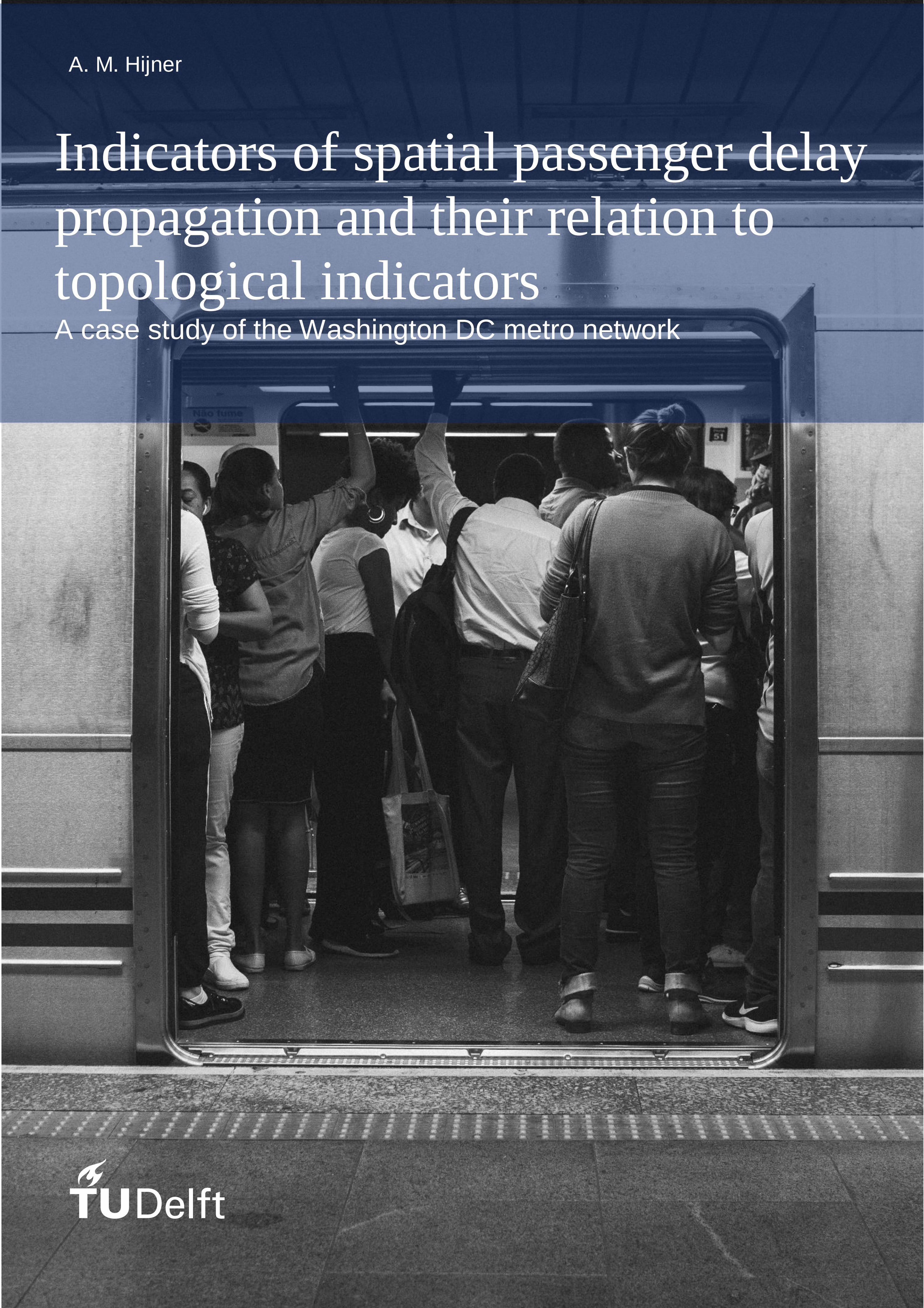
A. M. Hijner

Indicators of spatial passenger delay propagation and their relation to topological indicators

A case study of the Washington DC metro network



Delft University of Technology

Transportation, Infrastructure and Logistics

Thesis

Indicators of spatial passenger delay propagation and their relation to topological indicators

A case study of the Washington DC metro network

Author:
Anne Mijntje Hijner

Student number:
4389964

To be defended on:
September 27, 2019

Daily supervisor:
Dr. O. Cats

Committee chair:
Prof. Dr. S.P. Hoogendoorn

Committee members:
Dr. M.Y. Maknoon
Dr. M. Schmidt

September 9, 2019

Preface

This thesis is my final product for the Master of Science program: Transport, Infrastructure and Logistics at the Delft University of Technology. It describes the process and results of the study into informativity indicators, which can be used to gain information on the spatial propagation of passenger delay in a transport network. The study was performed using real data, made graciously available by the Washington Metropolitan Area Transit Authority. I hope they will find the results of this study to be valuable.

I would like to express my gratitude to my supervising committee: Oded Cats, thank you for your extensive feedback, our fruitful discussions, and the time you made for listening to my ideas; Marie Schmidt, thank you for your insights into the field of public transportation research; Yousef Maknoon, thank you for your critical view and challenging my mind; Serge Hoogendoorn, thank you for providing excellent guidance in the process and the collective discussions.

I would also like to thank Panchamy Krishnakumari, for processing the data and providing me with the data type that I needed, as well as for answering all my questions.

Lastly, I would like to thank my friends and family for their continued and unwavering support, and for always believing in me. Sanne and Jordi, thank you for your excellent feedback. Rogier, thank you for your feedback, and your support during dreary days in the library. Diederik, thank you for always giving me your insightful opinions, even when I don't agree with them, and for keeping me company during long days of hard work.

Annemijn Hijner
Delft, September 2019

Contents

Abbreviations

APE	Absolute Percentage Error
BN	Bayesian Network
CPD	Conditional Probability Distribution
DAG	Directed Acyclic Graph
MDL	Minimum Description Length
MI	Mutual Information
OD	Origin-Destination
PT	Public Transport
PTN	Public Transport Network
RMSE	Root Mean Squared Error
WMATA	Washington Metropolitan Area Transport Authority

Symbols

Sets

V	Set of stations $< s_1, s_2, ... >$
V_r	Set of stations differentiated by the direction of travel $< s_{1,a}, s_{1,b}, s_{2,a}, s_{2,b}, ... >$
E	Set of ordered pair of stations that define a track segment between the stations $< (s_1, s_2), (s_2, s_3), ... >$
I	Set of transfer stations where a transfer between lines can occur $< s_i, s_{i+1}, ... >$, where $I \in V$
I_r	Set of transfer stations differentiated by the direction of travel $< s_{i,a}, s_{i,b}, s_{i+1,a}, s_{i+1,b}, ... >$
D	Set of all days in the data set $< d_1, d_2, ... >$
T	Set of all time slices of 30 minutes in a day $< t_1, t_2, ..., t_{48} >$
K	Set of all time slices in the data-set, ordered by day and time-of-day $< k_{d_1,t_1}, k_{d_1,t_2}, ... >$
N	Set of all the nodes in the Bayesian Network $< n1, n2, ... >$, where $N = V_r + I_r + E$
A	Set of ordered pair of nodes that define the edges between the nodes in the Bayesian Network $< (n1, n2), (n2, n3), ... >$
$\delta_{x,y,a}$	Set of all the arcs in path a from node $x \in N$ to node $y \in N$, like $< (x, i), (i, y) >$, where (x, i) and $(i, y) \in A$
$\Delta_{x,y}$	Set of all the paths from node $x \in N$ to node $y \in N$, like $< \delta_{x,y,a}, \delta_{x,y,b} >$
B	Set of all the bins in which the data is made discrete defined by the lower and upper values of the bin, like $< (t_i, t_{i+1}), (t_{i+1}, t_{i+2}), ... >$

Parameters

s_o	Origin station; where $s_o \in V$
s_d	Destination station; where $s_d \in V$
$\widetilde{\tau}_{s_o,s_d,k}$	Scheduled travel time between s_o, s_d, for departure time k

$\tau^{n}_{s_o,s_d,k}$ Observed travel time of passenger n between s_o and s_d, departing at period k

$b^{nT}_{s_x,s_{x+1}}$ Indication of whether track-segment $(s_x, s_{x+1})\forall(s_x, s_{x+1}) \in E$, is part of the estimated route of passenger n, where the value is 1 if it is part of the route, and 0 if it is not

$b^{nI}_{s_{i,r}}$ Indication of whether transfer station $s_{i,r}\forall s_{i,r} \in I_r$, is part of the estimated route of passenger n, where the value is 1 if it is part of the route, and 0 if it is not

$\delta^{n}_{s_o,s_d,k}$ Estimated total delay of passenger n, between s_o and s_d, having departed at period k

$\delta^{T}_{s_x,s_{x+1},k}$ Estimated on-board delay on track-segment (s_x, s_{x+1}) during a period k; where $(s_x, s_{x+1}) \in E$, and $s_x, s_{x+1} \in V$

$\delta^{W}_{s_r,k}$ Estimated initial waiting time delay at station s_r during period k, where the passenger is travelling in the direction r; where $s_r \in V_r$

$\delta^{I}_{s_i,k}$ Estimated transfer delay at station $s_{i,r}$ during period k; where $s_{i,r} \in I_r$

$p^{T}_{s_x,s_{x+1},k}$ Estimated number of passengers travelling over track-segment (s_x, s_{x+1}) during a period k; where $(s_x, s_{x+1}) \in E$, and $s_x, s_{x+1} \in V$

$p^{W}_{s_r,k}$ Number of passengers starting their journey at station s_r during period k, and travelling in direction r; where $s_r \in V_r$

$p^{I}_{s_{i,r},k}$ Estimated number of passengers transferring at station $s_{i,r}$ during period k and travelling in direction r; where $s_{i,r} \in I_r$

$d(u, v)$ The distance between nodes u and v

$\sigma_{u,v}$ The number of shortest paths between stations u and v

$w_{(i,j)}$ Weight of the arc (i, j) from node i to node j, where $(i, j) \in A$ and $i, j \in N$

$a_{(i,j)}$ Binary variable that is 1 if there is an arc between nodes i and j, and 0 if there is not, where $i, j \in N$

b_i Bin b with a lower value of t_i and an upper value of t_{i+1}, where $b_i \in B$

Pop_{b_i} The number of data points in bin b_i, where $b_i \in B$

Lpop_{b_i} The logarithm of the number of data points in bin b_i, where $b_i \in B$

Lpop^{T} The sum of the logarithms of the number of data points $\forall b_i \in B$

OND_x The Outgoing Node Degree of node $x \in N$

ADI_x The Average Direct Informativity of node $x \in N$

TI^{l}_{x} The lower bound of the Total Informativity of node $x \in N$

TI^{u}_{x} The upper bound of the Total Informativity of node $x \in N$

Functions

$b(s)$	The normalized betweenness centrality of station s
$c(s)$	The normalized closeness centrality of station s
$P(X = i)$	The probability of finding variable X in state i
$P(X = i \mid Y = j)$	The probability of finding variable X in state i, when variable Y has been observed to be in state j
$MI(X, Y)$	The mutual information between variable X and Y
$LS(X \rightarrow Y)$	The link strength of the arc from variable X to variable Y, independent of any other parents of variable Y

Chapter 1

Introduction

1.1 Problem definition

Public transportation is an important part of modern society, for its social, economic and environmental benefits over private motorized transportation (Schmöcker et al., 2004; Redman et al., 2013). It remains a topic of interest, as improvements of the quality of public transportation keep being necessary in the face of increased urbanization and to keep up ridership numbers. An important aspect of the quality of public transport is reliability (Schmöcker et al., 2004; Bates et al., 2001; Redman et al., 2013). Here, two aspects can be distinguished: the reliability of the information provided and the reliability of the service provided, while both aspects should be of sufficient quality. For both of these aspects, information and knowledge is required. For the former this is in the form of travel time estimates, estimates of the likelihood and length of disruptions, disturbances and delays. For the latter, this is in the form of how to decrease the likelihood of a disruption or disturbance of a certain type occurring in a certain place, and what measures can be provided to mitigate the effects of the disruption or disturbance in the case that it does occur.

Measures mitigating the effects of disruptions and disturbances, such as the bypass researched in Tahmasseby et al. (2008), are most effective when they are applied to the most important and most vulnerable areas of the network. Thus, methods must be developed to identify these areas in transport networks. To be able to do this, it must be known how delays behave in a network. Some research has been done into this subject (Berger et al., 2011; Kirchhoff and Kolonko, 2015), however, most of this research has approached the problem from the perspective of the operator, with vehicle delays being the source of information.

It is, however, important to take the passenger's perspective into account as well, as due to transfers and vehicle capacity constraints, the passenger's experience might be very different than would be inferred from only the vehicle delay. For example, if someone just misses their transfer due to a relatively small delay, their eventual delay could be much larger than the initial vehicle delay. It is also possible that after an initial delay has been resolved, many passengers are still displaced, making the capacity of the vehicles insufficient to allow all passengers to board the vehicles, resulting in passengers being delayed, despite there being no more vehicle delays at all (Nielsen et al., 2009). Furthermore, from a societal perspective, it is often the passengers that are most affected by delays, in terms of time loss and its economic implications. Thus prioritizing the reliability of public transport in terms of passenger delay, might have a greater societal benefit, than if these resources were used in an optimal way from the operator's perspective.

The increased awareness of the importance of the passenger perspective, aided by the increasing availability of both gps and smart-card data, results in an increase in research adopting the passengers perspective (Hendren et al., 2015; Pelletier et al., 2011). This is also the case for reliability studies (Li-Jun et al., 2011; Sun and Jin, 2015). Some research has been able to identify vulnerable areas in a network, based on passenger simulations (Rodríguez-Núñez and García-Palomares, 2014; Malandri et al., 2018), however these studies do not make use of empirical data, but only of simulations.

No empirical method of determining the areas of the network that are most relevant for the propagation of passenger delay, is available. Furthermore, not enough knowledge is available on how passenger delay propagates throughout a transport network, to make simulations and theoretical modelling an accurate enough method to assess a network in terms of passenger delay. This study thus aims to develop a method that can fill this knowledge gap. The exact objective of the study will be elaborated upon in the next section.

1.2 Objective and research questions

The overall objective of this study is to gain more insight into the phenomenon of passenger delay. In order to do so, two particular topics are researched. Firstly, the goal is to develop a method that can determine how passenger delay is spatially related and to represent this in a set of informativity indicators. This set of indicators should be similar to centrality indicators, but whereas centrality indicators are regarding the physical relation of nodes in the network, informativity indicators should say something about a nodes capability of providing information on the state of the rest of the network. Meaning that by observing the state of the node in question, the states of other nodes in the network can be inferred with a certain probability. By state, the amount of delay being incurred by passengers, is meant.

The second aim of this study is to relate this set of informativity indicators to a network's topological indicators. This is done in order to make future assessments of transport networks easier, even when little or no delay data of passengers is available.

This objective can be condensed into one overarching research question:

- **How well do topological indicators approximate informativity indicators?**

In order to answer the the overarching research question, several sub-questions are relevant:

- **What are the delay characteristics of the network?**
  To fully understand the network and the delays experienced on it, an evaluation should be done of the delays on the network overall. This includes things such as, how often delays are experienced at any station, how much delay is experienced at any station, etc.

- **How can the passenger delay experienced in different locations, be related to each other?**
  In order to gain more information about the passenger delay phenomenon, the spatial relations of delays at different areas of the network should be evaluated. So from the available data, which merely contains the delays experienced at different locations, it must be estimated how the incurred delays in different locations relate to each other statistically.

- **How can informativity indicators best be measured?**
  Informativity indicators should approximately describe the informational position of a node in the network. Meaning it indicates how well the delay at other areas of the network can be predicted knowing the delay at the node in question, as well as in how many areas of the network, such a delay can be predicted. This information can be gathered from the spatial relations of delay as determined in the previous question.

- **What type of graph representation of the transport network is best suited to the problem?**
  How the network is represented has great consequences for how the network is analyzed in terms of topological indicators (Derrible and Kennedy, 2011). It is thus important to evaluate which type of graph representation(s) of the transport network are best suited to the problem, so that the indicators are meaningful, especially with respect to the informativity indicators.

- **Are informativity indicators related to any topological indicators?**
  It is possible that some topological indicators are related to the informativity indicators of nodes. If this were to be the case, it could be very useful to determine these relationships, so that the informativity of nodes of other networks could be determined even when large data sets are unavailable or impractical to use.

1.3 Approach

In order to answer the research questions, a model is created to discover the relation between incurred passenger delay in one node (station), to the incurred delay in other nodes in the network. This is done by using historical data from the network in question, where delays are known per station and link, where a distinction was also made with respect to the direction in which the passengers were travelling when they incurred the delay. The information from the model can then be used to determine how the delays at nodes are related to each other, which in turn can be used to calculate the informativity of nodes.

Considering that delays in a network are related both over space and time, it is important to make the distinction. For the purpose of this study, the focus is on spatial relationships. To also consider temporal relationships would drastically complicate the problem, due to the added dimension and the complications for the representation of the network, which in time-space is dynamic due to the service layer. Still, the temporal aspect of the problem can't simply be disregarded, instead, a certain time-frame will be chosen. Within this time-frame, the delays found in the data are considered to be related, but beyond this time-frame, the delays will be considered unrelated.

1.4 Thesis structure

The structure of this thesis is as follows: In Section 2 this study will be placed in a scientific context by analyzing related literature; In Section 3 the methodology applied, is discussed. This includes how the delays in all locations will be related to each other, how the informativity and centrality indicators are calculated, and how these are compared; In Section 4, the application of the method is discussed in further detail. This includes information on the case study, as well as the methods of calculation used for the implementation; In Section 5 the results are discussed. This includes an exploratory analysis of the data, and the network. As well as the delay relationships, the informativity indicators and the centrality indicators; In Section 6 the results and findings are discussed, the limitation of the study are mentioned, as well as recommendations for future research.

Chapter 2

Literature review

In this section, we will place this study in a scientific context. First, related research on vehicle and passenger delay prediction and estimation will be discussed in Section 2.1. Then, we will look more closely into other reliability studies and how centrality indicators play a role in network analysis with the aim at improving reliability, in Section 2.2. Finally, all works are synthesized in Section 2.3, where a comparison is made of the most closely related research and this study.

2.1 Delay propagation and prediction

It is particularly important to discuss the efforts that have been made into researching vehicle delay, its causes, how it propagates over a network, and how predictions of this delay can be made, in order to understand what aspects can be applicable to this study (Section 2.1.1). To also understand what aspects are inherently different when considering passenger delay instead of vehicle delay, studies regarding passenger delay must also be analyzed (Section 2.1.2). This is also important to do, in order to understand what has already been researched, and what research gaps still exist.

2.1.1 Vehicle delay

Vehicle delay has long been of interest, but recent advances in big data and an increase in data sources, has advanced this area of research significantly. Wang and Work (2015), for example, uses delay data as input for a regression model that predicts arrival times. They show an increased prediction accuracy, especially when real-time information is provided.

Such a regression model is not the only type of data-driven method that aims to predict arrival times of vehicles. Yaghini et al. (2013) used a more complex neural network to predict train delays on the Iranian Railways, to eventually conclude that the neural network can generate accurate delay predictions in a reasonable time.

Bayesian networks have also been used for such predictions for both railways (Lessan et al., 2018; Corman and Kecman, 2018) and in other sectors such as the aviation sector (Laskey et al., 2012; Xu et al., 2007). These models aim to capture the complexities of vehicle delay, such as their dependence on each other due to the usage of the same infrastructure by different vehicles, and waiting policies where vehicles sometimes must wait for other, delayed, vehicles. Both of which are methods through which delay can propagate from the initially delayed vehicle, to many other vehicles. Bayesian networks can be used to model some of these relationships, as well as to incorporate other circumstances that can cause, influence or exacerbate delay (Lessan et al., 2018; Corman and Kecman, 2018; Laskey et al., 2012; Xu et al., 2007).

Besides these data-driven methods, attempts have been made at modelling the complex nature of vehicle delay through mathematical models. For example, Berger et al. (2011) describes a stochastic model to estimate driving time profiles, based on driving times (average driving times over a track), catch up potential (how much faster a vehicle could drive over a track) and waiting policies (how long a vehicle should wait for a delayed vehicles if transfer passengers are expected). It also has functionality as an online model, as real-time information significantly improves estimates of arrival and departure times. The output of the model by Berger et al. (2011) is similar to the output of the models mentioned above, namely estimates of departure and arrival times, while the method is very different.

Kirchhoff and Kolonko (2015), however, apply a relatively similar method to Berger et al. (2011), while producing a different type of output. As Kirchhoff and Kolonko (2015) also apply a mathematical model that uses delay distributions of a delay source, to calculate the delay distributions over the rest of the network, effectively estimating where to and to what extent delay propagates on a rail-network. Another model making use of the theory of delay propagation is the back-tracking approach by Manitz et al. (2017), who designed this specifically for estimating delay sources in railway networks. This method performs quite well, and under certain circumstances even better than established source-estimation methods based on epidemiological theory, proving that information on delay propagation can be quite valuable.

2.1.2 Passenger delay

What all these studies have in common, is that they approach the problem from a vehicle perspective, which is usually the most useful perspective for an operator. Furthermore, despite vehicle delay and vehicle delay propagation being a complex issue, how it works and the causes for propagation are quite well known, especially when compared to the knowledge on passenger delay. This means that the approach used in the studies mentioned above, won't be immediately applicable to passenger delay. Still, even a model with a vehicle perspective, can be applied to gain some information on passenger delays as well, as is done by Dollevoet et al. (2018). In that study, the management of delay is researched, which has implications for things such as waiting policies, by researching the effect of delay and delay propagation on passengers. Still, such work does not incorporate exact information, through data, on how delays are experienced by passengers.

Nielsen et al. (2009) did attempt to create a model to estimate total passenger delay. They did so by combining two possible systems, one in which people are completely unadaptive when faced with disruptions and will not reroute, and one in which people have perfect information and always choose the shortest route. These two systems were combined by adopting a certain threshold of delay, where if people would experience a delay larger than the threshold, they would consider rerouting.

Recently, more sophisticated passenger assignment methods are being applied to simulate transport systems in case of disruptions, such as that by Sun and Jin (2015), who used smart-card data to analyze passenger behaviour and subsequently model passenger flows. Information from these methods can be used to assess many qualities of the transport network, such as it's capacity (Malandri et al., 2018). These types of studies can generally be described as reliability studies, and will be discussed in more detail in the next section.

2.2 Reliability studies and the use of indicators

Reliability studies can be done in several ways, such as by using simulation (Rodríguez-Núñez and García-Palomares, 2014), by applying available data (Raghothama et al., 2016), or a combination of both. However, their goals are generally the same, to find out which areas in a network are vulnerable to disruptions and critical for the functioning of the entire network. This study aims to do a similar thing, and so it is important to evaluate what efforts have already been made regarding reliability studies.

Studies exist that use passenger flow distributions, such as that by Rodríguez-Núñez and García-Palomares (2014). In that study they used these flows to determine which links where most critical in the network, by evaluating how many alternative routes exist if a link was removed from the network. A link's importance was measured by the number of trips that could not take place if it was taken out, as well as by the increase in average travel time over the entire network. Cats et al. (2017) carried out similar research, but instead of analyzing the network when a link was taken out, they only reduced the capacity of certain links to see how this would influence the network. Besides only checking how badly the network would be influenced in case of a disruption on a certain link, studies exist that evaluate the added value of improving infrastructure at certain locations with respect to possible disruptions (Tahmasseby et al., 2008).

The use of sophisticated indicators is of interest in many reliability studies. Malandri et al. (2018), for example, used a volume over capacity ratio to evaluate the impact of a disruption both on the increase in travel time, but also on the quality of the service. Using this ratio they were able to determine which links in the network where most influential on the overall quality of service. Topological indicators are also often used to describe critical links and nodes in a network. They are often used in combination with passenger flow assignment methods (Sun et al., 2018), but not always. Chen et al. (2012), for example, attempted to develop purely topological indicators more sophisticated than node-degree, but less computationally intensive than betweenness centrality, to identify influential nodes in extremely complex networks. Still, passenger flows are important to take into account (Cats et al., 2017), as PT networks are neither static nor deterministic. Therefore Cats and Jenelius (2014) studied the usage of topology indicators while considering a stochastic and dynamic network, where real-time information could be used. Further work by Cats et al. (2016) also showed the importance of not only taking into account the effect of a disruption on a certain link or node on the rest of the network, as is the case by the studies mentioned above, but also the exposure of a certain link or node. He thus effectively combined the chance of a disruption occurring, with the effect this would have on the network, into a measure of criticality of links and nodes.

2.3 Synthesis

In Table 2.1 an overview of seven papers, and this study, can be found. The seven papers are a selection of the above discussed studies, that represent the work related to this study sufficiently. In the table, information on the purpose, approach and perspective of all studies can be found, as well as the required input and resulting output of the models used or created in the studies.

As can be seen in the table, this study has things in common with all other seven studies, but is not quite the same as any other study. For example, it aims to determine the relationships between delays at the different areas of the network, just like the studies by Berger et al. (2011) and Kirchhoff and Kolonko (2015). But it also comprises an evaluation to determine specific indicators for links and nodes, just like the studies by Rodríguez-Núñez and García-Palomares (2014), Malandri et al. (2018) and Cats et al. (2016), with whom it also shares a similar perspective, namely the passenger perspective. Yet the approach of this study is more similar to that of Corman and Kecman (2018), who also use a Bayesian network. Even the study by Yaghini et al. (2013) applies a more similar method, namely the data-driven neural network method. As opposed to the other studies who apply either simulation or mathematical modelling methods, this study attempts to reach the objective from an empirical approach.

It is thus clear what the contribution of this study is, namely to discover a data-driven method to identify delay dependencies in a network, which in turn can be used to determine the informativity of nodes in a PT network.

	Purpose	Approach	Perspective	Input	Output
Corman and Kecman (2018)	Predicting arrival times	Bayesian network	Vehicle perspective	Network; Arrival and departure times; Timetable	Expected departure/arrival times
Yaghini et al. (2013)	Predicting arrival times	Neural network	Vehicle perspective	Arrival and departure times	Expected departure/arrival times

Berger et al. (2011)	Researching delay propagation;
	Predicting arrival times
		Historical data;
		Stochastic, mathematical model
			Vehicle perspective	Waiting policies;
				Driving time profiles;
				(Arrival and departure times)
					Travel time distributions

Kirchhoff and Kolonko (2015)	Researching delay propagation	Stochastic, mathematical model; Monte-Carlo simulation	Vehicle perspective	Distribution source delay	Distribution propagated delays
Rodríguez-Núñez and García-Palomares (2014)	Node/link indicator analysis	Passenger flow simulation	Passenger perspective	OD-demand	Critical links

Malandri et al. (2018)	Node/link indicator analysis	Passenger flow simulation	Passenger perspective	OD-demand	Passenger flows;
					Volume Over Capacity ratio;
					Link saturation;
					Recovery time

Cats et al. (2016)	Node/link indicator analysis	Passenger flow simulation; Historical data	Passenger perspective	Disruption data; OD-demand	Critical links
This study	Researching delay propagation; Node indicator analysis	Bayesian network	Passenger perspective	Delay per link and node	Delay dependencies between nodes; Informativity of nodes

Table 2.1: An overview of seven closely related papers that were discussed, as well as this study, in terms of their purpose, the used approach, whether the vehicle or passenger perspective was central in the study, and the input (where information on the network is assumed everywhere) and output of the methods. The purpose 'Node/link indicator analysis', this can refer to the usage of any indicators specific to nodes and/or links, such as centrality or criticality indicators.

Chapter 3

Methodology

In this section the applied method, is described. This involves several distinguishable steps, which have been undertaken in a certain order, which can be seen in Figure 3.1.

The structure of this section is as follows: firstly the method of determining the delay dependencies between different station is discussed in Section 3.1, this section will include details on the input and output data; one requirement of the input data is that it must be discrete, how continuous data is made discrete will thus be discussed in Section 3.2; the results of these calculations will be applied to calculate the informativity indicators, how this is done is discussed in Section 3.3; these indicators will be compared to the topological indicators of the PT network, how these indicators are calculated is discussed in Section 3.4; how the relation of these two sets of indicators are determined, will be discussed in Section 3.5.

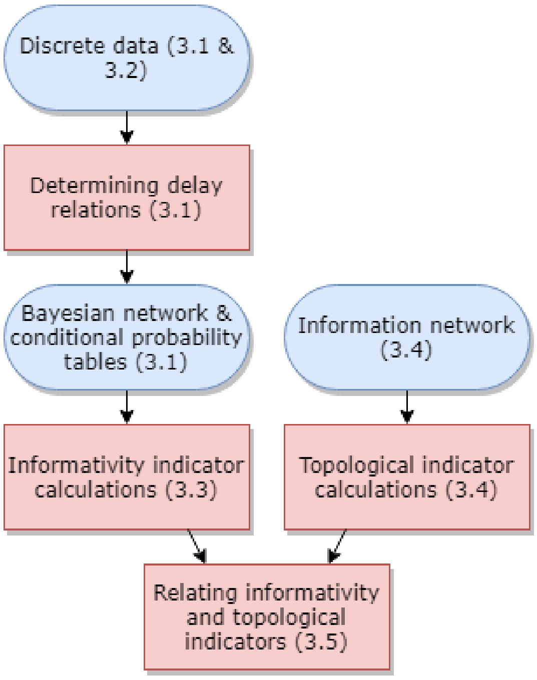

Discrete data (3.1 & 3.2)
Determining delay relations (3.1)
Bayesian network & conditional probability tables (3.1)
Information network (3.4)
Informativity indicator calculations (3.3)
Topological indicator calculations (3.4)
Relating informativity and topological indicators (3.5)

Figure 3.1: The methodology applied, where the round blue boxes indicate input/output and the red boxes indicate processes. The numbers in the boxes correspond to the sections in which that particular item is discussed.

3.1 Determining delay relations

In this section we discuss how it is determined how the delays experienced at different stations are related to each other. Firstly, in Section 3.1.1, it is discussed which method was chosen and why; how this method is implemented exactly is discussed in Section 4.2; the required input data will be described in Section 3.1.3; and the output data will be described in Section 3.1.4; how the output data can be extended to increase the possible applications of the output is described in Section 3.1.5; lastly, how the model is validated will be discussed in Section 3.1.6.

