Optimizing Public Transportation Efficiency in Vancouver: A Graph Theory Approach

Zhixin Zeng, Wei Song, Chanyuan Liu

Introduction

Public transportation is complicated in one of Canada's busiest and most crowded cities and plays a crucial role in people's daily lives. In 2024, approximately 2.6 million people will live in Metro Vancouver, making it the third-largest metro area in the country. An efficient transit system is a must to support the city's diverse and growing population. Thousands of people depend on Vancouver's public transportation system, which TransLink manages, for their daily travels to work, school, and other activities. Route inefficiencies and time gaps often result in an unhappy commuter and poor performance in the system. Recent customer feedback highlights several pain points in Vancouver's public transportation, including route efficiency and operational performance. Commuters frequently experience issues such as overcrowded trains and buses, inadequate peak-hour service, and inconsistent scheduling, leading to delays and extended travel times. Inefficiencies like these are more than just inconveniences. They affect the city's economy and residents' quality of life.

Research Question: "How can the Floyd-Warshall algorithm[1] be used to create a distance matrix, combined with clustering analysis[2] and Dijkstra's algorithm[3], be employed to optimize the positioning of new transit stops within Vancouver's public transportation network and assess the potential impacts of these new routes on overall network efficiency?"

In this question, we are trying to address the inefficiencies and improve transit for Vancouver's growing population. The Floyd-Warshall's all-pairs shortest path matrix will provide a comprehensive overview of current route efficiencies and inefficiencies by highlighting the shortest paths across all pairs of stops. Following that, a clustering analysis will help identify underserved regions within the network or could significantly benefit from adding new stops. Integrating Dijkstra's algorithm allows for precise modeling of potential new routes, offering a detailed overview of how these additions could enhance connectivity, reduce travel times, and improve service reliability across the network.

Rationale

The research question addresses the critical challenges Vancouver's public transportation system faces. These issues are especially pressing given the rapid population growth and urban expansion in Vancouver, which stress existing infrastructure and demand effective solutions. An inspiring response to such demands is TransLink's RapidBus Projects[4], which have enhanced the customer

experience through more widely spaced stops, all-door boarding, and extensive bus priority measures such as queue jumps or bus lanes. During rush hours, RapidBus routes accommodate up to 12,000 passengers per hour, significantly reducing road congestion by replacing 34 ferries' worth of single-occupancy cars. These routes operate frequently throughout the day, ensuring reliable service from early morning until midnight across six significant corridors in the region.[5]. Applying the Floyd-Warshall algorithm, combined with clustering analysis and Dijkstra's algorithm, addresses these challenges by harnessing comprehensive data-driven insights to propose optimal adjustments to the transit network.

Significance of the Question: This question is significant as it leverages advanced computational techniques to propose practical solutions for enhancing public transit efficiency. By optimizing the placement of new transit stops and assessing the impacts of new routes, this study aims to directly improve overall network performance, which could lead to shorter travel times, better service coverage, and improved passenger satisfaction. The success of the RapidBus project demonstrates the potential benefits of such optimized transit solutions, which are designed to effectively meet the demands of a growing urban population.

Gap in Knowledge: Despite the critical role of public transportation in urban sustainability and mobility, many existing networks operate based on outdated models that need to reflect current urban dynamics or future growth projections adequately. This project addresses this gap by integrating modern algorithmic approaches to reevaluate and enhance the network design. The combination of the Floyd-Warshall algorithm, Clustering algorithms, and Dijkstra's algorithm provides a novel approach to transit planning, offering a holistic view of the network's inefficiencies and potential improvements.

This research focuses on the above aspects to improve Vancouver's public transportation and our understanding of and skills in a real-world setting. We can apply complex theoretical concepts to real-life situations through this project, which offers a great learning opportunity. We seek opportunities to enrich our knowledge of graph theory, algorithm design, and data analysis, skills that are highly applicable to the rapidly evolving field of urban planning. The project could boost our academic and professional development and matters of public interest.

