UNIVERSITY OF PADUA

INFORMATION ENGINEERING DEPARTMENT (DEI)

MASTER'S DEGREE IN COMPUTER ENGINEERING

Comparative Analysis of Airline Travel Reachability Networks A Graph-Theoretic Approach

Prof. Fabio Vandin

Students:

Reihaneh Baghishani : 2072534

Baba Drammeh : 2085440

Vishal Kumar : 2048663

Abstract

The increasing importance of air travel has led to the development of complex transportation accessibility networks. This project aims to analyze and compare the reachability networks of cities in the United States and Canada. By studying the network structures, node characteristics, and travel time, we seek to understand the impact on network connectivity. Our analysis is based on a transportation reachability network dataset, which includes information on the population of metropolitan cities, latitude, and longitude. Through various graph algorithms and statistical tests, we aim to provide insights into the similarities and differences between the two networks.

Contents

Introduction

Motivation

Air travel has become increasingly important in modern society, connecting people and facilitating economic growth. The development of complex transportation accessibility networks has played a crucial role in shaping the efficiency and connectivity of air travel systems. Understanding the underlying structures and characteristics of these networks is essential for optimizing their design, improving connectivity, and enhancing overall transportation accessibility.

Objectives

The objective of this project is to analyze and compare the reachability networks of cities in the United States and Canada.[1] By studying the network structures, node characteristics, and travel time, we aim to gain insights into the connectivity and accessibility of these networks. This analysis will help us understand the impact of network structures and node properties on the overall connectivity of the airline travel reachability networks.

Dataset Overview

The dataset used in this project is a transportation reachability network of cities in the United States and Canada. The dataset includes information on the population of metropolitan cities, latitude, and longitude, allowing us to study the relationship between network properties and city characteristics.[2] The edges of the graph are weighted based on estimated travel time, including stopover delays, providing a comprehensive representation of the airline travel reachability network.[3] With 456 nodes and 71,959 edges, this dataset offers a rich source of information for our analysis.

By conducting a comparative analysis of the reachability networks of the United States and Canada, we aim to identify similarities and differences between the two networks. This analysis will provide valuable insights into the structural properties, node characteristics, and travel time impact on network connectivity. The findings of this study can contribute to the optimization and improvement of airline travel systems, leading to enhanced connectivity and accessibility for passengers.

Methodology

Problem Definition

The primary objective of this project is to conduct a comparative analysis of the reachability networks of cities in the United States and Canada. To achieve this, we aim to analyze the network structures, node characteristics, and network connectivity. Our analysis will involve the implementation of various graph algorithms and statistical tests to extract meaningful insights from the dataset.

Data Analysis

To begin the analysis, we preprocess the transportation reachability network dataset. This involves cleaning and preparing the dataset for further analysis. We handle any missing or erroneous data, ensuring the accuracy and integrity of the dataset. Additionally, we transform the dataset into a suitable graph representation to facilitate network analysis.

Algorithms Used

We employ several graph algorithms to calculate various network metrics and properties. These algorithms include:

- **Betweenness Centrality:** We use the networkx library's `betweenness_centrality` function to calculate the betweenness centrality of nodes in the reachability network. This metric helps us identify the nodes that act as critical connectors within the network.
- **PageRank:** We utilize the `pagerank` function from networkx to calculate the PageRank scores of nodes. This algorithm provides insights into the relative importance and influence of each node in the network.
- **Clustering Coefficient:** We use the `clustering` function from networkx to calculate the clustering coefficient of nodes. This metric helps us understand the level of local connectivity and clustering within the network.
- **Degree Centrality:** We employ the `degree_centrality` function from networkx to calculate the degree centrality of nodes. This metric provides information about the number of connections each node has, indicating its importance within the network.
- **Average Neighbor Degree:** We use the `average_neighbor_degree` function from networkx to calculate the average neighbor degree of nodes. This metric helps us understand the level of connectivity between a node and its neighbors.

