Data Mining Analysis of SmartRider Data for Better Understanding of Travel Patterns – Journey Centered Analytics

Jordan Lilburne

Jianxin Li
Department of Computer Science and Software Engineering
The University of Western Australia

Sharon Biermann
CEED Client: PATREC

Abstract

The purpose of this research project is to design a tool capable of assisting key decision makers when evaluating transport models. To do this, the concepts of dense subnetwork discovery and transport network design are investigated. This analysis is done with the use of the SmartRider data set. This dataset contains all records of public transport system usage in Western Australia. Using this dataset, an algorithm is designed to identify the regions of the transport network with the most usage, or in mathematical terms, the dense subnetworks of the transport system. Through efficient and accurate identification of these dense subnetworks, key decision makers in network design can make more informed decisions about transport network design and hence create a much more optimised transport system. This research proposes a new dense subnetwork discovery algorithm optimised for transport analysis.

1. Introduction

1.1 Project Overview

Public transport is a key cog in any major city. Every day thousands of people in Perth utilise the public transport system to get from point A to point B. One of the key challenges of operating a public transport system is ensuring that the system is running as efficiently and effectively as possible. In order to achieve this, many different techniques and strategies have been developed to analyse public transport networks. One key area yet to be evaluated in much depth is the identification of dense subnetworks within the public transport network. This is effectively identifying the regions within the transport network with the highest levels of patronage.

1.2 Graph Theory

This research will also look into the area of dense subnetworks. This is a topic within graph theory that looks to find the set of nodes which maximises some score function. The concept of dense subnetwork discovery has been applied to many other fields such as biology and finance, however has not been applied extensively with the area of public transport. There are

two types of dense subnetwork discovery algorithms relevant to this research: the first being finding dense subnetworks of networks of limited size and the second, finding dense subnetworks of networks with no limits on size.

This concept stems from the pure mathematics topic graph theory. Graph theory in mathematics or computer science is the study or graphs and networks. Graphs consist of a set of nodes connected through edges. Figure 1 shows a basic undirected unweighted graph with nodes 1,2,3,4,5 and 6. The edges in graphs can be either directed or undirected. Directed edges involve an edge beginning at one node and ending at another, while undirected edges do not begin or end at either node. The edges can also be either weighted or unweighted. Weighted edges contain scores (weights) for each edge. These scores can represent many different quantities depending on the context of the graph. Unweighted edges are just the case of having a graph containing edges all with weight 1.
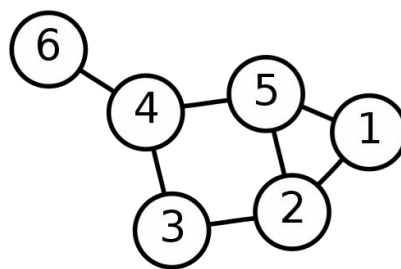

Figure 1: Graph Theory Example

2.   Process

2.1   Problem Definition

The aim of the project is to create a program capable of giving greater insight into the journeys made using the transport network within Perth. In mathematical terms, the aim of this project is to be able to answer the question: Given a time period T, and a user group U, can dense subnetworks be identified efficiently and effectively? To solve this problem, we model the transport network as a graph G=G (V, E, W), where V represents the vertices of the graph (in this case the suburbs), E, the edges in the graph (in this case the transfers between suburbs) and W, the weights associated with each edge (in this case the number of transfers between the two suburbs). Through the use of this graph, we attempt to identify the k-dense subnetworks of minimum size j within the graph network. In the context of transport networks, the identification of dense subnetworks will identify the regions within the transport system with the most usage in the time period of interest. To do this, a new algorithm is proposed, using key assumptions relevant to transport networks to improve the time efficiency from existing dense subnetwork discovery algorithms.

2.2   Dense Subnetwork Discovery Algorithm

In order to identify and validate the dense subgraphs of the Perth SmartRider network the following methodology will be used.

1.   Import Data

In order to produce the dense subnetworks data containing the SmartRider data for a certain period is required. The data will be entered into a pandas data frame. Currently, all of the

Smart Rider data is stored and collected by netBI, a Brisbane based data company. This data can be exported into a csv file and then used for analysis.

