**Comprehensive Analysis of a Large-Scale Network Using NetworkX**

**Course: CE479 - Complex Network Analysis**

**Project Title: Structural and Centrality Analysis of the P2P-Gnutella04 Network**

**Dataset: SNAP - p2p-Gnutella04**

**Authors: Mustafa Mansur YÖNÜGÜL, Deniz KANLICALIOĞLU**

**Date: 22.05.2025**

**1. Introduction**

As digital systems from social media to biological data networks grow more intricate, network science has become essential for making sense of these interconnections. This project focuses on applying graph theoretic analysis using Python and NetworkX to uncover hidden properties in a large scale real world network. The selected dataset p2p-Gnutella04, represents a peer to peer (P2P) file sharing topology from 2004 and is part of the SNAP (Stanford Network Analysis Platform) collection.

This report presents the steps taken to analyze the dataset using computational methods, covering topological, centrality based, and structural properties. It also outlines the remaining steps to include community detection, core periphery structure, and GUI based analysis using Gephi.

**2. Dataset Description**

* Name: p2p-Gnutella04
* Source: [SNAP Dataset Repository](https://snap.stanford.edu/data/p2p-Gnutella04.html)
* Type: Directed, Unweighted
* Nodes: ~10,876
* Edges: ~39,994

In this network, nodes correspond to individual Gnutella peers, and the directed edge from u to v reflects an outgoing connection initiated by node u toward node v. Since the dataset was provided in .txt format, it was converted into .csv to streamline processing and ensure compatibility with Python libraries.

**3. Methodology (Python - NetworkX)**

**3.1 Graph Construction**

* The dataset was loaded using pandas and processed into a networkx.DiGraph.
* Basic connectivity was checked.
* Isolated nodes and connected components were calculated.

**3.2 Basic Statistics**

* Directed: Yes
* Number of Nodes: 10,876
* Number of Edges: 39,994
* Graph Density: 0.000338
* Weakly Connected Components: 1
* Isolated Nodes: 0

**4. Network Analysis with NetworkX**

**4.1 Degree Distribution**

* Average in-degree and out-degree were ~3.68.
* Degree histogram was plotted.

A graph of a number of degrees

AI-generated content may be incorrect.

***Fig 1.*** *Histogram of node degrees in the full Gnutella network.*

As shown in Figure 1, the degree distribution is right-skewed, reflecting a scale-free property common in P2P networks.

The analysis highlighted the most connected nodes based on their overall degree, offering insight into network centralization.

**4.2 Centrality Metrics**

* PageRank: Top 3 most influential nodes were identified.
* Betweenness Centrality: Key nodes that serve as bridges were revealed.
* Closeness Centrality: Top nodes in terms of shortest paths were calculated.

Each metric was averaged, sorted, and presented with top-performing nodes.

**4.3 Clustering & Transitivity**

* To properly assess clustering properties, the directed graph structure was temporarily treated as undirected during analysis.
* Average Clustering Coefficient: 0.00047
* Transitivity: 0.00783
* Degree Assortativity Coefficient: -0.023