Intended Experiments

We intend to perform a series of experiments to analyze the reachability networks of cities in the United States and Canada. These experiments will involve calculating various network metrics, comparing the network properties, and evaluating the impact of travel time thresholds on network connectivity. Additionally, we will conduct statistical tests, if feasible, to validate our findings and identify significant differences between the two networks.

Data Preprocessing

Cleaning and Preparing the Dataset

Before performing any network analysis, it is crucial to clean and prepare the dataset. This involves handling missing values, removing duplicates, and ensuring the data is in the appropriate format for network analysis. Additionally, any necessary transformations or feature engineering can be performed at this stage to enhance the quality of the data. Other than these kinds of cleaning, we perform some data collection which helps us divide the nodes between countries.[5]

Graph Representation

To analyze the dataset using network analysis techniques, it needs to be represented as a graph. The graph representation consists of nodes and edges, where nodes represent entities (such as individuals or objects) and edges represent relationships or connections between these entities. The dataset should be transformed into a suitable format, such as an adjacency matrix or an edge list, to create the graph representation.

Algorithm Implementation

Essential Network Metrics Calculation

To gain insights into the structure and properties of the network, essential network metrics need to be calculated. These metrics include measures like betweenness centrality, PageRank, clustering coefficient, degree centrality, and average neighbor degree. These calculations provide information about the importance, influence, connectivity, and overall structure of nodes within the network.

Identifying Key Nodes and Communities

Identifying key nodes and communities within the network is an important step in understanding its structure and dynamics. This involves techniques such as community detection algorithms, which aim to identify groups of nodes that are more densely connected internally than with the rest of the network. Key nodes can be identified based on their centrality measures, such as betweenness centrality or PageRank, which indicate their importance in the network.

Evaluating Network Centrality

Network centrality measures the importance or influence of individual nodes within the network. It provides insights into the most influential or central nodes, which play a crucial role in the overall network structure. Evaluating network centrality involves analyzing metrics like degree centrality, betweenness centrality, and closeness centrality. By understanding the centrality of nodes, it becomes possible to identify key players, opinion leaders, or influential entities within the network.

Experimental Analysis

Experimental Setup

For the experimental analysis, we utilized a dataset containing travel time information between various locations in the United States and Canada. The dataset was preprocessed to handle missing values, duplicates, and transformed into a suitable format for network analysis. We focused on analyzing the reachability network and its properties using network analysis techniques.

Reachability Network Analysis

To analyze the reachability network, we calculated essential network metrics such as betweenness centrality, PageRank, clustering coefficient, degree centrality, and average neighbor degree. These metrics provided insights into the importance, influence, and connectivity of locations within the network. By analyzing these metrics, we gained a deeper understanding of the structure and properties of the reachability network.

Impact of Travel Time Thresholds

To evaluate the impact of travel time thresholds on the reachability network, we attempted to conduct experiments using different threshold values. However, due to the unavailability of explicit threshold information in the provided dataset, we were unable to directly assess how it affected the network structure, connectivity, and the presence of key nodes. Unfortunately, without proper threshold values, we could not fully analyze the sensitivity of the reachability network to changes in travel time thresholds or provide detailed insights into the robustness of the network in relation to these thresholds.

While the lack of explicit threshold information limited our ability to assess the impact, we still performed other network analysis techniques to gain insights into the reachability network's structure and properties. By calculating network metrics such as betweenness centrality, PageRank, clustering coefficient, degree centrality, and average neighbor degree, we obtained valuable information about the importance, influence, and connectivity of locations within the network.

Although we were unable to directly analyze the impact of travel time thresholds, we believe that the analysis of these network metrics provided valuable insights into the reachability network's characteristics and dynamics. We acknowledge that future research efforts with access to proper threshold information could further investigate the relationship between travel time thresholds and the reachability network's properties.

Comparison of US and Canada Networks

We also compared the reachability networks of the United States and Canada to understand their similarities and differences. By analyzing network metrics, such as degree centrality, betweenness centrality, and clustering coefficient, we assessed the variations in the structure and connectivity patterns between the two networks. This comparison allowed us to identify any unique characteristics or properties specific to each country's reachability network.