2. Clean and Merge Data

For every log in the SmartRider data log, the On Location and Off Location columns containing stop IDs, must be converted to an On and Off Suburb. This is done by merging the Stop data frame and the data itself. Any data found to not have an On and Off location is then removed from analysis. All dates and times are then converted to a time date format.

3. Query the Data

A prompt is given to the end user asking for the specific query to be made. The end user is given 4 possible options to query the data:
- Start Time (given in 15-minute intervals)
- End Time (given in 15-minute intervals)
- Smart Rider type (for example Tertiary Student, Senior, etc)
- Day of Week (Weekends, Weekdays, Monday, Tuesday, etc)

Once the query has been given, a subset of the dataset is created based on the end users' inputs. This subset will be used for the rest of the algorithm.

4. Aggregate Data by Edge ID

Every pair of unique origin and destinations suburbs is given a unique ID. The data is then grouped by this unique ID into a new data frame containing the edge ID, origin suburb, destination suburb and count for the number of trips generated between the two suburbs in the query.

5. Identify Dense Subnetwork

The following process is used to identify the dense subnetworks:
- Identify the relationship (transfer) in the graph with the most passengers using the query specifications
- Add the origin and destination of the edges to a list of potential start points
- Find the next highest connected edge (an edge with a node currently in the dense subnetwork) and add it to the subnetwork
- Continue steps b) and c) until the size of the graph exceeds the minimum size of the subgraph and there exists no connected edge such that the average density of the subgraph will be increased.

The density function used in this process is given below. This density function is a standard formula used for weighted graph networks.

$$Density = \frac{2 * \sum W\left(E_{i,j}\right)}{|H|}$$

Where:
- $W\left(E_{i,j}\right)$ is the weight of the edge between nodes i and j.
- $|H|$ is the total number of nodes inside the network (Anderson, 2009)

6. Remove Connected Nodes to Subnetworks

To reduce the overlap of dense subnetworks and increase the amount of usable information generated from the algorithm, a rule is put in place to restrict the overlap of the dense subnetworks returned. To do this, any connected edges to complete dense subnetworks (subnetworks returned in step 5), will be removed from the list of eligible edges. By doing this, there is no overlap in the dense subnetworks returned by the algorithm. This is a key difference from most other subnetwork discovery algorithms mentioned earlier. Most other dense subnetwork discovery algorithms are designed to identify the densest subnetworks regardless of overlap. Application of this assumption improves the efficiency of the algorithm and also increases the usefulness of the findings.

7. Return n-Subnetworks

Repeat step 5 and 6 until the required number of dense subnetworks have been identified or there are no sufficient edges remaining to produce a dense subnetwork.

3. Results and Discussion

3.1 Algorithm Efficiency

One of the key components of this research is the speed of the algorithm. Ideally, the algorithm will be capable of returning returning results instantly, however as the calculations become more difficult, the algorithm will take longer to run. In the context of transport networks it is important to calculate the run time of the algorithm compared to the size of the network of interst as the algorith must be capable of returning results within a reasonable time frame for subnetworks of all sizes found within the Perth transport network.

In order to investigate the time efficiency of the algorithm, the time to compute the subgraphs in the queries above (and some additional queries) is compared with the size of the graph associated. The results are shown below in the table. A plot of the run time of the dense subnetwork algorithm against the number of edges in the graph is shown below.

| Query Name	Number of Edges in Graph	Graph Creation Time (seconds)	Dense Subnetwork Discovery Time (seconds)	Visualisation Time (seconds)
All Users	4403	1.459	2.023	4.406
50c Users	3387	0.434	0.922	1.615
Tertiary Students	2575	0.252	0.695	1.502
Senior	2623	0.205	0.680	1.377
Health Care	2208	0.104	0.601	1.007
6AM – 11AM	3514	0.624	1.262	1.790
11AM – 3 PM	2950	0.289	0.860	1.802