3.1.1 Method used to determine relations

As discussed in Section 2.3, several methods have been applied to studies closely related to this one. Considering the purpose of this study, it would make sense to use a similar method to the one used in studies of delay propagation: mathematical modelling (Berger et al., 2011; Kirchhoff and Kolonko, 2015), or to the one used in studies of indicator analysis: simulations (Rodríguez-Núñez and García-Palomares, 2014; Malandri et al., 2018; Cats et al., 2016). However, these methods are purely theoretical, and not enough knowledge about the theory behind passenger delay is available to construct an accurate and sophisticated model or simulation. Therefore, the empirical methods such as those applied by Corman and Kecman (2018) (Bayesian Network) or Yaghini et al. (2013) (Neural Network), are more appropriate for this work as they are data-driven and don't require extensive knowledge about the underlying phenomenon, while still being capable of uncovering prior unexpected relationships and using a holistic view to consider all possible dependencies (Kjaerulff and Madsen, 2008; Koller et al., 2009; Zilko et al., 2016; Lievens, 2014; Lessan et al., 2018). From these possible data-driven methods, it was decided to use the Bayesian Network method, as it is less complex than the Neural Network method and the results are easier to work with, making it more appropriate for a short study. Additionally the Bayesian network has the benefit of providing information through the structure, which is meaningful on its own, as opposed to Neural Network, whose structure is not. Furthermore, this method can be used when limited data is available, or when not all observations are complete. However, it does require enough data for the results to be meaningful and robust (Zuk et al., 2012), but it is expected that the data available for this study is sufficient, while not being so large as to lead to computational time or storage demand issues, which can also occur for the Bayesian Network method (Zilko et al., 2016). Below, the theory behind Bayesian networks will be discussed shortly.

Introduction to Bayesian Networks

A Bayesian Network (BN) is by definition a Directed Acyclic Graph (DAG) - $G(N, A)$ where N are the nodes and A are the arcs - which represents the dependencies of its variables on each other graphically. It also contains information on the amount of dependence of the variables, by means of a conditional probability table of a variable's state (Kjaerulff and Madsen, 2008; Koller et al., 2009; Zilko et al., 2016). The nodes of the graph are the variables, in this case the

nodes are the initial stations and the transfer stations (respectively set V_r and I_r - as described in the Section Symbols - are the sets that comprise the nodes of the BN, N). The edges of the graph represent the flow of influence between the nodes (the edges are comprised in the set A as described in the Section Symbols). The nodes not connected by an edge are assumed to be conditionally independent. Other nodes influencing the node in question, are called parents. While the nodes influenced by the node in question are called children. Any nodes previous to the node in question are called ancestors, while all nodes after the node in question are called descendants, as can be seen in Figure 3.2. The ancestors beyond the parents do indirectly influence the node in question, but as long as knowledge on the state of the parents is available, knowledge on the other ancestors holds no additional value when estimating the state of the node in question (Kjaerulff and Madsen, 2008; Koller et al., 2009; Lievens, 2014).
As the name suggests, a BN stems from Bayes' Theorem. This theorem describes the probability of an event occurring, based on knowledge of related events. Hereby previous knowledge or a belief about the probabilities (prior probability distribution of the event) is adapted when new knowledge (data) becomes available. This results in the posterior probability distribution. Then, when the values of the parents of a node are known, the probability of a state of the node in question can be calculated by using Bayes' theorem, as (Kjaerulff and Madsen, 2008; Koller et al., 2009; Pearl, 1988):

$$P(X = i|Y = j) = \frac{P(Y = j|X = i) \cdot P_{pr}(X = i)}{P_{pr}(Y = j)}, \tag{3.1}$$

where $P(X = i|Y = j)$ is the probability of observing node X in state i, if it is known that node Y is in state j; and $P_{pr}(X = i)$ the prior probability of observing node X in state i, independent of the state of Y. If these calculations are done for all states of nodes X and Y, the conditional probability table of node X can be constructed.

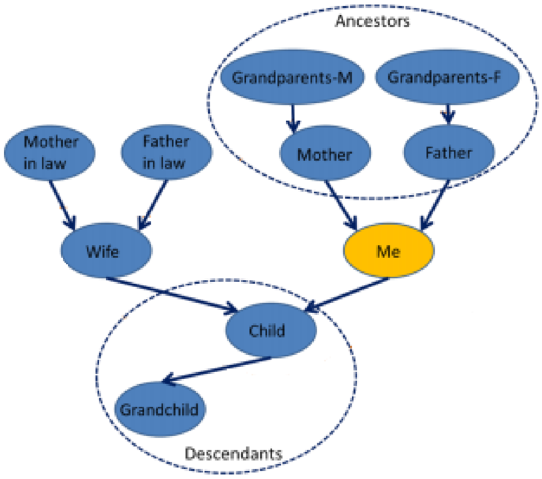


Figure 3.2: An example of the structure of a Bayesian network where relations to the node 'me' are indicated (Lievens, 2014).

3.1.2 Creation of the Bayesian Network

The creation of a BN consists of two steps. Firstly, it must be checked which nodes are related to each other, to create a skeleton of the network. The skeleton holds information on the dependence of the nodes, and the direction of the flow of influence, it is a directed, acyclic graph (DAG) (Kjaerulff and Madsen, 2008; Koller et al., 2009). The dependencies between the nodes are determined by comparing the states (amount of delay) of each node-pair at every point in time, to see if they are related to each other with a certain significance. The exact manner in which this is done depends on the method that is used, and will be discussed in more detail in Section 4.2. Once the dependencies are determined, the directions of the arcs of the graph are determined under the constraint that the graph must be acyclic, limiting the possible configurations. How this constraint can be used to determine the directions of the arcs of the graph, is discussed in Section 4.2.2.

After this, the labels of the nodes, in the form of conditional probability distributions (CPD), are determined. The conditional probabilities are determined by checking the combinations of the parent states present in the data, and in which child-state these result (Koller et al., 2009). How exactly this is done depends on the method used, which will be discussed in Section 4.2. Both of these processes are accomplished by inference from the existing data (Koller et al., 2009; Zilko et al., 2016). The implementation of this process can be tedious, but much software and applications exists to accomplish these two steps (Raiko et al., 2007; Luttinen, 2016).

3.1.3 Input

Creating a BN requires data. The more nodes need to be modelled, the more data is required (Koller et al., 2009; Kjaerulff and Madsen, 2008). The data required for a BN is in the form of multiple observations of the states of the different nodes, where per observation these states are related to each other. Usually this means that the states of the nodes are observed at the same time, or that observations are made per subject or event. In the context of this study, the observations are made of different subjects (delays at the initial station or at a transfer station) during a short time-frame. Each observation thus contains information of the states of the subjects (the amount of delay per station) averaged over a short time frame. How exactly this data is constructed and averaged is discussed in Section 4.1. Due to the nature of the method, it is not required that each observation contains information on each subject, however the more data is available the more reliable the outcome of the method (Koller et al., 2009; Kjaerulff and Madsen, 2008).

An important requirement of the input data, is that it is either discrete or normally distributed (also called Gaussian distributed) (Zilko et al., 2016; Koller et al., 2009; Kjaerulff and Madsen, 2008). In case the available data is continuous but not normally distributed, as is often the case, a discretization method must be applied. The method applied in this study will be discussed in more detail in Section 3.2.

3.1.4 Output

The BN will have several different outputs. First of all, it will output the network, meaning all of the nodes and the directed arcs between the nodes that are related. Secondly, it will give the conditional probability tables. For nodes with no parents, these tables will give the probability per state of finding the node in that state during an observation. For nodes with one or more parents, the table will have a dimension of s^n, where n is the number of parents

and s is the number of possible states. Here, the probabilities of finding the node in a state is given for each possible combination of parent states (Koller et al., 2009).
Clearly, due to the nature of the BN, the arcs between the nodes are not labelled in any way. However, ascribing a weight to these arcs could be very useful. How this can be done is described in the next section.

3.1.5 Labelling the arcs of the Bayesian Network

Several methods exist that aim to label the arcs between two nodes in a BN, in a meaningful way. These labels indicate the the strength of the dependence between the two nodes. For example, one such method compares the distributions of binary nodes to identify the connection strength (Boerlage, 1992; Ebert-Uphoff, 2009), while another compares the probability distributions of one node with or without information on its parents and comparing this to the Bhattacharyya distance (Jitnah and Nicholson, 1997; Ebert-Uphoff, 2009). However, the most common approaches are those based on Mutual Information (MI) (Pearl, 1988; Nicholson and Jitnah, 1998; Ebert-Uphoff, 2009; Kjaerulff and Madsen, 2008), which is a measure of the dependence of two variables, that quantifies how much information can be gained on one variable, if the state of the other variable is observed. It is defined as

$$MI(X,Y) = \sum_i P_{pr}(X=i) \sum_j P(Y=j|X=i) log_e(\frac{P(Y=j|X=i)}{P_{pr}(Y=j)}) \tag{3.2}$$

where i, j represent all possible states of nodes X, Y, respectively; $MI(X,Y)$ is the mutual information between nodes X and Y; $P(Y=j|X=i)$ is the probability of finding variable Y in state j, given that variable X is in state i; $P_{pr}(X=i)$ is the prior probability of finding variable X in state i, so independent of the observed state of Y.
Regarding the exact calculation of MI, there is some debate as to whether the logarithm of 2 (Ebert-Uphoff, 2009) or of e (Nicholson and Jitnah, 1998; Pearl, 1988) should be taken. As the works by Pearl (1988) and Nicholson and Jitnah (1998) are more rigorous and well established in the scientific community, the logarithm of e will be used during this study.
The concept of MI can only be applied to nodes that have only a single parent, because if it is applied to nodes with multiple parents, the dependencies between the parents can influence the MI between either parent and the child, resulting in an overestimation of the MI. Therefore, an extension of the concept is necessary, that can take into account the states of all parents, so possible dependencies between parents will not influence the other arc weights (Ebert-Uphoff, 2009; Nicholson and Jitnah, 1998). This extended concept can be described in an equation as:

$$LS(X \rightarrow Y) = \sum_k P_{pr}(Z=k) \sum_i P_{pr}(X=i) \sum_j P(Y=j|X=i,Z=k) log_e(\frac{P(Y=j|X=i,Z=k)}{P_{pr}(Y=j|Z=k)}) \tag{3.3}$$

Here, $LS(X \rightarrow Y)$ is the link strength, meaning the strength of the dependency, of variable X on variable Y, independent of the value of the set of other parents of Y (which is the set Z);

$P(Y = j|X = i, Z = k)$ is the probability of finding variable Y in state j, given that variable X is in state i and Z is in state k; $P_{pr}(Z = k)$ is the prior probability of finding variable Z in state k. If variable Z has no parents, this information can be directly retrieved from the BN. However, if Z does have parents, the prior probability must be approximated by averaging all the probabilities of the states of Z over all states of it's parent-nodes (Nicholson and Jitnah, 1998).

The meaning of $LS(X \to Y)$ can be interpreted as the amount of information flow along the arc $X \to Y$, but also as how much uncertainty in the state of Y is reduced by knowing the state of X, if all other parent-states of variable Y are already known (Ebert-Uphoff, 2009).

It should also be noted that

$$LS(X \to Y) = LS(Y \to X) \tag{3.4}$$

meaning this method does not distinguish between the direction of an arc, and can thus not contain any information about the direction of a connection between two nodes (Nicholson and Jitnah, 1998).

3.1.6 Validation

Just as for any model, the outcome of the BN method must be tested for its validity. One way to do so is by using the Root Mean Squared Error (RMSE) (Lessan et al., 2018). This metric checks the difference of the observed data and the expected outcome (which is based on the average of all the data points in the training set), using the equation

$$RMSE = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(E(i) - O(i))^2}, \tag{3.5}$$

where N is the total number of observations checked, E(i) is the weighted average of the possible states of observation i, and O is the observed state of observation i. The RMSE is then given per node. The RMSE of every node can be combined into a single value for the entire model, by simply averaging the RMSE over all nodes. With the RMSE we can then say which nodes can be predicted with a high accuracy, and which cannot. Furthermore, if the available data is split into a training set, used to construct the model, and a test set, the RMSEs of both sets can be compared to see how well the BN performs for the test set. If the RMSEs of the test set are much higher, that would mean the model is overfitted, if the RMSEs are only slightly higher, the model can be stated to be a good predictor (Lessan et al., 2018). This method is commonly known as k-fold cross-validation.

Besides the RMSE, the Absolute Percentage Error (APE) can be used (Waller, 2003; Sakauchi, 2011). This metric gives the percentage of error of any node. This metric is calculated by using equation

$$APE = \frac{100\%}{N}\sum_{i=1}^{N}\frac{|E(i) - O(i)|}{E(i)}, \tag{3.6}$$

where N is the total number of observations checked, E(i) is the weighted average of the possible states of observation i, and O is the observed state of observation i. Whereby the estimations are based on the training set of the BN, and the observed values are taken from the test set. The APE then gives an indication of the accuracy of predictions made by using the BN.

3.2 Data discretization

As was mentioned in Section 3.1, the required data input for the model must be discrete. Data used for studies like this one, however, is often continuous, and must thus be made discrete. In this case the data is of delays, and can thus take on any positive number. There are several general methods that can be applied to discretize continuous data. It is common to discretize data in partitions with equal intervals, or in partitions with equal populations (Clarke and Barton, 2000). For the former the partitions would be chosen by dividing the difference of the maximum and minimum value in the data set, by the number of partitions desired, to find the interval sizes. For the latter, the total number of data points is divided by the number of partitions desired, to find the population per partition. The data can then be assigned to the partition it belongs to.

During the discretization of continuous data, much information can be lost (Clarke and Barton, 2000). Therefore, alternative methods have been developed that limit the loss of information, such as the Minimum Description Length (MDL) method (Clarke and Barton, 2000; Fayyad and Irani, 1993), and the discretization by entropy loss minimization (Clarke and Barton, 2000). Both of these methods perform better than the previously mentioned methods of dividing the data set by equal intervals or equal populations (Clarke and Barton, 2000). For this study, the entropy loss minimization method is of particular interest, due to its advantage of being able to discretize a variable completely independently of other variables or knowledge of the Bayesian Network structure that is to be constructed using the data, making it efficient and easy to apply (Clarke and Barton, 2000).

This entropy loss minimization method (Clarke and Barton, 2000) is based on equation (Pearl, 1988)

$$H(X) = -\sum_i P(X = i) log_e P(X = i) \tag{3.7}$$

which gives the entropy of a discrete variable X, where $P(X = i)$ is the probability of finding variable X in state i.

The method is initialized by assigning every possible value of X in the data set it's own partition. The number of initial partitions is thus equal to or smaller than the number of data points. Partitions are then merged, based on which merge would result in the smallest change in entropy. Effectively this means that the partitions with the smallest difference between $p(i)$ and $p(i+1)$ are merged (Clarke and Barton, 2000). Partitions are merged until a certain point is reached. This point is stated to be when the change in X (meaning the number of partitions) becomes bigger than the change in the entropy of X, so that the number of partitions remains manageable, while not too much information is lost (Clarke and Barton, 2000). This can be described in an equation like (Clarke and Barton, 2000)

$$x_{\max} y_i - y_{\max} x_i > x_{\max} y_{i-1} - y_{\max} x_{i-1} \tag{3.8}$$

where $x_{\max}$ indicates the maximum number of partitions (the initial number); $y_{\max}$ the maximum entropy (the initial entropy); x_i the number of partitions at iteration i; and y_i the entropy at iteration i. The merging of partitions is finalized when the equation above is no longer true.

One disadvantage of the entropy loss minimization method that should be noted, is the inability to specify a maximum amount of partitions. It could thus be that the method results in relatively many partitions.

3.3 Informativity indicators

From the output of the BN, we want to determine the informativity of the nodes. From extensive research, it became apparent that no indicators of informativity, or any other measure of it, has been used before. Therefore, three novel indicators are suggested: outgoing node degree, average direct informativity, and total informativity. These will be discussed respectively in Sections 3.3.1, 3.3.2 and 3.3.3.

3.3.1 Outgoing node degree

One possible and simple informativity indicator is the node degree of all the nodes in the BN. This is an indication of the number of nodes related to the node in question, meaning the number of other nodes for which the node can provide information. A distinction can be made between incoming and outgoing node degree. Most interesting is the outgoing node degree, which indicates how many nodes, the node in question can provide information on.

However, it should be noted that the directions of the arcs in the original BN do not imply causation and are ambiguous. Therefore, expert knowledge is necessary to determine the directions of the arcs to make them meaningful, making the outgoing node degree a relevant indicator. If it is not possible to assign a meaningful direction to the arcs, the arcs should be bi-directional and the value of the outgoing node degree would be the same as the value of the incoming node degree.

This indicator can be described by

$$\mathrm{OND}_x = \sum_{i \in N} a_{(x,i)} \tag{3.9}$$

where OND_x is the outgoing node degree of node $x \in N$; and $a_{(x,i)}$ is the arc between node x and node i.

3.3.2 Average direct informativity

The outgoing node degree indicator mentioned above, cannot measure the exact informational influence of a node, as it cannot distinguish the size of the dependencies in the BN, as it does not take into account the weights of these arcs. These arc weights can then be used to give an indication of the average influence of a node on its neighbours in the BN. This can be done by adding the weights of all the outgoing arcs from the node in question, and dividing it by the number of outgoing arcs of the node, to get the average influence on its neighbours. The average is taken instead of the total, to get an indication of the expected information a node can provide on any of its direct neighbours, regardless of the amount of neighbours it has. The total is better encapsulated in the indicator discussed in Section 3.3.3. The calculation can be written out like

$$\mathrm{ADI}_x = \frac{1}{\mathrm{OND}_x} \sum_{(x,i) \in A} w_{(x,i)} \tag{3.10}$$

where ADI_x is the average direct influence of node $x \in N$; (x, i) is the arc from node x to any other node $i \in N$; $w_{(x,i)}$ is the weight of this arc; and OND_x the outgoing node degree of node x.

3.3.3 Total informativity

The above mentioned indicators only take into account the direct information held by a node on it's immediate descendants. However, any node also indirectly holds information on it's further descendants. This can be incorporated through multiplying the arc weights along the path to the descendant, and adding these numbers for each descendant. Naturally, these numbers get increasingly small as the descendant is further removed from the node in question. So to approximate the true value of this indicator, while keeping the calculation times limited, the calculation can be done for only a few generations, rather than the entire network.

In case there are multiple possible paths from the node in question to a descendant, two things can be done: only the path with the highest value after multiplying all the arc weights can be used, as this is the most informative path; or all paths can be taken into account. The former will result in a lower estimate of total informativity, and the latter in a higher estimate. Neither is completely accurate as the former neglects some additional paths that could be informative, while the latter can overestimate the information contained in node n_x as some paths could have overlap. Thus, the former method will result in a lower bound of the total informativity, while the latter will result in an upper bound. In practice, these two values should be relatively similar, as most node pairs with many possible paths between them, will most likely have multiple generations between them, resulting in only a small contribution to the total informativity.

The upper bound of this indicator can be described as

$$\mathrm{TI}_x^u = \sum_{y \in N} \sum_{\delta_{x,y,a} \in \Delta_{x,y}} \prod_{(i,j) \in \delta_{x,y,a}} w_{(i,j)}, \tag{3.11}$$

where TI_x^u is the upper bound of the total informativity of node x; $w_{(i,j)}$ is the weight of the arc from node i to node j; $\Delta_{x,y}$ is the set of all paths from node x to node y; and $\delta_{x,y,a}$ is the set of all arcs that form path a.

The lower bound of this indicator can be described as

$$\mathrm{TI}_x^l = \sum_{y \in N} \max_{\delta_{x,y,a} \in \Delta_{x,y}} \prod_{(i,j) \in \delta_{x,y,a}} w_{(i,j)}, \tag{3.12}$$

where TI_x^l is the lower bound of the total informativity of node x.

3.4 Topological indicators

In order to compare the informativity and topological indicators, which is described in Section 3.5, some topological indicators must be calculated first. As informativity indicators are specific to stations in the network, it would make most sense to compare them to station-specific topological indicators; thus centrality indicators. Therefore, from here on out, only centrality indicators are referenced.

Several different, possibly useful, centrality indicators can be distinguished. The indicators that will be used for comparison to the informativity indicators, are discussed in Section 3.4.1. For these indicators it is important to distinguish in which space the graph of the transport network is represented, as the indicators have different meanings for the different spaces. The different spaces deemed relevant for this study are discussed in Section 3.4.2.

3.4.1 Indicators

Several indicators of graphs could contain useful information. The calculations of all the indicators are the same in every space and are discussed below.

Degree centrality

The degree of a node indicates the size of the neighborhood. For every node, the degree centrality is the number of edges connecting to it (Chen et al., 2012). In case of a directed network, an in-going and an out-going degree can be distinguished (Lin and Ban, 2013).

Closeness centrality

The closeness centrality indicates how far all other nodes are from the node in question. It is calculated by (Chen et al., 2012; Lin and Ban, 2013)

$$c(s) = \frac{|V| - 1}{\sum_t^V d(s,t)}, \tag{3.13}$$

where $c(s)$ is the normalized closeness centrality of node s; $|V|$ is the total number of nodes in the graph; and $d(s,t)$ is the distance between node s and t, either by time or distance for a weighted network, or number of nodes passed for an unweighted network. The $|V| - 1$ factor is a normalization factor, so that the indicator does not depend on the number of nodes in the network.

Betweenness centrality

The betweenness centrality of a node gives an indication of how many shortest paths of all shortest paths (meaning all shortest paths between all possible node pairs excluding the node in question), pass through the node in question. It is calculated by (Barthélemy, 2011; Von Ferber et al., 2009; Lin and Ban, 2013)

$$b(s) = \left(\sum_{u \neq v \neq s}^{V} \frac{\sigma_{u,v}(s)}{\sigma_{u,v}} \right) / \frac{(|V|-1)\cdot(|V|-2)}{2}, \tag{3.14}$$

where $b(s)$ is the normalized betweenness centrality for node s; $\sigma_{u,v}$ is the number of the shortest paths between node u and v; $\sigma_{u,v}(s)$ are all the shortest paths between node u and v that pass through node s; and V is the total number of nodes in the graph. The sum is normalized by the division, so that the value is meaningful when comparing different graphs. When the graph is directed the factor two in the normalization factor must be omitted.

3.4.2 Network representation in different graph spaces

Three different graph space representations of the network are considered potentially interesting for this study (letters correspond to those in Figure 3.3): the L-space(b), B-space(c) and P-space(d). The other commonly used space, the C-space, where the lines are the nodes, and the arcs represent direct transfer possibilities between the lines (Von Ferber et al., 2009; Derrible and Kennedy, 2011), was not chosen to be represented. This space is less relevant for this work, as all the delay data is available per node and direction, but not per line, and so it would be difficult to compare it to the delay data.

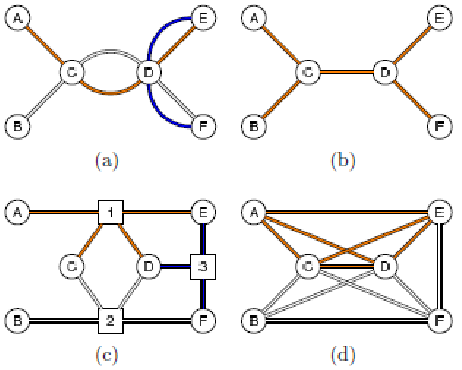


Figure 3.3: Four different graph representations of a network. (a) is a simple map of the services in a transport network, (b) is the L-space graph, (c) is the B-space graph and (d) is the P-space graph (Von Ferber et al., 2009).

Additionally, a representation as in (a) in Figure 3.3, whereby each service line has their own arc on each service section (Von Ferber et al., 2009; Derrible and Kennedy, 2011), was not used. This as only information per station is available in the delay data, not per service per station, making the comparison difficult. Furthermore, this space is very similar to the L-space, and using both could be superfluous. The other three spaces are discussed below. Their characteristics are mentioned, and how these spaces can be created from knowledge on the network is discussed.

L-Space representation

The L-space representation corresponds to (b) in Figure 3.3. This space considers all stations to be nodes, and all stations immediately adjacent on any number of services are connected by an edge (Von Ferber et al., 2009; Derrible and Kennedy, 2011), this results in a representation

that resembles the physical network, where no distinction is made between different line services. This graph can be directed or undirected, and weighted or unweighted, but is usually weighted in terms of either travelling time or distance.
The meaning and relevance of the indicators discussed in Section 3.4.1 in L-Space is as follows:

- **the degree centrality** indicates the number of adjacent nodes - a higher number of adjacent nodes could mean a higher potential for influence on the rest of the network;
- **the closeness centrality** indicates how close all other nodes are to the node in question in terms of time or distance - a higher proximity to the rest of the network could mean a higher potential for influence as there is less potential for delays to be deluded over time and space;
- **the betweenness centrality** indicates how many shortest paths pass through the node in question - more shortest paths passing through a node means more potential for delaying passengers, resulting in a potentially more critical node.

B-Space representation

The B-space representation corresponds to (c) in Figure 3.3. The nodes in this space are both all the stations, as well as all the lines. Station nodes can only be connected to line nodes, and line nodes can only be connected to station nodes. Edges represent the service of a line to a station (Von Ferber et al., 2009; Derrible and Kennedy, 2011). This graph can only be undirected and unweighted, as here the edges have no physical meaning and can thus not have any weight or direction.
The meaning and relevance of the indicators discussed in Section 3.4.1 in B-Space is as follows (it should be noted that the meaning of the indicators for line and station nodes differs, and that the indicators for line nodes are not relevant in this study as there is no corresponding delay data for them):

- **the degree centrality for station nodes** indicates the number of lines serving a station - if a station is serviced by more lines, there are more potential causes for delay, and chances of delay propagating (through e.g. vehicles waiting for other delayed vehicles from another line), this could lead to the station having a high influence;
- **the degree centrality for line nodes** indicates how many stations are served by the line in question;
- **the closeness centrality for station nodes** indicates how many lines need to be used in order to reach all other station nodes - if this value is low, people travelling from this station potentially need to transfer many times, this can increase the chance of incurring a delay;
- **the closeness centrality for line nodes** indicates how easily any station node is reached when this line can be accessed;

- **the betweenness centrality for station nodes** indicates whether this station is a transfer station, and for how many shortest paths this transfer station could be used - the relevance of this indicator is similar to the betweenness centrality in the L-space: if more shortest paths pass through a node, there is a higher potential for delaying more people;
- **the betweenness centrality for station nodes** indicates how many shortest paths make use of this line.

P-Space representation

The P-space representation corresponds to (d) in Figure 3.3. The nodes in this space are all the stations, similar to the L-Space representation. Edges are present between all stations that can be reached without having to transfer (Von Ferber et al., 2009; Derrible and Kennedy, 2011). This graph can be directed in case some service is only directional. It can also be weighted if distances or travel times between all stations on a line are known. However, this space is usually represented as undirected and unweighted.
The meaning of the indicators discussed in Section 3.4.1 in P-Space is as follows:

- **the degree centrality** indicates how many other stations can be reached without having to transfer - if few stations can be reached without having to transfer, this implies that people travelling from this location potentially need to transfer often, increasing their chances of incurring a delay;
- **the closeness centrality** indicates how accessible the network is in terms of few transfers - if this value is low, people travelling from this station potentially need to transfer many times, this can increase the chance of incurring a delay;
- **the betweenness centrality** indicates how many shortest paths have a transfer at the station in question - the relevance of this indicator is similar to the betweenness centrality in the L-space: if more shortest paths pass through a node, there is a higher potential for delaying more people.

3.5 Relating informativity and topological indicators

Calculating the informativity indicators, as discussed in Section 3.3, requires the BN, which requires much data. It would be very useful if the informativity of stations could be determined even when little data is available. Centrality indicators can be calculated with only knowledge on the network. If these indicators can be used as an approximation for informativity indicators, then the informativity of stations could be determined without requiring much data. This requires that the informativity indicators, as described in Section 3.3, and the centrality indicators, as described in Section 3.4, are related to each other in some way, although not necessarily causally. It must thus be determined whether some of the centrality and informativity indicators are correlated.

Some correlation between topological and informativity indicators is certainly expected. To understand why, the main causes of passenger delay propagation must be kept in mind. Some of these are: delayed vehicles remaining delayed and causing people at different stations to be delayed; and vehicles having to wait for delayed vehicles so people can transfer, or until the infrastructure becomes available, and thus becoming delayed vehicles themselves; and a disturbance in the infrastructure causing multiple vehicles to be delayed.

A station with, for example, a high degree centrality in the L-space (meaning it is directly connected to many stations) can be expected to be informative over many of these stations (resulting in a high score for the informativity indicator Outgoing Node Degree), as delays occurring at this station can spread to many of the nearby stations through a transfer of vehicle delay. This would thus result in a strong correlation between the degree centrality indicator in L-space, and the OND informativity indicator.

Moreover, a high closeness centrality in L-space means that many stations are relatively nearby. This would lead to a limited ability of delay to be diminished or mitigated, thus it can be expected that delay from such a station would easily spread to many other stations. This means that this station could provide much information on many other stations, which can be expressed in the total informativity indicator, resulting in a high correlation between this indicator and the closeness centrality indicator in L-space.

Indicators from B- and P-space could also be very relevant. For instance the degree centrality in P-space, gives an indication of how many other stations can be reached directly from the station in question (meaning without having to make transfers). Therefore delay can propagate to these stations easily through a single delayed vehicle or due to line-specific disturbances, both of which are common causes of delay. Therefore, a high degree centrality in the P-space could be correlated to any of the informativity indicators, as it would potentially be informative of many stations. In B-space, the betweenness centrality indicator could be very useful, as this indicates how many shortest paths pass through a particular transfer node. If many shortest paths pass through it, it is an indication of many lines merging at the station, increasing the potential of delay spreading from one line to other lines, and thus making this station informative over large parts of the network (which can be indicated by the total informativity). The betweenness centrality in L-space could result in a strong correlation to the total informativity for similar reasons.

Besides these expected correlations, other correlations could be observed as well, some of which for similar reasons as mentioned above. But potentially also for different reasons, as not all passenger delay propagation means are understood extensively. The approach taken in this study has the unique advantage to find these unexpected correlations as well.

Chapter 4

Application

In this section, the application of the methodology of Section 3, is discussed. First we will look at the case study that is used in Section 4.1. Afterwards, the exact implementation of the model is discussed in Section 4.2

4.1 Case Study

This section is devoted to discussing the available data, this data is on the Washington DC metro network, an American metro network that enables approximately 180 million trips per year (WMATA, 2017). The data was provided by the Washington Metropolitan Area Transit Authority (WMATA). An overview of the network can be seen in Figure 4.1. The metro network has 91 unique stations, of which 9 are considered transfer stations, and 93 unique service links. The metro is currently serviced by 6 unique lines, most of which have some overlap.

The rest of this section is organized as follows: a description of the data is given in Section 4.1.1; and how and which parts of the data were selected to be used is explained in Section 4.1.2.