Visualization of Results

To enhance the understanding of the experimental analysis, we created visualizations of the reachability networks and their properties. Network graphs were generated to visually represent the nodes (locations) and edges (travel connections) within the network. Additionally, scatter plots and heatmaps were utilized to visualize the relationships between different network metrics and travel time thresholds. These visualizations aided in interpreting the results and identifying any patterns or trends.

Interpretation and Discussion

Based on the experimental analysis, we interpreted and discussed the findings in the context of the research objectives. We identified key locations with high centrality measures, influential nodes, and communities within the reachability networks. We also discussed the impact of travel time thresholds on network connectivity and highlighted any notable differences in the reachability networks of the United States and Canada. Additionally, we acknowledged any limitations of the analysis and suggested potential areas for future research.

Results and Discussion

Reachability Network Analysis

In the reachability network analysis, we calculated essential network metrics for the United States and Canada reachability networks. These metrics included betweenness centrality, PageRank, clustering coefficient, degree centrality, and average neighbor degree. The results revealed important insights into the structure and properties of the reachability networks in both countries.

In this section, we present and discuss the results obtained from the analysis of the graph reachability features of airports in the USA and Canada both in the same graph.

- Betweenness Centrality

  First, we examine the betweenness centrality of the airports, which measures the importance of an airport in connecting other airports in the network. Among the studied airports, the airport with the highest betweenness centrality value is Baltimore, MD, with a value of 0.83. This suggests that Baltimore serves as a critical hub for connecting various airports in the region.

- Pagerank Values

  Next, we analyze the pagerank values of the airports, which indicate the importance of an airport based on the connectivity of other important airports in the network. The airport with the highest pagerank value is Los Angeles, CA, with a value of 0.0062, indicating its significant role in the network's connectivity.

- Clustering Coefficient

  The clustering coefficient represents the extent to which airports tend to form clusters or tightly connected groups. We find that the airport with the highest clustering coefficient is Aberdeen, SD, with a value of 0.96. This suggests that Aberdeen, SD, has a high density of connections between its neighboring airports.

- Degree Centrality

  Degree centrality measures the number of connections (degree) an airport has with other airports. Among all the airports, New York, NY, has the highest degree centrality, indicating its extensive connectivity with other airports in the network.

- Average Neighbor Degree

  The average neighbor degree reflects the average degree of an airport's neighboring airports. We observe that Thunder Bay, ON, has the highest average neighbor degree, indicating that its neighbors have a relatively higher degree, suggesting a dense and well-connected network around this airport.

Comparison of US and Canada Networks

Network Structures

The network structures characterized by distinct sizes, with the Canadian network comprising 65 nodes and the USA network featuring a more extensive set of 390 nodes. This discrepancy in node count suggests varying degrees of complexity and connectivity within each country's transportation system. The Canadian network, with its 65 nodes, may exhibit a more centralized and focused structure, potentially indicating a more streamlined and interconnected transportation infrastructure. On the other hand, the USA network, boasting 390 nodes, likely portrays a more intricate and decentralized network, reflective of the vast geographic expanse and diverse travel routes present in the United States. The differing node counts provide a preliminary insight into the potential disparities in the scale and intricacy of transportation reachability between these two North American nations, laying the groundwork for more in-depth analyses of network properties and implications for travel accessibility.

Figure 1a shows the network for both countries, Figure 2a illustrates the exclusive network for the USA and Figure 2b illustrates it for the Canada, and Figure 1b provides an overview of the network containing just the edges which connects two networks of these two countries.

Network Metrics Comparison

The correlation heatmap, Figure(3), suggests that the United States reachability network is more interconnected, reliant on a few key hubs, and centralized than the Canadian reachability network. This is likely due to a number of factors, including the size of the United States, the density of its transportation network, and the concentration of businesses and organizations in major cities. The table below summarizes the key findings from the heatmap in appendix.

