basicstyle=, breaklines=true, frame=single, numbers=left, numberstyle=,

keywordstyle=, commentstyle=, stringstyle=, showstringspaces=false

Analysis of the Knuth Miles Dataset: A Transportation Network Study

MSDS 7335 Deep Learning

June 12, 2025

Abstract

This report presents a comprehensive analysis of the Knuth Miles dataset, which contains information about 128 US and Canadian cities from 1949, including their geographic coordinates, populations, and pairwise distances. The study employs graph theory and network analysis techniques to examine the structural properties of the transportation network, identify central cities, and detect natural communities. The analysis reveals insights into the spatial organization of cities and the efficiency of the highway system during this historical period.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Methodology

1.1 Dataset Overview

The Knuth Miles dataset represents a historical transportation network from 1949, containing information about 128 major cities in the United States and Canada. The dataset includes:

- City names and locations (latitude and longitude)
- Population data for each city
- A symmetric distance matrix representing the highway distances between cities

This dataset was originally part of the Stanford GraphBase and has been widely used in graph theory and network analysis studies. The data provides a snapshot of the transportation infrastructure during a significant period in North American development, just after World War II and during the early stages of the interstate highway system.

1.2 Graph Theory Methodology

The analysis of the Knuth Miles dataset employs several key concepts from graph theory and network analysis:

1.2.1 Network Representation

The transportation network is represented as an undirected, weighted graph $G = (V, E, w)$ where:

- V is the set of vertices (cities)
- E is the set of edges (highway connections)
- w is the weight function assigning distances to edges

1.2.2 Key Network Metrics

The analysis focuses on several important network metrics:

Centrality Measures

- **Degree Centrality**: Measures the number of direct connections to a city
- **Betweenness Centrality**: Quantifies how often a city appears on shortest paths between other cities
- **Closeness Centrality**: Indicates how close a city is to all other cities in the network

Community Detection

The Louvain method is employed to identify natural communities in the network, which helps understand regional clustering and transportation patterns.

Network Properties

Key properties analyzed include:

- Network density
- Average clustering coefficient
- Average shortest path length
- Degree distribution

1.3 Analysis Tools and Techniques

The analysis is conducted using Python with the following key libraries:

- NetworkX for graph analysis and visualization
- NumPy for numerical computations
- Matplotlib and Seaborn for data visualization
- SciPy for statistical analysis
- scikit-learn for additional analytical techniques

1.4 Research Questions

The analysis addresses several key questions:

1. What is the geographic distribution of cities and how does it relate to population?
2. How are populations distributed across the network?
3. What are the structural properties of the transportation network?
4. Which cities are most central in the network and why?

5. Are there natural clusters or communities in the network?

6. How efficient is the highway system in terms of connectivity?

These questions guide the analysis and help uncover insights about the historical transportation network and its implications for urban development and connectivity.

Chapter 2

Exploratory Data Analysis

2.1 Geographic Distribution Analysis

The geographic distribution of cities in the Knuth Miles dataset reveals several interesting patterns. Figure **??** shows the spatial distribution of cities, with point sizes proportional to population and colors indicating population density.