4.1.1 Data description

Originally, a years worth of data from WMATA was available on: rail trips and movements, indicating the journey from an origin to destination as well as all the movements between adjacent stations; passenger journeys, indicating the passenger's origin and destination, which were known from smart card data; passenger movements, indicating an inference of the route travelled by the passenger based on the known origin and destination, and the known rail movements; disruption events; the timetable; and the locations of the stations and tracks.

This data was then further processed by Krishnakumari et al. (2019), all steps mentioned in this subsection (4.1.1) were implemented by her and were done previous to this study. After processing, the information took on a network-wide view, rather than an individual view. First, it was calculated which passengers were delayed and by how much. This was done by

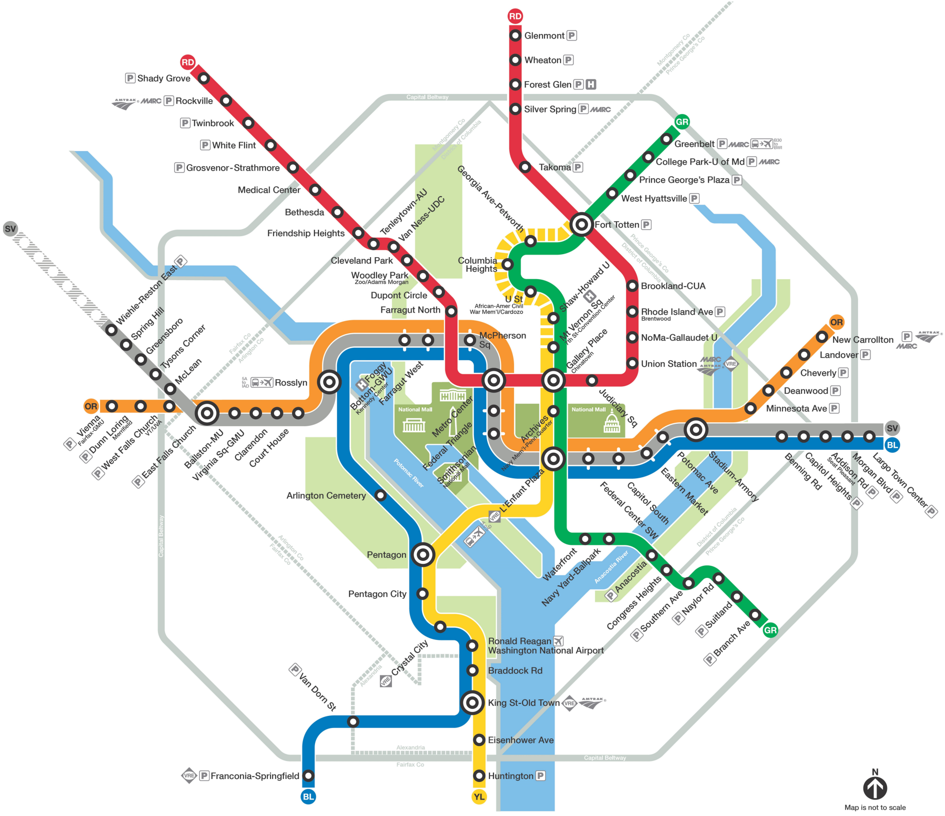


Figure 4.1: Map of the Washington DC metro network (Washington Metropolitan Area Transit Authority, 2017)

comparing the scheduled and observed travel times, like

$$\delta^n_{s_o,s_d,k} = \tau^n_{s_o,s_d,k} - \widetilde{\tau}_{s_o,s_d,k}, \tag{4.1}$$

where $\delta^n_{s_o,s_d,k}$ is the difference between the scheduled and actual travel time from origin station s_o to destination station s_d during time slice k (the time slice in which the passenger checked in at s_o); $\tau^n_{s_o,s_d,k}$ is the observed travel time; and $\widetilde{\tau}_{s_o,s_d,k}$ is the scheduled travel time. This equation will result in a positive number if the passenger was delayed, and a negative number if the passenger arrived earlier than the scheduled travel time. The observed travel time was simply determined by comparing the check-in and check-out times of a passenger. The scheduled travel time is stated to be the in-vehicle time of the estimated route, plus the transfer time if a transfer is included, plus the headway of each line travelled on (Krishnakumari et al., 2019). The headways are added to account for a random arrival of passengers with respect to the schedule, a reasonable assumption for a metro network where most headways are short.

The goal was then to estimate the delay incurred per link and station. Here, the links and stations are all directed. Furthermore, a distinction was made between the delay incurred at the origin station of a passenger, and the delay incurred at the transfer station of a passenger. An equation was thus created with the delay on the initial station, delay on the transfer stations, and the delay on the links, which should be equal to the delay at the final destination, for a single passenger journey, like (Krishnakumari et al., 2019)

$$\delta^n_{s_o,s_d,k} = \delta^W_{s_o,k} + \sum_{(s_x,s_{x+1})}^{E} (b^{nT}_{(s_x,s_{x+1})} \cdot \delta^T_{(s_x,s_{x+1}),k} + \sum_{s_i}^{I} (b^{nI}_{s_i} \cdot \delta^I_{s_i,k}), \tag{4.2}$$

where $\delta^n_{s_o,s_d,k}$ is the estimated delay of passenger n traveling between s_o and s_d, having departed at period k; $\delta^W_{s_o,k}$ is the initial waiting time; $(b^{nT}_{(s_x,s_{x+1})}$ is a binary value indicating whether the track between stations s_x and s_{x+1} is part of the estimated path; $\delta^T_{(s_x,s_{x+1}),k}$ is the delay on the track between stations s_x and s_{x+1}; $b^{nI}_{s_i}$ is a binary value indicating whether transfer station s_i is path of the path; and $\delta^I_{s_i,k}$ is the delay at transfer station s_i.
A set of equations was then created be setting up such an equation for each passenger. As it would be impossible to solve this set of equations in a continuous manner, the data is made discrete over time, where time slices of 30 minutes were taken. This was done as all headways of all lines are shorter than 30 minutes, meaning a service is always provided within the time slice. Passenger data was assigned to the time slice in which they checked-in. It should be noted that during their journey they can cross into the next time slice, time wise. Therefore, there are some dependencies between the information in time slices, which should be taken into consideration when further using this data (Krishnakumari et al., 2019).
This method of inferring link and station specific delay from passenger delay delay works, as there are no services in any of the stations, otherwise delays could be related to things other than the service of the transport network, making the equations meaningless. This is an important limiting factor for other networks, which might not be able to infer the link and station specific delay in this same manner.
Eventually, data is available on the delay at each directed link, initial station and transfer station. For this study, this information was available for an entire year, from September 2017 till August 2018. A time slice is available for each time of the day, however, during the night no services are available and so there is no meaningful delay data. Furthermore, during service times it is possible a certain station is not used during a certain time or by very few people (less than 10), leading to unavailable or unreliable data (Krishnakumari et al., 2019). However, effectively the delay incurred here is then 0 minutes, and thus these unavailable data points have been set to 0 minutes.

4.1.2 Data selection

Considering demand patterns and services are very different on weekends compared to weekdays, data from weekends should not be taken into account. It is possible that relations between the delay at the different stations could be influenced by different timetables or demand patterns, so including all the data could muddle the results. Considering weekdays usually see higher passenger counts due to people travelling to their work, these days are socially more relevant, so it was chosen to discount the weekends rather than the weekdays.
Moreover, some days might see irregular demand patterns in case of a national holiday, big maintenance or due to some check-in system failure. The demand pattern of all days must

thus be checked, and days with any abnormal patterns should be discounted, for similar reasons as to why weekends were discounted. For example, days with ten times lower passenger counts than average and no discernible morning and evening peak pattern can be discounted as they point to an irregular day like a holiday. Furthermore, dates were over half of the data shows zero passengers in the system, were either special holidays, subject to maintenance or check-in system or data processing failures. It is hard to distinguish when the earlier data in such a day is still relevant or when it might have been influenced as well, so these days can be discounted as well. An overview of these days can be found in Appendix C. As originally an entire year's worth of data is available, discounting some of it should still leave enough data for valid results.

4.2 Implementation

In this section, several of the specifics of the implementation of the model will be discussed. Firstly the implementation of the eventual discretization method is elaborated upon in Section 4.2.1. After this, the general implementation of the structure learning (Section A.1) and CPD table learning (Section 4.2.3) are discussed, while any software specific details such as function names can be found in Appendix A. Lastly, some issues and solutions regarding computational demands are evaluated in Section 4.2.4.

4.2.1 Discretization

For the purpose of this study, the method of entropy loss minimization turned out to be intractable, as it resulted in far too many partitions to make the creation of the Bayesian network possible in an appropriate amount of time. Furthermore, an abundance of partitions would make it very difficult to interpret the results in a meaningful and complete manner.
Since the focus of the study is not on discretization methods, it was decided that the less accurate but far more tractable approach of discretizing data by equal population, was more appropriate. However, considering almost all the data in the data set is of delays of approximately 0 minutes, all but some of the bins were regarding 0-values. Therefore it was decided to slightly adjust this method by using a logarithm. The approach used was to divide the data set into very small initial bins (a size of 0.01 minute was taken), and to then calculate the logarithm of the number of data points in this bin. An example of this process is visualized in Figure 4.2, where the blue line resembles the original population per bin of size 0.01, and the orange line the logarithmic value of this population.

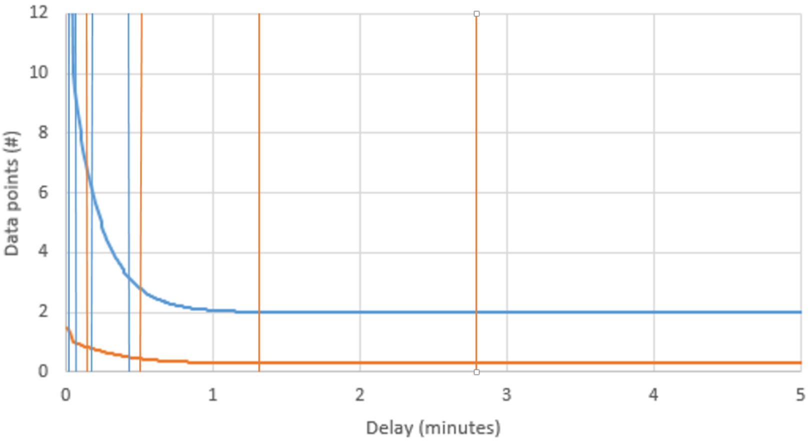


Figure 4.2: An example of the number of data points or population per bin of size 0.01 (blue line) and the logarithm of this population (orange line), and the resulting dividers of the final 5 bins, for an imaginary data set.

These logarithmic numbers where then taken to determine the total population, as in:

$$\mathrm{LPop}^T = \sum_{b_i}^{B} \log(\mathrm{Pop}_{b_i}), \tag{4.3}$$

where LPop^T is the logarithmic total population, B is the total number of bins, and Pop_{b_i} is the number of data points in bin b_i. Dividing the value of LPop^T by the number of final bins desired, will result in the desired logarithmic population per final bin. To then determine the sizes of the final bins, the logarithmic populations in the bins of size 0.01 can be added, until the desired population for the final bin is reached, this indicates the upper limit in terms of delay of the final bin. This process is then continued for the rest of the final bins.

The difference between this logarithmic method, and the standard equal population discretization method, can clearly be seen in Figure 4.2. Here the blue line represents the normal population, and the blue vertical lines, the bin dividers of the bins of the equal population method. The orange line represents the logarithm of the populations, and the orange vertical lines, the bin dividers of the equal logarithmic population method - note that the area below the line, between the dividers represents the population of the final bin, and should thus be equal. It can clearly be seen that the orange method results in a more interesting division of bins in this case, as they are not almost all near the end of the spectrum.

Thus, eventually it was chosen to use this equal logarithmic population to create the bins that are used when creating the BN, rather than the entropy loss minimization method.

4.2.2 Structure learning

After the data has been appropriately configured, the BN can be created. The first step to do this, is determining its structure, which is a directed acyclic graph (DAG), without any labels for either nodes or edges (Koller et al., 2009; Kjaerulff and Madsen, 2008). This requires input: the data-set, as well as two parameters: the p-value and the in-degree, which will be discussed below. The output of this step is then the directed, acyclic graph, the structure of the BN (Cyberpoint International, LLC, 2012). The exact approach of determining the DAG is discussed below.

Parameters

The first input parameter is the p-value. The p-value is used as a measure of the significance of any dependence in the network (Koller et al., 2009). If the significance of a dependence is lower than the p-value, it is not added to the structure of the BN. The p-value used in this study is 0.05, this is the standard significance threshold (Koller et al., 2009), and there is no reason to diverge from this. The second input parameter is the in-degree. The in-degree is the upper-bound of the size of the witness set, what this does exactly will be explained in more detail below. If some nodes in the resulting DAG are expected to have an in-degree (the number of incoming edges) of more than 2, an in-degree of more than 1 should be considered for accuracy, if at all possible. However, increasing the in-degree to consider for a network increases computational times. Furthermore, increasing this parameter when a small data set is available, could lead to issues such as a division by zero, so it must be considered carefully (Koller et al., 2009; Cyberpoint International, LLC, 2012). For this project an in-degree of 1 was chosen. A higher in-degree would have been desirable but lead to errors caused by divisions by zero, due to the relatively small data set when comparing the amount of data points to that of the number of nodes in the BN.

Approach

For this study, a constraint-based approach is used to determine the DAG. There are effectively goes two steps to this approach. Firstly, all of the dependencies between the variables in the data-set provided, are determined. This is done by testing the dependency of every pair of variables X, Y. This dependency is checked given the set of variables U. The in-degree discussed earlier, gives the maximum size of the set of U (Cyberpoint International, LLC, 2012). The null-hypothesis of this test is that the variables X, Y are conditionally independent on each other, given the set U. Meaning that if the set U is known, X and Y are independent, thus no edge exists between them (Cyberpoint International, LLC, 2012; Koller et al., 2009; Kjaerulff and Madsen, 2008). In an equation, this is described as

$$P(X = i, Y = j, U = k) = P(U = k) \cdot P(X = i|U = k) \cdot P(Y = j|U = k). \tag{4.4}$$

where $P(X = i, Y = j, U = k)$ is the combined probability of finding variable X in state i, Y in state j and the set of U in state k; $P(U = k)$ is the probability of finding variable k in state k; and $P(X = i|U = k)$ is the probability of finding variable X in state i if it is known that variable U is in state k. This equation is true if X and Y are independent of each other. If it is found that X and Y are not independent at a significance level of the p-value for every possible set U, an edge is added to the DAG between these two variables.

When all significant edges are determined, the DAG structure is completed by assigning directions to all edges. Firstly the structure is checked for any immoralities. An immorality is the structure $X \rightarrow Z \leftarrow Y$, as here X and Y are considered independent, even if Z is not observed (meaning it is not part of the witness set when checking X and Y for their independence). It can thus be distinguished from the three other possible structures $X \rightarrow Z \rightarrow Y$, $X \leftarrow Z \leftarrow Y$ and $X \leftarrow Z \rightarrow Y$, each of which can only show X and Y to be independent if Z is part of the witness set (Koller et al., 2009; Kjaerulff and Madsen, 2008). Any sub-structure in the form of $X - Z - Y$ is a potential immorality if there is not also an edge between X and Y (meaning it has a V-structure, rather than a triangle structure) (Koller et al., 2009; Kjaerulff and Madsen, 2008). Checking for which witness sets X and Y are independent, can confirm whether there is an immorality or not, thus determining part of the directions of the arcs of the BN.
After all immoralities have been determined, the rest of the structure is determined based on the constraint that the graph must be acyclic. Three rules can be distinguished, as can be seen in Figure 4.3. On the left hand side possible structures are shown that can be encountered during the process. On the right hand side, the structures are shown that are the result of the requirement that the final graph must be acyclic. The first rule , makes use of the fact that all immoralities have been determined before this step. The fact that the the edge from Z to Y is not yet present, indicates that the structure cannot be that of an immorality, and must thus be as shown. The second rule is simply based on the fact that cycles are not allowed. The third rule is the result of a combination of the determination of all immoralities and the acyclic nature of the graph: if the edge from X to Y had been reversed, inevitably either an immorality or cycle would come into existence (Koller et al., 2009; Kjaerulff and Madsen, 2008).

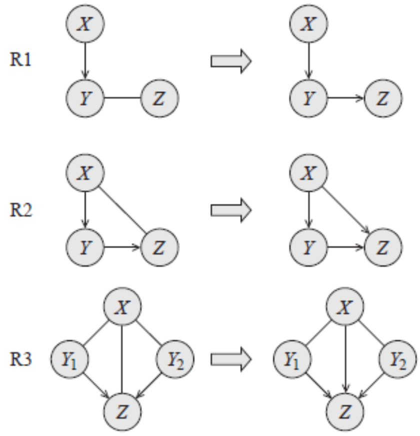


Figure 4.3: The three rules that help determine the directions of the edges in a DAG (Koller et al., 2009).

As the rules mentioned above are applied, more structures fitting the structures on the left hand side in Figure 4.3 are created. Eventually all edges are assigned a direction, and the DAG is completed (Koller et al., 2009; Kjaerulff and Madsen, 2008). It is possible that the graph found contains a cycle. In this case, an error is generated and the process cannot continue.

4.2.3 CPD learning

The second step in learning the BN, is determining the conditional probability distributions (CPDs) of all variables. This requires requires two inputs, namely the DAG of the BN, as discussed in the previous section, as well as the data-set on which the DAG is based.

The output is then a BN including DAG and CPDs. Below the approach to achieve this is discussed.

Approach

First, information about the parents and children of each variable is gathered from the DAG input. Then, information about the possible states of the variables is gathered from the data-set, which is used to initialize CPD tables for all possible parent and state combinations for each variable, with all probabilities initially being zero. Then, the number of times a state is observed, for a certain combination of parent states, is counted and normalized. This results in complete CPD tables, where for every parent-state combination, the sum of the probabilities of the variable being in state x, is exactly one (Cyberpoint International, LLC, 2012; Koller et al., 2009).

If a parent-state combination is not present in the data, then every state of the child node is assigned an equal probability (so for five possible states, the probability of finding each would be 20%, for the given parent-state combination). This is a disadvantage of the BN, which cannot make an accurate prediction of situations not observed in the data set. It might be more realistic to assign the probabilities depending on the prior distribution of the child-node (so the probability of finding each state independent of any parent states). However, it is expected that the influence of this limitation is not extensive, as clearly the probability of the parent-states occurring in that particular pattern is also low enough for it to not have been observed (meaning it is unlikely to happen and therefore less relevant).

4.2.4 Decreasing computational demands

It was tried to create a BN for the entire network in one go. However, due to the large number of nodes used when creating the BN, no results were found after a full month for even one fold. Due to the large combination of node pairs to check, it could turn out that such computations for the entire network could take more than months using the processing power available for this study. Two potential solutions to this problem of large computational times were tried: dividing the network into sectors and combining nodes to form super nodes. Both methods will be shortly discussed below.

Sector method

Using this method, the network is divided into seven sectors. Where for each sector a separate Bayesian Network is created. The first and biggest sector encompasses the center of the network (all the nodes not in a sector in Figure 4.4), as well as the first two stations as seen from the center of the radial sectors, and also all transfer stations whether they are in another sector or not. The transfer stations have been added as it might be expected to find relations among transfer stations even when they are further apart. Why the first two stations of each radial sector are added, is explained below. The other six sectors each encompass one of the radial parts of the network, as seen in Figure 4.4 where the sectors are also numbered.

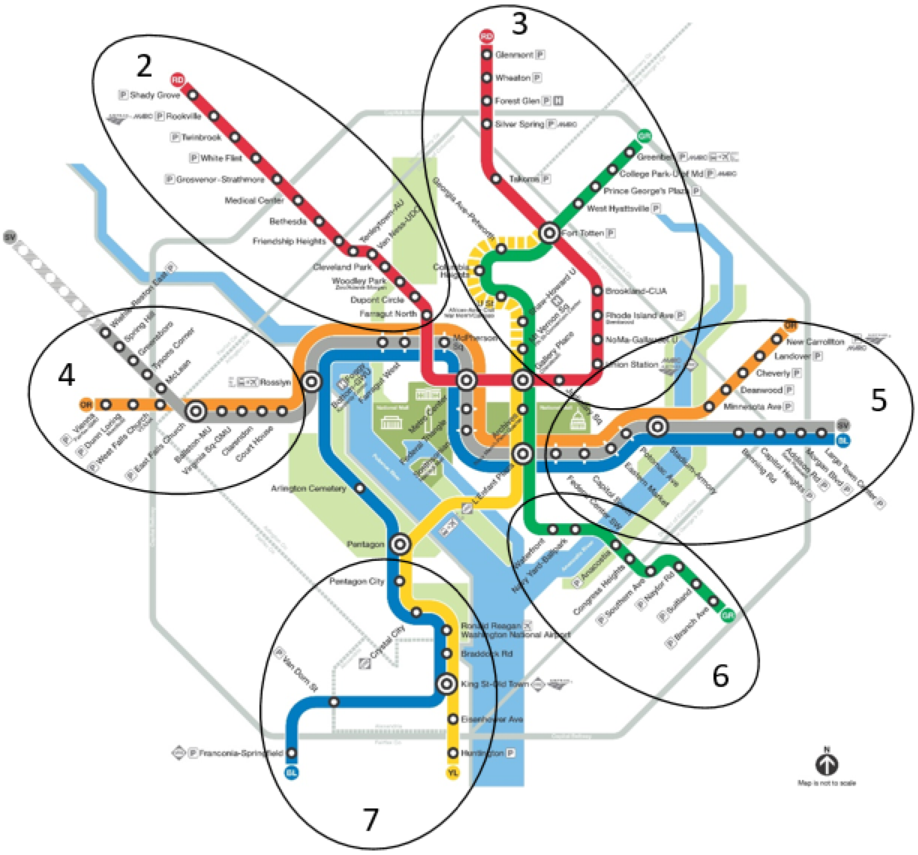


Figure 4.4: An overview of the network with an indication of the six radial sectors.

This method assumes that stations that are further apart (separated by the center of the network), cannot be directly related in terms of information. This is a reasonable assumption, as such stations most likely only indirectly hold information on each other, if at all, through the stations that lie in between.

As this method divides the network into sectors, each sector will have significantly fewer nodes when creating the BN, making the computational time much less. Eventually the BN for all sectors can be created within several hours, depending slightly on the data set used (some data sets might result in more densely connected networks which would increase computational times slightly).

Another potential downside of this method is that the result will be the seven separate BNs, which is not optimal when they must also be used for further processing for the informational indicators. This can be resolved by recombining the sectors after the creation of the BNs. This can be done by having some overlapping nodes in the different sectors. Nodes that can be used for this are the transfer stations, which are all present in the first sector, as well as in the sector in which they are physically located. Furthermore, the first two stations from each radial sector, are added to the first sector as well, for extra connection possibilities. It was chosen to add two stations to limit the increase in computational time for the first sector, while still having multiple possibilities to reconnect the BNs after their creation.

Super Node method

The Super Node method works by aggregating multiple nodes into a single node. The delay in the aggregated node is then the average of the delays of all its underlying nodes. The average is chosen instead of simply the sum, as some super nodes could contain more sub-nodes than others, meaning the summed would be influenced by the number of sub-nodes resulting in a seemingly higher delay, which would not lead to accurate comparisons. This method assumes that the set of nodes that are part of one aggregated node, are closely related in terms of delay. Furthermore, due to the averaging of delays, information is lost. However, as opposed to the Sector method, it can check the relation of stations that are further apart.

For this study, all nodes on sections where no transfers were possible, were aggregated, whereby a distinction is still made in terms of the direction of the delays. So most of the radial lines were aggregated into several node. It was decided to do this, as on these sections, less complicated relations would be expected due to limited possible causes for delay. Furthermore, transfer delays are slightly different from initial waiting time delays, so it would not be reasonable to aggregate transfer delays together with initial waiting time delays.

A distinction between the directions is still made, as there is no reason to believe that the delay in direction x, will always or even often, be related to the delay in direction y, so it would not be reasonable to aggregate these delays.

Chapter 5

Results

In this section, the results of the study are discussed. Firstly, some exploratory data analysis is done of the delays in the network in Section 5.1, this is done in order to understand the data in relation to the network, so interesting areas or features can be discovered. Then the results of the Bayesian Network compilation are discussed in Section 5.2, this will then lead to the results of the informativity indicators which are discussed in Section 5.3. Afterwards, the analysis of the network, and its centrality indicators, is discussed in Section 5.4. Finally, this results in the correlations of the informativity and centrality indicators, which are discussed in Section 5.5.

5.1 Exploratory data analysis

In this section the available data is explored. The focus is on delay and discovering peculiarities of this delay in the network. In order to do so, the passenger counts will also be analyzed, in order to find not only high delays, but also high average passenger delays.

For the purpose of this study, the focus is on regular weekdays, therefore holidays and weekends are excluded from the data set, as discussed in Section 4.1.2. Furthermore, only data from the service hours is taken into account, so the hours between 4:30 and 24:00.

In Section 5.1.1 some data on the network as a whole is discussed. In Section 5.1.2 data per node and link is analyzed.

5.1.1 Network-wide analysis

In this section the data is analyzed on a network wide level, meaning there is no differentiation between individual stations or links. This is done to check the general passenger flow and delay patterns to see if anything stands out, and to understand the characteristics of the network, such as peak-hours in terms of both demand and delay. Lastly, a distribution of all the delay data points is given, so it becomes clear what exactly the input data of the model is.

In Figure 5.1 the total number of passengers entering the metro system per time slice on an average day can be seen. This graph is obtained by adding the number of passengers entering the system each day per time slice, and dividing by the number of days taken into account. The demand clearly follows a standard morning and evening-peak pattern, with much higher passenger counts between approximately 7:00 and 9:00 (morning peak) and 16:00 and 18:00 (evening peak), where at peak demand over 40000 passengers on average enter the system during 30 minutes.

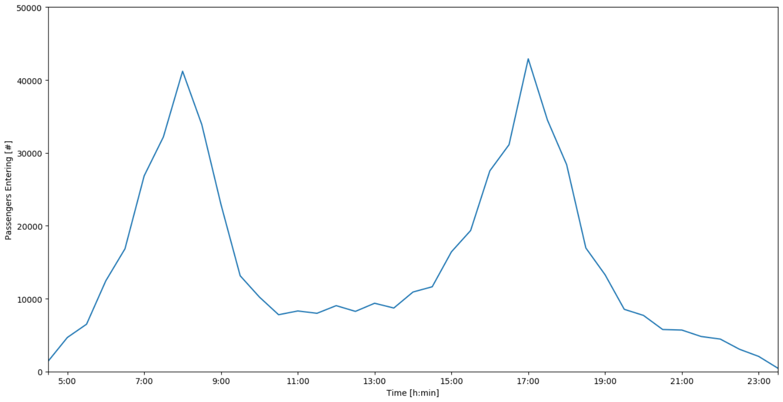


Figure 5.1: Total number of passengers entering the system per time slice on an average day.

In Figure 5.2 the average passenger delay per time slice can be seen per transfer and initial station and link. It might have been expected that the average delays experienced would be higher during rush hour, as overcrowding in vehicles can result in denied boarding for some passengers, and thus extra delays. However, this clearly is not the case for the Washington DC metro network, indicating that the vehicle capacity is not an issue for this network.

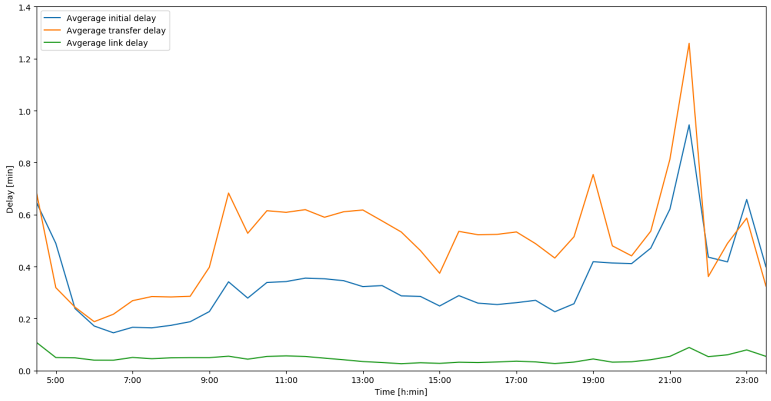
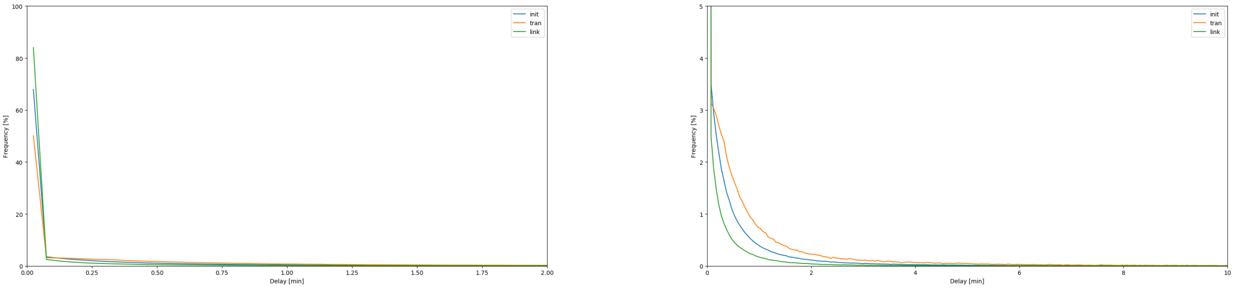


Figure 5.2: Average passenger delay per time slice per initial station, transfer station or link. Averaged over all days and passengers, and respectively the number of initial stations, transfer stations and links.

It is also notable that the average delay experienced on a link is rather low, compared to the delay experienced at initial and transfer stations. However, this does not mean the link delay is irrelevant, as a passenger journey usually consists of a multitude of links while only consisting of one initial station and, at maximum, a few transfer stations. The delay experienced on links thus still adds up to a significant proportion of a passengers delay.

The delay patterns of the initial and transfer stations are very similar to each other but inconsistent throughout the day, as opposed to the link delay which is rather stable. Relatively low delays are experienced during the morning rush hour (7:00-9:00), while big delays are experienced around 21:30. The reason for this latter peak could be a change in scheduled headways, which increase during off-peak hours. Such an increase can exacerbate delays. For example, if the headway on some line initially is 10 minutes, and a passenger just misses the vehicle they had intended to get and the next vehicle is cancelled, they can be up to 20 minutes delayed. When then the headways are 20 minutes off-peak, this potential delay in the same situation is increased to 40 minutes; a rather substantial increase. Moreover, the pattern of the initial and transfer stations also shows some severe peaks and lows, such as those at 19:00 and 21:00. As the delays are averaged over an entire year of data, it would be expected that the delay pattern is more smooth, similar to the pattern of the link delay. There are two possible reasons that it is not as smooth as might be expected: there are not enough data points, so random fluctuations are not averaged out; or there is a reason for these specific peaks, examples of such reasons are the earlier mentioned change in headways, or systemic delay issues due to bad timetabling.