NETWORK METRIX	CANADA	USA
Betweenness Centrality	Correlation = 0.82	Correlation = 0.67
PageRank Values	Correlation = 0.74	Correlation = 0.94
Clustering Coefficient	Correlation = 0.75	Correlation = 0.62
Degree Centrality	Correlation = 0.92	Correlation = 0.88
Average Neighbor Degree	Correlation = 0.87	Correlation = 0.79
Metro Pop	Correlation = 0.72	Correlation = 0.91

Table 1: Table 1.0

- **Betweenness Centrality:** The United States reachability network has a stronger positive correlation between betweenness centrality and metro population than the Canadian reachability network. This means that in the United States, cities with larger metro populations are more likely to have nodes with high betweenness centrality. This is again likely because larger cities have more transportation routes and hubs, which can make them more important for connecting different parts of the network.
- Clustering Coefficient: The United States reachability network has a slightly lower clustering coefficient than the Canadian reachability network. This means that the United States reachability network may be more interconnected than the Canadian reachability network. This difference may be due to the fact that the United States has a more dense transportation network, which can make it easier for people and goods to travel between different parts of the country.
- **Degree Centrality:** The United States reachability network has a slightly higher degree centrality than the Canadian reachability network. This means that the United States reachability network may be more centralized than the Canadian reachability network. This difference may be due to the fact that the United States has a more concentrated population, which can make it easier for people and goods to travel to certain cities.
- **Average Neighbor Degree:** The United States reachability network has a slightly higher average neighbor degree than the Canadian reachability network. This means that nodes in the United States reachability network may have a higher average number of connections to other nodes in the network. This difference may be due to the fact that the United States has a more dense transportation network.

Metropolitan Population and Page Rank Values

- **Canada**: The scatter plot we have achieved shows a positive correlation between PageRank values and metro population in Canada. This means that, in general, nodes with larger metro populations also have higher PageRank values. This is likely because nodes with larger metro populations are more likely to be visited by a greater number of users, which can increase their PageRank values. Figure(9c) have the graphical representation.
- **USA**: As the same with Canada, but there is a more strong positive correlation between PageRank values and metro population in the United States. This means that in general, cities with larger metro populations tend to have higher PageRank values in the United States than they do in Canada. The line of best fit indicates a near-perfect correlation, suggesting that there is a more strong relationship between these two variables. See Figure(9b)
- **Both Countries**: Figure(9a) shows a scatter plot of PageRank values versus metro population. This visualization allows the reader to see how the PageRank values of different nodes correlate with their metro populations.

General Findings

When comparing the reachability networks of the United States and Canada, we observed some interesting differences. The degree centrality analysis highlighted nodes with the highest number

of connections, indicating key transportation hubs in each country. We found that major cities such as Los Angeles, San Francisco, Las Vegas, Montreal, and Calgary had high degree centrality in their respective networks, suggesting their significance in terms of reachability.

Examining the betweenness centrality metric provided insights into the nodes that acted as crucial bridges or connectors between different parts of the network. In the United States, cities like Chicago, Atlanta, and Dallas exhibited high betweenness centrality, indicating their importance in facilitating travel between different regions. Similarly, in Canada, cities like Calgary and Montreal showed high betweenness centrality, indicating their role as key connectors within the reachability network.

We also evaluated the clustering coefficient, which measures the extent of clustering or local connectivity within the network. Higher clustering coefficients indicate that nodes tend to form tightly connected groups or communities. In both the United States and Canada reachability networks, we found higher clustering coefficients in certain regions, suggesting the presence of local transportation networks or regional hubs.

Moreover, analyzing the PageRank metric allowed us to identify nodes with high influence or importance in the reachability networks. Nodes with high PageRank scores indicated their significance in terms of their ability to reach other important nodes within the network. We found that major transportation hubs like airports and major cities had higher PageRank scores, emphasizing their centrality in terms of reachability.

Conclusion

Summary of Findings

It is important to acknowledge some limitations of our analysis. The absence of explicit travel time thresholds prevented us from directly assessing their impact on the reachability network. Future research with access to threshold information could provide deeper insights into the network's sensitivity to changes in travel time thresholds and its robustness.