Personal Importance

Zhixing Zeng: As a Computer Science student living in Coquitlam and commuting to downtown Vancouver, I heavily rely on the SkyTrain system operated by TransLink. The SkyTrain is my primary mode of transportation for attending classes, exploring the city, and occasionally traveling to and from Vancouver Airport. For instance, traveling from Coquitlam to YVR is akin to traversing two sides of an equilateral triangle instead of taking the direct path between the two points. This circuitous route makes me wonder if there might be a more optimal solution.

Wei Song: After a year of studying Computer Science at NEU, I was lucky to participate in the Translink-hosted hackathon last year. inspired my passion for applying computer science concepts to address complex real-world problems. The task we faced was inspiring: How could we leverage data, existing or newly proposed, to enhance the customer journey, ensuring it is safe, pleasant,

easy to use, timely, and reliable? Participating in this hackathon, I became aware of the real-world impact of seemingly abstract concepts like route planning and graph theory. It dawned on me that delivery trucks routes or the efficient mapping of city public transportation arent just problems on paper; when solved, they can make life better.

Chanyuan Liu:As a student living alone in Vancouver for half a year, good public transportation is extremely important to me. I need to take the SkyTrain to school every week, and occasionally to and from Vancouver Airport. However, even though I live near a very convenient station Metrotown, I still find it headache-inducing to transfer to some very common places from here. For example, going from Metrotown to Vancouver Airport requires transferring at downtown, which is like going from point A to point B of an equilateral triangle, but not taking the direct line between A and B. Instead, it goes from A to C, and then from C to B. I hope to apply the path algorithms Ive learned at school to find possible better routes. Ultimately, this will help me better integrate into this city and discover more of its beauty.

Methodology

Research methods, Algorithms and Techniques

Data Collection and Data Analysis: Our project leverages the General Transit Feed Specification (GTFS) data[6], a standard format for public transportation schedules and associated geographic information. Provided by TransLink, this data includes detailed descriptions of stops, routes, trips, and other attributes relevant to Vancouver's public transportation network. We ensured the data's accuracy and relevance by sourcing it directly from official and authorized transit databases. Additionally, we filtered the data only to include stops within specific geographic coordinates encompassing parts of Metro Vancouver, focusing on the areas most pertinent to our study.

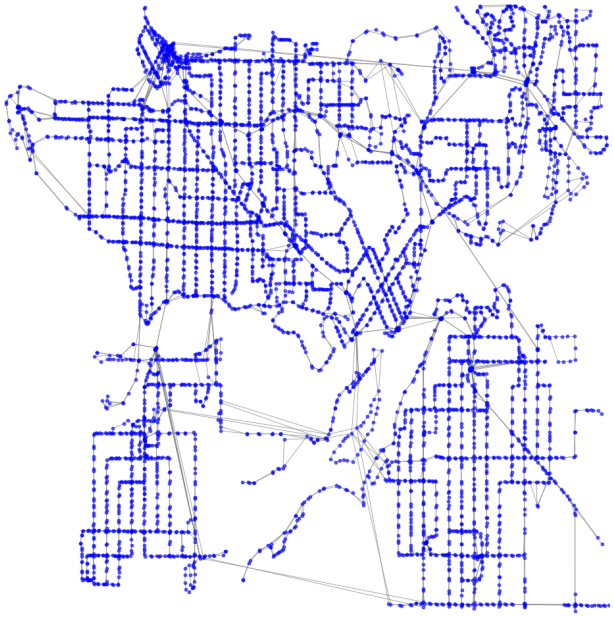

Figure 1: Visualization of the Transit Network

Tools and Software: We selected Python for the project because of its comprehensive support for data processing and analysis, as well as its extensive ecosystem of libraries tailored to all aspects of computational work.
In our project, we employed several essential Python libraries to handle different aspects of data manipulation, analysis, and visualization. We used **Pandas** for data processing and management, which enabled us to organize and preprocess the GTFS data efficiently. **NetworkX** was instrumental in analyzing graph-based representations of Vancouver's transit network, allowing us to explore connections and flow within the system. We utilized **NumPy** to ensure high-performance operations for numerical computations, particularly those involving arrays and matrices. Data and network structures were visualized using **Matplotlib**, providing us with the tools to generate insightful graphical representations of our findings. Lastly, **Scikit-learn (sklearn)** was applied to implement machine learning algorithms, helping us to predict and optimize transit stop placements based on patterns derived from the data.