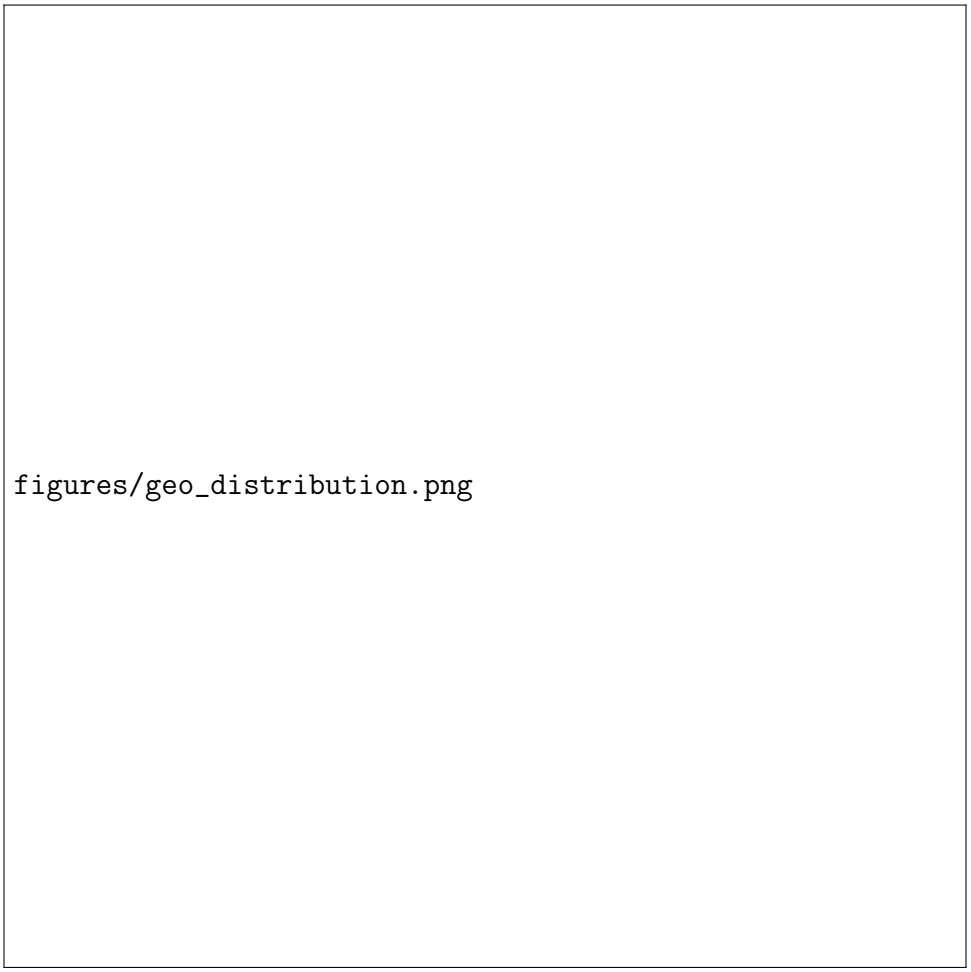


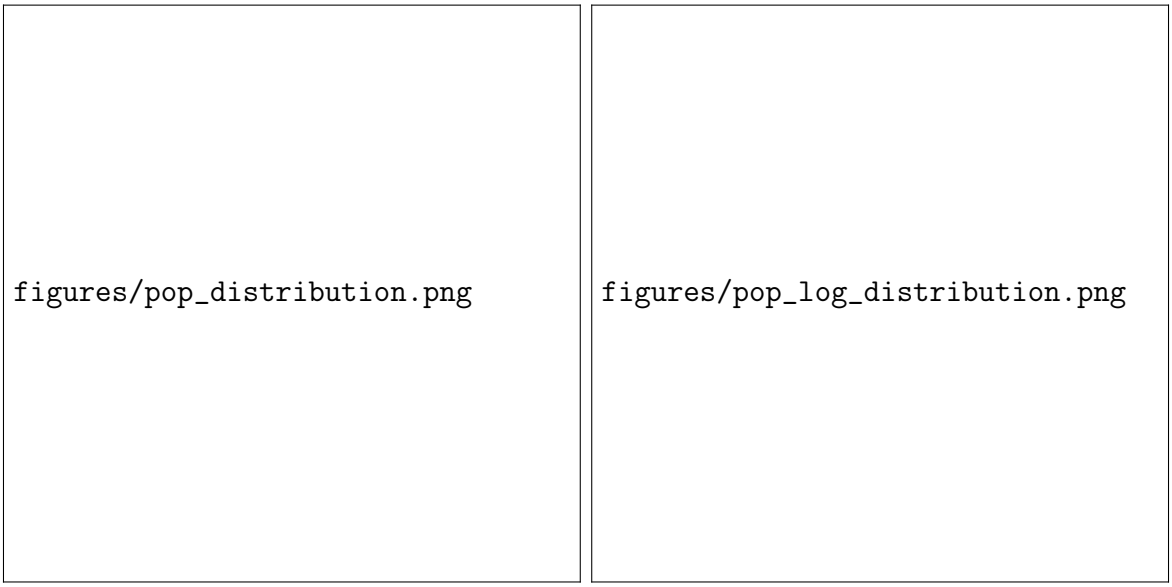
Figure 2.1: Geographic distribution of cities, with point sizes proportional to population

