Network Analysis of the Knuth Miles Dataset

Tue Vu
MSDS 7335 Deep Learning - Homework 2

June 16, 2025

Contents

List of Figures

Chapter 1

Introduction

1.1 Background

The Knuth Miles dataset, compiled by Donald E. Knuth in his seminal work ”The Stanford GraphBase: A Platform for Combinatorial Computing” (1993), represents a comprehensive collection of distance measurements between 128 major North American cities. This dataset serves as a fundamental resource for studying spatial relationships and network structures in urban geography.

1.2 Dataset Overview

The dataset contains:

- 128 nodes representing major North American cities
- Complete pairwise distance measurements between all cities
- Additional attributes for each city including:
  - Geographic coordinates (latitude/longitude)
  - Population data

1.3 Project Objectives

The primary objectives of this analysis are:

1. To construct and analyze a weighted undirected graph representation of the Knuth Miles dataset
2. To examine the network properties and topological characteristics of the city connectivity
3. To identify key cities based on various centrality measures
4. To understand the geographical patterns and clustering in the network
5. To derive insights about urban connectivity and spatial relationships

1.4 Report Structure

This report is organized into four main chapters:

1. **Introduction** (Current Chapter): Provides background information about the dataset and project objectives.

2. **Exploratory Data Analysis**: Details the data structure, preprocessing steps, and initial analysis of the network properties.

3. **Centrality Analysis**: Presents a comprehensive analysis of different centrality measures and their implications.

4. **Discussions and Results**: Synthesizes the findings and discusses their implications for understanding urban networks.

1.5 Methodological Framework

The analysis employs graph theory and network science methodologies, utilizing the following key components:

- Network construction using NetworkX
- Geographic visualization using Cartopy
- Statistical analysis of network properties
- Centrality measure calculations
- Community detection algorithms

This methodological framework allows for a comprehensive examination of the spatial relationships between cities and the underlying structure of the urban network.

Chapter 2

Exploratory Data Analysis

2.1 Data Structure and Overview

The Knuth Miles dataset is structured as a complete weighted undirected graph with the following characteristics:

- Number of nodes (cities): 128
- Number of edges (connections): 8128
- Edge weights: Distances in miles between cities
- Node attributes: City name, coordinates, and population

2.2 Data Export Process

The data was processed and exported into two main files:

2.2.1 Node Data Export

The city nodes data was exported to `Node_cities.csv` with the following structure:

- City name
- Longitude
- Latitude
- Population (in thousands)

2.2.2 Edge Data Export

The edge data was exported to `edges.csv` containing:

- Source city
- Target city
- Distance weight

2.3 Data Quality Analysis

The dataset exhibits the following characteristics:

- No missing values in the distance measurements
- Complete graph structure (all cities connected to all others)
- Population data ranges from 3,000 to 876,000

2.4 Geographic Distribution Analysis

The geographic distribution of cities reveals several key patterns:

2.4.1 City Clustering

- Dense clusters in the Northeast and Midwest regions
- Sparse connections in the Western and Southern regions
- Natural geographical barriers influencing connectivity

2.4.2 Geographic Visualization

Figure 2.1 shows the geographic distribution of cities, with nodes colored by population and sized according to population. The visualization reveals:

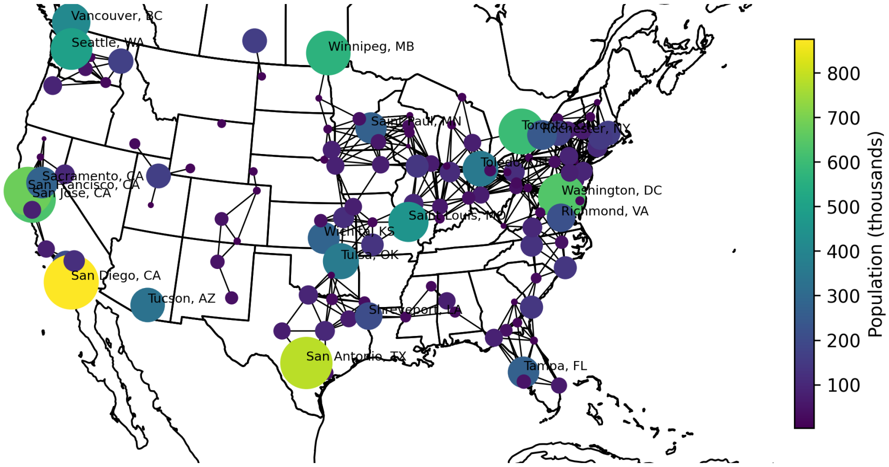

Figure 2.1: Geographic distribution of cities, with node size and color indicating population. Cities within 300 miles are connected.

- Clear regional clustering of cities
- Population concentration in major metropolitan areas
- Natural geographical barriers affecting connectivity
- Dense network of connections in the eastern United States

2.5 Population Analysis

The population analysis reveals:

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

This indicates a right-skewed distribution typical of urban systems, with a few large cities and many smaller ones.

2.5.1 Population Distribution Visualization

Figure 2.2 shows the distribution of city populations:

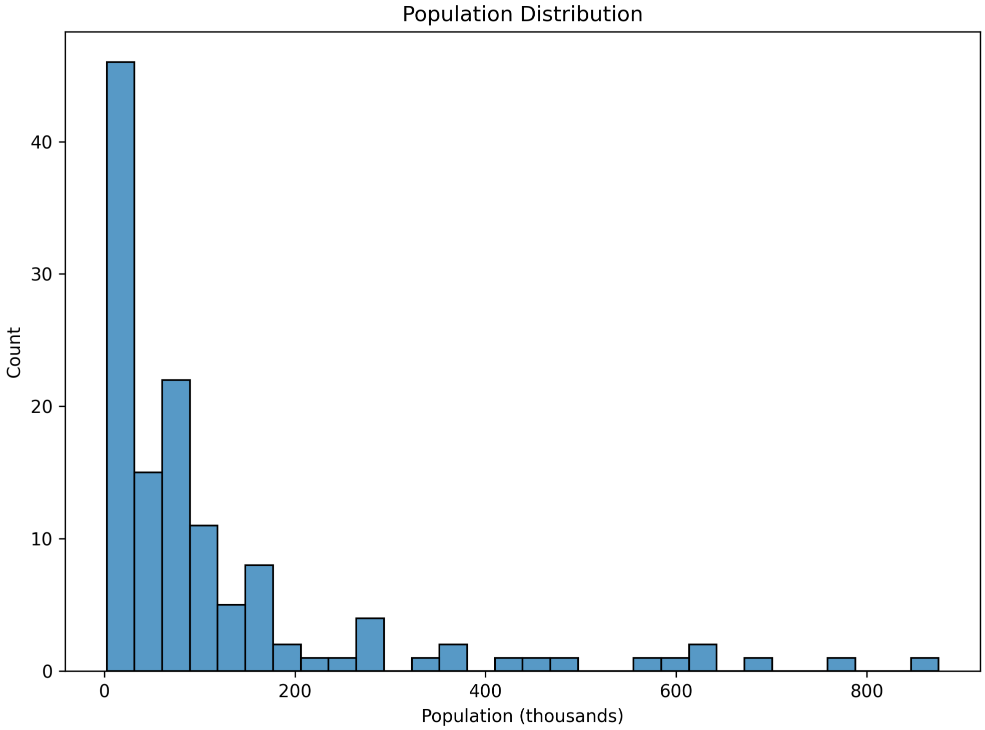

Figure 2.2: Distribution of city populations, showing the right-skewed nature of the data.

