

Assignment-01

Question 1: For each node compute the degree and identify top 10 hub nodes.

Solution:

Number of Nodes: 27848

Number of Edges: 1134670

From the Analysis, Top 10 Hub Nodes (Highest Degree) are:

Protein 43740578: Degree 12000

Protein 5071: Degree 5954

Protein 7157: Degree 5398

Protein 1956: Degree 4889

Protein 4609: Degree 4510

Protein 54764: Degree 4413

Protein 8452: Degree 4049

Protein 3845: Degree 4037

Protein 899: Degree 3695

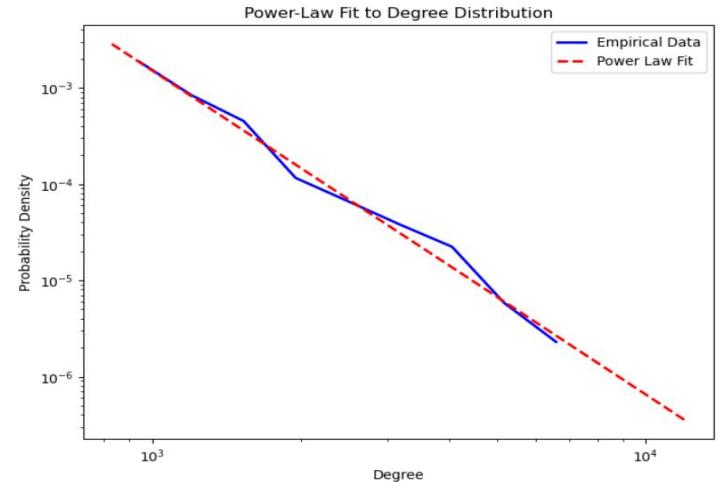
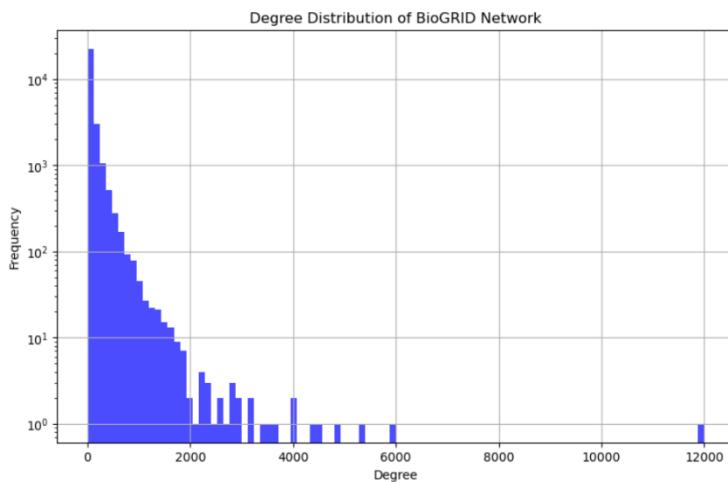
Protein 7316: Degree 3585

Question 2: Draw the graph of degree distribution to see if it is a scale free network or not?

Solution:

The degree distribution analysis confirms the network is scale-free:

- The histogram shows characteristic power-law distribution
- Power-law exponent (gamma): 3.367103700361787
- Statistical comparison ($R=26.371099419923468$, $p=0.030227429899136852$) strongly supports scale-free structure
- The power-law fit graph shows excellent alignment between empirical data (blue line) and theoretical fit (red dashed line)
- Yes, The BioGRID network is **scale-free**, indicated by the power-law distribution of node degrees.

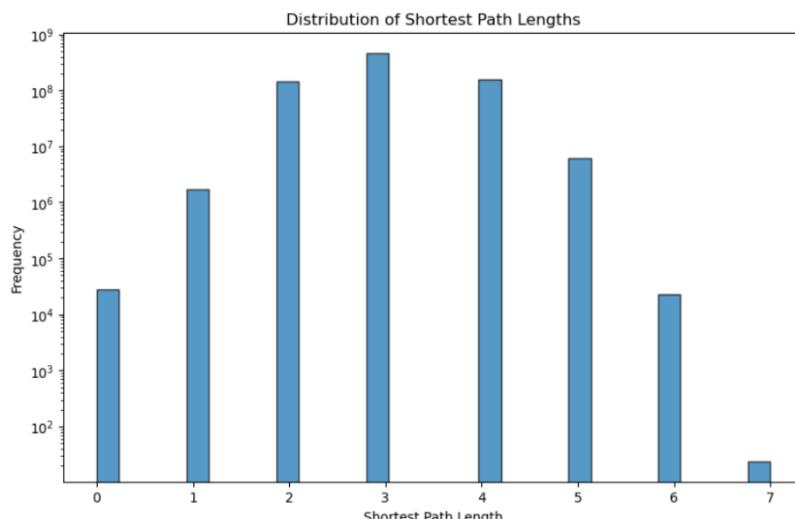


Question 3: Compute the shortest path for all node pairs and plot the distribution of shortest path length.

Solution:

The shortest path distribution shows:

- Path lengths range from 0 to 7
- Peak frequency occurs maximum at path length 3
- Distribution follows a bell-shaped curve on log scale
- Total paths analyzed: 775,344,028
- The network shows efficient connectivity with most paths being relatively short



Question 4: For each node, compute its betweenness and closeness, and identify top 10 nodes with highest betweenness and closeness respectively.

Solution:

Top 10 Nodes by Betweenness Centrality:

[(24833, 22744355.794518538), (1988, 18075501.634862106),
(8254, 17208990.83000888), (147, 16578083.77738637),
(2260, 11924420.185736349), (101, 10340120.092920352),
(6611, 9897384.060843017), (437, 9745672.422829254),
(673, 8891632.72300223), (17, 8662558.25781677)]

Top 10 Nodes by Closeness Centrality:

[(8199, 0.4912231180424466), (24833, 0.4891777933942375),
(2260, 0.4858404145800981), (101, 0.4829751435361052),
(739, 0.47891296869625044), (1077, 0.4767889861127759),
(5943, 0.4759495401866603), (437, 0.4750968314365178),
(10196, 0.473472996871174), (6611, 0.4688331368917326)]

Dataset and Tools Used:

Dataset: BioGRID Human PPI dataset (version 4.4.218)

Programming Language Used: Python

Libraries Used: pandas, igraph, matplotlib, seaborn, powerlaw

GitHub Link: https://github.iu.edu/msiddhe/Systems-Biology_Assignment---01.git