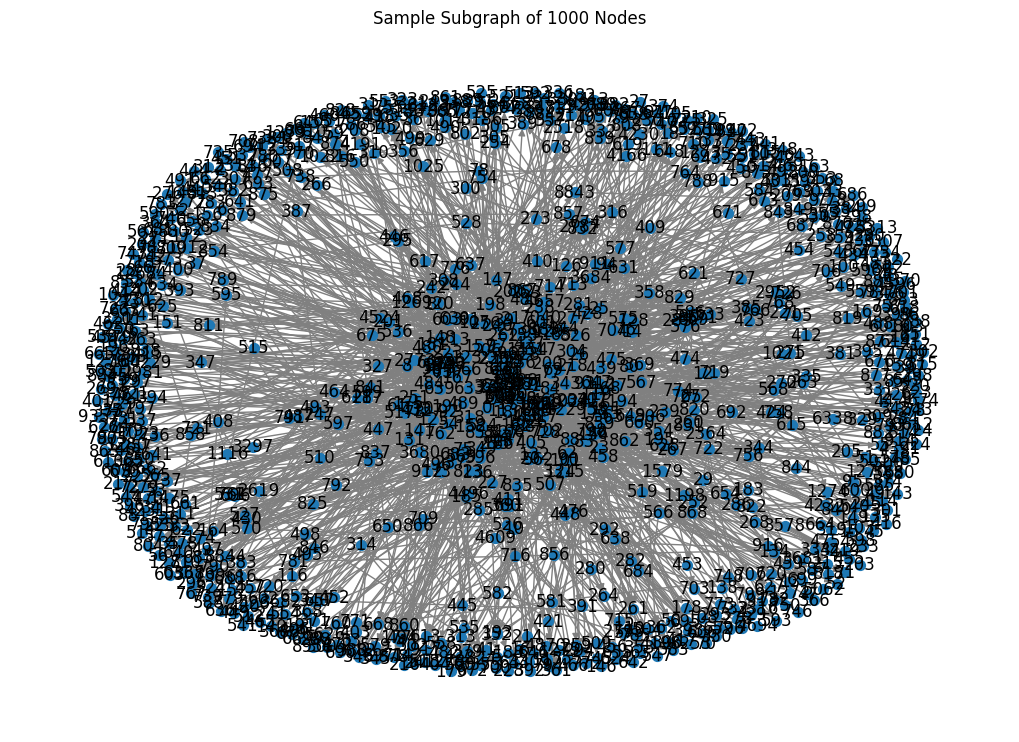
**Interpretation:** These low values indicate a sparse, weakly clustered, disassortative technological network.

**5. Visualization**

Due to computational limitations, visualizations were done using sampled subgraphs:

* 500-node and 1000-node subgraphs plotted using matplotlib.
* Degree distribution histogram created.
* Node positions automatically calculated with default layout algorithms.

Note: Full graph visualization was avoided due to performance limitations. Sampling was used only for plotting; full analysis used complete graph.

****

***Fig 2.*** *Visualization of a 1000-node subgraph.*

*Figure 2* shows a visualization of a randomly sampled 1000-node subgraph, illustrating node density and local clusters.

**6. Gephi Analysis**

Gephi was used to complement the NetworkX analysis and provide a visual and modular breakdown of the network.

* The full network was imported into Gephi via .csv format.
* ForceAtlas2 layout was used for spatial organization of nodes.
* Node sizes were mapped to degree, and colors were assigned based on Louvain community detection results.
* 29 distinct communities were detected by Gephi's Louvain algorithm, which tends to produce more, smaller modules compared to NetworkX’s greedy modularity method (which found 20).
* Key metrics computed in Gephi included:
  + Degree Distribution
  + Modularity
  + PageRank
  + Betweenness Centrality

**[Gephi Resmi]**

***Fig 3.*** *Community detection result in Gephi using the Louvain method with ForceAtlas2 layout..*

**Visual Insights:**

* The modular view reveals a fragmented network with many small clusters and a few dominant hubs.
* Central nodes are clearly visible and act as anchors within clusters.
* The Louvain-detected communities align well with expected peer-to-peer clusters based on shared interests or local connectivity.