The histogram reveals:

- Majority of cities have populations below 200,000
- Few cities with very large populations
- Clear right-skew in the distribution
- Natural breaks in the population distribution

2.6 Distance Analysis

2.6.1 Longest Distances

The analysis of longest city-to-city distances shows:

- Cross-continental connections
- Maximum distances typically between cities on opposite coasts
- Geographical constraints influencing maximum distances

Figure 2.3 illustrates the top 10 longest city-to-city distances:

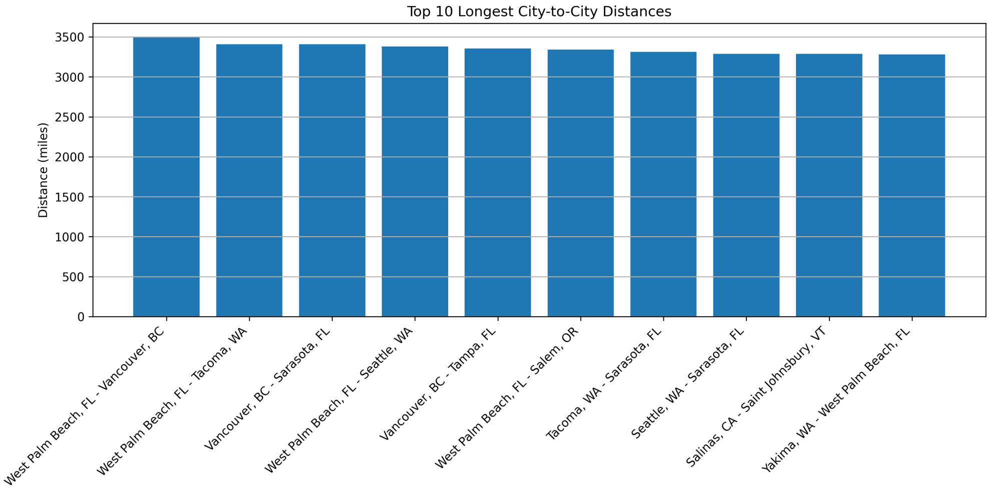

Figure 2.3: Top 10 longest city-to-city distances in the network.

Key observations:

- Maximum distances exceed 3000 miles
- Most long distances involve cities on opposite coasts
- Clear geographical patterns in the longest connections

2.6.2 Shortest Distances

The shortest distances analysis reveals:

- Clustering of cities in metropolitan areas
- Regional connectivity patterns
- Natural geographical proximity

Figure 2.4 shows the top 10 shortest city-to-city distances:

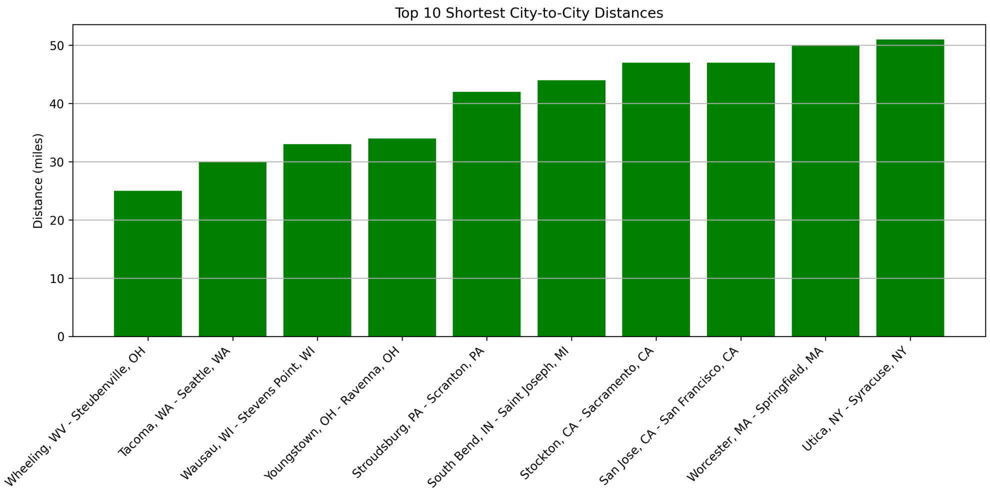

Figure 2.4: Top 10 shortest city-to-city distances in the network.