In Figure 5.3 a distribution of the frequency of delays can be seen for the initial stations, transfer stations and links. A zoomed in overview is provided as well. Here, it is clear that transfer delays are most often higher, while link delays usually are lower, with 98% of all data points (where every link is a separate data point in each day and time slice) having a delay of under 1 minute, and only 0.09% of data points having a delay of over 5 minutes. For initial stations these percentages are 94% and 0.55% respectively. Transfer stations clearly experience more delay on average, at 86% of data points showing a delay of under 1 minute, and 2.2% of data points showing a delay of over 5 minutes.



(a) Full overview

(b) Zoomed overview

Figure 5.3: The percentage of times a delay of certain size is present in the data set for transfer, initial and link delays. The label init indicates initial stations, tran indicates transfer stations and link indicates links.

Clearly, most data points are of near-zero minutes delay. This does not necessarily mean that few delays are experienced by passengers, as delays might be more common in areas with large passenger counts. Furthermore, the number of data points with a larger delay is still substantial, especially when it is considered that the delays experienced at different stations and links stack up per passenger. Clearly, in an absolute sense there are relatively little data points of much interest, but there should still be enough points of interest for meaningful results.

5.1.2 Node and link analysis

This section discusses the delay data per node and link. To visualize this data, maps were created colored by different indicators regarding passenger counts or delay, per directed initial station (nodes), directed link (arcs between the nodes) and transfer station. The transfer stations are represented by multiple arcs between nodes of the corresponding initial station. For example, the delay of passengers transferring at Metro Center into the direction of Gallery Place, is represented by an arc to the node of Metro Center in the direction of Gallery Place, from all the other nodes of Metro Center.

Passenger counts

In Figure 5.4 the number of passengers making use of the stations and links on average weekdays can be seen, this figure can be used to find hot-spots in the network, and to check whether the data is as expected.

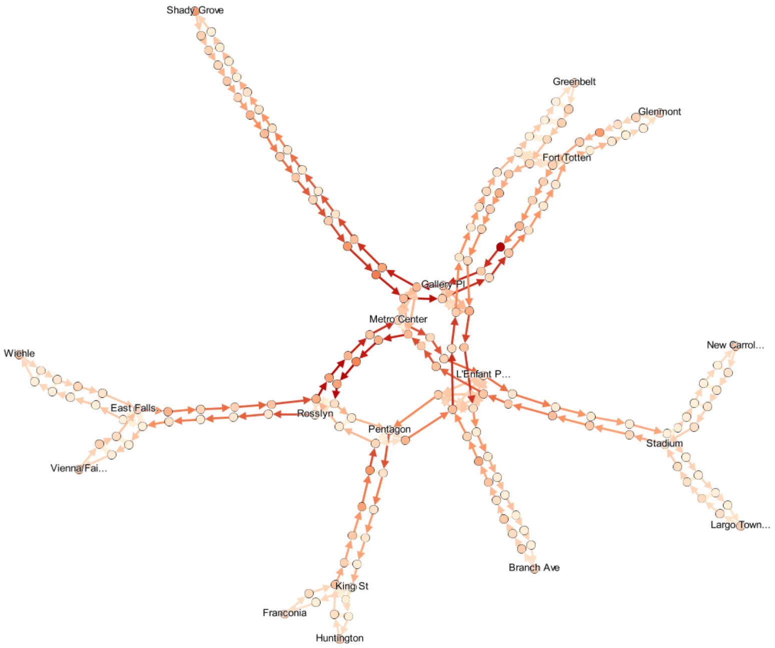


Figure 5.4: Passenger counts per day per directed initial station, transfer station and link. The maximum value is 22000 passengers/day while the minimum is 0 passengers/day.

It can be noted that along the radial lines, the passenger counts increase towards the center of the network. This makes sense, as passengers accumulate along the line. Furthermore, the links between the same physical station, but going in opposite directions, all have approximately the same passenger count. This is to be expected as it can be assumed that most people travelling from some place, will go back to the same place at another time. Therefore, the data can be said to adhere to expectations.

The minimum value of 0 passengers/day for some of the initial station nodes might seem very low, as a station with an average passenger service of 0 passengers is not very useful. However, it should be taken into account that the nodes are differentiated per direction. So it is only natural that the second to last station on a line, would not see any passengers entering it to travel to the last station in the line, especially if these stations are geographically very close, as then it would be easier to use other means of transportation such as walking.

Another noteworthy feature is the very high maximum value of 22000 passengers served daily. This is the value of the one initial station node with an extremely dark color compared to the other nodes (on the right line between Fort Totten and Gallery Place). This is Union Station, and it is the only station connected to all three rail systems (VRE, MARC and AMTRAK). It is also an important leisure destination in Washington D.C. and one of the US's larger transportation hubs. This makes this metro station an important feeder station for passengers coming from further away by rail, as well as an important station for the population of Washington D.C. due to the nearby leisure hot-spot (Washington D.C., 2019). This explains the high passenger counts, and indicates this might be a station of interest regarding delays as well, as many people would be affected by a delay at this station.

Average delay

In Figure 5.5 the average delay can be seen. This value is calculated by simply adding the delay occurring at any station or link during every time slice available, and then dividing by the total number of time slices.

It can be noted that there are relatively high initial delays throughout the network, while the link delays are relatively low. Especially at Metro Center and Gallery Place some high delays can be observed. Furthermore, the transfer arc at Pentagon in the direction of Rosslyn has a relatively high delay. Interestingly enough, the other transfer delays do not seem very high at all, which was expected due to the fact that in Figure 5.3 it can be seen that high transfer delays occur more often than high link or initial delays. The reason they don't stand out much in this figure, could be that although the initial waiting delays on average are not often very high, few stations account for many of the higher delays.

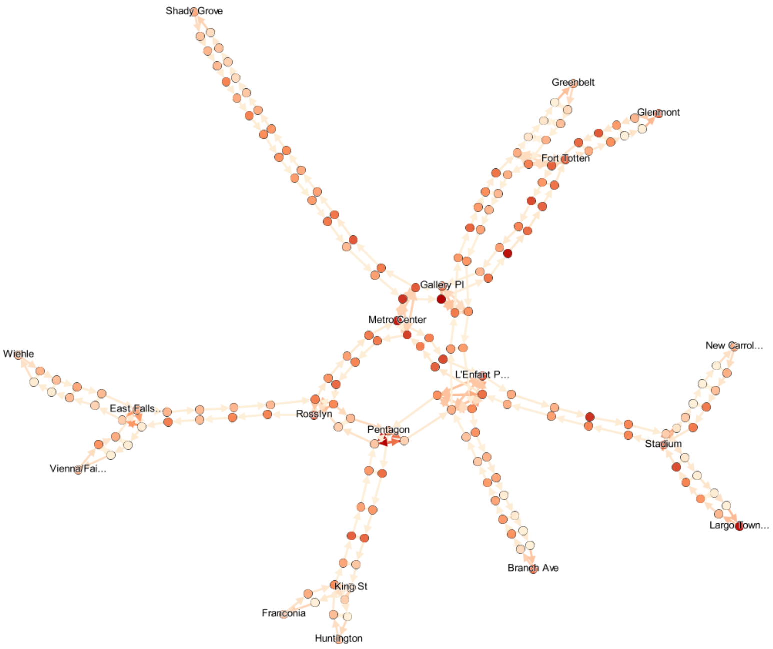


Figure 5.5: Average delay per directed initial station, transfer station and link. The maximum value is 2.69 minutes while the minimum is 0 minutes.

Total passenger delay

In Figure 5.6 the total passenger delay per day can be seen for the entire network. This delay is calculated by multiplying the number of passengers experiencing a delay by the delay, and summing this value for every data point, and then dividing by the number of days.

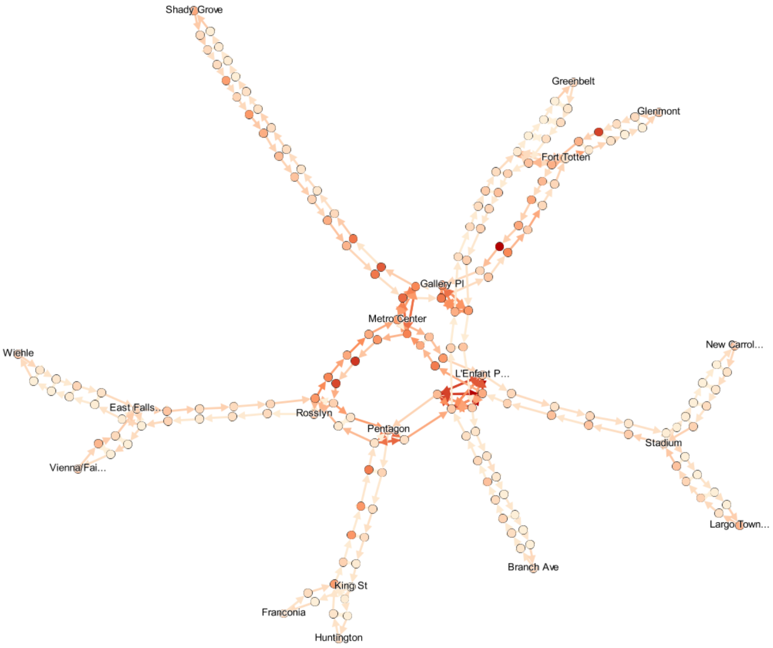


Figure 5.6: Total passenger delay per day per directed initial station, transfer station and link. The maximum value is 8400 passengers-minutes/day while the minimum is 0 passenger-minutes/day.

In this figure, again, Union Station stands out. Clearly, this is not because of the high delays experienced here, but rather due to the high number of passengers making use of this station, so that any delay occurring here leads to a high number of passenger-minutes of delay. Still, this does not take away from its relevance, as any improvements in diminishing delays here would have big impacts. The high passenger delays seen at the transfer links of L'Enfant Plaza follow a similar reasoning: they are mostly caused by high passenger counts, rather than high delays. Yet this would then still be a relevant place for improvements as many people would be affected by it.

What also stands out is that most passenger-minutes of delay are occurring in the center ring of the network. Here, the figure is not quite the same as the map of passenger counts in Figure 5.4, meaning that the high number of passenger-minutes of delay cannot be explained completely by the high number of passengers travelling in this center ring. This must mean that relatively high or many delays occur in the center ring, which could mean there is greater potential here for improvements when trying to reduce delays, than on any of the radial sections of the network.

Average passenger delay

In Figure 5.7 the average delay for any passenger can be seen. These values are calculated by taking the value of the total passenger delay for any location, and dividing it by the total number of passengers.

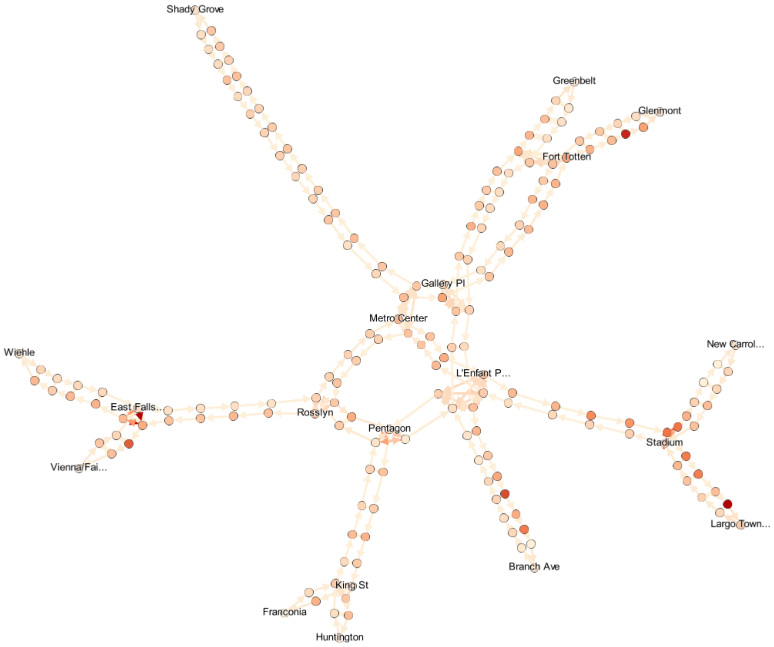


Figure 5.7: Average passenger delay per directed initial station, transfer station and link. The maximum value is 4.63 minutes while the minimum is 0 minutes.

This figure does not resemble any of the other figures and the numbers seem generally rather uniform over the entire network, meaning there are not many areas (such as the center or one of the lines) where a relatively high delay can be expected for the average passenger. However, there are some standout stations, such as a station close by the end station Glenmont, and one by Largo Town Center, and a transfer at East Falls Church. Why these stations would see relatively high expected delays, is unclear and could be due to local circumstances such as capacity constraints specifically during times when most people are travelling from there. What should be noted about all of these figures, is that all the maximum values for the transfer stations are on average very high compared to those of the initial stations and links. It can be said that more delays and also more severe delays are incurred during transfers than during initial waiting times and any in-vehicle delay on a single link. However, only 36% of all trips have a transfer, and the average number of legs per trip is 1.43, indicating most trips don't have a transfer at all, and very few have more than one. Meaning not necessarily very many people are affected by these transfer delays, though their size and frequency still make them very relevant.

5.2 Bayesian network

In this subsection the results of the Bayesian Network calculations are discussed. As mentioned in Section 3, from here on out only transfer and initial station delays are considered, while link delays are neglected due to their different nature.

The results discussed here are those created by using the Sector method (see Section 4.2.4 for details on the Sector method). The Super Node method in the end was not able to reduce computational times enough to be a viable method. After two weeks the calculations were still not finished and it was thus decided to abandon this method and use only the Sector method. The Sector method is able to reliably do the calculations with computational times between a few minutes to three hours per sector per fold. However, it should be noted that initial tests of the Sector method on a small section of the network resulted in a significantly different Bayesian Network as compared to the Bayesian Network created when the section was not divided into sectors. The overlap between the two BNs was only 44% in terms of arcs present, while the Sector method had an additional 122% new arcs. This indicates the Sector method is not quite representative of the reality, although the specifics could be different on a larger scale. It was decided to continue using the Sector method regardless, as no viable alternative was available. Therefore, it should be kept in mind that the results are not necessarily an entirely accurate representation of reality.

In the results in this section, all nodes that are not connected by any arc, are left out of the figures in favor of clarity. These results are created using all the available data for regular weekdays during service hours. The nodes are given special codes based on their line, location and direction (line code in two letter, number based on position in line, and a letter between a-f to indicate direction), so that they can easily be related to each other. How these codes are determined and can be read is described in Appendix B. Due to the chosen representation of the BNs, it is difficult to immediately compare the BN to the physical network, however, the codes serve to make this easier as approximate location and direction can easily be seen.

Furthermore, due to the inaccuracy of the Sector method, the specific relations in the BN are not reliable and should not be regarded as such. Therefore, only general trends in types of relations are of interest. Moreover, to show the general relations as compared to the physical network, in the end the BNs are mapped onto the physical network so some comments can be made on this as well.

The structure of the rest of this section is as follows: first the results of the discretization method, in terms of bin sizes, are discussed in Section 5.2.1; then the results of the BN for the different sectors are discussed in Sections 5.2.2 through 5.2.7; these results are synthesized in Section 5.2.8; lastly, the errors of the results are commented on in Section 5.2.9.

5.2.1 Discretization

For this study, a final number of five bins is adhered to. Initially, several numbers of bins were used (four, six and eight bins) and it was found that the density of connections in the BN was similar for all numbers of bins. It might have been expected that a higher number of bins would lead to fewer connections, as slight randomness of the data would more easily result in the data ending up in a different bin and thus correlations being smaller. But this is not the case. Furthermore, a larger number of bins did not necessarily lead to higher calculation times due to the larger conditional probability distribution tables, and so neither a higher nor lower number of bins would be advantageous. Therefore an average number of bins, five, is used. The resulting bin sizes can be found in Table 5.1. The bin sizes of the initial and transfer delays are clearly very similar. Furthermore, they are meaningful, as all bins are clearly distinguishable (e.g. no multiple bins within the same range of minutes like 0-0.01 and 0.01-0.02), nor are they excessive (all bins comprise only a few minutes, with the exception of the very last which simply contains all the very long delays).

	Initial delay (min)		Transfer delay (min)	
	Lower	Upper	Lower	Upper
Bin 1	0	1.6	0	1.6
Bin 2	1.6	3.75	1.6	3.75
Bin 3	3.75	6.55	3.75	6.45
Bin 4	6.55	10.5	6.45	10.15
Bin 5	10.5	-	10.15	-

Table 5.1: A table with the upper and lower limit in minutes of the five bins for both the initial and transfer station delays.

5.2.2 Sector 1

In Figure 5.8 the results for sector 1 can be seen. This sector is the most dense in terms of both connected nodes and arcs. This makes sense as this sector also consists of the most nodes used when creating the BN. Clearly, the transfer stations are very well connected, especially to each other, even when they are further apart. However, many of the connections are not very strong, such as the connection between *T: EP-YL* (Transfer station L'Enfant Plaza in the direction of Pentagon) and *T: OR(SV)-5-d* (Transfer station East Falls Church in the direction of West Falls Church). This connection of only 0.014 indicates only 1.4% of the delay at either station can be determined by the delay at the other station. However, these stations are in fact quite far apart, and their directions are also not related, therefore a large or even any connection at all would not be expected anyways.

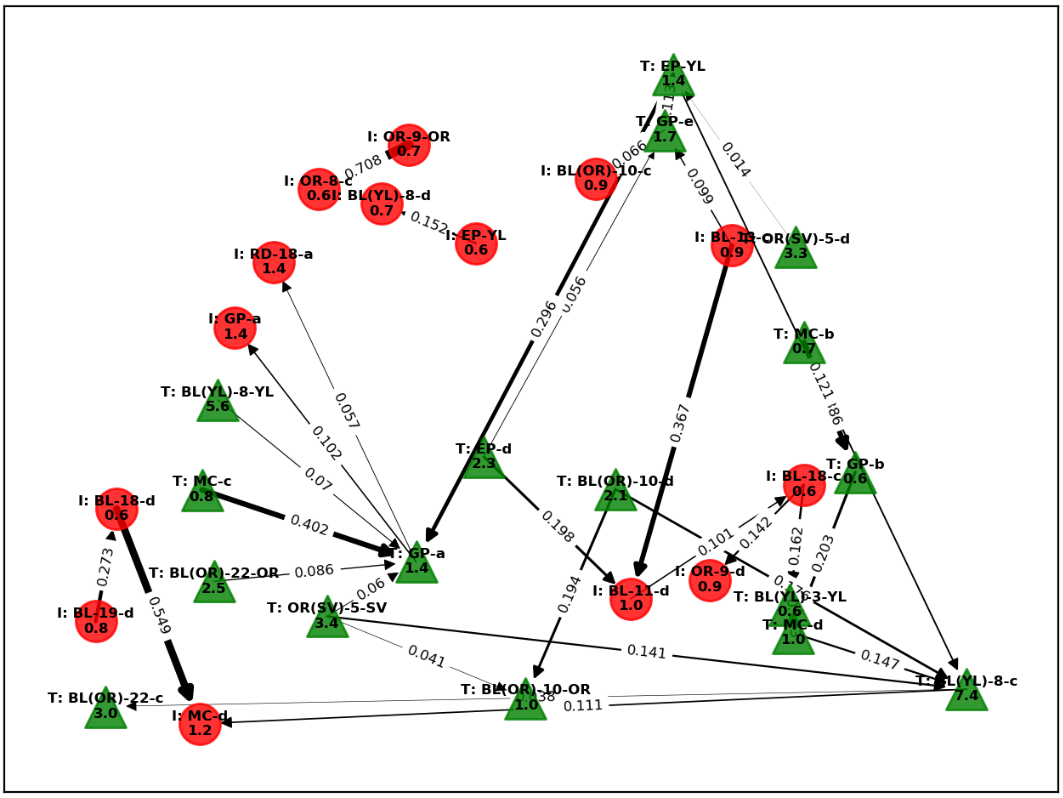


Figure 5.8: The resulting BN graph of sector 1, where round circles indicate an initial station and green triangles a transfer station. The numbers below the station codes indicate the percentage error of that node. The thickness and labels of the arcs correspond to the arc strength. How the codes of the nodes correspond to stations and directions, can be found in Appendix B.

Some stronger connections can be seen between the adjacent stations *I: OR-8-c* and *I: OR-9-OR* (strength 0.708), and *T: MC-c* and *T: GP-a* (strength 0.402), although the parts of these pairs do both have different directions compared to each other. Contrarily, *I: BL-18-d* and *I: MC-d* are also closely connected (arc strength of 0.549), and going in the same direction, but are not adjacent. Clearly, strong connections are more common between either adjacent stations or nodes going into the same direction.

5.2.3 Sector 2

In Figure 5.9 the results for sector 2 can be seen. This sector is also quite well connected, with many relatively strong connections. The BN consists of two separate parts, the first consists of connections between nodes in direction a, the second revolves around *I: RD-13-b*, which is connected to other nodes in direction b as well as *I: RD-7-a*. This last connection is rather surprising as the nodes are neither adjacent in the metro network, nor are they in the same direction. Such an odd connection might be explained through some peculiarity in the network relating to e.g. the timetable, but could also easily be a coincidence or an indication of the method not being robust.

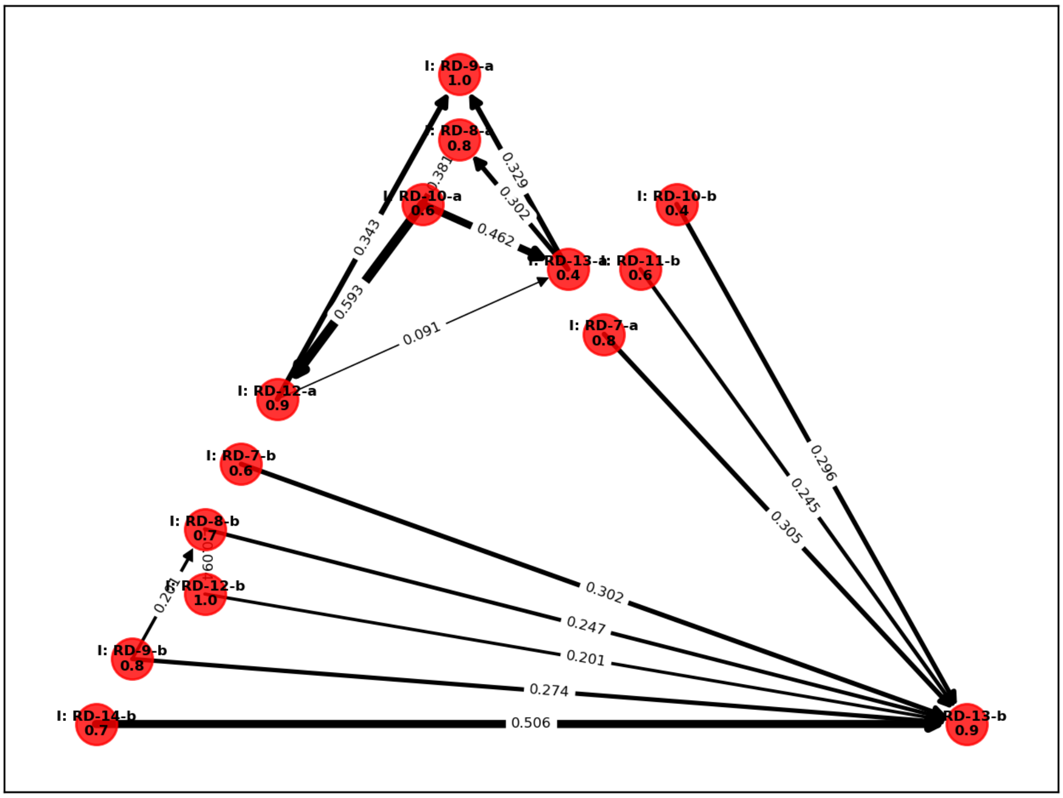


Figure 5.9: The resulting BN graph of sector 2, where round circles indicate an initial station and green triangles a transfer station. The numbers below the station codes indicate the percentage error of that node. The thickness and labels of the arcs correspond to the arc strength. How the codes of the nodes correspond to stations and directions, can be found in Appendix B.

It is also interesting to see that many nodes are connected to nodes other than the one's they are close to, e.g. *I: RD-7-b* is only connected to *I: RD-13-b*, but not to the adjacent *I:RD-8-b* or the same station in the opposite direction *I: RD-7-a*, but both these stations are also connected to *I: RD-13-b*. The fact that these connections are not logical, but logical node pairs are indirectly connected, could indicate that the BN method is not entirely accurate, and should be used in combination with expert knowledge.

Another important thing to note is the number of incoming arcs for node *I: RD-13-b*, which is eight. This means that the conditional probability table of this node will have over 1.9

million entries, since for each parent state combination (8 parents with all 5 possible states), 5 child states are possible, so there are $5 \cdot 8^5 = 1953125$ entries in the CPD. Clearly, this CPD table will not be very meaningful as there are fewer data points than entries, and the BN method cannot predict information for states that are not observed. However, the arcs connected to this node are still quite strong, and the CPD will not be used further, so this limitation should not have a big influence on the results of further calculations.

5.2.4 Sector 3

In Figure 5.10 the BN of sector 3 can be seen. Here, nodes are exclusively connected to other nodes in the same direction, and to nodes that are not necessarily adjacent but are still fairly close to each other. The strengths of the arcs are rather average, without any extremely high or low values. This is more or less what is expected in the results, although in this sector, very many nodes aren't connected at all. Especially nodes along the green/yellow line are not present in the connected Bayesian Network. why exactly only stations along the red line are connected, is not obvious. Potentially the green line is more robust against delays, meaning delays do not propagate over it, so that the stations cannot provide information on each other.

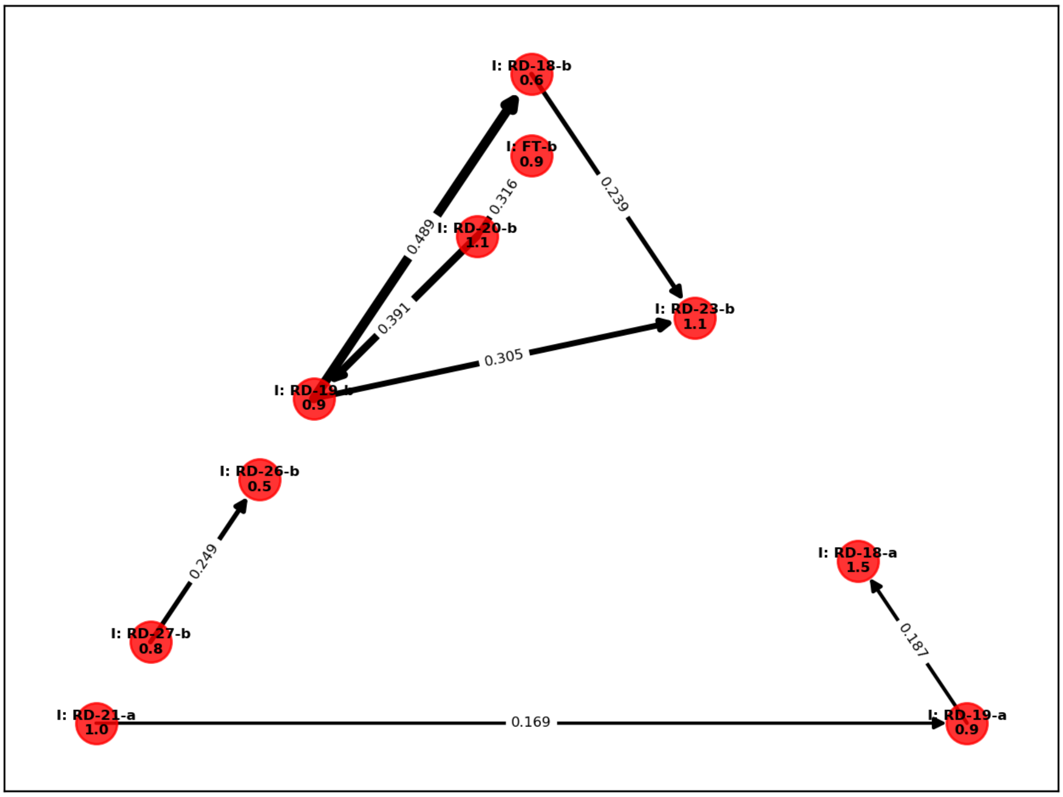


Figure 5.10: The resulting BN graph of sector 3, where round circles indicate an initial station and green triangles a transfer station. The numbers below the station codes indicate the percentage error of that node. The thickness and labels of the arcs correspond to the arc strength. How the codes of the nodes correspond to stations and directions, can be found in Appendix B.

5.2.5 Sector 4

In Figure 5.11 the results of sector 4 can be seen. One arc stands out, namely the one between *I: OR-9-OR* and *I: OR-8-c* with a strength of 0.695. This same connection is also strongly present in sector 1 (with a strength of 0.702), which contained both of the same nodes for the purposes of recombining the sectors for the calculation of the informativity indicators. This similar finding, with similar arc strengths in both cases, gives an indication that the method is robust and reliable, yielding similar results in different circumstances. As the arc strengths are very close to each other, the average of the arc strengths can be taken when creating the recombined network.