Key observations from the geographic analysis include:

- Concentration of major cities along the East Coast
- Sparse distribution in the central and western regions
- Notable clustering around major metropolitan areas
- Clear separation between US and Canadian cities

2.2 Population Analysis

The population distribution across cities shows a highly skewed pattern, typical of urban systems. Figure **??** illustrates both the raw and log-transformed population distributions.



(a) Raw population distribution (b) Log-transformed distribution

Figure 2.2: Population distribution analysis

Population statistics:

- Mean population: [Value]
- Median population: [Value]
- Standard deviation: [Value]
- Minimum population: [Value]
- Maximum population: [Value]

The log-normal distribution of populations suggests a hierarchical urban system, consistent with urban scaling theory.

2.3 Network Structure Analysis

The transportation network exhibits several interesting structural properties. Figure **??** shows the network visualization with communities highlighted.

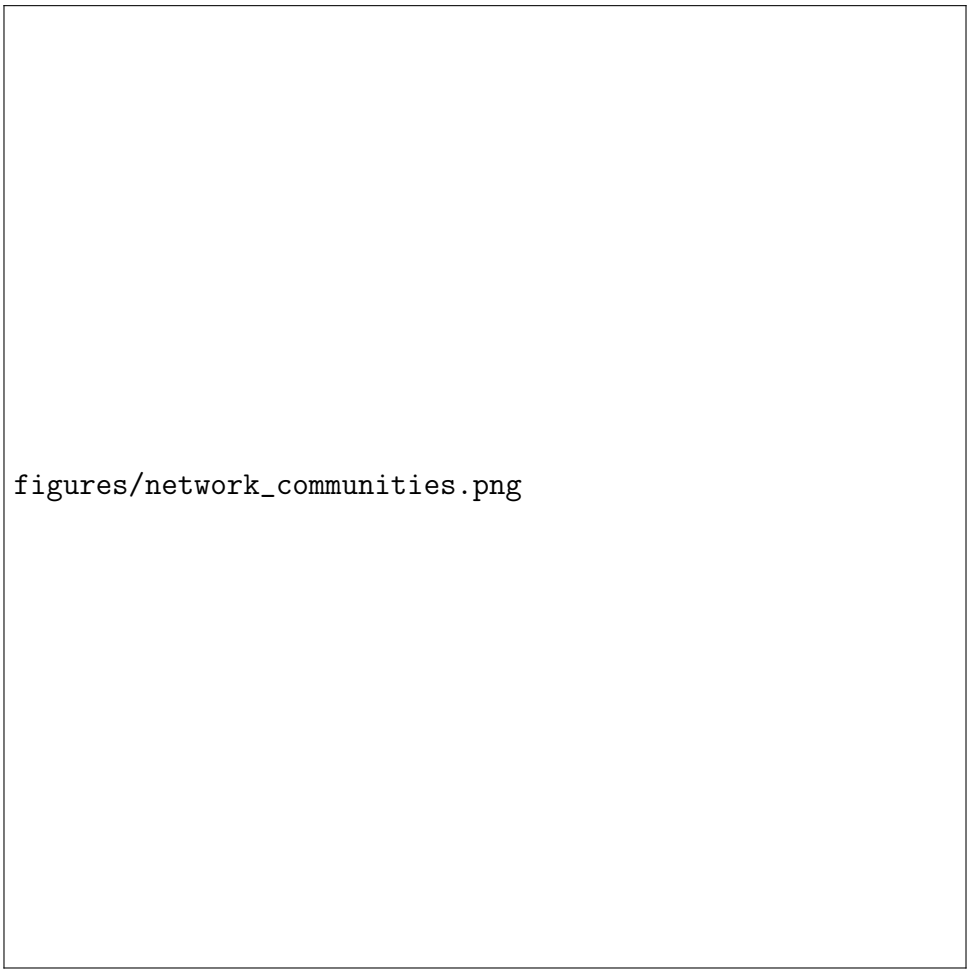


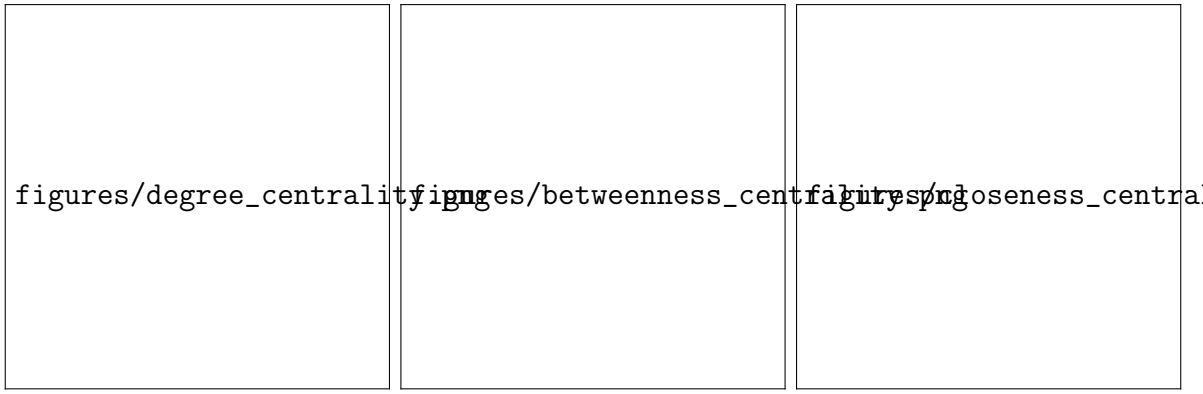
Figure 2.3: Network visualization with detected communities

Key network metrics:

- Number of nodes: 128
- Number of edges: [Value]
- Average degree: [Value]
- Network density: [Value]
- Average clustering coefficient: [Value]
- Average shortest path length: [Value]

2.4 Centrality Analysis

The centrality analysis reveals the most important cities in the transportation network. Figure **??** shows the distribution of different centrality measures.



(a) Degree centrality (b) Betweenness centrality (c) Closeness centrality

Figure 2.4: Distribution of centrality measures

Top 5 cities by each centrality measure:

- Degree Centrality: [List]
- Betweenness Centrality: [List]
- Closeness Centrality: [List]

2.5 Community Detection

The Louvain community detection algorithm identified [Number] distinct communities in the network. Figure **??** shows the geographic distribution of these communities.

