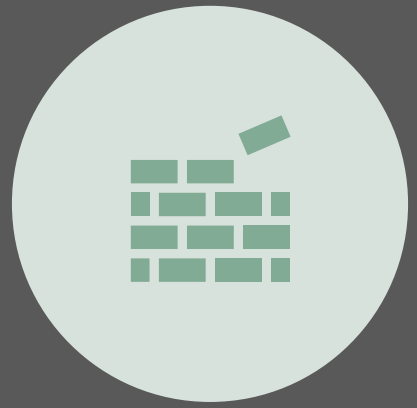
Amtrak Rail Station Network

Tanner Kogel

Overview

OBJECTIVE



GRAPH GENERATION

ANALYSIS &
DISCUSSION

Objective

Study the interconnectedness of populous U.S. cities by use of Amtrak rail travel

Three Networks:

- Amtrak Rail Stations
- Populous U.S. Cities (Induced)
- Populous U.S. Cities (Non-Induced)

Graph Generation

Amtrak Rail Station Network

- Node definition: Rail station utilized by Amtrak
- Edge exists if: Amtrak route between two stations exists
- Weighted and Undirected

Graph Generation

Amtrak Rail Station Network

OBTAIN .TXT FILE FROM AMTRAK WEBSITE

```
Eugene, OR - Amtrak Station (EUG)
Albany, OR (ALY)
Salem, OR - Amtrak Station (SLM)
Oregon City, OR (ORC)
Portland, OR - Union Station (PDX)
Vancouver, WA (VAN)
Kelso-Longview, WA (KEL)
Centralia, WA (CTL)
Olympia-Lacey, WA (OLW)
Tacoma, WA (TAC)
Tukwila, WA (TUK)
Seattle, WA - King Street Station (SEA)
Edmonds, WA (EDM)
Everett, WA (EVR)
Stanwood, WA (STW)
Mount Vernon, WA (MVW)
Bellingham, WA (BEL)
Vancouver, BC - Pacific Central Station (VAC)

Springfield, MA (SPG)
Windsor Locks, CT (WNL)
Windsor, CT (WND)
Hartford, CT (HFD)
Berlin, CT (BER)
Meriden, CT (MDN)
Wallingford, CT (WFD)
New Haven, CT - State Street Station (STS)
New Haven, CT - Union Station (NHV)
```

Route #1

Break

Route #2

Graph Generation

Amtrak Rail Station Network

OBTAIN .TXT FILE FROM AMTRAK WEBSITE

PARSE DATA TO WRITE GML FILE

530 Nodes

592 Edges

Graph Generation

Populous U.S. City Network

- Node definition: City estimated to have a population of 100,000 or greater by July 1, 2021, as estimated by the 2020 U.S. Census

Graph Generation

Populous U.S. City Network

331 Nodes

Induced

Non-Induced

OBTAIN .TXT FILE OF CENSUS DATA

PARSE DATA TO OBTAIN NODES

MAP AMTRAK NETWORK TO CITY NETWORK

7	San Antonio	Texas	1,451,853	1,434,625	+1.20%	498.8 sq mi	1,291.9 km2	2,876/sq mi	1,110/km2	29.46°N 98.52°W
8	San Diego	California	1,381,611	1,386,932	−0.38%	325.9 sq mi	844.1 km2	4,256/sq mi	1,643/km2	32.81°N 117.14°W
9	Dallas	Texas	1,288,457	1,304,379	−1.22%	339.6 sq mi	879.6 km2	3,841/sq mi	1,483/km2	32.79°N 96.77°W
10	San Jose	California	983,489	1,013,240	−2.94%	178.3 sq mi	461.8 km2	5,683/sq mi	2,194/km2	37.30°N 121.81°W
11	Austin	Texas	964,177	961,855	+0.24%	319.9 sq mi	828.5 km2	3,007/sq mi	1,161/km2	30.30°N 97.75°W

Graph Generation

Populous U.S. City Network (Induced)

