it allows complex networks to be divided into smaller, more manageable units for detailed analysis. Many biological interaction networks exhibit modular structures [36], and the choice of algorithm for partitioning the network is critical for accurate analysis.

Hierarchical Network

Hierarchical networks build upon scale-free networks by exhibiting both a power-law degree distribution and a hierarchical structure. Hierarchical organization within networks is an active area of research, particularly in biology. For instance, studies have investigated hierarchical structures in metabolic networks [37], which reveal complex layers of interaction and organization within biological systems.

3 Function Prediction

4 Evolutionary Study

5 Metabolic Network Analysis

6 Discussion

7 Conclusion

Acknowledgment

We would like to express our sincere gratitude to Professor Can Cenik for introducing us to the study of biological networks. We also extend our thanks to Professor Kevin Tian for his course and lecture notes on graph-based algorithms, which contributed to this work. The writing of this manuscript was carried out solely by the author, with a language model (LLM) providing suggestions for clarity and grammar.

References

[1] L R Foulds. *Graph Theory Applications*. en. 1st ed. Universitext. New York, NY: Springer, Jan. 1995.

[2] J L Gross, J Yellen, and M Anderson. “Graph theory and its applications”. In: (2018).

[3] Abdul Majeed and I Rauf. “Graph theory: A comprehensive survey about graph theory applications in computer science and social networks”. In: *Inventions* (Feb. 2020).

[4] E Alm and A P Arkin. “Biological networks”. In: *Curr. Opin. Struct. Biol.* (2003).

[5] R D Luce and A D Perry. “A method of matrix analysis of group structure”. en. In: *Psychometrika* 14.2 (June 1949), pp. 95–116.

[6] Tore Opsahl and Pietro Panzarasa. “Clustering in weighted networks”. en. In: *Soc. Networks* 31.2 (May 2009), pp. 155–163.

[7] E W Dijkstra. “A note on two problems in connexion with graphs”. In: *Edsger Wybe Dijkstra*. New York, NY, USA: ACM, July 2022, pp. 287–290.

[8] R Bellman. “ON A ROUTING PROBLEM”. In: *Quarterly of Applied Mathematics* 16 (Apr. 1958), pp. 87–90.

[9] M E J Newman and M Girvan. “Finding and evaluating community structure in networks”. en. In: *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 69.2 Pt 2 (Feb. 2004), p. 026113.

[10] M E J Newman. “Modularity and community structure in networks”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 103.23 (June 2006), pp. 8577–8582.

[11] Aaron Clauset, M E J Newman, and Cristopher Moore. “Finding community structure in very large networks”. en. In: *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 70.6 Pt 2 (Dec. 2004), p. 066111.

[12] Georgios A Pavlopoulos et al. “Using graph theory to analyze biological networks”. en. In: *BioData Min.* 4.1 (Apr. 2011), p. 10.

[13] Asuda Sharma, H Ali, and D Ghersi. “Cluster analysis of biological networks”. In: *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics* (2019), pp. 1036–1046.

[14] Fan R K Chung. *Spectral Graph Theory.* CBMS regional conference series in mathematics. Providence, RI: American Mathematical Society, Dec. 1996.

[15] Bogdan Nica. “A brief introduction to Spectral Graph Theory”. In: *arXiv [math.CO]* (Sept. 2016).

[16] Alex Bavelas. “Communication Patterns in Task-Oriented Groups”. In: *Journal of the Acoustical Society of America* 22 (Nov. 1950), pp. 725–730.

[17] M R da Silva, Hongwu Ma, and An-Ping Zeng. “Centrality, network capacity, and modularity as parameters to analyze the core-periphery structure in metabolic networks”. In: *Proc. IEEE Inst. Electr. Electron. Eng.* 96.8 (Aug. 2008), pp. 1411–1420.

[18] L Freeman. “A set of measures of centrality based upon betweenness”. In: *Sociometry* (Mar. 1977).

[19] Mahendra Piraveenan, Kishan Wimalawarne, and Dharshana Kasthurirathn. “Centrality and composition of four-node motifs in metabolic networks”. en. In: *Procedia Comput. Sci.* 18 (2013), pp. 409–418.

[20] Maliackal Poulo Joy et al. “High-betweenness proteins in the yeast protein interaction network”. en. In: *J. Biomed. Biotechnol.* 2005.2 (June 2005), pp. 96–103.

[21] E Landau. “Zur relativen wertbemessung der turnierresultate”. In: *Deutsches Wochenschach* (1895).

[22] Rainer Sinn and Günter M Ziegler. “Landau on chess tournaments and Google’s PageRank”. In: *arXiv [math.HO]* (Oct. 2022).

[23] Elena Zotenko et al. “Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality”. en. In: *PLoS Comput. Biol.* 4.8 (Aug. 2008), e1000140.

[24] Ernesto Estrada and Juan A Rodríguez-Velázquez. “Subgraph centrality in complex networks”. en. In: *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 71.5 Pt 2 (May 2005), p. 056103.

[25] P Erdős and A Rényi. “On the strength of connectedness of a random graph”. In: *Acta Mathematica Hungarica* 12.1 (1961), pp. 261–267.

[26] P Erd6s and A R nyi. “On the evolution of random graphs”. In: *Publ. Math. Inst. Hungar. Acad. Sci* (1960).

[27] D Watts and S Strogatz. “Collective dynamics of ‘small-world’ networks”. In: *Nature* 393 (June 1998), pp. 440–442.

[28] A Barrat and M Weigt. “On the properties of small-world network models”. en. In: *Eur. Phys. J. B* 13.3 (Jan. 2000), pp. 547–560.

[29] Duncan J Watts. “Networks, dynamics, and the small-world phenomenon”. en. In: *Am. J. Sociol.* 105.2 (Sept. 1999), pp. 493–527.

[30] A Wagner and D A Fell. “The small world inside large metabolic networks”. en. In: *Proc. Biol. Sci.* 268.1478 (Sept. 2001), pp. 1803–1810.

[31] S N Dorogovtsev, A V Goltsev, and J F F Mendes. “Pseudofractal scale-free web”. en. In: *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 65.6 Pt 2 (June 2002), p. 066122.

[32] D J Price. “Networks of scientific papers: The pattern of bibliographic references indicates the nature of the scientific research front”. en. In: *Science* 149.3683 (July 1965), pp. 510–515.

[33] A L Barabasi and R Albert. “Emergence of scaling in random networks”. en. In: *Science* 286.5439 (Oct. 1999), pp. 509–512.

[34] H Jeong et al. “Lethality and centrality in protein networks”. en. In: *Nature* 411.6833 (May 2001), pp. 41–42.

[35] Hawoong Jeong et al. “The large-scale organization of metabolic networks”. In: *Nature* 407 (Oct. 2000), pp. 651–654.

[36] G Wagner, M Pavličev, and J Cheverud. “The road to modularity”. In: *Nat. Rev. Genet.* 8 (Dec. 2007), pp. 921–931.

[37] Erzsébet Ravasz. “Detecting hierarchical modularity in biological networks”. en. In: *Methods Mol. Biol.* 541 (2009), pp. 145–160.