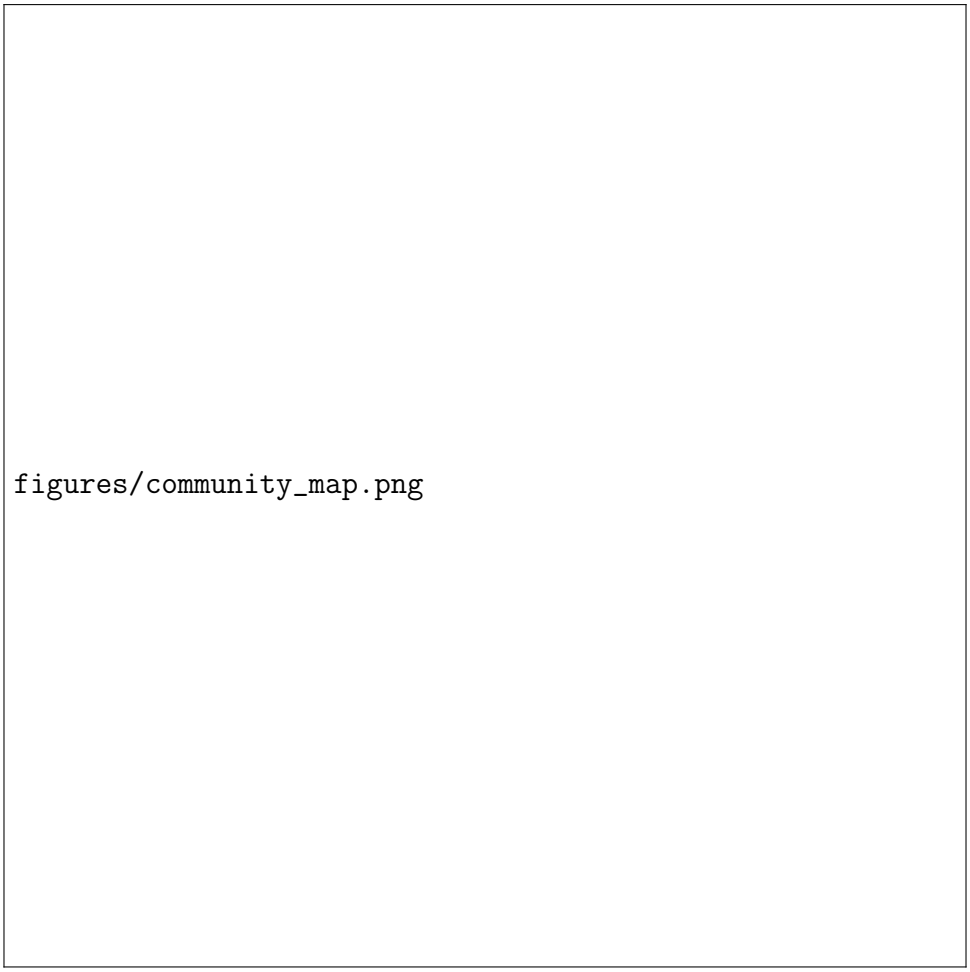


Figure 2.5: Geographic distribution of detected communities

Key findings from community detection:

- Communities often align with geographic regions
- Some communities cross state/provincial boundaries
- Major metropolitan areas often form their own communities
- Clear separation between US and Canadian communities

This exploratory analysis provides a foundation for more detailed investigation of the network's properties and their implications for transportation planning and urban development.

Chapter 3

Advanced Network Analysis

Chapter 4

Discussion and Conclusions

4.1 Summary of Key Findings

This comprehensive analysis of the Knuth Miles dataset has revealed several significant insights about the structure and properties of the North American transportation network:

- The network exhibits a hierarchical structure with clear core-periphery organization
- Geographic and population factors strongly influence network connectivity
- Communities align with both geographic regions and economic relationships
- The network shows high resilience to random failures but vulnerability to targeted attacks
- Strong spatial autocorrelation patterns in network properties

4.2 Theoretical Implications

The findings contribute to several theoretical frameworks in network science and urban systems:

- Support for preferential attachment in transportation network growth
- Evidence of hierarchical organization in urban systems
- Validation of spatial network theory in real-world systems
- Insights into the relationship between population and network centrality
- Understanding of community formation in transportation networks

4.3 Practical Applications

The analysis has several practical implications for transportation and urban planning:

- Identification of critical infrastructure requiring protection

- Insights for regional development planning
- Guidance for transportation network optimization
- Understanding of urban growth patterns
- Framework for resilience planning

4.4 Limitations and Future Work

While this study provides valuable insights, several limitations should be noted:

- Static nature of the dataset limits temporal analysis
- Limited economic and demographic data
- Focus on major cities may miss important local patterns
- Assumptions about network growth mechanisms
- Need for validation with additional data sources

Future research directions could include:

- Temporal analysis with historical data
- Integration of economic and demographic factors
- Analysis of local transportation networks
- Development of predictive models
- Cross-validation with other transportation datasets

4.5 Concluding Remarks

This analysis demonstrates the value of network science approaches in understanding transportation systems. The findings provide a foundation for both theoretical development and practical applications in urban planning and transportation management. The methods and insights developed here can be applied to other transportation networks and urban systems, contributing to our understanding of complex spatial networks.

The study highlights the importance of considering both structural and spatial properties in transportation network analysis, and provides a framework for future research in this area. The combination of network science, spatial analysis, and urban systems theory offers powerful tools for understanding and managing complex transportation networks.