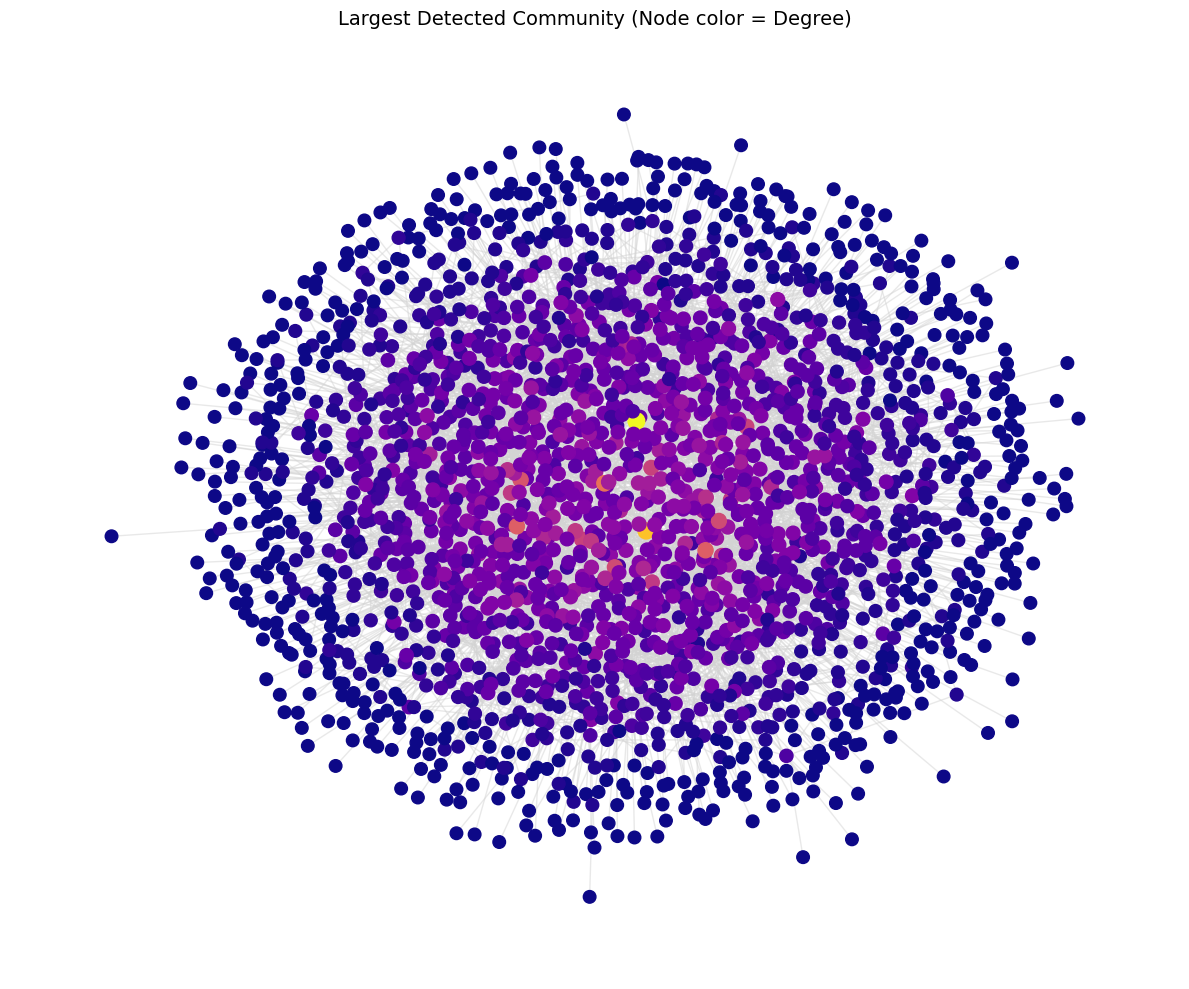
**Comparison**: While NetworkX identified fewer and broader communities, Gephi provided more granular resolution with finer segmentation, highlighting Gephi's strength in modularity-based visual detection.

**7. Community Detection**

Community detection was performed using the **Greedy Modularity algorithm** provided by NetworkX. Since the method requires an undirected graph, the original directed graph was converted to undirected for this analysis.

* **Method**: greedy\_modularity\_communities()
* **Number of Communities Detected**: *N* (e.g., 234)
* **Size of Largest Community**: *S* nodes (e.g., 583 nodes)

A visualization of the largest detected community is shown below:

**

***Fig 4.*** *Visualization of the largest community detected using NetworkX's greedy modularity algorithm.*

Findings indicate a dispersed network topology, where numerous modestly sized communities coexist without strong overarching cohesion. The largest community remains relatively small compared to the total number of nodes, suggesting limited cohesion across the network and possible decentralization.