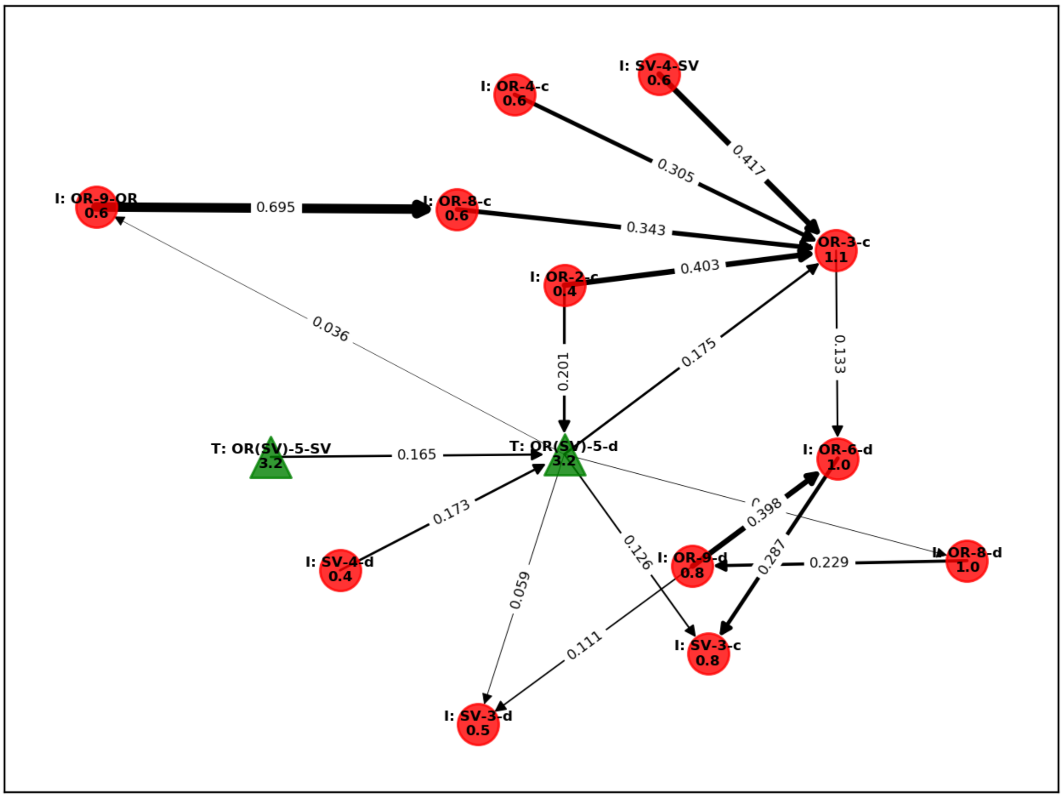


Figure 5.11: The resulting BN graph of sector 4, where round circles indicate an initial station and green triangles a transfer station. The numbers below the station codes indicate the percentage error of that node. The thickness and labels of the arcs correspond to the arc strength. How the codes of the nodes correspond to stations and directions, can be found in Appendix B.

5.2.6 Sector 5

In Figure 5.12 the results for sector 5 can be seen. Here, the arc strengths are not as high as for the previous figures, with the highest being 0.274 between *I: BL(OR)-22-c* and *T: BL(OR)-22-OR*. However, once again only nodes that are going in the same direction and are near each other are connected, where *T: BL(OR)-22-OR* is also going towards the east just like direction c, but is indicated differently to distinguish the split between the blue and orange line.

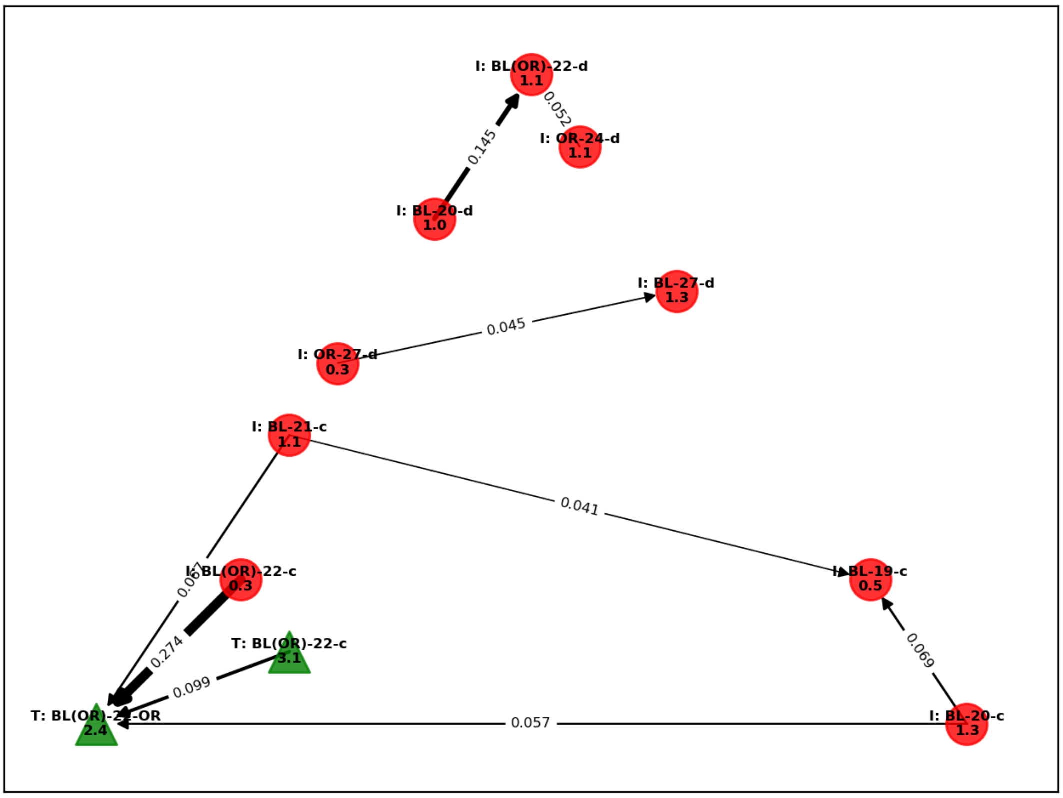


Figure 5.12: The resulting BN graph of sector 5, where round circles indicate an initial station and green triangles a transfer station. The numbers below the station codes indicate the percentage error of that node. The thickness and labels of the arcs correspond to the arc strength. How the codes of the nodes correspond to stations and directions, can be found in Appendix B.

5.2.7 Sector 6 & 7

In Figure 5.13 the results for sector 6 can be seen, and the results for sector 7 can be seen in Figure 5.14. These two BNs are very sparsely connected with only two and three arcs, and both have one arc that is not strong. Still, just as the overwhelming majority of the other results, the connected nodes are all going in the same direction, and are spatially close to each other.

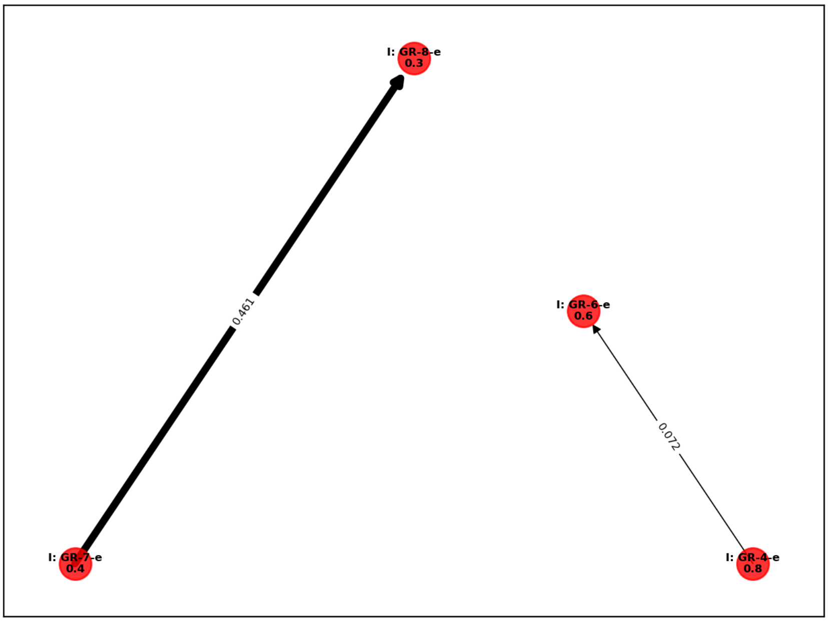


Figure 5.13: The resulting BN graph of sector 6, where round circles indicate an initial station and green triangles a transfer station. The numbers below the station codes indicate the percentage error of that node. The thickness and labels of the arcs correspond to the arc strength. How the codes of the nodes correspond to stations and directions, can be found in Appendix B.

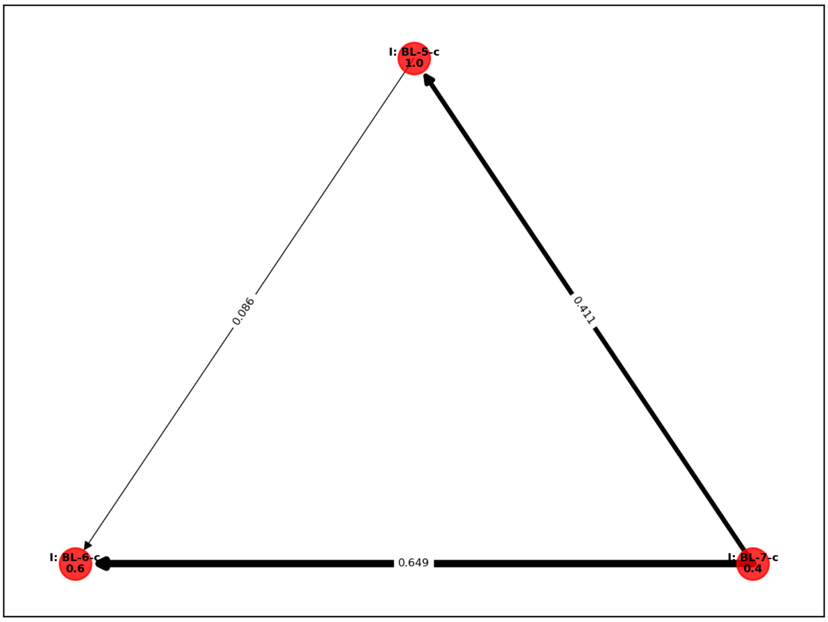


Figure 5.14: The resulting BN graph of sector 7, where round circles indicate an initial station and green triangles a transfer station. The numbers below the station codes indicate the percentage error of that node. The thickness and labels of the arcs correspond to the arc strength. How the codes of the nodes correspond to stations and directions, can be found in Appendix B.

5.2.8 Synthesis

When looking at all the results, several observations can be made. Namely that there are no nodes that can provide all information on any other node (there are no arcs of strength 1 or close to 1), meaning there will always be some unknown variability in the observed delays that can not be predicted or explained from delay observed at other stations. Furthermore, connections are almost exclusively between nodes that are going in the same direction, which is expected as delays in opposite directions are not usually related. This would only be in case of big physical infrastructure blockages such as an object on the tracks, or in case of a big disruption that affects the circulation of the rolling stock, neither of which seem to be problems for this particular network.
In Figure 5.15 the recombined BN can be seen. mapped onto the semi-geographical layout used in Section 5.1. Here, there is no differentiation between transfer and initial nodes, but each node with either a transfer or initial station connected in the BN is colored dark red, while all other nodes are light pink. Furthermore, all the arcs of the BN are mapped so that the relations can be seen in a geographical context.

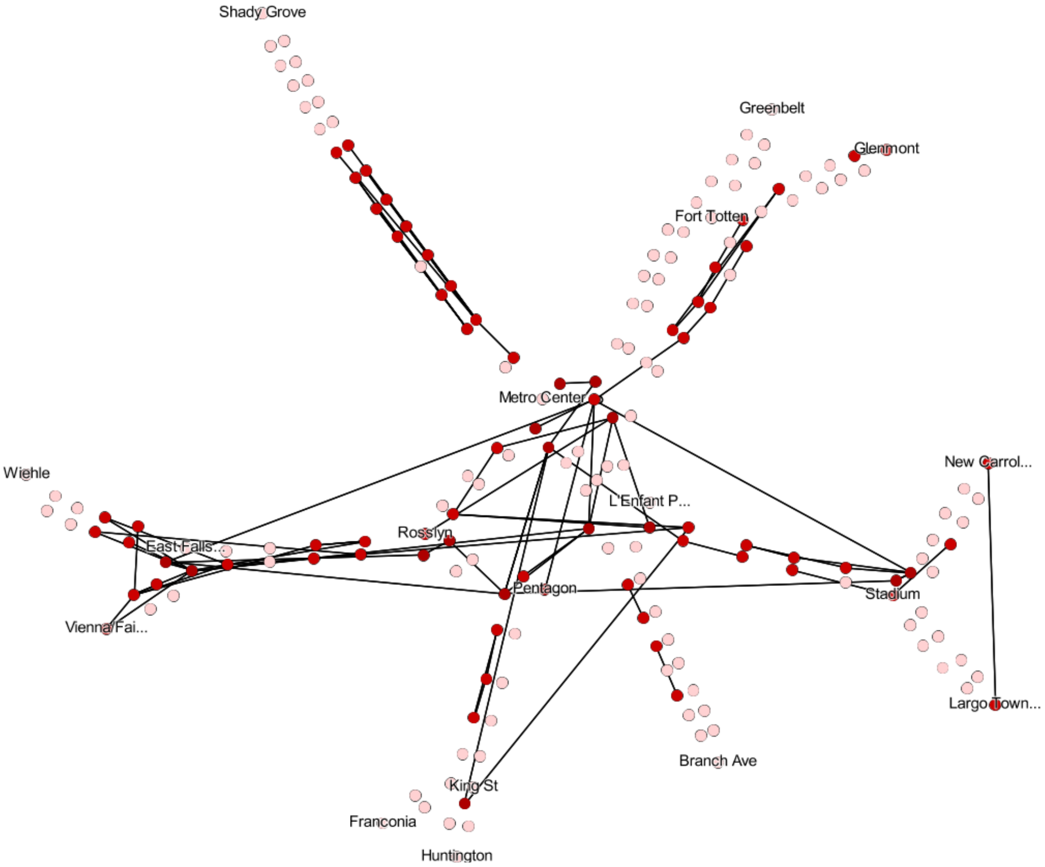


Figure 5.15: The recombined BN mapped onto the geographical map of the stations. No distinction is made between transfer and initial stations. Nodes that are not connected are a light pink, while nodes that are connected are a dark red.

It can be seen that many nodes are not connected in the BN. Only 75 of the original 186 nodes are connected, meaning for only just over 40% of the nodes in the network, some delay prediction can be made based on information from other nodes. With an arc density of 1.13 arcs per node, and an average arc strength of 0.22, the network is not very densely or strongly connected either. Whether this is normal or not cannot be said without comparing these numbers to similar statistics of other networks, something that does not exist yet.

When examining the figure in more detail, it becomes apparent that especially the green line (Greenbelt-Branch Ave) features very few relations, while the Orange/Silver section between Wiehle and Rosslyn is the most densely connected. This could mean that the green line is particularly robust against delays, meaning delays do not propagate much over this line. Conversely, the opposite would be true for the section east of Rosslyn.

Furthermore, it stands out that some of the transfer stations that are located further away from the center (East Falls Church, King Street and Stadium), have some connections over quite a long distance to transfer stations in the center (Pentagon, Metro Center and L'Enfant Plaza). Also the two end stations of the Orange (New Carrolton) and Blue/Silver (Largo Town Center) lines are connected while they are some distance apart. These connections over large distances indicate that the assumption on which the Sector method is based, namely that such long connections are unlikely, is not entirely accurate, meaning the Sector method is not optimal. Long connections (between stations on different radial lines) are not considered in the Sector method, meaning many connections that might exist are not found, which could lead to bad estimations of the informativity of nodes and underestimating the importance and informative power of some nodes.

5.2.9 Errors and accuracy

When looking at the MPEs of all the nodes, which are calculated using the 5-fold method, it can be seen that the errors for the transfer station nodes are generally higher than those for the initial station nodes. This makes sense as the transfer station nodes more often have high delays and the possible states of these stations is slightly more diverse, as can be seen in Figure 5.3. Still, all of the errors are relatively low, with the highest being only 7.4% for *T:BL(YL)-8-c*, meaning the delay at this node can still be predicted with an error of only 7.4% when the states of its parent nodes are known. This means that the model can estimate the expected delays with a high accuracy.

The RMSEs are also calculated using the 5-fold method. The training set resulted in an average RMSE of 0.0046 minutes, with no significant difference in the RMSEs of each of the different folds. The test set resulted in an average RMSE of 0.0092 minutes, again with no significant differences between the folds. The RMSEs of both the test set and training set are rather low, given that delays can range from 0 to over 5 minutes. The RMSE of the training set is clearly lower than that of the test set, as would be expected. However, the RMSE of the test set is not much higher than that of the training set, so it can be said that overfitting of the model is limited and not a concern.

Although it seems that both the MPE and the RMSE of the model are low, this does not necessarily mean that the model is an accurate representation of reality. It is possible that, due to the abundance of near-zero data points, the model is simply good at predicting these values, regardless of the delay state of the rest of the network. It is thus possible that the model is underfitted, and that the network structure is not necessarily meaningful. Whether this is the case cannot be determined for certain by only looking at the error values. Due to the specific implementation of the Bayesian Network as was done in this study, it is impossible to score the structure of the Bayesian Network, thus it is impossible to comment on the accuracy of the structure. Since it was discovered that slightly different data sets and different choices in terms of discretization method and number of bins, resulted in different network structures, it can be said that the structure as presented in this section, is not necessarily the most accurate structure. However, from observations of the different Bayesian Network structures, the main observations made in this section - namely that relations occur most often between nearby nodes and nodes going into the same direction - are still accurate, as the same observations were made for the differing structures.

5.3 Informativity indicators

After the BNs are created for all the seven sectors, the informativity indicators of all nodes can be calculated using the method as described in Section 3.3. The BNs from the sectors are combined through the overlapping nodes, although the final result is not a fully connected graph. Furthermore, many of the nodes in the graph do not have any arcs connecting them, meaning all of the informativity indicators of these nodes are automatically zero as they cannot provide any information on the rest of the network.

Another change is made to the BN before the informativity indicators are calculated, namely the direction of the arcs. As the direction assigned by the BN is not necessarily the direction of any causal relationship and the arc strength does not depend on the direction, these directions are removed entirely, and arcs are effectively made bi-directional. It is theorized that with expert knowledge, some directions could be inferred between certain pairs of nodes: those that are going in the same direction. It might be assumed that the causal flow of delay only moves upstream, since a delayed vehicle at station A might remain delayed along the rest of the line, so that all passengers boarding the vehicle will incur a delay. However, it is also possible that a delay travels downstream, as a blockage at station B could ensure passengers here are delayed, while also obstructing the free passage of vehicles at station A, ensuring passengers at station A are also delayed, although this scenario might occur less frequently. Still, the method whereby causal arc-direction is assigned in the upstream-direction was tried to see it's effect, whereby any node pair for which the nodes are not going in the same direction, or are on another line, is assigned a bi-directional arc. Eventually the correlations found between the informativity and centrality indicators are lower for this method than when using a completely undirected network, and this method is thus not deemed superior. Therefore all the following results are calculated using a completely undirected network.

The results will be discussed below per indicator (Sections 5.3.1 to 5.3.3) and finally the indicators are discussed in relation to each other in Section 5.3.4.

5.3.1 Outgoing node degree

In Figure 5.16 the resulting histogram and graph of the outgoing node degree of the connected nodes can be seen. In the graph, the transfer nodes are larger in size than the initial station nodes.

What can be seen is that only a few nodes have a relatively high degree, while most nodes only have a degree of one or two, resulting in a relatively exponential node degree distribution. This means that very few stations are directly informative for a wide range of other stations. For the most part, this makes sense as most stations in the physical network are only connected to two other stations, as is common for a predominantly radial network. So it would be expected that those stations can only give information on a limited number of stations as well.

5.3.2 Average direct informativity

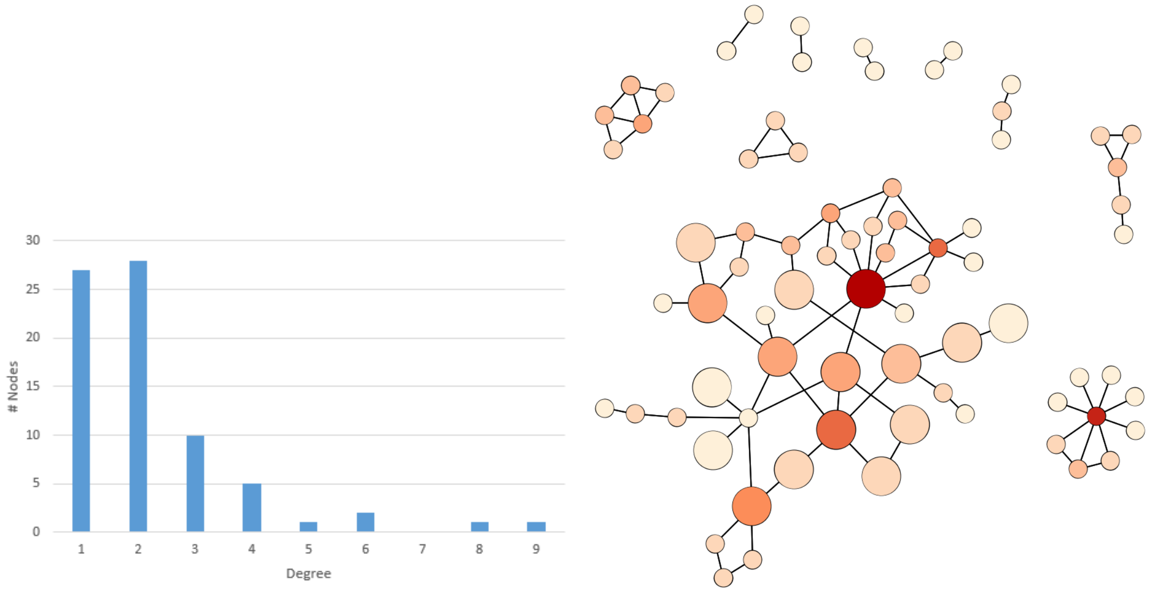
In Figure 5.17 the resulting cumulative probability distribution and graph of the average direct informativity of the connected nodes can be seen. In the graph, the transfer nodes are larger in size than the initial station nodes.
The cumulative probability distribution is nearly linear, meaning neither higher nor lower scores are more likely. However, the maximum lies at a value of 0.53, whilst theoretically a score of 1 could be possible, which would imply a complete informational relation between at least two nodes. A truly high score has a low probability of occurring in any network, which is obvious here as such a score is not observed for any node.

5.3.3 Total informativity

In Figure 5.18 the resulting cumulative probability distribution and graph of the upper bound of the total informativity of the connected nodes can be seen, while in Figure 5.19 the same information can be found for the lower bound. In the graph, the transfer nodes are larger in size than the initial station nodes.
The cumulative probability distribution is very similar for both results, where they are somewhat linear for lower values of total informativity, but the probability of having a high value is slim.
The graphs of both results are also very similar, with only a big increase in total informativity for the five nodes in the small sub-graph in the top left of the Figures 5.18b and 5.19b. This is in line with the prediction that the upper and lower bound will have similar values, which is also beneficial as this narrows down the range in which the an accurate prediction for the total informativity can be made.

5.3.4 Synthesis

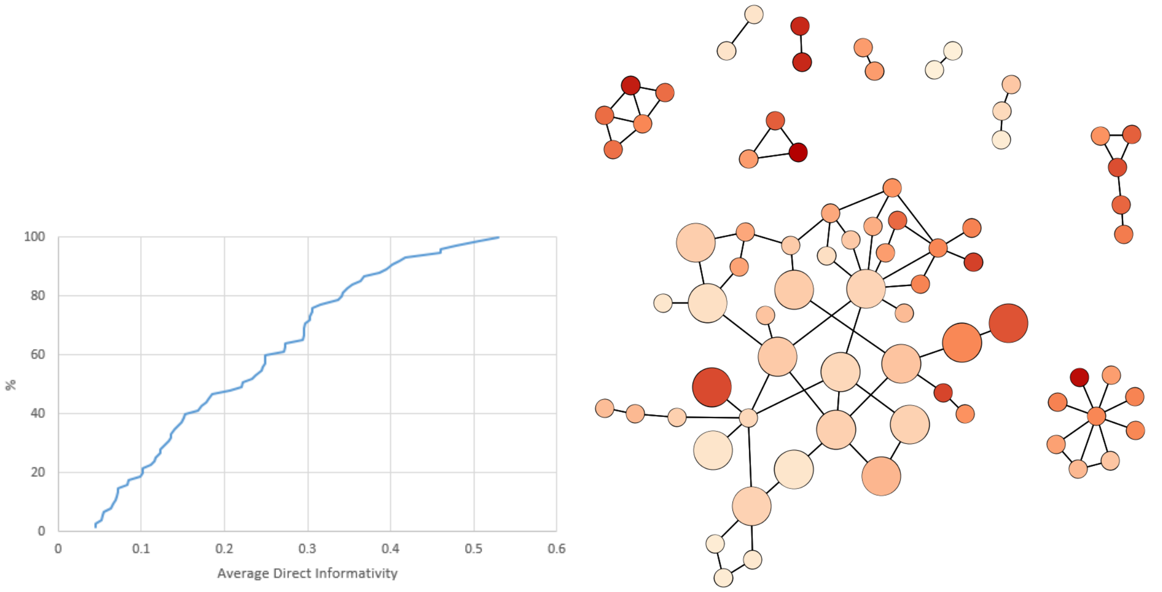
When comparing the graphs of all the informativity indicators (Figures 5.16b, 5.17b, 5.18b and 5.19b) it can be seen that especially the average direct informativity indicator stands out from the rest. Nodes along the periphery of the BN have higher values than for the other indicators, while some centrally located nodes have relatively low values. This is because the value is an average of only the surrounding arcs, and thus not quite representative of the overall informativity of the node. The node degree is a better representation of the quantity of information that can be gathered from a node, as it indicates on how many nodes information can be provided, and of course the total informativity is an even better representation. Still, the average direct informativity can still be a useful indicator, as it does not represent the overall informativity, but it can indicate the accuracy or usefulness of the information provided.



(a) A histogram of the outgoing node degree indicator for the connected nodes.

(b) A graph of the outgoing node degree indicator for the connected nodes, colored by the value. Here, transfer nodes are larger in size than initial station nodes. The maximum value is 9 and the minimum is 1.

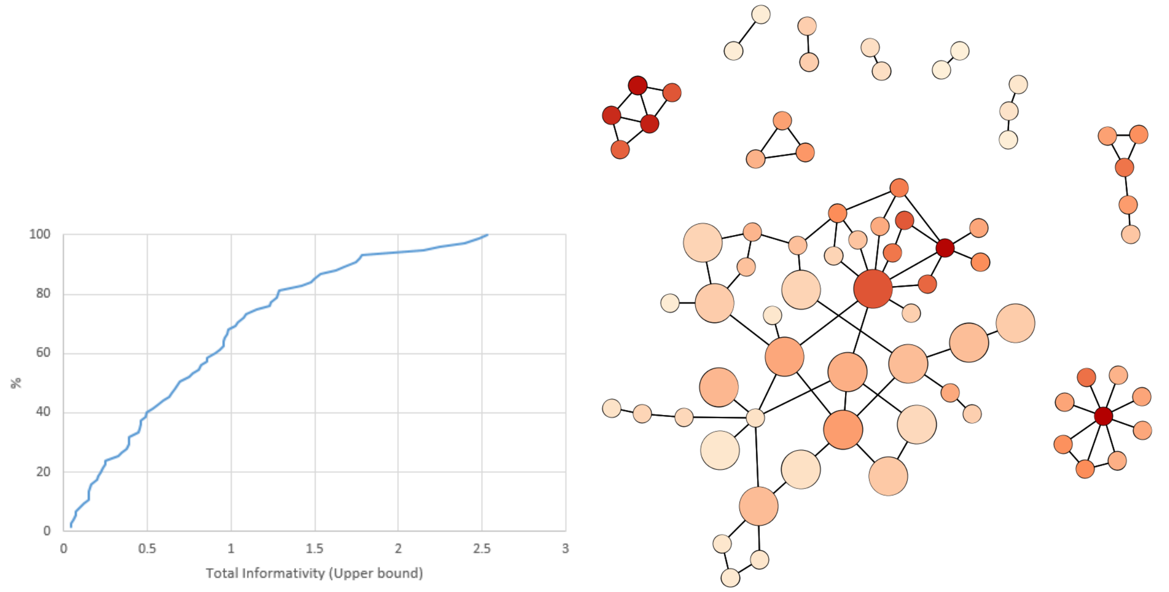
Figure 5.16: Results for the outgoing node degree indicator



(a) A histogram of the average direct informativity indicator for the connected nodes.

(b) A graph of the average direct informativity indicator for the connected nodes, colored by the value. Here, transfer nodes are larger in size than initial station nodes. The maximum value is 0.53 and the minimum is 0.045.

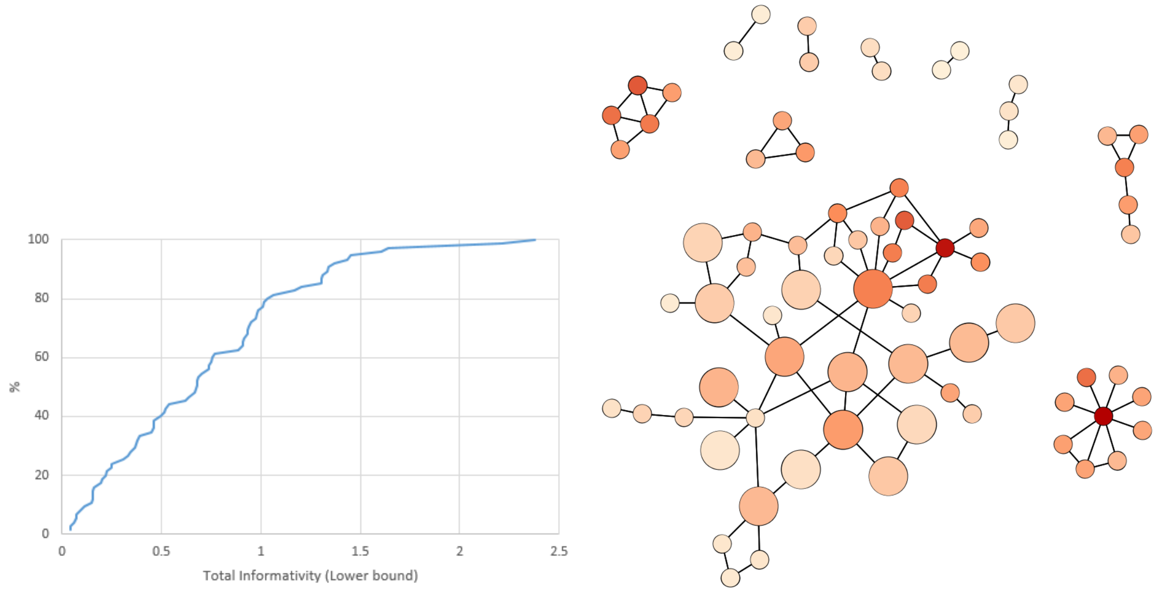
Figure 5.17: Results for the average direct informativity indicator



(a) A histogram of the upper bound of the total informativity indicator for the connected nodes.

(b) A graph of the upper bound of the total informativity indicator for the connected nodes, colored by the value. Here, transfer nodes are larger in size than initial station nodes. The maximum value is 2.53 and the minimum is 0.045.

Figure 5.18: Results for the upper bound of the total informativity indicator



(a) A histogram of the lower bound of the total informativity indicator for the connected nodes.

(b) A graph of the lower bound of the total informativity indicator for the connected nodes, colored by the value. Here, transfer nodes are larger in size than initial station nodes. The maximum value is 2.37 and the minimum is 0.045.