Algorithms and Techniques

Graph Construction:

- We modeled Vancouver's public transportation network as a directed graph using GTFS data. Each transit stop is a node, with direct routes between stops as edges, where travel times are incorporated as edge weights. We used Pandas for data manipulation and NetworkX for graph construction. The process involved filtering GTFS data files for stops within designated geographic coordinates of Metro Vancouver to ensure focus on relevant areas. Stops were filtered based on latitude, longitude, and ID constraints, and corresponding stop times data were integrated with trip and route details to construct the network. This directed graph, built using NetworkX, encapsulates nodes representing stops with attributes including names

and coordinates and edges that signify routes with travel time-derived weights, facilitating precise path finding and network analysis critical for optimizing transit system efficiency.1

Floyd-Warshall Algorithm for All-Pairs Shortest Path:

- This algorithm is applied to compute the shortest paths between every pair of graph nodes. This comprehensive matrix of shortest paths serves as the foundation for both the analysis of network robustness and the optimization of the network.

Network Robustness Analysis

- **Average Shortest Path Length Calculation:** Initially, the average shortest path length across the entire network is calculated. This measure establishes a baseline for assessing the overall efficiency of the network. The calculation considers all viable paths, using finite values from the shortest paths matrix to compute the mean.

- **Node Removal Simulation:** To evaluate network robustness, the effect of each node (stop) on network connectivity is tested by simulating its removal. The shortest paths matrix is modified for each node by removing the node and recalculating the average shortest path length. This step assesses the impact of each node's removal on the overall network connectivity and efficiency, quantified by changes in the average shortest path length compared to the baseline.

- **Identification of Top 100 Critical Nodes:** Nodes are ranked based on their impact on the network, measured by the average shortest path length change resulting from their removal. The nodes causing the most significant increases in path lengths are identified as critical. The top 100 nodes with the highest impact scores are selected for further analysis, prioritizing those whose absence would most disrupt the network's operational efficiency.

Analysis of Critical Nodes Using Clustering

- **Extraction and Clustering of Critical Nodes:** We extract key features for the top 100 critical nodes identified from the robustness analysis, such as the node's connectivity degree and geographical coordinates (latitude and longitude). These features are standardized to facilitate a uniform analysis. The nodes are then subjected to a cluster analysis using the K-means algorithm, where the optimal number of clusters is determined using the Elbow Method. This clustering segregates nodes into groups based on their connectivity and spatial distribution, aiming to discern patterns and common characteristics that could inform strategic transit planning.

- **Visualization and Detailed Cluster Analysis:** After clustering, the nodes are visually represented on a map to illustrate their spatial distribution across different clusters. Each cluster is further analyzed to identify the most centrally located node based on its degree of connectivity. This analysis includes reverse geocoding to fetch detailed location information (such as street and community names) for these central nodes, enhancing our understanding of the structural and spatial characteristics of critical nodes within the network. Such insights facilitate targeted interventions and optimizations in the transit network.

Network Optimization Using Dijkstras Algorithm

- **Network Expansion and Optimization:** Utilizing the identified critical nodes, new routes are strategically proposed to improve network connectivity. This step involves using Dijkstras algorithm to simulate potential changes in the network by adding new routes between critical nodes and other under-connected nodes. The goal is to model and evaluate how these additions could reduce travel times and enhance overall network efficiency. The new routes are specifically designed to link critical nodes that are not directly connected, optimizing the networks structural robustness and ensuring better service coverage.

- **Impact Evaluation of New Routes:** The effectiveness of these new routes is quantitatively assessed by calculating the percentage improvement in travel times across the network. Dijkstras algorithm is applied again to determine the shortest paths after integrating the new routes, comparing these results against the baseline travel times established before the additions. The optimization aims to significantly decrease overall travel times, enhance connectivity, and facilitate quicker, more efficient travel across the network, focusing on integrating strategically important nodes that were previously under-connected.