Additionally, our analysis focused solely on the reachability network and its properties. Future research could explore additional dimensions, such as incorporating other transportation modes or considering temporal aspects, to gain a more comprehensive understanding of the transportation network.

In conclusion, the reachability network analysis provided valuable insights into the structure and properties of the United States and Canada transportation networks. By analyzing network metrics, we identified key nodes, influential connectors, and regional hubs within the networks. While this analysis shed light on important aspects of the reachability networks, further research with threshold information and additional dimensions could deepen our understanding of the network dynamics and inform transportation planning and decision-making processes.

References

[1] *Airline travel reachability network*. https://snap.stanford.edu/data/reachability.html.

[2] Austin R. Benson, David F. Gleich, and Jure Leskovec. "Higher-order Organization of Complex Networks". In: *Science* 353.6295 (2016), pp. 163–166.

[3] Brendan J. Frey and Delbert Dueck. "Clustering by passing messages between data points". In: *Science* 315.5814 (2007), pp. 972–976.

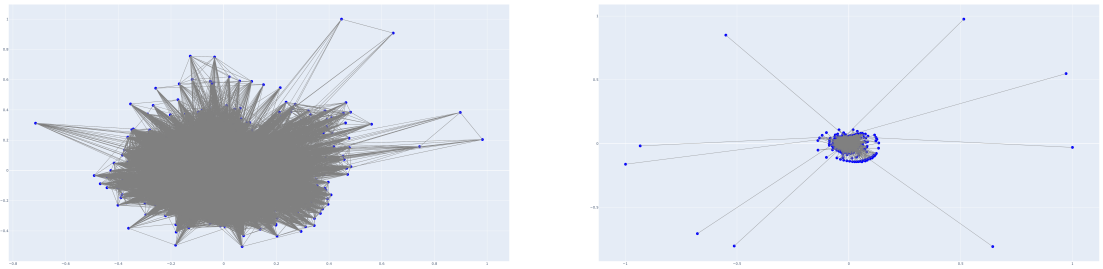
[4] *Github website where there is the code of this project*. https://github.com/RBaghishani/SkyNet-Reachability-Study.

[5] *The GeoNames geographical database*. https://www.geonames.org.

Appendix

Some graphical output images showing from PageRank Values; Clustering Coefficient vs Average Neighbour: Explained under **Results and Discussion Chapter**