Figure 5.19: Results for the lower bound of the total informativity indicator

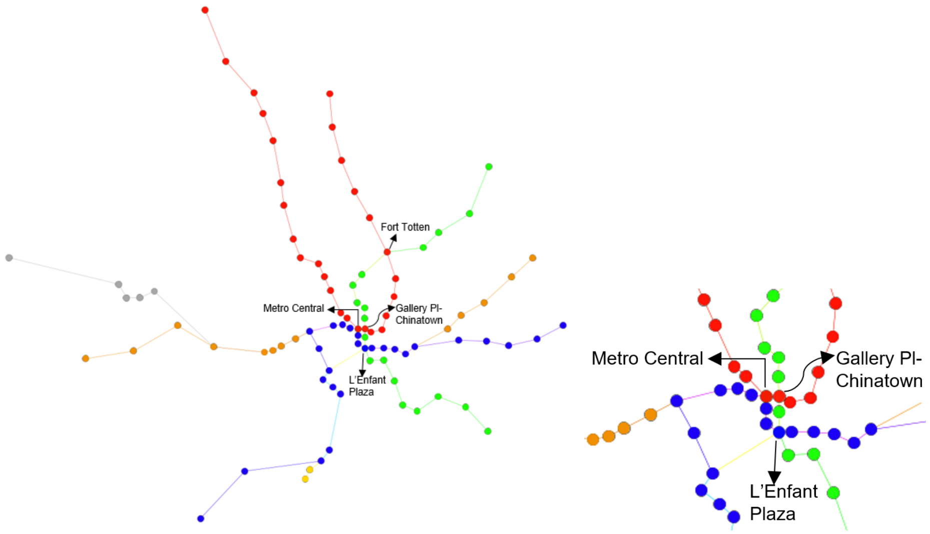
5.4 Centrality indicators

In this section the centrality indicators are presented, and a network analysis is done using these indicators. These indicators are necessary so they can be related to the informativity indicators. After this, it can also be checked which indicators and graph spaces are most relevant with respect to delay analysis, so that the third research question can be answered. Three different spaces are considered when calculating the centrality indicators: L-space, B-space and P-space. The results for these spaces will be discussed respectively in Sections 5.4.1, 5.4.2 and 5.4.3. The four stations Metro Center, Fort Totten, Gallery Place-Chinatown and L'Enfant Plaza will be discussed in more detail, as compared to the other stations, due to their unique positions as major transfer stations in the network.

5.4.1 L-space

The L-space is created in Gephi by entering all the unique nodes and edges. It is chosen to create an undirected network, as there are no node-pairs only connected in a single direction, and so the directed network would not contain any different information than the undirected network.

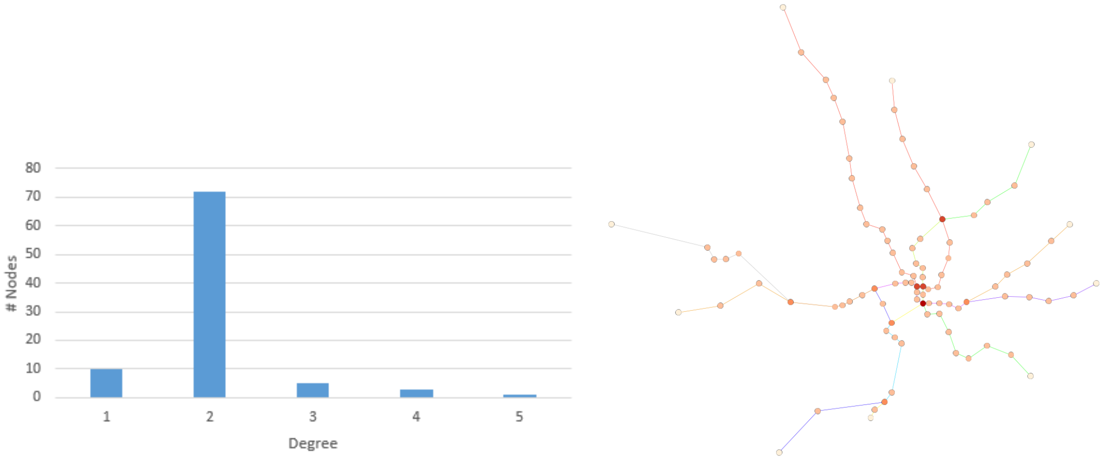
In Figure 5.20 an overview of the network in L-space can be seen, with an extra zoomed-in version of the center. Here, the nodes are colored by the line by which they are served. This is done in the order of Red - Blue - Green - Orange - Silver - Yellow, so if a node is served by both the Red line and the Blue line, it will be colored red.



(a) Total network **(b)** Zoomed in

Figure 5.20: An overview of the network in L-space. The nodes have been colored by the line by which they are served, in order of Red - Blue - Green - Orange - Silver - Yellow.

In Figure 5.21 the results for the degree centrality indicator in L-space can be seen. Clearly, the majority of the nodes have a degree of two, which is to be expected for a mainly radial network with few transfer stations, as radial stations naturally have few adjacent nodes. While the four transfer stations clearly stand out in the graph for having a higher number of connections.

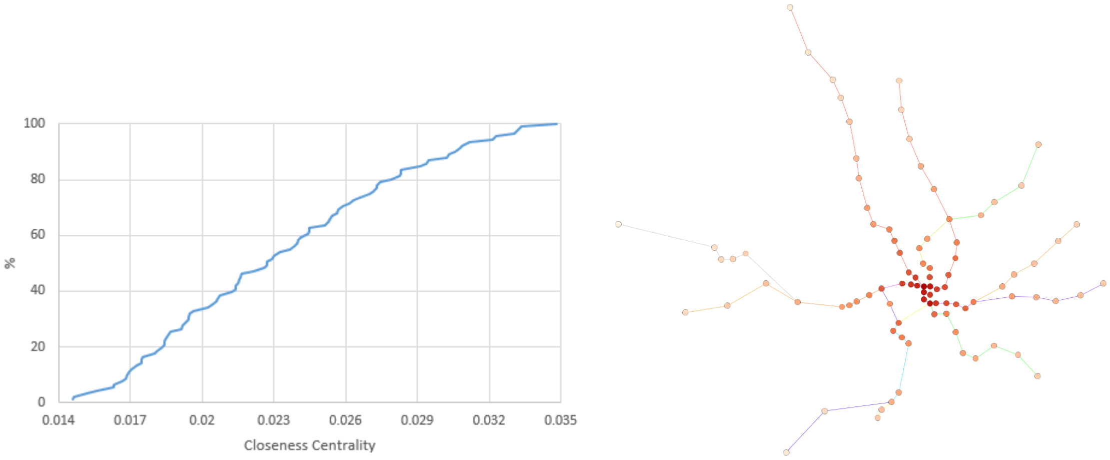


(a) Histogram

(b) Graph colored by node degree where the minimum value is 1 and the maximum is 5.

Figure 5.21: Results for the L-Space degree centrality indicator.

In Figure 5.22 the results for the closeness centrality indicator in L-space can be seen, these values indicate how close all other nodes are. Clearly, the cumulative probability distribution is fairly linear, meaning neither a high nor a low value is more likely to occur. This is to be expected as the network is fairly radial, without any large densely connected sectors, or extremely long lines.



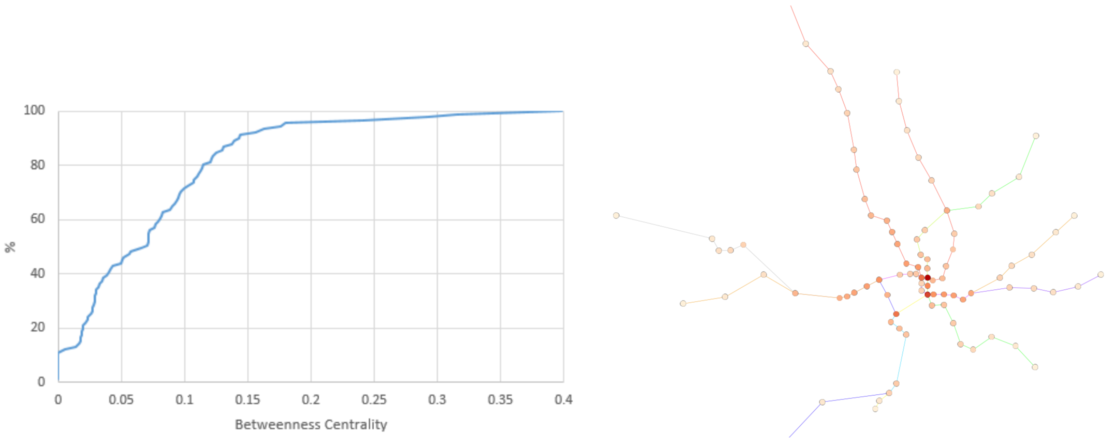
(a) Cumulative probability distribution

(b) Graph colored by closeness centrality where the minimum is 0.015 and the maximum is 0.035

Figure 5.22: Results for the L-Space closeness centrality weighted by the maximum travel time

In Figure 5.23 the results for the betweenness centrality indicator in L-space can be seen, these values indicate how many shortest paths out of all shortest paths pass through a node. Here, the shortest paths are calculated based on minimizing the travel time. However, not the entire network is weighted, along the radial lines where only one shortest path is possible, the weights are set to 1. However, in the center where multiple paths could be possible, the weights were retrieved from Washington Metropolitan Area Transit Authority (2019). Furthermore, to accurately determine the amount of shortest paths passing through each node, for each OD-pair, the shortest path is multiplied with the daily demand between the origin and destination.

From the cumulative probability distribution it can be seen that very few stations have a very high relative betweenness centrality. From the graph it is clear that especially Gallery Place-Chinatown has a high value, making it an incredibly important transfer station, and a critical node in the network.



(a) Cumulative probability distribution

(b) Graph colored by betweenness centrality where the minimum is 0 and the maximum is 0.4

Figure 5.23: The results for the L-Space Betweenness centrality indicator, where passenger counts have been taken into account during the calculation.

In Table 5.2 an overview of the centrality indicators in L-space of the four distinct stations can be seen, where the highest values have been marked red. No single station has the highest value for more than one indicator, while Fort Totten stands out for having relatively low values for both the closeness and betweenness centrality. This is to be expected as it is located much less centrally in the network than the other three distinct stations, meaning it is not as important for the functioning of the network.

	Centrality		
Station	**Degree**	**Closeness**	**Betweenness**
Metro Center	4	0.035	0.293
Fort Totten	4	0.024	0.139
Gallery Pl-Chinatown	4	0.033	0.399
L'Enfant Plaza	5	0.033	0.315

Table 5.2: Centrality indicators of the four distinct stations in the L-Space graph. The cells marked red indicate the maximum values of each indicator for the entire network.

5.4.2 B-space

The B-space is created in Gephi by entering all unique stations as nodes, and all lines as nodes. Then the edges can be created by checking which station-nodes are served by which line-nodes, and entering this information in Gephi.

In Figure 5.24 an overview of the network can be seen. Here, the line nodes are colored, where it becomes clear that the blue, orange and silver lines are the most intertwined, while the red line clearly stands apart. The unique positions of the four distinct transfer stations also becomes very obvious in this figure.

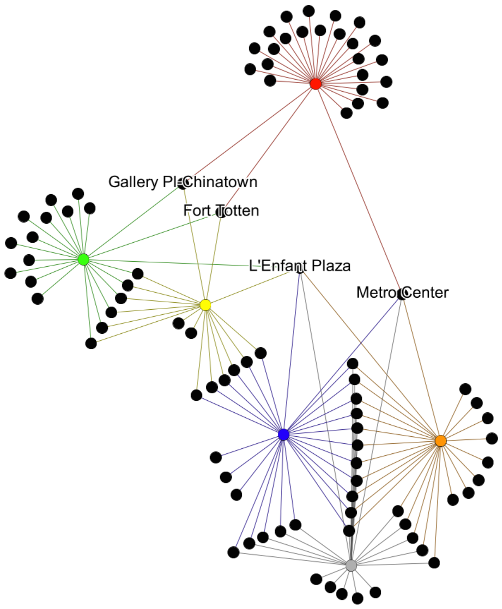


Figure 5.24: An overview of the network in B-Space

In Table 5.3a the B-space centrality indicators for the four distinct nodes can be seen, where the red cells indicate the highest value of the entire network, where the values of the line-nodes have been excluded as their meanings are entirely different.

It can be seen that Metro Center has high values, although not the top-score for the degree centrality, which indicates the number of servicing lines. However, the lines by which it is served (Red, Blue, Orange and Silver), are more distinct from each other than is the case for L'Enfant Plaza, which is served by Blue, Orange, Silver, Green and Yellow, lines that have

	Centrality		
	Deg.	Clos.	Bet.
Metro Center	4	0.417	0.241
Fort Totten	3	0.361	0.076
Gallery Pl-Ch.	3	0.361	0.076
L'Enfant Plaza	5	0.407	0.190

(a) Centrality indicators of the four distinct nodes in the B-Space graph.

Station	Centrality		
Service	Deg.	Clos.	Betw.
Red	1	0.296	0.000
Blue	1	0.296	0.000
Green	1	0.286	0.000
Orange	1	0.294	0.000
Silver	1	0.298	0.000
Yellow	1	0.279	0.000
Bl+Or+Si	3	0.338	0.004
Bl+Si	2	0.318	0.001
Or+Si	2	0.316	0.001
Bl+Ye	2	0.318	0.005
Gr+Ye	2	0.302	0.003

(b) Centrality indicators of the nodes per the lines by which they are served in the B-space graph.

Table 5.3: Centrality indicators of the B-Space graph. The cells marked red indicate the maximum values of each indicator for the entire network, excluding the line nodes. Here, Deg. is degree, Clos. is closeness and Betw. is betweenness.

much overlap. Clearly, the uniqueness of the Red line ensures that Metro Center has by far the highest Closeness and Betweenness centrality, which indicate respectively the amount of transfers needed to reach all other stations and the amount of shortest paths passing through it as a transfer station, making it an important node in the network when looking at the accessibility of the entire network.

In Table 5.3b the centrality indicators of all the other station nodes can be seen. A station falls under the label of the combination of the lines it is served by. Nodes that are served by multiple stations naturally have slightly higher values.

5.4.3 P-space

The P-space is created in Gephi by entering all the unique nodes and adding edges between all stations that are serviced by the same line.

In Figure 5.25 an overview of the network in P-space can be seen. Here, nodes are colored by the line by which they are served, in the order of Red - Blue - Green - Orange - Silver - Yellow. From the figure it becomes clear that especially the Yellow, Silver and Blue lines have very few nodes unique to their line.

In Table 5.4a the P-space centrality indicators for the four distinct nodes can be seen, where the red cells indicate the highest value of the entire network. It can be seen that Metro Center has by far the highest values for all three indicators, meaning that the Red line, together with the combination of the Blue, Orange and Silver lines, provides the most unique destinations that can be reached without transfer (as indicated by the degree centrality and closeness centrality), and that Metro Center is the most important transfer station in the network (as indicated by the betweenness centrality).

In Table 5.4b the centrality indicators of all the other station nodes can be seen. A station falls under the label of the combination of the lines it is served by. Nodes that are served by multiple stations naturally have slightly higher values.

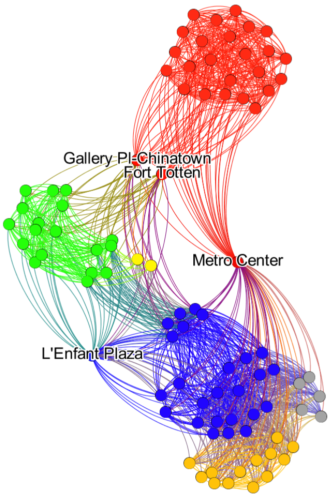


Figure 5.25: An overview of the network in P-Space

	Centrality		
Station	**Deg.**	**Clos.**	**Betw.**
Metro Center	70	0.818	0.245
Fort Totten	53	0.709	0.078
Gallery Pl-Ch.	53	0.709	0.078
L'Enfant Plaza	66	0.789	0.166

(a) Centrality indicators of the four distinct nodes in the P-Space graph.

Station	Centrality		
Service	**Deg.**	**Clos.**	**Betw.**
Red	26	0.584	0.000
Blue	26	0.584	0.000
Green	20	0.563	0.000
Orange	25	0.581	0.000
Silver	27	0.588	0.000
Yellow	16	0.549	0.000
Bl+Or+Si	44	0.662	0.004
Bl+Si	36	0.625	0.001
Or+Si	35	0.621	0.001
Bl+Ye	36	0.625	0.007
Gr+Ye	28	0.592	0.003

(b) Centrality indicators of the nodes per the lines by which they are served.

Table 5.4: Centrality indicators of the P-Space graph. The cells marked red indicate the maximum values of each indicator for the entire network. Here, Deg. is degree, Clos. is closeness and Betw. is betweenness.

5.5 Relation Between Informativity and Centrality indicators

With the results from Section 5.3 and 5.4, it is possible to create a correlation matrix between the informativity indicators and centrality indicators, where each node not connected in the BN has informativity indicator values of zero. This is done in Excel using the correlation analysis tool.

In Table 5.5 the results of the correlation analysis can be seen, with on the left side the informativity indicators and on the top the centrality indicators.

Only one negative correlation is found, namely between the upper bound of the Total Informativity and the B-space betweenness centrality. However, this correlation is so low it can be neglected, meaning all significant correlations are positive. Therefore, a higher centrality, in any way, indicates a higher informativity. This is to be expected, as more 'central' nodes would be expected to give information on more of the network.

When looking at the centrality scores, it is the betweenness centrality in the L-space which seems to be most strongly correlated with all of the informativity indicators. The other centrality indicators of the L-space also correlate significantly with the informativity indicators, while the indicators in the B-space do not correlate as strongly.

When looking at the informativity indicators, it can be seen that especially the outgoing node degree, as well as the average direct informativity, correlate better with the centrality indicators than the total informativity indicator. This latter would arguably be the best representation of a node's true informativity, meaning that this concept is harder to approximate using centrality indicators, than the indicators that only represent part of the concept (the outgoing node degree and average direct informativity).

	L-space			B-space			P-space		
	Deg.	Clos.	Betw.	Deg.	Clos.	Betw.	Deg.	Clos.	Betw.
OND	0.343	0.280	0.390	0.217	0.249	0.144	0.273	0.251	0.143
ADI	0.203	0.280	0.334	0.093	0.221	0.236	0.224	0.230	0.244
TI (u)	0.108	0.176	0.250	-0.042	0.045	0.053	0.055	0.051	0.054
TI (l)	0.154	0.214	0.292	0.020	0.107	0.099	0.117	0.113	0.102

Table 5.5: A table with the correlation between the informativity and centrality indicators. OND is the Outgoing Node Degree, ADI the Average Direct Informativity, TI (u) the upper bound of the Total Informativity, and TI (l) the lower bound of the Total Informativity. Deg. is degree centrality, Clos. is closeness centrality and Betw. is betweenness centrality. The cells with a darker color indicate a higher correlation.

Chapter 6

Conclusion

This study aimed to fill the research gap regarding the empirical knowledge on the propagation of passenger delay over a transport network, and to construct a set of indicators that could be used to analyze the network, gain information and find areas where delay mitigation measures can be applied most effectively.

In order to do so, the relations between the stations in terms of observed delay were determined through a Bayesian Network approach. From these relationships the newly defined informativity indicators were calculated. These indicators say something about the ability of a station to provide information on the state of the rest of the network. Three indicators were introduced: the outgoing node degree, which indicates how many nodes the node in question is informative on; the average direct information, which indicates the extent of information the node in question is expected to provide on any node it is connected to; and the total informativity, which indicates the total amount of information the node can provide on the state of the rest of the network.

As informativity indicators require large amounts of data to be calculated, it would be very beneficial if networks with little data available could approximate the indicators through some other means. It was expected that the informativity indicators are related to some of the topological indicators of a network, therefore a correlation analysis was done to test this hypothesis. This also results in the main research question of this study:

- **How well do topological indicators approximate informativity indicators based on a data driven approach?**

The goal of this chapter is to answer the research question, discuss the implications of this study, evaluate how this study has contributed to the scientific community, to discuss the limitations of this study and how these affect the credibility of the findings, and to evaluate which topics are interesting for further research.

The conclusions of the study are thus presented in terms of: the key findings (Section 6.1); the implications of these findings (Section 6.2); the key scientific contributions (Section 6.3; and lastly the limitations and topics for further research (Section 6.4).

6.1 Key findings

6.1.1 Findings related to the network structure with respect to delays

This study started out by analyzing the available data: the delay experienced by passengers at all stations and links differentiated per travel direction at different times throughout a year. From this analysis it was found that for this particular network, that of the Washington DC metro, there were no standout stations in terms of the amount of delay experienced, or the frequency of delay experienced. Another important finding was that most of the data consisted of near-zero values. These values are not of as much interest as higher delay values. The conclusion that can be drawn from this observation is that this type of study requires an extensive data set, so that there are enough valuable data points to use.

This data on the delay experienced by passengers was then used as input for the Bayesian Network, where the relationships between delays at different nodes - meaning stations differentiated by direction and type (initial station or transfer station) - were determined and quantified. Here it was found that relationships existed predominantly between nearby nodes and nodes going into the same direction, which was expected. Despite some instability in the Bayesian Network result due to the limited data set (which will be discussed in Section 6.4), these type of relationships were predominant in all results, meaning that although the exact relationships could not be determined with much certainty, it can be said that relationships between nearby nodes and nodes going into the same direction are most common. However, on occasion relationships over large distances (when compared to the geographical layout of the network) and between nodes going into different directions were observed as well, indicating that occasionally such unexpected relations can occur. They must thus not be neglected, although their meaning is unclear, and should be examined further.

6.1.2 Findings related to the informativity indicators

Using the information from the Bayesian Network relations, several informativity indicators could be calculated for each node. These informativity indicators were first introduced in this study, and ascribe three values to a node indicating how much information observing the state of the node in question can provide on the state of the rest of the network (total informativity, with an average upper bound value of 0.83 and lower bound value of 0.71), as well as the quality of this information (average direct informativity, with an average score of 0.226), and the amount of nodes the node in question can provide information on (outgoing node degree, with an average score of 2.23). However, only 40% of the nodes were connected in the Bayesian network, leaving 60% of all nodes with scores of 0 for all indicators. Furthermore, due to the instability of the Bayesian Network results, and other uncertainties regarding some decisions made in the process (such as the choice of discretization method), the exact values of the informativity indicators for each node, are uncertain as of yet.

It was found that the informativity indicators, as calculated during this study, did not clearly correspond with any peculiarities observed in the data analysis of the network. Meaning nodes at which higher delays or more frequent delays are observed, are not necessarily more capable of providing more or less information on the rest of the network, nor is this the case for stations that see high passenger flows. This means that many delays do not necessarily propagate at all from these hot spots, resulting in those stations not being able to provide information on the rest of the network. Why there is no excessive propagation here can be due to local circumstances such as already existing delay mitigation measures, but it might also be an important insight into how passenger delay behaves in a network in general. However, these observations are not necessarily accurate, as the informativity indicators of specific nodes might be somewhat different when using a different data set or making different choices during the process.

When the informativity indicators were compared to the centrality indicators, several significant correlations were found, in the order of 0.2 and 0.3. This validates part of the hypothesis that some centrality indicators can approximate informativity indicators. However, not all correlations were found to be as strong as expected, relative to the other correlations. Furthermore, these correlations must be treated carefully, as the informativity indicator values could turn out to be different when more data is available, or if different methods of applying the Bayesian Network (such as using expert knowledge to determine the structure) are used. Furthermore, correlations in this range are not necessarily sufficient to approximate the informativity indicators from the centrality indicators. To be able to do this, it must first be examined in more detail how these correlations differ for different networks, and the meanings of the informativity indicators and their relation to other indicators should also be determined in more detail.

Still, from the correlations that were found in this study, several observations can be made, although the significance of these observations is limited. Firstly, the indicator regarding the information contained in a single node on the rest of the network (total informativity), does not result in the highest correlations. Rather, the indicators for the accuracy of the information (average direct information) and the number of nodes the node in question can provide information on (outgoing node degree), are more strongly related to the centrality of a node. Secondly, the amount of shortest paths that pass through a node (the betweenness centrality), is closely related to the informativity of a node, more so than any other measure of the centrality of a node. Furthermore, the centrality indicators of the geographical representation of the network (L-space), are the most relevant with respect to the informativity indicators. Conversely the representation of the network that indicates which stations are served by which lines (B-space), leads to centrality indicators with the least significant representations. From this it can be concluded that the geographical representation of the network is the most relevant when analyzing or trying to estimate the informativity of a node. While each of the representations can be used to some degree when estimating the informativity of a node.

6.2 Implications

The results of this study, and its conclusions, have several potential implications for the operator of the metro network of Washington DC (WMATA), as well as for other transport network operators.

WMATA can apply the results of this study in three major ways to improve the quality of their network in terms of the delay experienced by passengers. Firstly, the informativity of stations can be used to determine where in the network it would be most fruitful to systemically measure delays. These systemic measurements in turn can be useful for, e.g. informing passengers of potential delays and alternate routes, or determining the influence of a new delay mitigation measure. Secondly, the informativity of nodes can potentially be an indication of the node's relevance with respect to delay creation or propagation (this does not necessarily mean the node experiences many delays in an absolute sense, but rather that it is the cause of delays or the cause of delay propagation), meaning stations with a high informativity could potentially be the most lucrative stations to apply delay mitigation measures. Lastly, WMATA can use the information on the relations of delay between the different stations, to try to determine the most common or most likely causes of certain systemic delays, so they can effectively apply delay mitigation measures. However, to successfully and effectively do these things, the calculations of the informativity indicators should be done again for a larger data set, for more reliable results.

The applications mentioned above are potentially also interesting for other transport network operators with the necessary input data available, meaning they can apply the method used in this study to their own data. This does not have to be limited to metro or even public transportation networks, but this method is also potentially interesting for other transport networks, such as road networks or shipping networks. However, if such extensive data is not available or unreliable, which is often the case, the results of this study could still be valuable. As the calculation of centrality indicators does not require extensive data, the correlations between the centrality and informativity indicators can be used to estimate the informativity of a node when its centrality is known. This estimation can then be used to determine the most relevant areas for systemic measurements and the for delay mitigation measures. However, before such a thing would be possible, it must first be determined exactly how the correlations translate from one network to another, by determining the correlations of more and different networks. It is hereby important to transfer correlations from similar networks in terms of type of transport network (e.g. rail-bound public transport, shipping), as well as the structure of the network (e.g. radial, grid), and possibly the size of the network.

6.3 Key contributions

One of the major scientific contributions of this study, is the used methodology. This study applied well established methods, namely the Bayesian Network method and correlation analysis, in a novel way. This was done by assigning weights to the arcs of the Bayesian Network, and using these to calculate the set of newly defined informativity indicators and correlating these with the well established centrality indicators.
This novel methodology lead to several findings. Firstly, the Bayesian Network method is a versatile method that is easily interpretable, making it an excellent choice for discovering the delay relations between stations. Combined with the method that assigns weights to the arcs of the Bayesian Network, it is an excellent basis for calculating the informativity indicators. However, the Bayesian Network does require several decisions to be made, such as on the discretization method, number of bins, and the amount of expert knowledge to incorporate. To use the method optimally, the influence of these decisions must be examined carefully. Secondly, correlating the centrality indicators with the informativity indicators lead to significant results, indicating this can lead to meaningful and useful results. These correlations can be used to approximate informativity indicators for networks with no or limited data, or to gain a better understanding of the informativity indicators, their meanings, and their possible applications. Therefore, such a correlation analysis is a useful tool to consider when creating new indicators and to further develop existing indicators.
Lastly, the most novel contribution of this study is the creation of the informativity indicators, which had not been used before. During this first study of these indicators, they were found to have some inherent meaning, as they correlate significantly with the well established centrality indicators. However, their exact meanings and possible applications are complex and could not be fully examined during this study due to certain limitations. These will be discussed in the next section, where possible solutions and other topics for further research are also evaluated.

6.4 Limitations and Further research

While this study did lead to several interesting conclusions, certain limitations must be mentioned. These discussion points have been divided into three parts: issues and possible solutions regarding the accuracy of the model (Section 6.4.1); changes or extensions of the model that could either improve the results or increase the knowledge that can be gained from them (Section 6.4.2); and possible applications of the model that need to be explored further (Section 6.4.3).

6.4.1 Accuracy and sensitivity analyses

One of the limitations of this study was the amount of data available. Although an entire year's worth of data would seem to be enough, it turned out that the results of the Bayesian Network were not stable. Meaning that for different partial data sets (as is necessary for the k-fold method of error calculation), different graphs were found in some cases in terms of which nodes were connected to each other. The Bayesian Network method, although overall very effective for this study, does require a lot of data when there are many nodes present in the graph. For this study, the amount of data needed to be larger still for stable results. Although this didn't hinder the continuation of the study, it does make the results less robust and reliable. For example, when only 80% of the data was used, correlations could differ by as much as 100% compared to when the full data set was used (e.g. a correlation between two indicators would be 0.13 for the data set of 80% and would be 0.34 for the results from the full data set), which clearly indicates unstable results. In the future, if more data becomes available, the model should be run again to obtain more reliable results, and to possibly determine the minimum size of the data set required to obtain stable results.

Furthermore, the data available to this study was given only in time slices of 30 minutes. However, from the results it is clear that some relations could exist between stations that are further apart than 30 minutes. Longer time slices could thus be more appropriate for this type of study, so that these relationships can be uncovered more accurately. However, longer time slices also have disadvantages, as a greater degree of averaging could lead to a loss of information. Therefore, different sizes of time slices should be used, and the results should be analyzed for their sensitivity, so finally an optimal time slice might be determined. Here it should also be noted that a data set over the same year, with longer time slices, will result in less individual data points, so first the issue of unstable results in the face of limited data should be resolved.

Besides these limitations regarding the input data, the processing of the data was also not entirely clear cut. For instance, the discretization of the data could have been done in several different ways. However, only one was used to obtain the final results. Although the method used was the most promising and logical, to thoroughly understand the results a sensitivity study should be done of the informativity indicators depending on the discretization method used, so the exact influence of the discretization method - which always leads to a loss of information - can be determined.

A similar case is that of the link strength calculations. For this study, the mutual information method created by Boerlage (1992) and Jitnah and Nicholson (1997) was used. However, it might be possible other, more accurate, methods exist. To determine the dependence of the informativity indicators on the method chosen, different methods should be applied, so a sensitivity analysis can be done using the different results. Furthermore, it might be possible to create a new method of using the conditional probability tables to determine the informativity indicators, which could be more fitting to this study.