- Edge exists if:
  - Edge exists in Amtrak Network
  - Both adjacent stations are within a defined city
- "Direct-Route" Graph

331 Nodes

74 Edges

Populous U.S. City Network (Non-Induced)

- Edge exists if:
  - Amtrak route travels between adjacent cities without passing through any other defined cities
- "Destinations" Graph

331 Nodes

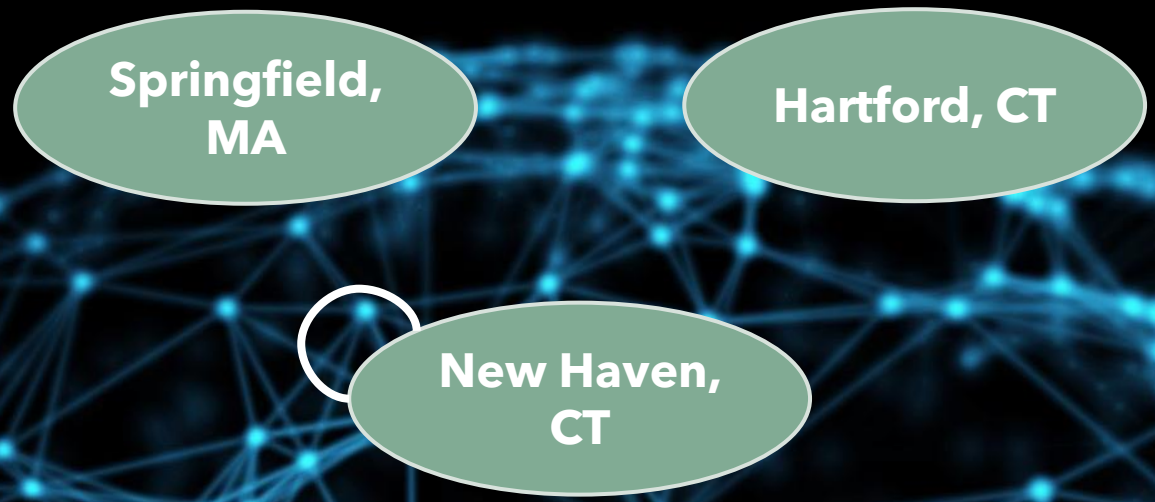
163 Edges

Graph Generation

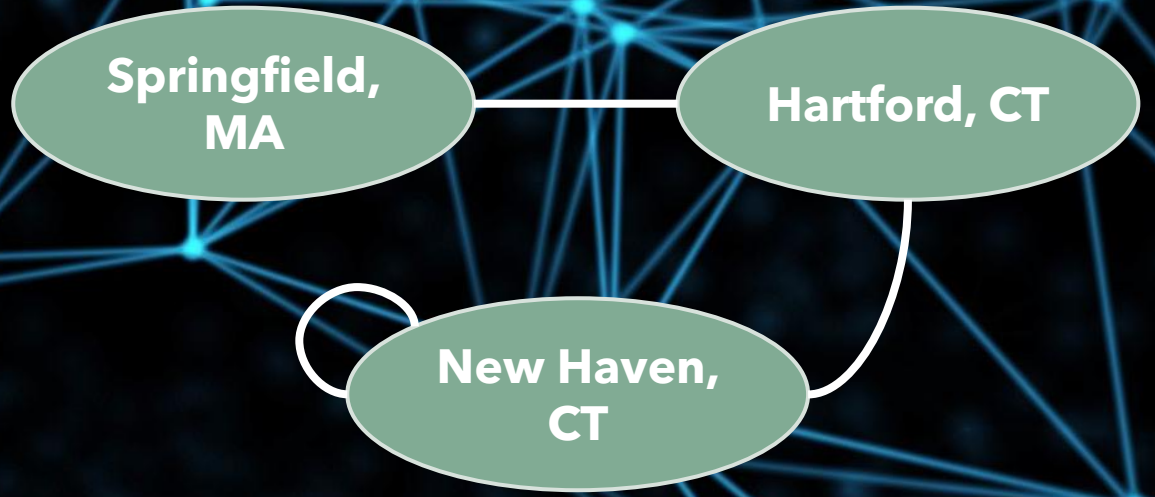
Populous U.S. City Network Mappings

Springfield, MA (SPG)	☑
Windsor Locks, CT (WNL)	☒
Windsor, CT (WND)	☒
Hartford, CT (HFD)	☑
Berlin, CT (BER)	☒
Meriden, CT (MDN)	☒
Wallingford, CT (WFD)	☒
New Haven, CT - State Street Station (STS)	☑