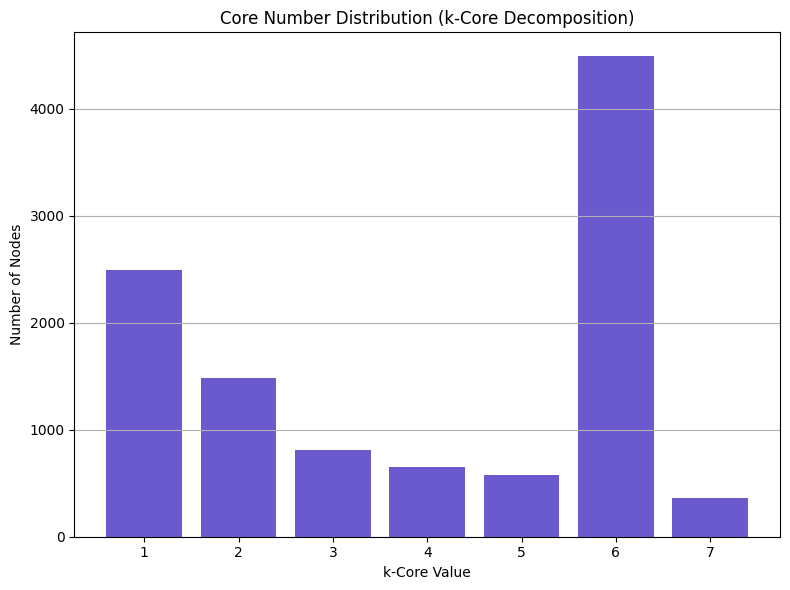
**Interpretation**: The modularity-based partitioning indicates weak global community structure, which aligns with the decentralized nature of peer-to-peer networks such as Gnutella.

**8. Core-Periphery & Path-Based Analysis *(Pending)***

To better understand structural depth and resilience:

**Core-Periphery (k-core Decomposition):**

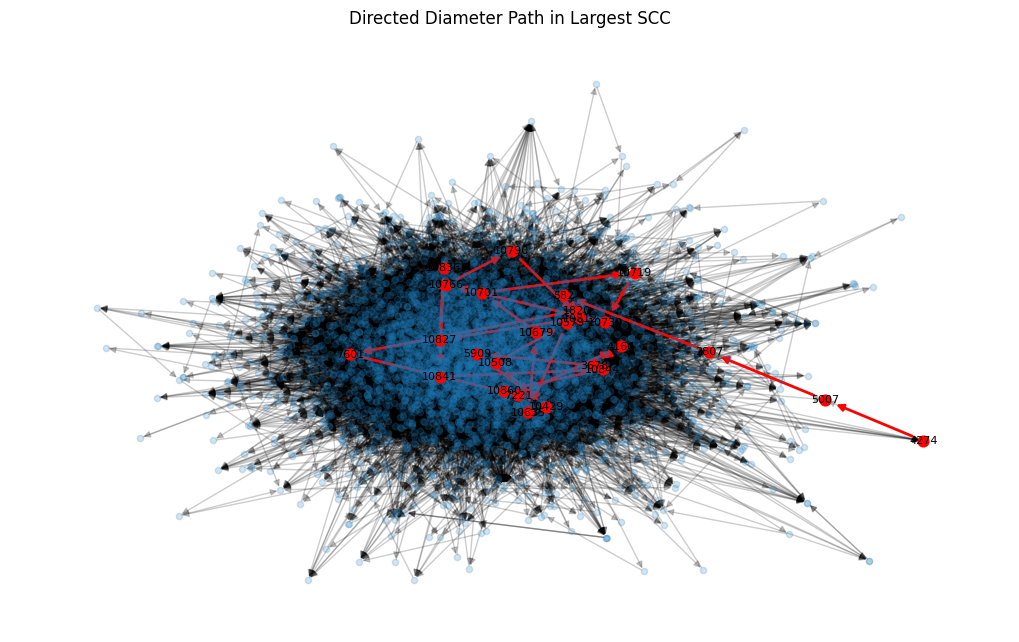
* The network was decomposed into **k-cores**, identifying a deep inner core where nodes are highly interconnected.
* Nodes in **high k-cores** represent the structural backbone of the network.
* Nodes in **lower k-cores** are more loosely attached and vulnerable to disconnection.