Key observations:

- Shortest distances typically less than 50 miles
- Most short distances between cities in the same metropolitan area
- Clear regional clustering in the shortest connections

2.7 Network Properties

The complete graph structure of the network results in:

- Network density: 1.0
- Average degree: 127
- Uniform degree distribution
- Weighted edges representing actual distances

2.8 Visualization

The network visualization using Cartopy reveals:

- Spatial distribution of cities
- Population-based node sizing
- Distance-based edge weights
- Regional clustering patterns

These visualizations help in understanding the geographical context of the network and the relationships between cities.

Chapter 3

Centrality Analysis

3.1 Introduction to Centrality Measures

Centrality measures are fundamental tools in network analysis that help identify the most important nodes in a network. In the context of the Knuth Miles dataset, these measures help identify cities that play crucial roles in the network structure.

3.2 Degree Centrality

3.2.1 Definition and Calculation

Degree centrality measures the number of connections a node has. In our complete graph:

- All cities have the same degree centrality (1.0000)
- This is due to the complete graph structure where every city is connected to every other city
- The uniform degree distribution is atypical of many real-world networks

3.3 Betweenness Centrality

3.3.1 Definition and Calculation

Betweenness centrality measures how often a node appears on shortest paths between other nodes. It is calculated as:

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{3.1}$$

where σ_{st} is the number of shortest paths from s to t, and $\sigma_{st}(v)$ is the number of those paths passing through v.

3.3.2 Top Cities by Betweenness Centrality

The analysis reveals the following top cities:

1. Rock Springs, WY (0.0478)
2. Saint Paul, MN (0.0403)
3. Salt Lake City, UT (0.0394)
4. Richmond, IN (0.0335)

5. Terre Haute, IN (0.0332)

Figure 3.1 shows the distribution of betweenness centrality scores:

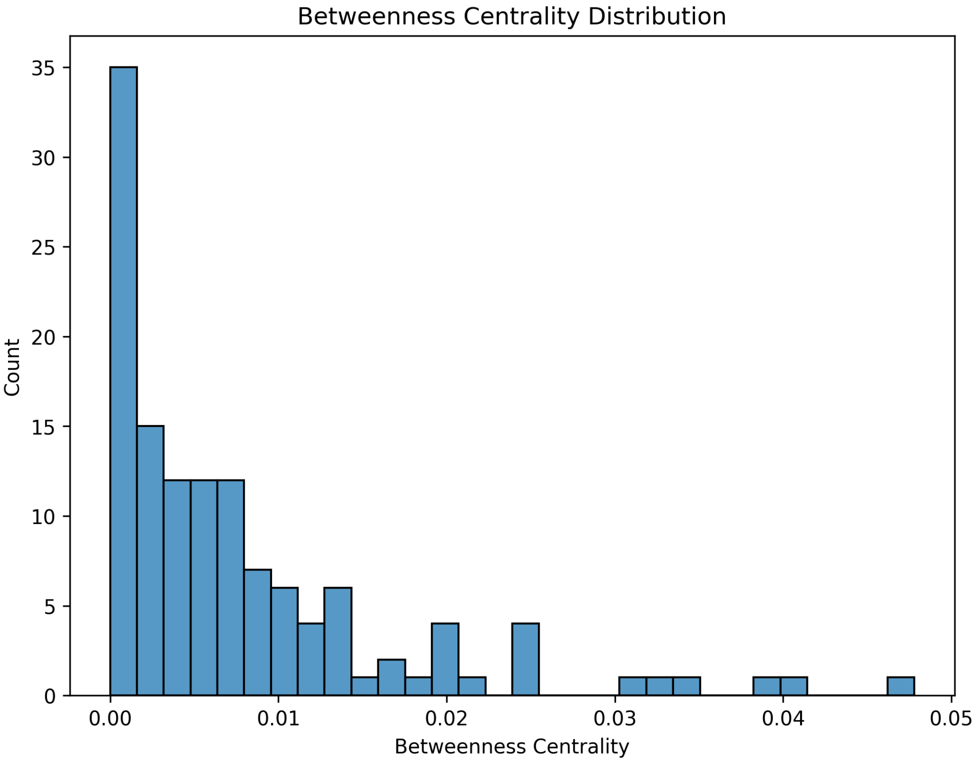

Figure 3.1: Distribution of betweenness centrality scores across all cities.