6.4.2 Improvements

One way this model could be improved is by trying out the Super Node method (where some nodes are aggregated into super nodes), rather than the Sector method that was used (where the network was divided into sectors so the Bayesian Network calculations could be done separately), as well as simply calculating the Bayesian Network for the entire network at once. This wasn't done for the full network during this study, as not enough computational power was available to complete such calculations in an appropriate amount of time. However, the Sector method was based on an assumption (namely that relations between stations that are far apart, will not occur) that turned out to be unjustifiable, as several relations over large distances were observed in the Bayesian Network. Therefore, circumventing this method could greatly improve the knowledge that can be gained from the results.

Another improvement could be made by applying expert knowledge. An advantage of the Bayesian Network method is the possibility to use expert knowledge, for example to create the structure of the BN. It was tried to apply some knowledge by ascribing directions to the arcs of the Bayesian Network, but eventually it was decided to neglect this step as some of the reasoning behind the directions wasn't correct and it did not lead to superior results. However, there might exist other methods of ascribing these directions that should be tried. Furthermore, it could be interesting to create the structure of the Bayesian Network purely based on expert knowledge instead of machine learning, to see if those results would potentially be superior.

A big extension of the use of the model could also be made by incorporating the temporal element. Currently, only spatial propagation was taken into account. However, in doing so it limits the results, as it is impossible to fully separate the two components. Furthermore, temporal delay propagation is also very relevant in gaining knowledge on passenger delays and insight into how to diminish these. One possible way to incorporate the temporal element, is by comparing the states of nodes in different time slices. For example, the state of station A at time i, would be compared to that of station B at time $i + 1$. This can be extended to multiple time slices apart $(i + x)$, possibly depending on the travel times between stations, by creating a big matrix of data points dependent on both time and location. However, adding the temporal component to the problem complicates it very much, meaning it might require a completely different method.

6.4.3 Applications

Before the results of this study can be applied, it is important to determine the robustness of the correlations between the informativity and centrality indicators, especially if they were to be used for networks with limited data. This should be done by carrying out similar studies to this one, for different transport networks that have access to similar data as was used here. It can then be determined how well the correlations transfer to other networks. It is important to carry out these studies for several differently structured transport networks, such as grid networks, and networks of a slightly different nature such as train or bus networks. The current study was of a ring-radial metro network, but it cannot be known if the correlations of such a network can also be applied to slightly different networks, unless these other types of networks are also examined. Nor can it be said that the results of this study can be applied to other ring-radial metro networks before more of these networks are examined as well. Furthermore, if informativity indicators become available for more and different networks, they can be compared to one another, to see if there are similarities or distinct differences for different network structures or types of transport networks. These similarities or differences could also lead to more knowledge about the meaning of the informativity indicators, or they could be used to define certain network characteristics.

Another topic that is important to look into more, are the applications of the informativity indicators. As they have been defined and found to be meaningful in this study, it is important to find the different ways in which they can be of use. For example, they - possibly in combination with the Bayesian Network - could potentially be used to find common causes of delay. How exactly this could be done needs to be researched in more detail.

Moreover, it might be possible to calculate informativity indicators for networks other than transport networks. For instance, energy networks could potentially make use of the indicators, whereby they are not indicative of delay but of capacity used or of supply (something that is becoming more relevant with the rise of green energy which does not always deliver a constant or controllable supply). This should be researched by performing similar studies on these types of networks. It is expected that the correlations for such different networks will also be very different, but their comparisons could potentially also lead to interesting findings.

Bibliography

Barthélemy, M. (2011). Spatial networks. *Physics Reports*, 499(1-3):1–101.

Bates, J., Polak, J., Jones, P., and Cook, A. (2001). The valuation of reliability for personal travel. *Transportation Research Part E: Logistics and Transportation Review*, 37(2-3):191–229.

Berger, A., Gebhardt, A., Müller-Hannemann, M., and Ostrowski, M. (2011). Stochastic delay prediction in large train networks. In *OASIcs-OpenAccess Series in Informatics*, volume 20. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Boerlage, B. (1992). *Link strength in bayesian networks*. PhD thesis, University of British Columbia.

Cats, O. and Jenelius, E. (2014). Dynamic vulnerability analysis of public transport networks: mitigation effects of real-time information. *Networks and Spatial Economics*, 14(3-4):435–463.

Cats, O., Koppenol, G.-J., and Warnier, M. (2017). Robustness assessment of link capacity reduction for complex networks: Application for public transport systems. *Reliability Engineering & System Safety*, 167:544–553.

Cats, O., Yap, M., and Van Oort, N. (2016). Exposing the role of exposure: Public transport network risk analysis. *Transportation Research Part A: Policy and Practice*, 88:1–14.

Chen, D., Lü, L., Shang, M.-S., Zhang, Y.-C., and Zhou, T. (2012). Identifying influential nodes in complex networks. *Physica a: Statistical mechanics and its applications*, 391(4):1777–1787.

Clarke, E. J. and Barton, B. A. (2000). Entropy and mdl discretization of continuous variables for bayesian belief networks. *International Journal of Intelligent Systems*, 15(1):61–92.

Corman, F. and Kecman, P. (2018). Stochastic prediction of train delays in real-time using bayesian networks. *Transportation Research Part C: Emerging Technologies*, 95:599–615.

Cyberpoint International, LLC (2012). libpgm. URL: https://pythonhosted.org/libpgm/index.html.

Derrible, S. and Kennedy, C. (2011). Applications of graph theory and network science to transit network design. *Transport reviews*, 31(4):495–519.

Dollevoet, T., Huisman, D., Schmidt, M., and Schöbel, A. (2018). Delay propagation and delay management in transportation networks. In *Handbook of Optimization in the Railway Industry*, pages 285–317. Springer.

Ebert-Uphoff, I. (2009). Tutorial on how to measure link strengths in discrete bayesian networks.

Fayyad, U. M. and Irani, K. B. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. *Proceedings of the International Joint Conference on Uncertainty in AI*, 334(571):1022–1027.

Von Ferber, C., Holovatch, T., Holovatch, Y., and Palchykov, V. (2009). Public transport networks: empirical analysis and modeling. *The European Physical Journal B*, 68(2):261–275.

Hendren, P., Antos, J., Carney, Y., and Harcum, R. (2015). Transit travel time reliability: shifting the focus from vehicles to customers. *Transportation Research Record: Journal of the Transportation Research Board*, 2535:35–44.

Jitnah, N. and Nicholson, A. (1997). treenets: A framework for anytime evaluation of belief networks. In *Qualitative and Quantitative Practical Reasoning*, pages 350–364. Springer.

Kirchhoff, F. and Kolonko, M. (2015). Modelling delay propagation in railway networks using closed family of distributions. Technical report, Technical Report.

Kjaerulff, U. B. and Madsen, A. L. (2008). *Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis.* Springer.

Koller, D., Friedman, N., and Bach, F. (2009). *Probabilistic graphical models: principles and techniques.* MIT press.

Krishnakumari, P., Cats, O., and Van Lint, H. (2019). Day-to-day and seasonal regularity of network passenger delay for metro networks. In *Transportation Research Record.*

Laskey, K. B., Xu, N., and Chen, C.-H. (2012). Propagation of delays in the national airspace system. *arXiv preprint arXiv:1206.6859.*

Lessan, J., Fu, L., and Wen, C. (2018). A hybrid bayesian network model for predicting delays in train operations. *Computers & Industrial Engineering.*

Li-Jun, Q., Yan, L., Li-Nan, Z., and Xu, C. (2011). Evaluation of the reliability of bus service based on gps and smart card data. In *Quality and Reliability (ICQR), 2011 IEEE International Conference on*, pages 130–134. IEEE.

Lievens, R. A. (2014). Process data analysis: Using a Bayesian Network approach to model processes in the Marine Contracting practice. Master's thesis, Delft University of Technology.

Lin, J. and Ban, Y. (2013). Complex network topology of transportation systems. *Transport reviews*, 33(6):658–685.

Luttinen, J. (2016). Bayespy: variational bayesian inference in python. *The Journal of Machine Learning Research*, 17(1):1419–1424.

Malandri, C., Fonzone, A., and Cats, O. (2018). Recovery time and propagation effects of passenger transport disruptions. *Physica A: Statistical Mechanics and its Applications*, 505:7–17.

Manitz, J., Harbering, J., Schmidt, M., Kneib, T., and Schöbel, A. (2017). Source estimation for propagation processes on complex networks with an application to delays in public transportation systems. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(3):521–536.

Nicholson, A. E. and Jitnah, N. (1998). Using mutual information to determine relevance in bayesian networks. In *Pacific rim international conference on artificial intelligence*, pages 399–410. Springer.

Nielsen, O. A., Landex, O., and Frederiksen, R. D. (2009). Passenger delay models for rail networks. In *Schedule-Based Modeling of Transportation Networks*, pages 1–23. Springer.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference.* Morgan Kaufman Publishers, Inc.

Pelletier, M.-P., Trépanier, M., and Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, 19(4):557–568.

Raghothama, J., Shreenath, V. M., and Meijer, S. (2016). Analytics on public transport delays with spatial big data. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*, pages 28–33. ACM.

Raiko, T., Valpola, H., Harva, M., and Karhunen, J. (2007). Building blocks for variational bayesian learning of latent variable models. *Journal of Machine Learning Research*, 8(Jan):155–201.

Redman, L., Friman, M., Gärling, T., and Hartig, T. (2013). Quality attributes of public transport that attract car users: A research review. *Transport policy*, 25:119–127.

Rodríguez-Núñez, E. and García-Palomares, J. C. (2014). Measuring the vulnerability of public transport networks. *Journal of transport geography*, 35:50–63.

Sakauchi, T. (2011). Applying bayesian forecasting to predict new customers' heating oil demand.

Schmöcker, J.-D., Bell, M. G., and Lam, W. H. (2004). Importance of public transport. *Journal of Advanced Transportation*, 38(1):1–4.

Sun, L., Huang, Y., Chen, Y., and Yao, L. (2018). Vulnerability assessment of urban rail transit based on multi-static weighted method in beijing, china. *Transportation Research Part A: Policy and Practice*, 108:12–24.

Sun, L. and Jin, J. G. (2015). Modeling temporal flow assignment in metro networks using smart card data. In *Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on*, pages 836–841. IEEE.

Tahmasseby, S., van Oort, N., and van Nes, R. (2008). The role of infrastructures on public transport service reliability. In *2008 First International Conference on Infrastructure Systems and Services: Building Networks for a Brighter Future (INFRA)*, pages 1–5. IEEE.

Waller, D. L. (2003). *Operations management: a supply chain approach.* Cengage Learning Business Press.

Wang, R. and Work, D. B. (2015). Data driven approaches for passenger train delay estimation. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pages 535–540. IEEE.

Washington D.C. (2019). Guide to Union Station in Washington D.C.

Washington Metropolitan Area Transit Authority (2017). System map.

Washington Metropolitan Area Transit Authority (2019). Trip planner.

WMATA (2017). Metro facts 2017.

Xu, N., Laskey, K. B., Chen, C.-H., Williams, S. C., and Sherry, L. (2007). Bayesian network analysis of flight delays. In *Transportation Research Board 86th Annual Meeting, Washington, DC.*

Yaghini, M., Khoshraftar, M. M., and Seyedabadi, M. (2013). Railway passenger train delay prediction via neural network model. *Journal of advanced transportation*, 47(3):355–368.

Zilko, A. A., Kurowicka, D., and Goverde, R. M. (2016). Modeling railway disruption lengths with copula bayesian networks. *Transportation Research Part C: Emerging Technologies*, 68:350–368.

Zuk, O., Margel, S., and Domany, E. (2012). On the number of samples needed to learn the correct structure of a bayesian network. *arXiv preprint arXiv:1206.6862.*

Appendix A

Software

Several applications and libraries for programming languages exist that allow for the learning of Bayesian Networks. As the author was most familiar with Python, and as most libraries and applications for Python are open source, it was chosen to work in Python. Here the libpgm library was used, a library for Python 2.7 (Cyberpoint International, LLC, 2012). This library was chosen as it was accessible, easy to learn and had the required functionality. The libpgm library is capable of doing several things, such as sampling, learning and inference for both discrete and Gaussian Bayesian Networks. Not all of the capabilities were used in this study, so only the ones relevant for the implementation of the model described in 3.1 are discussed here in terms of functions, and input and output structures: learning the structure of a discrete BN (Section A.1), and learning the conditional probability distributions of a discrete BN (Section A.2).

A.1 Structure learning

The structure learning can be done with the function, *discrete_constraint_estimatestruct (data, pvalparam = 0.05, indegree=1)*, which is part of the *PGMlearner* class. It has one required input (the data-set), as well as two optional input parameters (the p-value and the in-degree). The output of the function is an instance of the class *GraphSkeleton*, which contains the directed, acyclic graph, the structure of the BN, which will be discussed in more detail below (Cyberpoint International, LLC, 2012).

Input

The required input of the function, the data-set (called *data*), must have the structure of an array of dictionaries, where the variable name is followed by the delay, as ('variable name' : delay). It is important to note that all variable names must be unique (Cyberpoint International, LLC, 2012). Every dictionary in the array represents a new time slice, and dependencies are only checked within each time slice/dictionary.

The first optional input parameter is the p-value (called *pvalparam*), if it is not given, a value

of 0.05 is assumed (Cyberpoint International, LLC, 2012).
The second optional input parameter is the in-degree (called *indegree*), if it is not given, a value of 1 is assumed (Cyberpoint International, LLC, 2012).

Output

As stated, the output of the function is an instance of the *GraphSkeleton* class, which contains information on the DAG of the BN. Relevant are its two attributes: *V*, a list of all vertices in the graph; and *E*, a list of lists containing information on the edges in the graph, where the sub-lists have a structure of ['origin', 'destination'] (Cyberpoint International, LLC, 2012).

A.2 CPD learning

The CPD learning can be done with the function *discrete_mle_estimateparams (graphskeleton,data)*, which is a function of the class *pgmlearner* (Cyberpoint International, LLC, 2012). It has two required inputs, the DAG of the BN (called *graphskeleton*), which has the same structure as the output of the function discussed in Section A.1. And the data-set (called *data*), which has the same structure as the data-set as discussed in Section A.1. The output of the function is an instance of the class *DiscreteBayesianNetwork*, which will be discussed in more detail below.

Output

The output of the function is an instance of the class *DiscreteBayesianNetwork*. This class contains three sets of information: *V*, a list of all the vertices in the BN; *E*, a list of lists containing information on all the edges in the BN, where the sub-lists have a structure of ['origin', 'destination']; and *Vdata*, a dictionary of dictionaries, containing information on all the variables, in the form of ('variable name' : information). The information dictionary contains 5 aspects (Cyberpoint International, LLC, 2012):

- *numoutcomes*: a single integer value indicating the number of possible states this variable can have;
- *vals*: a list of the different states the variable can have, in no particular order;
- *parents*: a list of the names of the parents of the variable in no particular order, the list is empty if the variable has no parents;
- *children*: a list of the names of the children of the variable in no particular order, the list is empty if the variable has no children;
- *cprob*: a dictionary with the conditional probability tables in the form of '['parent 1, value 1',...,'parent n, value 1']' : [probability of value 1,...,probability of value n],...,'['parent 1, value n',...,'parent n, value n']' : [probability of value 1,...,probability of value n], where the parents and values are in the same order as in the lists at the items 'parents' and 'vals' respectively.

Appendix B

Station Codes

In this section it will be explained how the nodes were coded. This was done in order to quickly see relations in the BN, such as between close nodes and nodes going into the same direction.

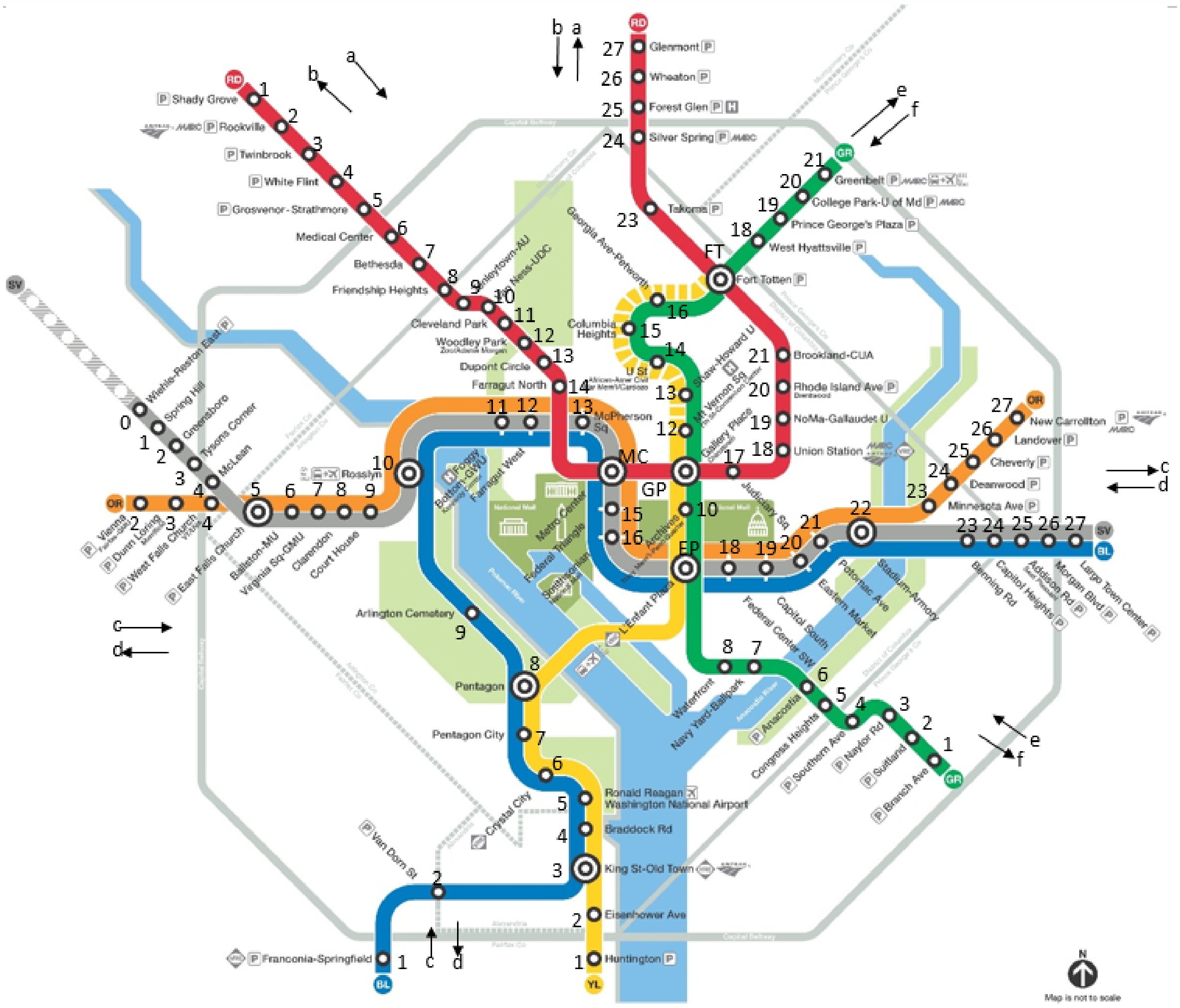


Figure B.1: Numbered network with directions indicated at end and start of each line.

The four distinct stations Metro Center (MC), L'Enfant Plaza (EP), Gallery Place-Chinatown (GP) and Fort Totten (FT) were given special codes, as stated between brackets behind their names. All other stations were given a code combination of the line color, and a number. The numbers can be seen in Figure B.1. The line color code is as follows, stations along the red line get code RD, stations along the blue line get a code BL, stations along the green line get code GR, stations along the orange line (but not the blue line) get code OR, stations along the silver (but neither the orange nor the blue line) get the code SI and stations along the yellow line (but neither the blue nor green line) get code YL. The station where a line merges, will get an extra color code. So for example Rosslyn will have the code *BL(OR)-10.* All nodes are also distinguishable by direction and must thus be ascribed a direction code. The codes per direction can also be seen in Figure B.1, where they are indicated at the beginning and ending of each line. The blue, orange and silver lines share most of the direction, and thus share the c/d direction specifiers. It should be noted that the yellow line follows both the green and blue line, which have different specifiers, so the stations along this line will follow the specifier of the other line. The first two yellow stations will use the specifier c/d.

At the place where lines split, the two regular specifiers are not enough, as there is a third direction. In these cases, the two main specifiers will be used for the direction of the major line color, while the third direction will have the specifier of the minor line color. So the node Rosslyn in the direction of Court House will have the code *BL(OR)-10-OR*, while Rosslyn in the direction of Pentagon will have the code *BL(OR)-10-d*, and L'Enfant Plaza in the direction of Pentagon will have the code *EP-YL*.

Lastly, every code will be prefixed by either an *I:* denoting it is an initial station or a *T:* denoting it is a transfer station.

Appendix C

Excluded Days

A year's worth of data was originally available, from September 2017 til August 2018. However, not all days during this year adhere to the expected morning/evening peak demand pattern and were thus excluded from the data set. This was done as it is expected those days could potentially have different delay relations, and they might thus muddle the results.
The days excluded are all weekends, as well as the dates listed in Table C.1, where also the reason for exclusion is given, this includes things such as holidays and irregular demand patterns without clear reason. These latter cases can fall into the category of maintenance that severely changed the timetables and thus demand, or some fault in the data gathering system.

Date(dd-mm-yyyy)	Reason
04-09-2017	Labor Day
23-10-2017	No afternoon data
23-11-2017	Thanksgiving
24-11-2017	Thanksgiving
08-12-2017	No afternoon data
13-12-2017	No morning data
25-12-2017	Christmas
01-01-2018	New Year's
15-01-2018	Martin Luther King Jr. day
19-02-2018	President's day
02-03-2018	Very little data
09-03-2018	No data
21-03-2018	Very little data
26-04-2018	No data
19-06-2018	No data
04-07-2018	Independence day
24-08-2018	No morning data

Table C.1: A table with the dates that were excluded and the reason why

Appendix D

Paper

INDICATORS OF SPATIAL PASSENGER DELAY PROPAGATION AND THEIR RELATION TO TOPOLOGICAL INDICATORS

Anne M. Hijner
Department
Delft University of Technology
Delft, the Netherlands
a.m.hijner@student.tudelft.nl

September 9, 2019

ABSTRACT

In order for public transportation to remain an attractive travel option, its reliability must be improved. To enable this, extensive knowledge on the passenger delay phenomenon is necessary. This study explores the possibility of using empirical data of passenger delay to determine the relationships between stations through a Bayesian Network approach. This approach is able to uncover any dependencies, regardless of known or unknown causes. The results from the Bayesian Network can be used to calculate a set of newly defined informativity indicators. These indicators give information on a station's capacity of providing information on the delay state of the rest of the network. Among the possible applications of these indicators are: providing accurate information to passengers and the operator on delays, and aiding in determining the most effective areas for delay mitigation measures. To increase the usefulness of these indicators, they are compared to centrality indicators, so that even for networks with little available data, the informativity indicators can still be approximated. It was found that the centrality and informativity indicators correlate to some extent, meaning the informativity indicators could be very meaningful and relevant in further gaining knowledge on and improving public transport reliability.

***K*eywords** Public transport · Reliability · Bayesian Network · Correlations · Informativity Indicators

1 Introduction

An important aspect of the quality of public transport is reliability (Schmöcker et al., 2004; Bates et al., 2001; Redman et al., 2013). Improving reliability can be done in several ways, e.g. by decreasing the frequency of delay, mitigating delays so the severity diminishes, or providing accurate information and travel time predictions.

To effectively improve reliability, it must be known how delays behave in the network. Some research has been done into this (Berger et al., 2011; Kirchhoff and Kolonko, 2015), but most of this previous work has approached the problem from the perspective of the operator of the network, rather than the passenger. Although complications with transfers or vehicle capacity constraints are not of direct importance to the operator, they can significantly influence the experience of a passenger, and should thus be taken into account (Nielsen et al., 2009). Moreover, these studies and some vulnerability studies that do approach the problem from the passenger's perspective (Rodríguez-Núñez and García-Palomares, 2014; Malandri et al., 2018), are all either based on theoretical models or simulations. However, not enough knowledge exists on passenger delay to accurately construct propagation models. Therefore, it is necessary to gain more insight into the issue, which can best be done by using an empirical approach, such as the Bayesian Network method. This is a data-driven method, which doesn't require extensive knowledge about the underlying phenomenon of passenger delay propagation, while still being capable of uncovering prior unexpected relationships and using a holistic view to consider all possible dependencies (Kjaerulff and Madsen, 2008; Koller et al., 2009; Zilko et al., 2016; Lessan et al., 2018).

Data on passenger delays at different stations can be used by the Bayesian Network method to determine the relationships with respect to delay between the different stations in the network. This would establish how and which station's state (meaning the passenger delay observed there) can provide information on the state of other stations. These relations are a way of representing the propagation of passenger delay, as such relations will only exist if stations experience delays from the same cause and at approximately the same time, meaning those delays have propagated. This knowledge on the relations between stations can then be condensed into a set of indicators, hereafter called informativity indicators. These indicators give information on a station's capability of providing information on the delay state of the rest of the network. They contain valuable information that can be used to determine critical areas for delay mitigation measures, and to provide information to both the operator and the passenger.

Such an empirical approach requires large amounts of data, something that is not available for many networks due to various reasons. Therefore, it must be established how the informativity indicators relate to indicators that can be calculated without requiring much data, such as centrality indicators. It is likely that these two sets of indicators are in some way related as it is only natural that stations which are centrally located or served by many lines (something that can be expressed by topological indicators), are more informative over the state of the network (expressed in the informativity indicators), than stations at the periphery of the network, served by a single line. This is due to the fact that a station is likely to experience the same delays, from the same cause, as other nearby stations and stations that it is directly connected to. Therefore a well connected station should be able to provide information on more other stations than one that is not well connected. The goal of this paper is thus to examine how to calculate the informativity indicators, and to uncover whether they are related to centrality indicators and how strongly.

The rest of this paper is structured as follows: the methodology is discussed in Section 2; then the application of this methodology is described in Section 3; afterwards the results are discussed in Section 4; finally, the main conclusions and limitations of this study are given in Section 5.

2 Methodology

In this section, the methodology of this study is discussed. This consists of the method for determining the relations between the stations with regards to delay (Section 2.1), the definition of the informativity indicators (Section 2.2), and how these informativity indicators will be compared to the centrality indicators (Section 2.3). The order of the necessary steps involved in completing the study, including their input information, is shown in Figure 1.

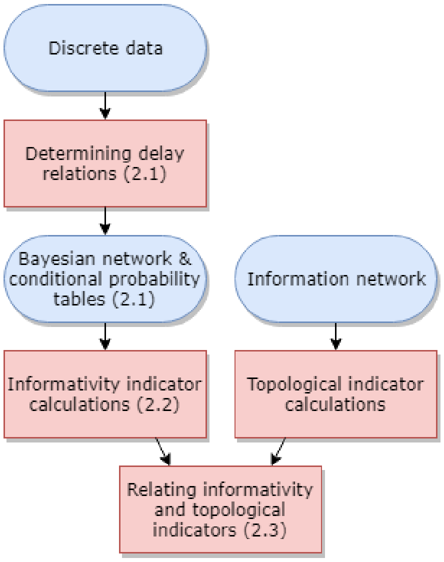

Discrete data
Determining delay relations (2.1)
Bayesian network & conditional probability tables (2.1)
Information network
Informativity indicator calculations (2.2)
Topological indicator calculations
Relating informativity and topological indicators (2.3)

Figure 1: The methodology applied, where the round blue boxes indicate input/output and the red boxes indicate processes. The numbers in the boxes correspond to the sections in which that particular item is discussed.

2.1 Determining Delay Relations

As mentioned in the Introduction, the Bayesian Network method is the most promising method for determining delay relations. A Bayesian Network (BN) is by definition a Directed Acyclic Graph (DAG) - $G(N, A)$ where N are the nodes and A are the arcs - which represents the dependencies of its variables on each other graphically. Here, a node that has an arc pointed to another node is called a parent, and the node the arc is pointed towards is called the child. A BN also contains information on the amount of dependence of the variables, by means of conditional probability tables of a variable's state, where the probability of a state occurring is dependent on the observed state of the parents (Kjaerulff and Madsen, 2008; Koller et al., 2009; Zilko et al., 2016). These conditional probability tables are learned through the input data.

The nodes of the BN graph are the variables, in this case the nodes are the initial stations and the transfer stations. The arcs of the graph represent the flow of influence between the nodes, and are unrelated to the physical connections of the nodes. These arcs are also learned from the data, whereby every pair of nodes is checked to see if they are dependent on each other with a certain significance. The nodes not connected by an arc are assumed to be conditionally independent. The arcs do not have any label or weight, although in certain cases this could be very useful. Therefore a method to ascribe weights to the arcs is described in the following section.

2.1.1 Labelling the arcs of the Bayesian Network

Several methods exist that aim to give a meaningful label to the arcs of Bayesian Networks, where the label of an arc should be an indication of the connection strength between the two nodes. The most common approaches are those based on mutual information (MI) (Pearl, 1988; Nicholson and Jitnah, 1998; Ebert-Uphoff, 2009; Kjaerulff and Madsen, 2008), which quantifies how much information can be gained on one node, if another node is observed. This concept can be extended to include the possibility of a node having multiple parents, so that inter-dependencies of the parent nodes are taken into account when determining the arc strength (Ebert-Uphoff, 2009; Nicholson and Jitnah, 1998). This can be described as

$$LS(X \to Y) = \sum_k P_{pr}(Z = k) \sum_i P_{pr}(X = i) \sum_j P(Y = j|X = i, Z = k) log_e\left(\frac{P(Y = j|X = i, Z = k)}{P_{pr}(Y = j|Z = k)}\right) \quad (1)$$

Here, $LS(X \to Y)$ is the link strength, meaning the strength of the dependency, of variable X on variable Y, independent of the value of the set of other parents of Y (which is the set Z); $P(Y = j|X = i, Z = k)$ is the probability of finding variable Y in state j, given that variable X is in state i and Z is in state k; $P_{pr}(Z = k)$ is the prior probability of finding variable Z in state k.

It should be noted that the link strength is independent of the direction of the arc, thus if the arc were switched the link strength value would still be valid.

2.2 Informativity Indicators

From the output of the BN (the structure of the relationships between the different nodes, as well as the strengths of these relationships), the informativity of the nodes should be determined. From extensive research, it became apparent that such indicators have not been used before, and are first used in this study. Therefore, three novel indicators are suggested: outgoing node degree, average direct informativity and total informativity. These will be discussed respectively in Sections 2.2.1, 2.2.2 and 2.2.3.