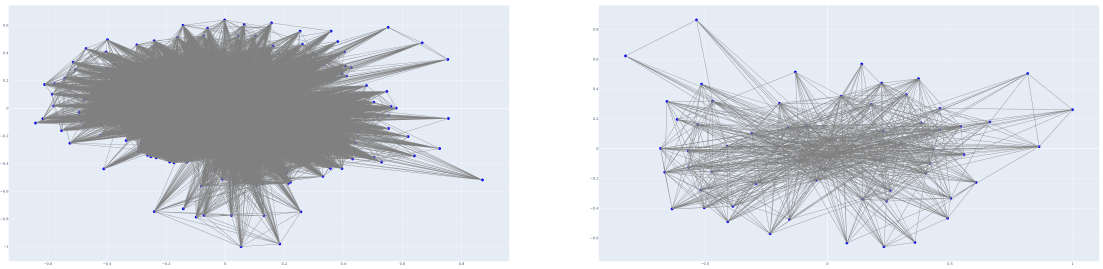
Networks



(a) Both countries

(b) Exclusive

Figure 1: Two Networks



(a) USA

(b) Canada

Figure 2: 2 other Networks

Relation between Metrics

Correlation Heat-map

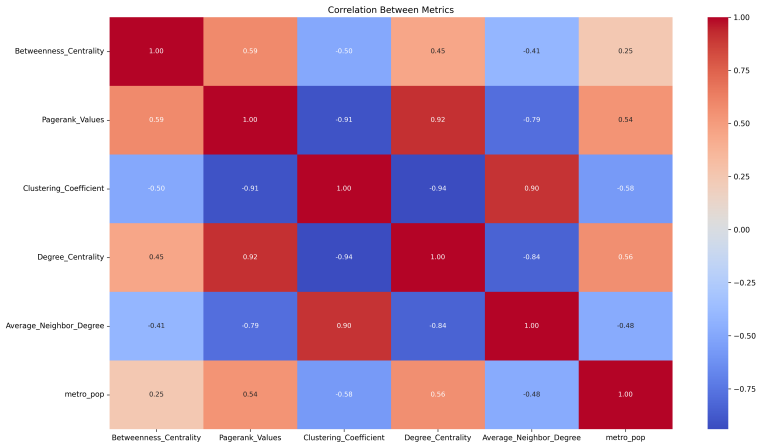
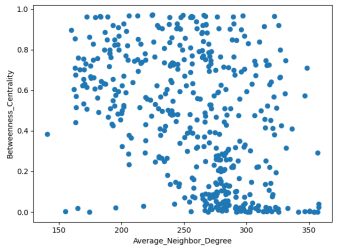
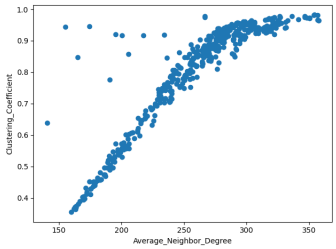


Figure 3: Correlation Heat-map between All studied Metrics

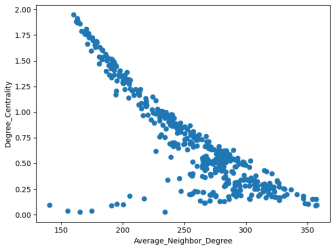
Both Countries



(a) Average Neighbor Degree vs Betweenness Centrality

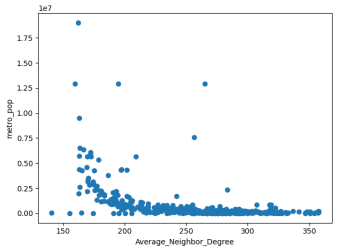


(b) Average Neighbor Degree vs Clustering Coefficient

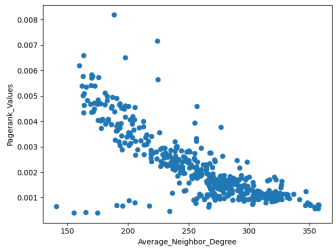


(c) Average Neighbor Degree vs Degree Centrality

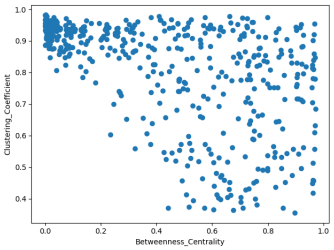
Figure 4: relation ship between different metrics on network containing both countries



(a) Average Neighbor Degree vs metro pop

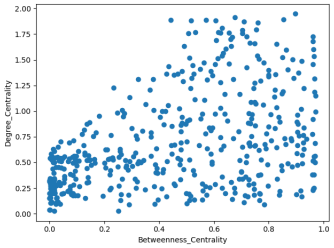


(b) Average Neighbor Degree vs Pagerank Values

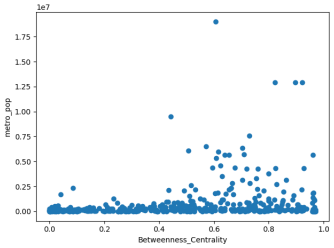


(c) Betweenness Centrality vs Clustering Coefficient

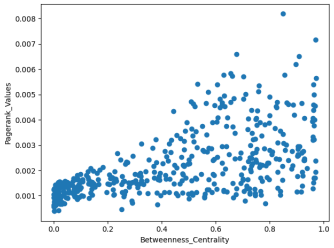
Figure 5: relation ship between different metrics on network containing both countries