Considerations and Limitations

Ethical Considerations:

- **Data Privacy:** While GTFS data is generally anonymized and public, careful consideration was taken to ensure no privacy invasions occur, particularly when handling data that could be traced back to individual travel patterns or behaviors if combined with other data sources.

Limitations in methodology:

- **Assumptions on Travel Times:** The Assumption operated for this project is that travel times provided by GTFS data accurately represent real-world conditions. However, this assumption may not always be true at times due to various factors such as traffic fluctuations, delays, and other operational challenges that can affect real-time bus transit time costs.

- **Static Data Use:** The analysis is primarily based on static scheduling data. It does not consider real-time traffic conditions or dynamic changes in commuter demand, which can significantly influence transit efficiency and service needs.

- **Geographical and Operational Constraints:** Practical constraints limit the recommendations for network enhancement. These include physical barriers, regulatory restrictions, and budgetary limitations, which were not fully integrated into the network modeling process.

Analysis & Findings

Network Robustness and Stability:

- Our initial analysis centered on calculating the average shortest path length throughout the transportation network. This metric established a baseline for understanding the network's efficiency under existing conditions. Our findings highlighted that specific nodes are crucial

for maintaining short travel times, indicating areas where the network is robust yet also exposing potential vulnerabilities.

- Subsequently, we assessed the network's robustness by simulating the removal of each node and recalculating the average shortest path length. This exercise revealed that removing specific nodes significantly increased overall travel times. These results identified a critical tier of nodes whose presence is essential for the network's optimal functionality.
- We generated a list of the top 100 most impactful nodes within the transportation network using computational methods. These nodes represent critical points whose potential disruption could significantly increase average travel times across the network. This identification highlights the nodes as prime candidates for service enhancements or infrastructural reinforcements, signaling strategic areas for improving network stability and efficiency.

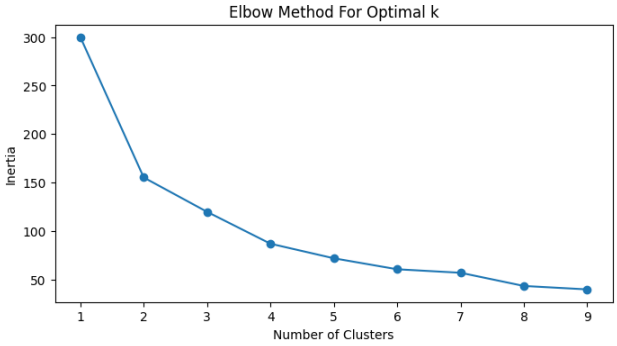

Figure 2: Elbow Method used to determine the optimal number of clusters

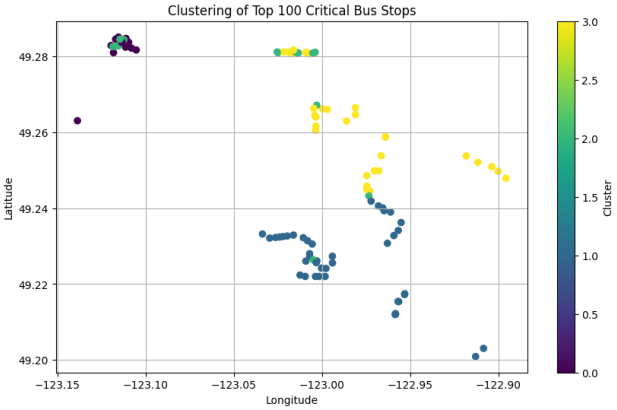

Figure 3: Clustering Top 100 nodes with highest impact on the network

Cluster Analysis of Critical Nodes: The clustering analysis has successfully delineated the characteristics and commonalities among critical nodes within the transportation network, revealing insights linked to geographical locations, node connectivity, and established travel patterns. Notably, Cluster 0 nodes, centered around West Pender Street in Central Vancouver, demonstrate a

moderate connectivity with an average degree of 4.07, suggesting a strategic hub of activity. In contrast, Cluster 2 stands out with the highest average degree of connectivity at 7.38, located on Seymour Street, also in Central Vancouver, indicating a critical node central to the network's efficiency. Meanwhile, Cluster 1, located in Metrotown with the lowest average degree of 2.81, and Cluster 3, in Burnaby with an average degree of 2.97, reveal areas with lower connectivity, highlighting potential opportunities for network expansion or strengthening to enhance overall service accessibility and efficiency.