New Haven, CT - Union Station (NHV)	☑

Induced Graph



Non-Induced Graph



Analysis

COMPONENTS
& DIAMETER

DEGREE
DISTRIBUTION

ASSORTATIVITY

CENTRALITY

Components & Diameter

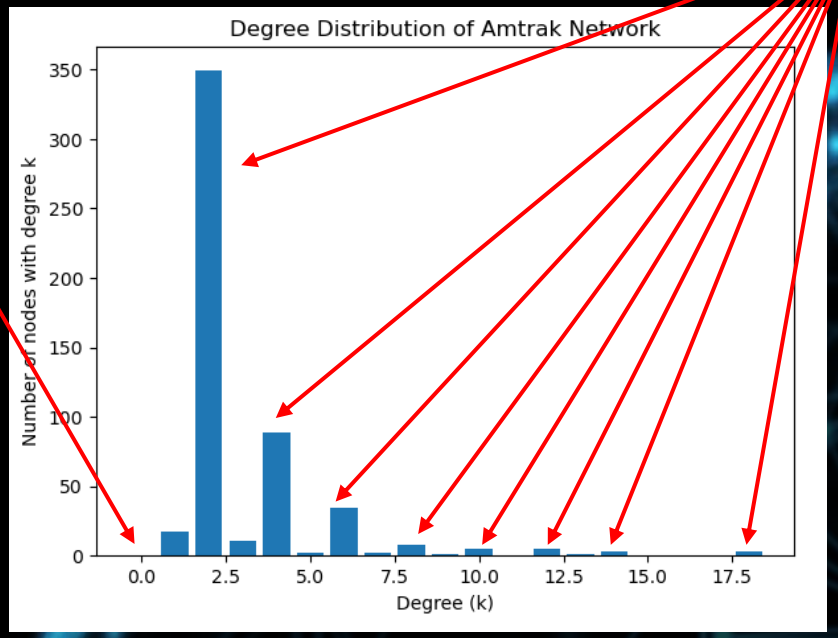
	Amtrak Network	Non-Induced Graph	Induced Graph
Components	3	204	272
Diameter	81	30	7
Path	San Diego…Miami	San Diego…Miami	Irvine, CA…Ontario, CA

Discuss

- Why do the sampled graphs have so many more components?
- Why is the diameter of the original graph so large?
  - Structure of transportation networks
- Why do the original graph and the non-induced graph's diameter span the continent, but the induced graph stays within California?