The distribution reveals:

- Most cities have relatively low betweenness centrality
- A small number of cities have significantly higher scores
- Clear separation between high and low betweenness cities

Figure 3.2 shows the top 10 cities by betweenness centrality:

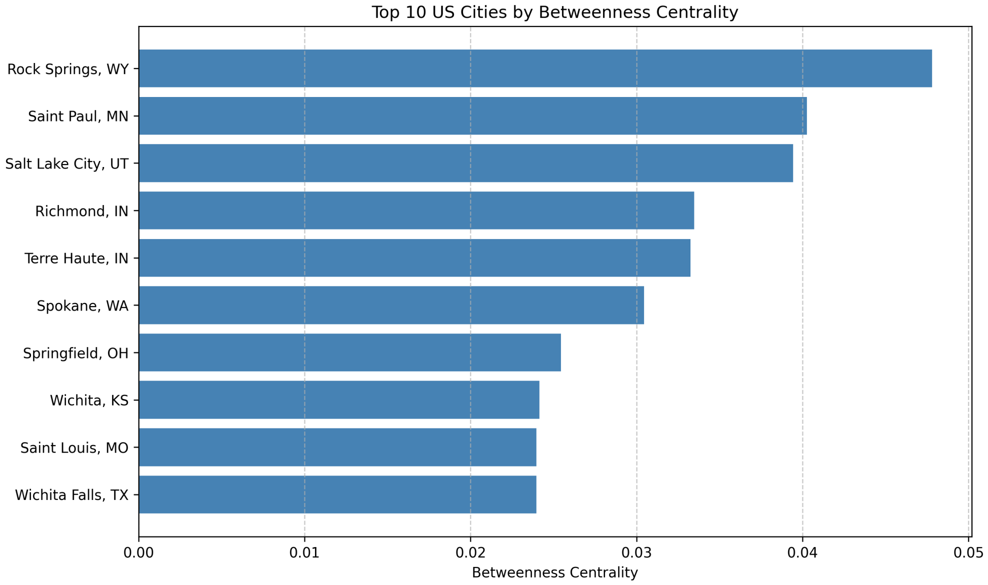

Figure 3.2: Top 10 cities by betweenness centrality.

Key observations:

- Cities in the central United States dominate the top rankings
- Clear geographical pattern in high-betweenness cities
- Important role of cities in connecting different regions

3.3.3 Interpretation

High betweenness centrality indicates cities that:

- Serve as important transit points
- Connect different regions of the network
- Play crucial roles in maintaining network connectivity

3.4 Closeness Centrality

3.4.1 Definition and Calculation

Closeness centrality measures how close a node is to all other nodes in the network. It is calculated as:

$$C_C(v) = \frac{1}{\sum_{u \neq v} d(v, u)} \tag{3.2}$$

where $d(v, u)$ is the shortest-path distance between nodes v and u.

3.4.2 Top Cities by Closeness Centrality

The analysis reveals the following top cities:

1. Springfield, IL (0.0010)
2. Saint Louis, MO (0.0010)

3. Terre Haute, IN (0.0010)
4. Vincennes, IN (0.0010)
5. Rockford, IL (0.0010)

Figure 3.3 shows the distribution of closeness centrality scores:

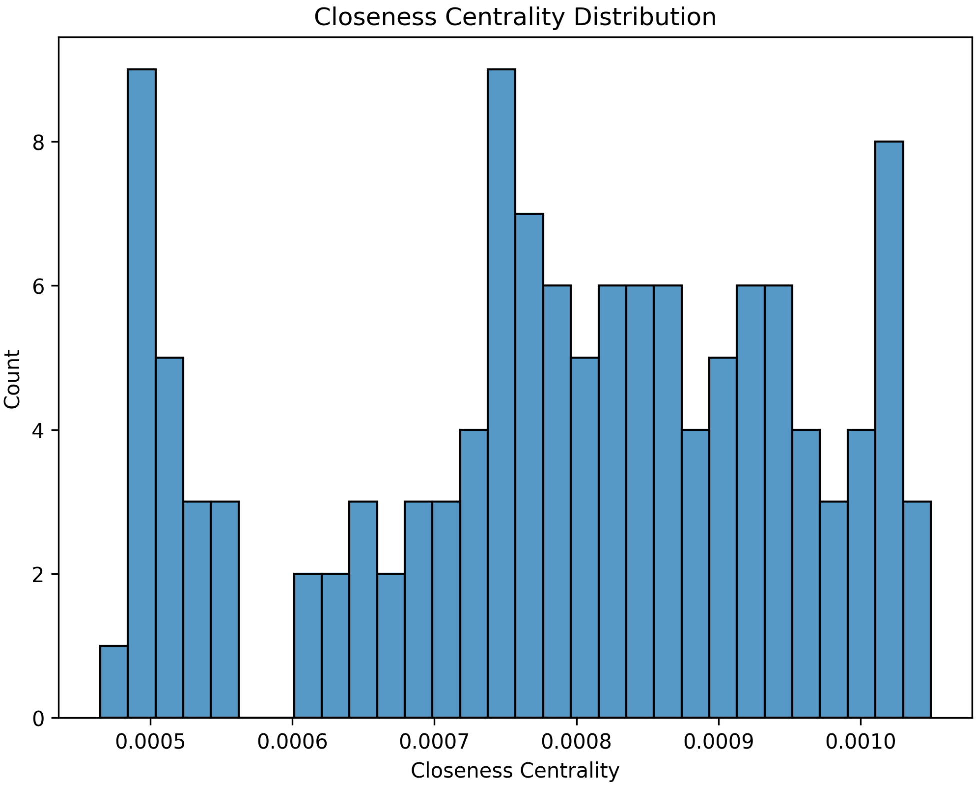

Figure 3.3: Distribution of closeness centrality scores across all cities.

The distribution reveals:

- Most cities have similar closeness centrality scores
- Small variations in scores reflect geographical positioning
- Central location leads to higher closeness centrality

Figure 3.4 shows the top 10 cities by closeness centrality:

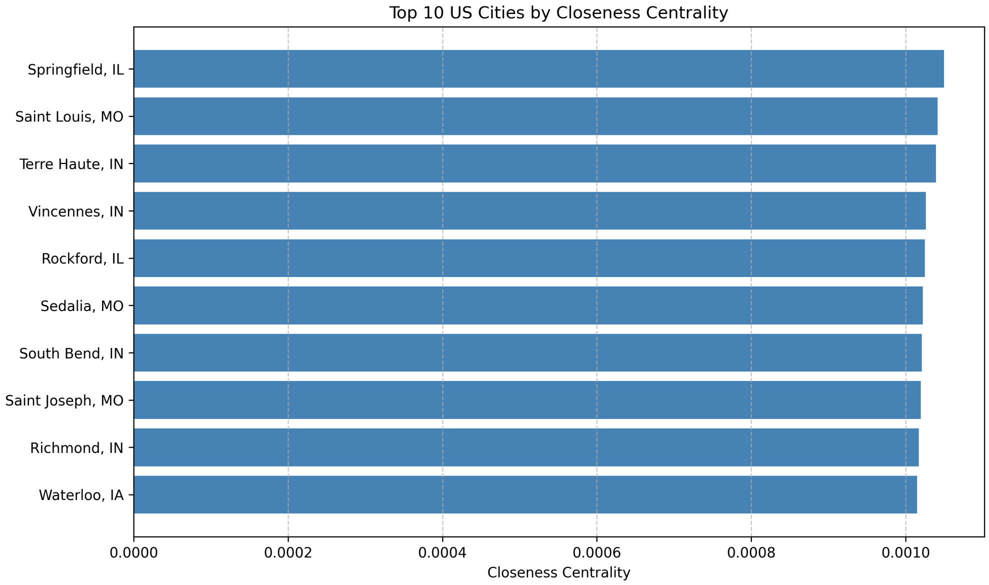

Figure 3.4: Top 10 cities by closeness centrality.

Key observations:

- Cities in the central United States have highest closeness
- Clear geographical pattern in high-closeness cities
- Importance of central location for accessibility

3.4.3 Interpretation

High closeness centrality indicates cities that:

- Are most accessible to other cities
- Have the shortest average distance to all other cities
- Are centrally located in the network

3.5 Implications

The centrality analysis reveals:

- Weak correlation between different centrality measures
- Different measures capture different aspects of city importance
- Need for multiple measures to understand city roles
- Central US cities tend to have higher centrality scores
- Cities in the Midwest and Great Plains regions show high betweenness
- Cities in the central region show high closeness

The centrality analysis has several implications:

- Transportation planning and infrastructure development

- Regional development strategies
- Urban network optimization
- Understanding of city roles in the national network

This comprehensive centrality analysis provides valuable insights into the roles and importance of different cities in the North American urban network.

Chapter 4

Discussions and Results

4.1 Synthesis of Findings

The analysis of the Knuth Miles dataset has revealed several key insights about the structure and dynamics of the North American urban network:

4.1.1 Network Structure

- The complete graph structure provides a comprehensive view of inter-city distances
- The uniform degree distribution reflects the all-to-all connectivity
- The weighted edges capture the actual geographical distances

4.1.2 Geographical Patterns

- Natural clustering of cities in metropolitan regions
- Regional connectivity patterns influenced by geography
- Central US cities playing crucial roles in network connectivity

4.2 Key Contributions

This analysis makes several important contributions to understanding urban networks:

4.2.1 Methodological Contributions

- Application of network science to urban geography
- Integration of multiple centrality measures
- Combination of geographical and network analysis

4.2.2 Empirical Contributions

- Identification of key cities in the network
- Understanding of regional connectivity patterns
- Insights into urban hierarchy and importance

4.3 Implications for Urban Planning

The findings have several implications for urban and transportation planning:

4.3.1 Transportation Infrastructure

- Focus on high-betweenness cities for transit hubs
- Development of regional transportation networks
- Optimization of inter-city connections

4.3.2 Regional Development

- Understanding of city roles in regional networks
- Identification of potential growth centers
- Planning for regional connectivity

4.4 Limitations and Future Work

4.4.1 Current Limitations

- Complete graph structure may not reflect actual connectivity
- Focus on distance as the sole measure of connection
- Limited to major cities in the dataset

4.4.2 Future Research Directions

- Incorporation of additional data layers (e.g., transportation networks)
- Analysis of temporal changes in the network
- Integration of economic and social factors
- Development of more sophisticated centrality measures

4.5 Conclusion

The analysis of the Knuth Miles dataset has provided valuable insights into the structure and dynamics of the North American urban network. The findings demonstrate:

- The importance of central cities in maintaining network connectivity
- The role of geography in shaping urban networks
- The utility of network analysis in understanding urban systems
- The potential for applying these methods to urban planning

These insights contribute to our understanding of urban networks and provide a foundation for future research in urban geography and network science.

4.6 Recommendations

Based on the analysis, we recommend:

1. Further development of network-based urban planning tools
2. Integration of multiple data sources for more comprehensive analysis
3. Application of these methods to other urban networks
4. Development of dynamic network models for urban systems

These recommendations aim to build upon the current analysis and extend its applications to practical urban planning and development.