Cluster	Avg Degree	Median Degree	Min Degree	Max Degree	Std Dev of Degree	Street Name	Community Name	Transit Stop Name
1	4.07	4.00	2.0	6.0	1.14	West Pender Street	Central Vancouver	Westbound W Pender St @ Seymour St
2	2.81	2.00	2.0	6.0	1.02	Unknown	Metrotown	EB Central Blvd @ 4500 Block
3	7.38	7.00	5.0	10.0	1.39	Seymour Street	Central Vancouver	Lougheed Station @ Bay 10
4	2.97	3.00	2.0	5.0	0.87	Unknown	Burnaby	Lougheed Stn

Table 1: Cluster Analysis Summary Table

Optimization Using Dijkstras Algorithm: We focused on enhancing connectivity between critical nodes by integrating new routes using Dijkstra's algorithm. This approach resulted in a substantial average improvement of 38.01% in travel times across all valid node pairs, significantly enhancing the network's overall efficiency.

Significance of Findings: Our network analysis and optimization results are impactful, offering quantifiable insights into the robustness of Vancouver's public transportation network and suggesting practical improvements. The identification of critical nodes and their clustering informs targeted enhancements, whether through service expansions or infrastructural investments. The notable average improvement in travel times directly aligns with our research objective to optimize the transit network, significantly enhancing commuter experiences and the system's overall functionality.

Conclusion & Future Work

Achievements and Project Impact

Our simulation results indicate that a straightforward optimization introduced into the network led to an average improvement of 38.01% in network efficiency. This underscores the effectiveness of the algorithms taught in our class, demonstrating their practical applicability in enhancing public transit systems, as supported by robust statistical evidence.

```
# Calculate average improvement
valid_improvements = [imp for imp in improvements.values() if imp is not None]
average_improvement = sum(valid_improvements) / len(valid_improvements) if valid_improvements else 0

print(f"Average percentage improvement across all valid node pairs: {average_improvement:.2f}%")
```

```
Average percentage improvement across all valid node pairs: 38.01%
```

Figure 4: Improvement after Integrating New Routes

Weaknesses and Limitations

While our project has shown promising results in optimizing Vancouver's public transit network, it is essential to acknowledge its weaknesses and limitations. These areas can provide valuable insights into the project's scope and areas for improvement.

1. Our analysis is based on GTFS data published on March 15, 2024. Since TransLink updates this data weekly, our findings may not reflect current transit conditions.

2. The Floyd-Warshall algorithm has a running time complexity of $O(n^3)$, which can lead to slow performance when processing large datasets.

3. Based on the limitations of algorithm speed as described above, our analysis focused on a specific region of Vancouver, including some parts of the Metro Vancouver area (Vancouver, Burnaby, Richmond, Coquitlam, Surrey), thus limiting the generalizability of our findings to the entire transit network.

4. Due to considerations of algorithm speed, we did not include all nodes above the threshold but artificially limited it to 100 critical nodes. Our target area has approximately six thousand nodes, with around 20% exceeding the threshold.

5. In real life, when establishing new stations and optimizing time, TransLink also needs to consider many environmental and human factors. For example, as described in the Design Guide for Bus Stops Adjacent to Cycling Infrastructure[7], based on the findings of the technical review and findings from stakeholder engagement, sixteen key issues and challenges with the design of bus stops adjacent to cycling infrastructure were identified. The various factors may impact how to design bus stops when located adjacent to protected cycling infrastructure.

6. In the GTFS datasets, the arrival_time and departure_time for each stop are exactly the same, meaning the data does not account for the time spent at each stop for boarding passengers. Different stations will inevitably have different boarding times, leading to the current data needing more accuracy.

7. Our robustness analysis only measured the impact of node removal on the network and did not include other significant factors such as user behavior, environmental impacts, or economic effects.

8. The current clustering method only considers the degree and geographical coordinates of the nodes, making the features rather limited. This approach does not comprehensively reflect the properties of each node.