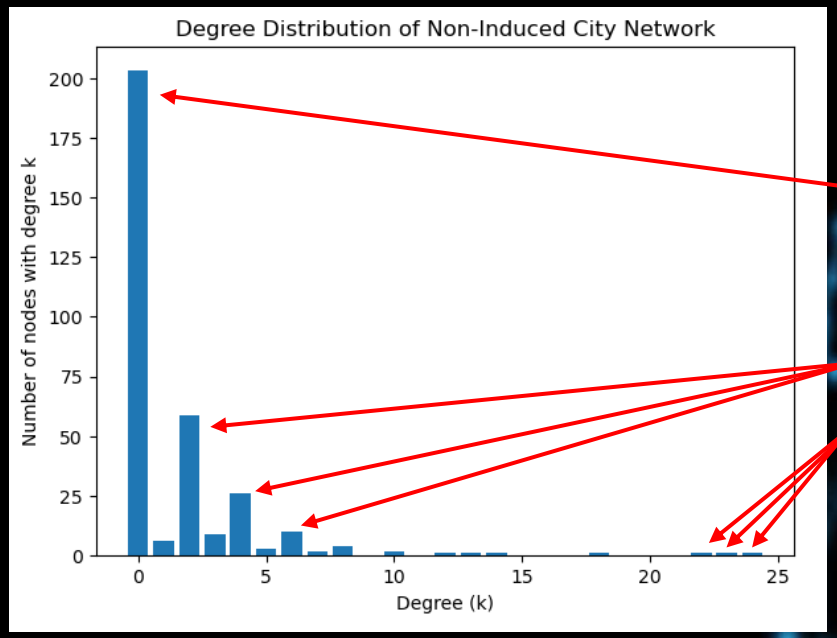
Degree Distribution

All nodes have degree > 0

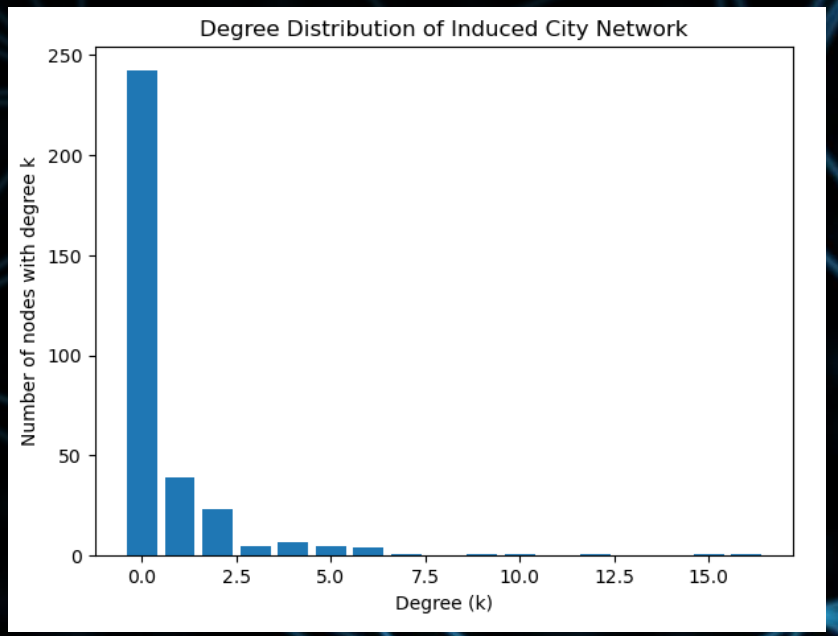


Even Degreed Nodes > Odd Degreed Nodes

- 242 nodes with degree 0
  - Nodes whose neighbors were not included in the sample
- Odd/Even pattern not shown



- 203 nodes with degree 0
  - Cities that do not have train stations
- Odd/Even Pattern Shown
- Higher Degrees?
  - Self Edges



Assortativity

Are cities more likely to be connected if they have similar _____?

STATE

POPULATION

CHANGE IN POPULATION

GEOGRAPHICAL AREA

Assortativity by State

	Non-Induced Graph	Induced Graph
Q	0.1469	0.1898
Q_{max}	0.9956	0.9927

Discuss

- Both are assortatively mixed
  - Trains travel between big cities in the same state more often than different states
- $Q_{max} \approx 1$
  - Graph could be almost perfectly assortatively mixed
  - Number of routes that cross a given states border is roughly even

Scalar Assortativity

Attribute	Non-Induced Graph r	Induced Graph r
Population	0.1033	0.0586
Change in Population	0.5949	0.6457
Geographical Area	0.5663	0.4500

Discuss

- All attributes are assortatively mixed. Why?
- Why is change in population the most assortatively mixed?
  - Geographical proximity?
- Why does it make sense for the non-induced graph to be more assortatively mixed by population than the induced graph?