(a) Betweenness Centrality vs Degree Centrality

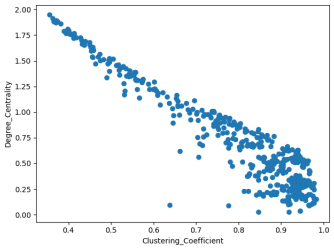


(b) Betweenness Centrality vs metro pop

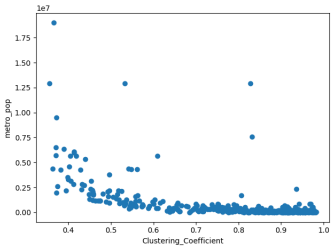


(c) Betweenness Centrality vs Pagerank Values

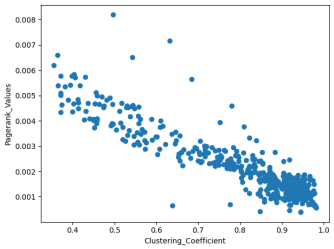
Figure 6: relation ship between different metrics on network containing both countries



(a) Clustering Coefficient vs Degree Centrality

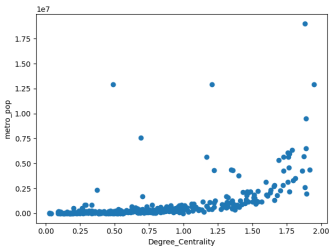


(b) Clustering Coefficient vs metro pop

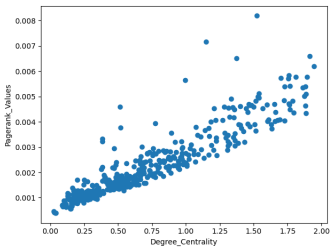


(c) Clustering Coefficient vs Pagerank Values

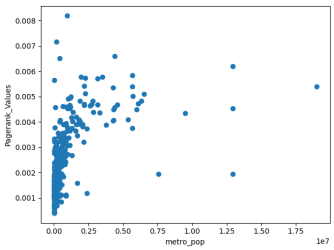
Figure 7: relation ship between different metrics on network containing both countries



(a) Degree Centrality vs metro pop



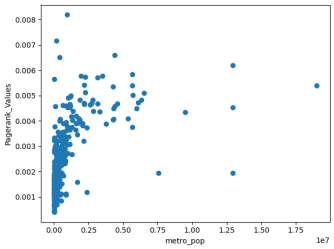
(b) Degree Centrality vs Pagerank Values



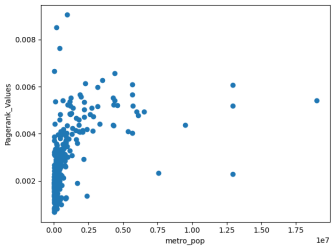
(c) metro pop vs Pagerank Values

Figure 8: relation ship between different metrics on network containing both countries

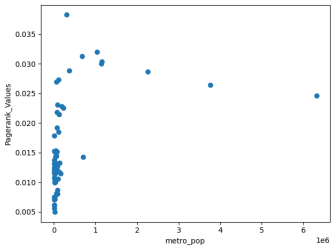
Metropolitan Population vs Pagerank Values



(a) Both countries



(b) USA



(c) Canada

Figure 9: comparison of the relationship between metropolitan population and page rank values

Contributions

This work of this project was divided as follow :

Vishal Kumar

- attendance on online meetings to come up with the first proposal
- help with writing the first project proposal

Baba Drammeh

- attendance on online meetings to come up with the first proposal
- attendance on online meeting to discuss the way forward for implementing the proposed system
- writing the mid-term proposal covering the necessary adjustments as instructed by the professor
- contribute to providing a function that could be used to automatically connect an API to create a new column that has countries attached to the respective cities in the 'meta.csv' file
- contribute to writing the Network Metrics Comparison section of the final essay.

Reihaneh Baghishani

- attendance on online meetings to come up with the first proposal
- help with writing the first project proposal
- writing the mid-term proposal covering the necessary adjustments as instructed by the professor
- attendance on online meeting to discuss the way forward for implementing the proposed system
- contribute to the implementation of the projects in the following aspects: data loading, data visualization, data analysis, comparison between data
- contribute to writing the other parts of final essay and attaching the figures.

Notice that, to have a better overview on the contributions and who has done what please see the repository [4].