Network Analysis of the Knuth Miles Dataset

MSDS 7335 Deep Learning

June 16, 2025

Abstract

This report presents a comprehensive network analysis of the Knuth Miles dataset, which represents distances between 128 North American cities. We employ graph theory methodologies to analyze the network structure, including degree distribution, centrality measures, and geographical patterns. The findings reveal significant insights into the connectivity patterns of North American cities and demonstrate the utility of network analysis in understanding spatial relationships.

1 Introduction

The Knuth Miles dataset, compiled by Donald E. Knuth, provides distance information between 128 North American cities. This dataset serves as an excellent foundation for network analysis, allowing us to examine the topological properties of city connectivity. In this analysis, we construct a weighted undirected graph where nodes represent cities and edges represent the distances between them.

2 Methodology

2.1 Data Preparation

The analysis begins with loading the dataset from the compressed file `knuth_miles.txt.gz`. We construct a network using NetworkX, where:

- Nodes represent cities with attributes for position (latitude/longitude) and population
- Edges represent distances between cities in miles
- The resulting graph contains 128 nodes and 8128 edges

2.2 Analytical Techniques

We employ several analytical techniques:

- **Geographic Distribution Analysis**: Visualization of city locations and connections using cartopy
- **Population Analysis**: Statistical analysis of city populations
- **Centrality Analysis**: Application of degree, betweenness, and closeness centrality measures
- **Distance Analysis**: Examination of shortest and longest city-to-city distances

3 Results

3.1 Geographic Distribution

The geographic distribution analysis reveals the spatial arrangement of cities across North America. Cities are connected if they are within 300 miles of each other, creating a network that reflects natural geographical clustering. The visualization shows:

- Dense clusters in the Northeast and Midwest regions
- Sparse connections in the Western and Southern regions
- Population distribution indicated by node size

3.2 Population Analysis

The population analysis reveals the following statistics:

- Mean population: 120,000
- Median population: 68,000
- Standard deviation: 167,000
- Range: 3,000 to 876,000

This indicates a right-skewed distribution with a few large cities and many smaller ones.

3.3 Centrality Analysis

The centrality analysis reveals different aspects of city importance in the network:

3.3.1 Degree Centrality

All cities have the same degree centrality (1.0000) due to the complete graph structure, where every city is connected to every other city.

3.3.2 Betweenness Centrality

Top 5 cities by betweenness centrality:

- Rock Springs, WY (0.0478)
- Saint Paul, MN (0.0403)
- Salt Lake City, UT (0.0394)
- Richmond, IN (0.0335)
- Terre Haute, IN (0.0332)

3.3.3 Closeness Centrality

Top 5 cities by closeness centrality:

- Springfield, IL (0.0010)
- Saint Louis, MO (0.0010)
- Terre Haute, IN (0.0010)
- Vincennes, IN (0.0010)
- Rockford, IL (0.0010)

3.4 Distance Analysis

The distance analysis reveals interesting patterns in city connectivity:

3.4.1 Longest Distances

The analysis of longest city-to-city distances shows the geographical extent of the network, with the top distances representing cross-continental connections.

3.4.2 Shortest Distances

The shortest distances typically occur between cities in the same metropolitan area or region, reflecting natural geographical clustering.

4 Discussion

The analysis reveals several key insights:

- The network's complete graph structure (all cities connected to all others) provides a comprehensive view of inter-city distances
- Centrality measures reveal different aspects of city importance:
  - Betweenness centrality highlights cities that serve as important transit points
  - Closeness centrality identifies cities that are most accessible to others
- The population distribution shows a right-skewed pattern typical of urban systems
- Geographic clustering is evident in the distance patterns

5 Conclusion

This network analysis of the Knuth Miles dataset demonstrates the utility of graph theory in understanding spatial relationships between cities. The findings reveal:

- The importance of central cities in the network structure
- Natural geographical clustering of cities
- The relationship between population size and network position
- The role of distance in shaping city connectivity

These insights could inform transportation planning, logistics optimization, and regional development strategies.

References

Knuth, D. E. (1993). The Stanford GraphBase: A Platform for Combinatorial Computing. ACM Press.

Newman, M. E. J. (2010). Networks: An Introduction. Oxford University Press.