Eigenvector Centrality

- Rank ‘importance’ of nodes based on the number of connections to other nodes
- But connections are more ‘important’ if the other node is ‘important’

Meaning for our Network:

Ranking of cities or train stations by number of route options available from the current city or train station as well as its neighbors

Eigenvector Centrality

Rank	Amtrak Network	Non-Induced Graph	Induced Graph
1	Philadelphia, PA - William H. Gray III 30th St. Station (PHL) - 0.4774	Philadelphia, PA - 0.5961	Newark, NJ - 0.7522
2	Trenton, NJ (TRE) - 0.4508	Newark, NJ - 0.5674	New York, NY - 0.6098
3	Newark, NJ - Penn Station (NWK) - 0.3694	Baltimore, MD - 0.4506	Yonkers, NY - 0.2293
4	Wilmington, DE (WIL) - 0.3638	New York, NY - 0.2817	Stamford, CT - 0.0956
5	New York, NY - Moynihan Train Hall (NYP) - 0.2574	Washington, DC - 0.1714	New Haven, CT - 0.0180
6	Baltimore, MD - Penn Station (BAL) - 0.2058	Yonkers, NY - 0.0714	Bridgeport, CT - 0.0171
7	Metropark, NJ (MET) - 0.1852	Alexandria, VA - 0.0443	Providence, RI - 0.0014
8	North Philadelphia, PA (PHN) - 0.1525	Stamford, CT - 0.0431	Baltimore, MD - 0.0005
9	Cornwells Heights, PA (CWH) - 0.1472	Pittsburgh, PA - 0.0386	Washington, DC - 0.0003
10	Newark, NJ - Liberty International Airport (EWR) - 0.1354	Boston, MA - 0.0206	Alexandria, VA - 0.0001

Betweenness Centrality

Rank nodes by how common it is for paths to pass through them

Meaning for our Network:

Ranking of cities or train stations by the likelihood that any trip planned will make a stop at that city or train station

Betweenness Centrality

Rank	Amtrak Network	Non-Induced Graph	Induced Graph
1	Chicago, IL - Union Station (CHI) - 0.5779	Chicago, IL - 0.0896	Fullerton, CA - 0.0007
2	Cleveland, OH (CLE) - 0.3454	Naperville, IL - 0.0707	Los Angeles, CA - 0.0006
3	South Bend, IN (SOB) - 0.2986	Alexandria, VA - 0.0474	Anaheim, CA - 0.0003
4	Elkhart, IN (EKH) - 0.2973	Indianapolis, IN - 0.0406	Riverside, CA - 0.0003
5	Waterloo, IN (WTI) - 0.2965	Greensboro, NC - 0.0400	Oakland, CA 0.0002
6	Toledo, OH (TOL) - 0.2947	Greensboro, NC - 0.0400	Santa Ana, CA - 0.0002
7	Sandusky, OH (SKY) - 0.2939	Omaha, NE - 0.0348	San Bernardino, CA - 0.0002
8	Elyria, OH (ELY) - 0.2930	Cleveland, OH - 0.0335	Pomona, CA - 0.0002
9	Sacramento, CA - Sacramento Valley Station (SAC) - 0.2827	Lincoln, NE - 0.0334	New York, NY - 0.0002
10	Naperville, IL (NPV) - 0.2810	Denver, CO - 0.0320	San Jose, CA - 0.0002

Conclusion

Does the Amtrak network do a good job of connecting major cities?

- No
  - Multiple Components
  - Large Diameter

Do either of the sampled networks accurately reflect the original?

- Non-Induced
  - Maybe, a lot of important information is missing, and it is usually better to use the original network since it is small enough to be analyzed easily, but it was accurate enough to give useful information for assortativity
- Induced
  - No, too much information is missing such that it does not reflect the behavior of the original graph in any of the analysis


Questions?