Table 1: Algorithm Efficiency
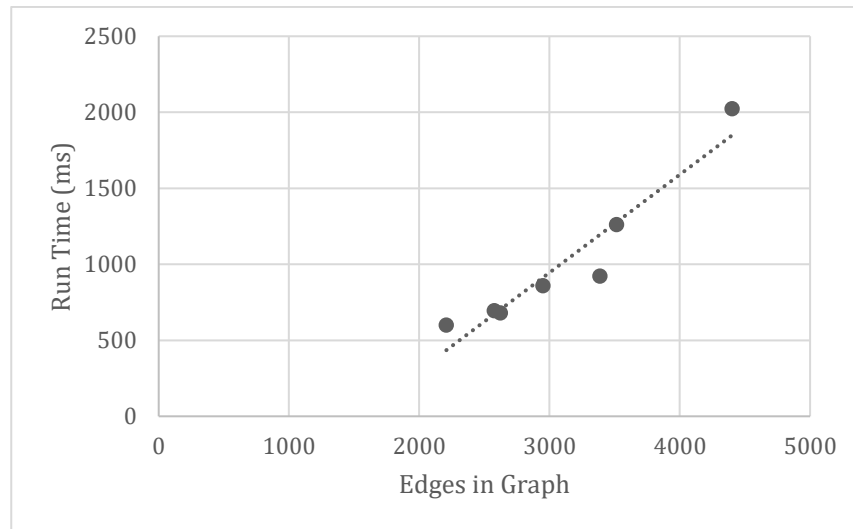
Figure 2: Algorithm Efficiency

As would be expected, as the number of edges in the graph increase, the run time of the dense subnetwork algorithm also increases. The average increase is approximately 0.064 seconds of run time per 100 edges in the graph. For the purpose of transport analysis, this is more than adequate.

3.2   Algorithm Results

This algorithm is designed to be used by the end user to uncover new information about the Perth public transport system. The results shown below are simply an example of the information that can be produced using the dense subnetwork discovery algorithm. These outputs are produced using all data collected in February 2017.
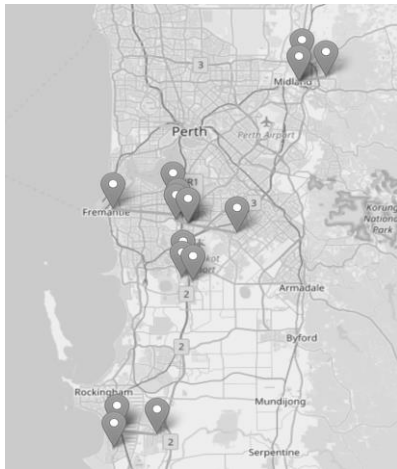
Query:
- Tertiary Smart Riders

Figure 3: Tertiary Student Results

Rank	Suburb 1	Suburb 2	Suburb 3
1.	Crawley	Perth	Mount Lawley
2.	Bentley	Como	Cannington
3.	Leeming	Murdoch	Canningvale
4.	Innaloo	Scarborough	Osbourne Park
5.	Claremont	Nedlands	Subiaco

Query:

- 50c Fare Users

Rank	Suburb 1	Suburb 2	Suburb 3
1.	Swan View	Midland	Middle Swan
2.	Leeming	Murdoch	Canningvale
3.	Atwell	Cockburn	Success
4.	Bateman	Fremantle	Booragoon

Figure 4: 50c User Results

4.   Conclusions and Future Work

The studies conducted in this paper attempted to find an efficient and effective way of producing information capable of helping key decision makers plan and design changes to a transport network. Through the use of dense subnetwork discovery algorithms and density plots, the end user of this program will be able to summarise and quickly extract key information from the SmartRider dataset. The algorithm proposed is optimised for the transport analysis and has shown a competitive efficiency with existing algorithms. This algorithm is capable of handling complex queries and returning powerful information to the end user of the program.

Future work in this area could include investigation of the scalability of this algorithm to other areas such as biology and social networks. Due to the size of the transport networks being used in this analysis, there was an upper limit of the size of the graph network being analysed. Finally, further work on this algorithm can be done to investigate different concepts of density and changing the density function.

5.   Acknowledgements

6.   References

Andersen. R, Chellapilla. K, 2009. *Finding Dense Subgraphs with Size Bounds,* s.l.: Microsoft Live Labs.