9. The time calculation for the added routes is purely theoretical. In actual operation, there will certainly be discrepancies, leading to less accurate conclusions.

Future Research

1. **Real-Time Data Analysis:** Investigate the impact of using more recent data to ensure our analysis reflects the current state of Vancouver's transit network.
2. **Optimization Algorithm Enhancement:** Explore alternative algorithms or optimizations to reduce the time complexity of the Floyd-Warshall algorithm.
3. **Expand Analysis Scope:** Extend the analysis to cover a broader region of Vancouver and the surrounding areas to generalize findings and capture the diversity of the entire transit network.
4. **Node Inclusion Strategy:** Develop a dynamic node selection method that adaptively includes nodes above the threshold based on computational feasibility and criticality to the network.
5. **Multi-Factor Consideration:** Integrate factors like environmental impacts, human behaviors, and stakeholder feedback when designing new transit routes and stops.
6. **Data Accuracy Improvement:** Enhance the GTFS datasets by incorporating actual boarding times at each stop, providing a more accurate representation of transit operations.
7. **Comprehensive Robust Analysis:** Implement a more nuanced robustness analysis that accounts for user behavior, environmental factors, and economic impacts to simulate real-world scenarios better.
8. **Advanced Clustering Techniques:** Adopt advanced clustering methods that consider multiple node features beyond degree and geographical coordinates to capture the full spectrum of node properties.
9. **Practical Route Time Estimation:** Develop algorithms or models that can estimate route time based on actual transit operations data rather than theoretical calculations, leading to more reliable results. Additionally, experimental run time analysis should be incorporated to calibrate the results.

Personal Reflections

Wei Song: Throughout the project, the application and understanding of algorithms have significantly increased. We got excellent results by implementing and integrating complex algorithms such as the Floyd-Warshall and Dijkstra algorithms to optimize Vancouver's public transit network. "By adding this simple optimization to the network, we can see an average of a 38% improvement in the network efficiency". If I had possessed this practical insight and applied such algorithmic strategies during last year's Hackathon, the results would have been even more exciting. Moreover, this project has served as an excellent platform to learn about and explore new video tools. The combination of technical knowledge with creative delivery skills has been useful in effectively conveying complex information and has opened up new avenues for future projects and presentations.

Zhixin Zeng: This project is a valuable experience for me to enhance my coding skills and deepen my understanding of complex graph algorithms and machine learning algorithms. When working with my two team members, I also improved my collaborative skills. Integrating code from multiple team members required careful coordination and peer review, ensuring our final result was robust and accurate. This project improved my technical abilities and reinforced the importance of effective teamwork in achieving complex goals.

Chanyuan: In this project, I had the opportunity to delve deep into graph theory, applying what I learned in this course to a real-world problem. The journey has been both challenging and enlightening. One key takeaway from this project is the consideration of data currency and scope. We need to balance the accuracy of results with our computing resource limitations. Secondly, identifying potential optimization directions and simulating hypotheses using code offered invaluable technical insights. Practicing this in the project improved my proficiency in it. Thirdly, implementing various graph algorithms, including Floyd-Warshall and Dijkstra, provided insights into network efficiency. I also realized the importance of algorithm time complexities, especially with large datasets.

References

[1] Robert W Floyd. Algorithm 97: shortest path. *Communications of the ACM*, 5(6):345–345, 1962.

[2] Geoffrey H Ball, David J Hall, et al. *ISODATA, a novel method of data analysis and pattern classification*, volume 699616. Stanford research institute Menlo Park, CA, 1965.

[3] Edsger W Dijkstra. A note on two problems in connexion with graphs. In *Edsger Wybe Dijkstra: His Life, Work, and Legacy*, pages 287–290. 2022.

[4] TransLink. Rapidbus projects. Report, TransLink, 2024.

[5] TransLink. Transport 2050: 10-year priorities for translink. Report, TransLink, 2022.

[6] TransLink. Gtfs static data. Report, TransLink, 2024.

[7] Urban Systems for TransLink, BC Ministry of Transportation and Infrastructure. Design guide for bus stops adjacent to cycling infrastructure. Guide, TransLink, 2024.