****

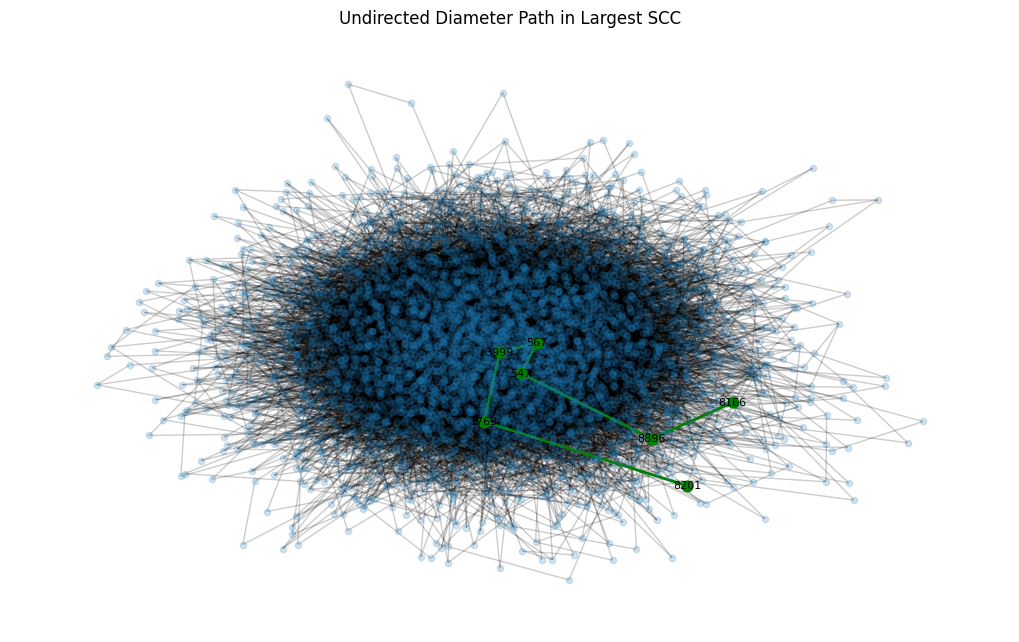
***Fig 5.*** *K-core visualization highlighting the central, densely connected core nodes.*

**Shortest Path & Diameter:**

* As the full graph is not strongly connected, we extracted the **largest strongly connected component (SCC)**.
* Within this SCC:
  + **Average shortest path length** represents typical hop distance between any two nodes.
  + **Diameter** indicates the longest minimal path between any node pair in the SCC.
* This gives a realistic insight into **functional reachability** and communication efficiency within the network’s active core.



***Fig 6.*** *Directed Diameter Path in Largest Strongly Connected Component*

****

***Fig 7.*** *Undirected Diameter Path in Largest Strongly Connected Component*

**9. Discussion**

From the results obtained so far:

* The Gnutella network is sparse, low in clustering, and disassortative.
* Centrality metrics reveal a small number of high-impact nodes likely super-peers or hub servers.
* The graph exhibits scale-free properties in its degree distribution, consistent with unregulated P2P networks.

Further interpretation regarding small-world phenomena, community strength, and core-periphery structure will be included after the pending sections are completed.

**10. Conclusion**

This report presented the initial analysis of the p2p-Gnutella04 network using NetworkX. We successfully extracted basic metrics, degree properties, and key centrality measures. Observed properties are consistent with typical peer-to-peer technological networks namely, sparse connectivity, low clustering levels, and the presence of high-degree hub nodes.

Future work will focus on:

* Community detection and modularity
* Core-periphery structure
* Shortest-path and diameter analysis
* Gephi-based visual exploration and comparative analysis

**11. References**

1. SNAP Datasets - <https://snap.stanford.edu/data/>
2. NetworkX Documentation - <https://networkx.org/>
3. Newman, M. E. (2010). Networks: an introduction.
4. Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), P10008.
5. Barabási, A. L. (2002). The new science of networks. *Cambridge MA. Perseus*.

**Appendix**

**A. Code Snippets**

*(Gerekli kod blokları, örnek çıktılarla birlikte buraya eklenecek)*

**B. Figures**

1. Fig 1. Histogram of node degrees in the full Gnutella network.
2. Fig 2. Visualization of a 1000-node subgraph.
3. Fig 3. Community detection result in Gephi using the Louvain method with ForceAtlas2 layout.
4. Fig 4. Visualization of the largest community detected using NetworkX's greedy modularity algorithm.
5. Fig 5. K-core visualization highlighting the central, densely connected core nodes.
6. Fig 6. Directed Diameter Path in Largest Strongly Connected Component
7. Fig 7. Undirected Diameter Path in Largest Strongly Connected Component