2.2.1 Outgoing Node Degree

The first proposed informativity indicator is the outgoing node degree. This indicator describes how many nodes a certain node can give direct information on, and can be calculated by

$$\text{OND}_x = \sum_{i \in N} a_{(x,i)} \quad (2)$$

where OND_x is the outgoing node degree of node $x \in N$; N is the set of all nodes; and $a_{(x,i)}$ is the arc between node x and node i.

2.2.2 Average Direct Informativity

The average direct informativity is the second indicator proposed. As opposed to the outgoing node degree it does not describe how many nodes information can be provided on, but rather the extent of the information. Due to this indicator being an average, it purely describes how informative this node is expected to be on any node it is connected to. The calculation can be described as

$$\text{ADI}_x = \frac{1}{\text{OND}_x} \sum_{(x,i) \in A} w_{(x,i)} \quad (3)$$

where ADI_x is the average direct influence of node $x \in N$; (x, i) is the arc from node x to any other node $i \in N$; $w_{(x,i)}$ is the weight of this arc; A is the set of all arcs in the BN; and OND_x the outgoing node degree of node x.

2.2.3 Total Informativity

The two above mentioned indicators only give an indication of the amount of information that can be provided for the immediate children of a node. However, a node can also indirectly provide information on its further descendants. This can be accounted for by multiplying the arc weights along the path to a descendant. When multiple possible paths exist, two things can be done: the most informative path can be taken into account, or all paths can be taken into account as different paths can provide novel information. In the former case there is the possibility of underestimating the informativity as not all informative paths are taken into account, while in the latter it can be overestimated as some paths could have overlap. Therefore, these two methods can be used to find the lower and upper bound of the informativity on a descendant.

In order to then determine the total informativity, meaning the total information a node can provide on the delay state of the rest of the network, the informativity on each descendant in the network should be added. The upper bound of this indicator can be described as

$$\mathrm{TI}_x^u = \sum_{y \in N} \sum_{\delta_{x,y,a} \in \Delta_{x,y}} \prod_{(i,j) \in \delta_{x,y,a}} w_{(i,j)}, \tag{4}$$

where TI_x^u is the upper bound of the total informativity of node x; $w_{(i,j)}$ is the weight of the arc from node i to node j; $\Delta_{x,y}$ is the set of all paths from node x to node y; and $\delta_{x,y,a}$ is the set of all arcs that form path a.

The lower bound of this indicator can be described as

$$\mathrm{TI}_x^l = \sum_{y \in N} \max_{\delta_{x,y,a} \in \Delta_{x,y}} \prod_{(i,j) \in \delta_{x,y,a}} w_{(i,j)}, \tag{5}$$

where TI_x^l is the lower bound of the total informativity of node x.

2.3 Relating Topological and Informativity Indicators

When little or no delay data is available, it might be useful to approximate informativity indicators by other means. One such possibility is using topological indicators, where especially the centrality indicators are of interest as they, just like informativity indicators, are specific to unique stations. It is expected that centrality indicators could approximate informativity indicators to some extent, whereby it is also important to consider the graph space in which the network is represented. Here, the L-, B- and P-space are considered relevant due to their unique representations, where stations are always present as nodes (Derrible and Kennedy, 2011; Von Ferber et al., 2009).

For example, in L-space, the indicator for the degree centrality indicates the number of adjacent nodes. This can logically correspond to the number of nodes information can be provided on (the outgoing node degree informativity indicator). While in P-space the same indicator evaluates how many other nodes can be reached without a transfer, which can correspond to the total informativity as delay could directly propagate to all these other stations. The betweenness centrality in all three spaces could also be relevant, as they all indicate in some way whether a station is a transfer station and how much transfers take place there. Such a station could be an important factor in delay propagating from one line in the network to other lines, possibly increasing its total informativity.

To evaluate whether the centrality indicators can approximate the informativity indicators, and to what extent, a correlation analysis can be done. This way, besides these expected correlations, other correlations could be observed as well, some of which for similar reasons as mentioned above. But potentially also for different reasons, as not all passenger delay propagation means are understood extensively. The approach taken in this study has the unique advantage to find these unexpected correlations as well.

3 Application

In this section the application of the model will be discussed shortly. This consists of information about the data used, in Section 3.1; as well as the explanation of some changes to this data and configurations of this data necessary for getting results, in Section 3.2.

3.1 Case Study: Washington DC metro

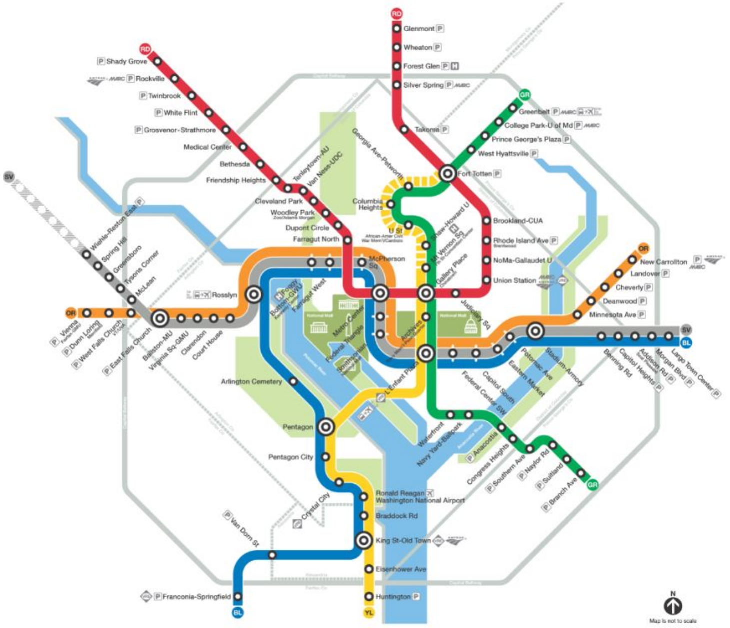


Figure 2: Map of the Washington DC metro network (Washington Metropolitan Area Transit Authority, 2017)

The data used in this study was provided by the Washington Metropolitan Area Transit Authority (WMATA) on the metro network of Washington DC (which can be seen in Figure 2). This network consists of 91 unique stations, of which 9 are considered transfer stations, and 93 unique service links. The metro is serviced by 6 unique lines, most of which have some overlap.

Originally a years worth of data on passenger journeys was available, meaning the tap-in station and time and the tap-out station and time were provided. From this data, Krishnakumari et al. (2019) calculated how much delay was experienced on average per 30 minute time slices, where the delays are differentiated by initial waiting times, transfer waiting times, and in vehicle delays. So eventually, the input data for this study is the average delay experienced at each initial station and each transfer station, during 30 minute time slices over an entire year. From this data set, a selection was made of weekdays that follow a regular demand pattern (morning peak and evening peak), meaning weekends, holidays and days with irregular demand patterns (e.g. due to maintenance) were not taken into account. This was done as it is expected that due to the different demand, and possibly different timetables, the delay relations could be different and should thus not be taken into account so as not to muddle the results.

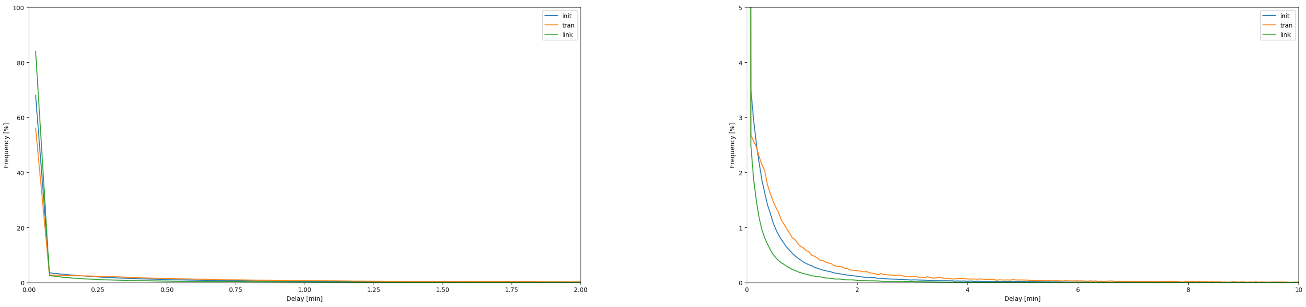
A distribution of delays found in the eventual data set can be seen in Figure 3. It can be seen that transfer delays are slightly higher than initial waiting delays and in-vehicle delays. However, only 34% of all trips contain a transfer, meaning the other delays are not necessarily less important, and in-vehicle delays are given per link which thus add up over an entire trip. It should also be noted that the overwhelming majority of data points show a delay of near-zero minutes.

In Figure 4 the average passenger delay can be seen per initial, link and transfer node. Here the stations are differentiated by direction into node pairs, which represent initial waiting delay, and transfer delays are represented on arcs between the corresponding initial nodes.

In this figure, it can be seen that, with the exception of some random nodes, the average delay is rather equal over the entire network, meaning there are no lines or sections with exceptional delays, and thus no special areas of focus in the network. The few nodes showing high delays could be caused by local circumstances, such as long lasting maintenance projects during the data gathering time.

3.2 Implementation

In this section, two subjects will be discussed, namely: discretization (Section 3.2.1), a necessary step to be able to apply the available data to the Bayesian Network; and division of the network into sectors (Section 3.2.2), which had to be done in order to obtain results in a reasonable time.



(a) Full overview

(b) Zoomed overview

Figure 3: The percentage of times a delay of a certain size is present in the data set for transfer, initial and link delays. The label init indicates initial stations, tran indicates transfer stations and link indicates links.

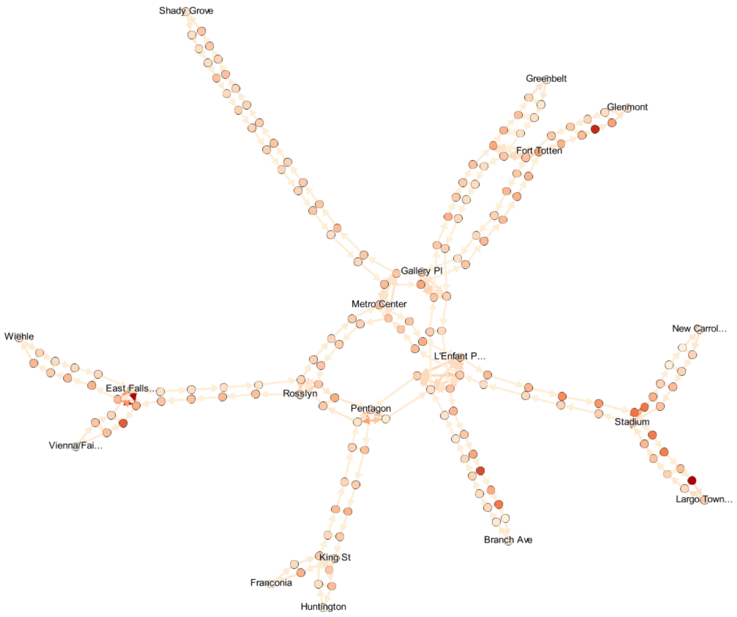


Figure 4: Average passenger delay per directed initial station, transfer station and link. The maximum value is 4.63 minutes while the minimum is 0 minutes.

3.2.1 Discretization

The input data for a Bayesian Network must be discrete as the conditional probability tables use discrete states, while the available data on delays is continuous. Therefore, the data must be discretized, a process in which information is lost (Clarke and Barton, 2000). In order to minimize this loss, it is tried to apply the entropy loss minimization method, a method that leads to less loss than the standard equal interval or equal population, which are often used to create discretization bins (Clarke and Barton, 2000). However, this method results in too many bins (over 1000), which would make the Bayesian Network impossible to construct or interpret. Therefore another method must be used, but neither the commonly used equal interval nor the equal population methods are deemed appropriate due to the unequal delay occurrence (as can be seen in Figure 3). This would lead to severely underrepresented bins when using the equal interval method, and too many near-zero bins when using the equal population method.

Therefore, an equal logarithmic population method is used. This method assigns each data point to equal interval bins of very small size, after which the logarithm is taken of the population per bin. These logarithmic values can then be used in a similar manner when using the standard equal population method, to find the sizes of the bins, and divide the data accordingly. This method results in bins of reasonable range, while all being populated by a significant number of data points.

3.2.2 Sector Division

The metro network is quite big, and differentiating the nodes by direction results in more than twice the number of nodes than stations in the network. Because of this, it is not feasible to compute the Bayesian Network for the entire network with the computational power available. Therefore, the nodes must be split into subsets. This is done by dividing the network into sectors, which is possible due to the radial nature of this specific network. The center of the network is the first sector, while the six radial directions of the network comprise six other sectors.

However, the information from the complete Bayesian Network is required for the further steps of this study. Therefore, it must be possible to re-combine the seven sectors of the Bayesian Network. In order to do so, the sectors will have some overlap, namely the first two stations on each radial line (as seen from the center) will also be part of the first sector (the center sector). This way, the arcs that are connected to these two nodes in each of the sectors it is present in, are able to reconnect all the sectors into one Bayesian Network. Furthermore, all transfer nodes are also added to the center sector, as well as the sector in which they are present, as these might have connections between them as well due to their slightly different nature.

The entire sector division method is based on the assumption that direct relations between nodes over large distances do not exist. This is a reasonable assumption as it is much more likely that direct relations exist between nearby stations, and that further away stations are only connected indirectly through the stations in between, since spatial delay propagation would first affect these in-between stations.

4 Results

In this sections, the results of the study are discussed. Firstly, the results of the Bayesian Network are presented in Section 4.1. Afterwards, the informativity indicators that were calculated from this are presented in Section 4.2. Lastly, the correlations between these informativity indicators and the network's centrality indicators are given in Section 4.3.

4.1 Bayesian Network

The Bayesian network is calculated for each of the seven sectors of the network, and recombined through the overlapping nodes. The result of this is mapped onto a semi-geographical representation of the network in Figure 5. This semi-geographical representation shows all stations as node-pairs, where the nodes are differentiated by direction. No distinction is made between transfer delay and initial delay nodes, but if either of the nodes is connected in the BN, the corresponding geographical node is colored dark. If a node is not connected to any arc, it is light in color.

Several things can be noted from the figure. It can be seen that the green line ('Greenbelt'-'Branch Ave') has very few connections, while the west part of the Orange/Silver line (Wiehle'-'Rosslyn' and 'Vienna Fairfax'-'Rosslyn') is the most densely connected, meaning there are many arcs between few nodes. There is no obvious reason for this, but it can indicate that delays propagate badly, respectively, well, over these sections. It can also be seen that most connections are between nodes that are going into the same direction, and are between nodes that are geographically near each other. However, especially the stations at which transfers are possible have arcs that cross large distances. Another arc over a large distance can be seen between the end stations 'New Carrolton' and 'Largo Town Center', indicating these longer arcs are not exclusive to transfer nodes. This means that the assumption on which the sector method is based - namely that direct relations over long distances will not occur - , is not justified.

Using the k-fold method with 5 folds, the errors for each node were calculated. A maximum error of 7.4% for one node is observed, with most nodes having errors of less than 2%. This means the states of those nodes can be predicted with an accuracy of over 98%, when the states of the other nodes in the network have been observed. This is a rather high accuracy and indicates the result is valid. Moreover, the average RMSEs of the training and test set are calculated. The training set results in an RMSE of 0.0046 minutes, and the test set in an RMSE of 0.0092 minutes. Clearly, the test set has a higher RMSE, which is expected. Still, both RMSEs are very low, especially when taking into consideration that the possible delays range from 0 to over 5 minutes. Therefore, it can be said the model is not overfitted significantly. However, with just the errors cannot say anything about underfitting and the accuracy of the model. For slightly different data sets (the 5 different folds), slightly different Bayesian Networks were found, in terms of where arcs were located.

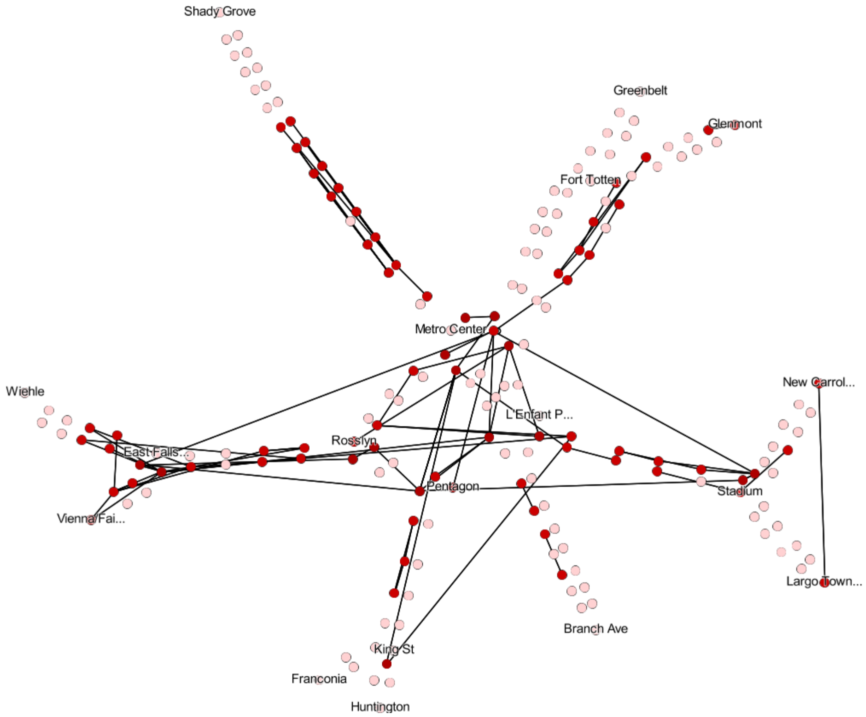


Figure 5: The recombined BN mapped onto the geographical map of the stations. No distinction is made between transfer and initial stations. Nodes that are not connected are a light pink, while nodes that are connected are a dark red.

This indicates that although the errors are low, the model is not necessarily accurate as it could be underfitted, possibly due to the large presence of near-zero values in the data set.

4.2 Informativity Indicators

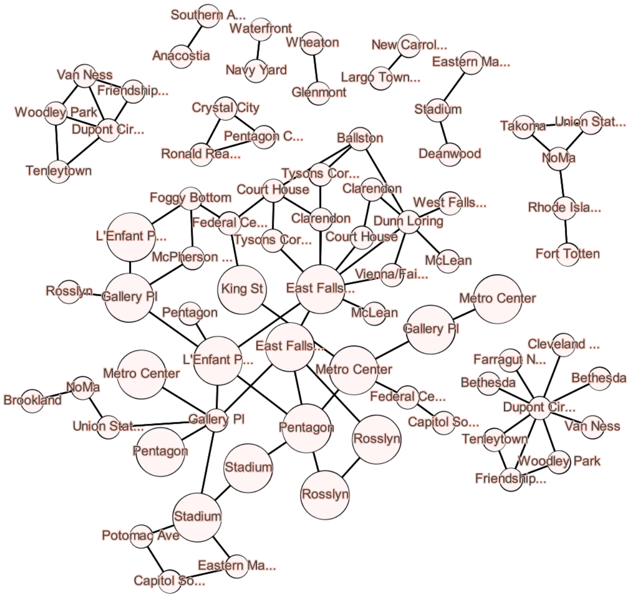


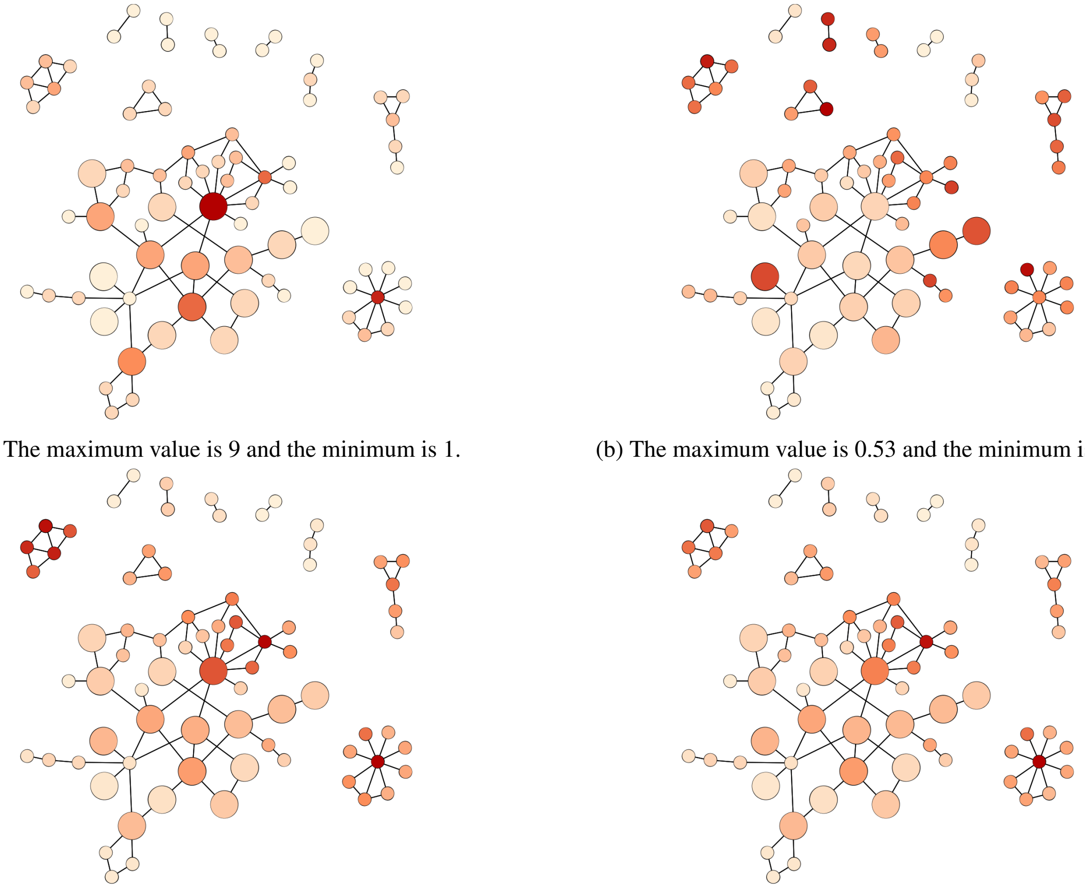
Figure 6: Graphs of the Bayesian Network with the names of the corresponding station indicated at each node.

Using the results from the Bayesian Network, the informativity indicators are calculated. The Bayesian Network structure, colored by the four values of the informativity indicators, can be seen in Figure 7. Here, transfer nodes have a bigger size than initial waiting nodes, to distinguish them. In these figures, it can also be seen more clearly that there are few small non-connected sub-graphs, none of which are connected to a transfer node.

For reference, in Figure 6, the names of the stations corresponding to the nodes can be seen, so that they can be related to the physical network. Here, the directions are not indicated for clarity reasons.

When looking at the four graphs in Figure 7, it is clear that those for the lower and upper bound of the total informativity are very similar. This is expected, of course, as these should be indications of the same thing. The fact that they adhere to this expectation is a sign that the calculations were accurate.

Besides this, the graph for the average direct informativity stands out most from the other three graphs. Here, nodes at the periphery are relatively high, while some nodes in the center have low values, as opposed to the graphs of the outgoing node degree as well as those of the Total



(a) The maximum value is 9 and the minimum is 1.

(b) The maximum value is 0.53 and the minimum is 0.045.

(c) The maximum value is 2.53 and the minimum is 0.045.

(d) The maximum value is 2.37 and the minimum is 0.045.

Figure 7: Graphs of the Bayesian Network, colored by the Node Degree (a), Average Direct Informativity (b), upper limit of the Total Informativity (c) and the lower limit of the Total Informativity (d), where transfer delay nodes are indicated by a larger size than the initial delay nodes. Below each figure the minimum and maximum values of the respective indicator are given.

Informativity. This does make sense, as this indicator ascribes a value to the expected informativity per connection, regardless of the number of connections. It is thus not a good indicator when trying to assess the quantity of information the node could provide regarding the entire rest of the network. The other two indicators would be a better indication for this. However, the average direct informativity could still prove to be a useful indicator for different purposes, for example when trying to discover the accuracy of information provided by a node.

4.3 Correlation Centrality and Informativity Indicators

The informativity indicator results, and the calculations of the three centrality indicators (node degree, closeness centrality, betweenness centrality) in three representations of the metro network (L-space, B-space, P-space), were used to construct a correlation matrix between the two sets of indicators. The result of this can be seen in Table 1.

When looking at the overall scores, the betweenness centrality in the L-space seems to be most strongly correlated with all of the informativity indicators. The other centrality indicators of the L-space also correlate significantly with the informativity indicators, while the indicators in the B-space correlate least strongly.

When looking at the informativity indicators, it can be seen that especially the outgoing node degree, as well as the average direct informativity, correlate better with the centrality indicators than the total informativity indicator. Although some strong correlations can be found for each indicator.

Table 1: A table with the correlation between the informativity and centrality indicators. OND is the Outgoing Node Degree, ADI the Average Direct Informativity, TI (u) the upper bound of the Total Informativity, and TI (l) the lower bound of the Total Informativity. Deg. is degree centrality, Clos. is closeness centrality and Betw. is betweenness centrality.

	L-space			B-space			P-space		
	Deg.	Clos.	Betw.	Deg.	Clos.	Betw.	Deg.	Clos.	Betw.
OND	0.343	0.280	0.390	0.217	0.249	0.144	0.273	0.251	0.143
ADI	0.203	0.280	0.334	0.093	0.221	0.236	0.224	0.230	0.244
TI (u)	0.108	0.176	0.250	-0.042	0.045	0.053	0.055	0.051	0.054
TI (l)	0.154	0.214	0.292	0.020	0.107	0.099	0.117	0.113	0.102

5 Discussion and Conclusion

From the Bayesian Network results it becomes apparent that connections mainly exist between nodes that are going into the same direction, and nearby nodes. However, some relations also exist between nodes that are further apart, meaning the assumption on which the sector method is based (namely that such long relations will not occur as it is more likely that far apart nodes would only be indirectly connected through in-between nodes), is not justified. Therefore, the results of this study are not entirely accurate representations, and somewhat different results might be found if the Bayesian Network had been created for the entire network at once. This should be kept in mind when using these results for further research.

It would thus also be very beneficial for future work to find a way to calculate the Bayesian Network for the entire metro network at once, eliminating the need for dividing it into sectors. This could for example be done by using more powerful computers, using better heuristic methods, or using some form of super node method whereby the number of nodes is decreased by aggregating certain nodes. Although especially this last option will be based on other assumptions that must be tested as well.

From the correlations between the informativity and centrality indicators, it can be said that the centrality indicators do approximate the informativity indicators to some extent, but not completely, as the maximum observed correlation is still just below 0.4. Especially the amount of shortest paths that pass through a node (the betweenness centrality), is closely related to the informativity of a node, more so than any other measure of the centrality of a node. Also, the centrality indicators of the geographical representation of the network (L-space), are the most relevant with respect to the informativity indicators.

It is possible that the correlations can be used to approximate the informativity of nodes in networks for which little data is available, but to do this properly it must first be determined if these correlations are robust by doing similar studies for other transport networks. Furthermore, the exact meaning of the informativity indicators and relations found in the Bayesian Network must be explored further before they can be accurately used to, e.g., determine common delay causes, find the most effective areas for delay mitigation measures, and generate accurate delay predictions for both transport authorities and passengers.

It should be noted that the data available for this study was given in time slices of 30 minutes. However, some stations in the network are more than 30 minutes apart, and given the possibility of relations over larger distances, the temporal nature of delay should be taken into account more. This could be done, for example, by increasing the size of the time slices. Which has its own disadvantages such as a loss of information due to an increase in averaging, but a sensitivity analysis to this parameter could still lead to meaningful insights. More difficult to do but possibly worthwhile, is trying to develop a method that would compare the delay states of nodes at different time slices, possibly depending on the travel time between the nodes. This way, temporal delay propagation could be studied in more detail, and the insights from such results could be very useful.

References

Bates, J., Polak, J., Jones, P., and Cook, A. (2001). The valuation of reliability for personal travel. *Transportation Research Part E: Logistics and Transportation Review*, 37(2-3):191–229.

Berger, A., Gebhardt, A., Müller-Hannemann, M., and Ostrowski, M. (2011). Stochastic delay prediction in large train networks. In *OASIcs-OpenAccess Series in Informatics*, volume 20. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Clarke, E. J. and Barton, B. A. (2000). Entropy and mdl discretization of continuous variables for bayesian belief networks. *International Journal of Intelligent Systems*, 15(1):61–92.

Derrible, S. and Kennedy, C. (2011). Applications of graph theory and network science to transit network design. *Transport reviews*, 31(4):495–519.

Ebert-Uphoff, I. (2009). Tutorial on how to measure link strengths in discrete bayesian networks.

Von Ferber, C., Holovatch, T., Holovatch, Y., and Palchykov, V. (2009). Public transport networks: empirical analysis and modeling. *The European Physical Journal B*, 68(2):261–275.

Kirchhoff, F. and Kolonko, M. (2015). Modelling delay propagation in railway networks using closed family of distributions. Technical report, Technical Report.

Kjaerulff, U. B. and Madsen, A. L. (2008). *Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis*. Springer.

Koller, D., Friedman, N., and Bach, F. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Krishnakumari, P., Cats, O., and Van Lint, H. (2019). Day-to-day and seasonal regularity of network passenger delay for metro networks. In *Transportation Research Record*.

Lessan, J., Fu, L., and Wen, C. (2018). A hybrid bayesian network model for predicting delays in train operations. *Computers & Industrial Engineering*.

Malandri, C., Fonzone, A., and Cats, O. (2018). Recovery time and propagation effects of passenger transport disruptions. *Physica A: Statistical Mechanics and its Applications*, 505:7–17.

Nicholson, A. E. and Jitnah, N. (1998). Using mutual information to determine relevance in bayesian networks. In *Pacific rim international conference on artificial intelligence*, pages 399–410. Springer.

Nielsen, O. A., Landex, O., and Frederiksen, R. D. (2009). Passenger delay models for rail networks. In *Schedule-Based Modeling of Transportation Networks*, pages 1–23. Springer.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufman Publishers, Inc.

Redman, L., Friman, M., Gärling, T., and Hartig, T. (2013). Quality attributes of public transport that attract car users: A research review. *Transport policy*, 25:119–127.

Rodríguez-Núñez, E. and García-Palomares, J. C. (2014). Measuring the vulnerability of public transport networks. *Journal of transport geography*, 35:50–63.

Schmöcker, J.-D., Bell, M. G., and Lam, W. H. (2004). Importance of public transport. *Journal of Advanced Transportation*, 38(1):1–4.

Washington Metropolitan Area Transit Authority (2017). System map.

Zilko, A. A., Kurowicka, D., and Goverde, R. M. (2016). Modeling railway disruption lengths with copula bayesian networks. *Transportation Research Part C: Emerging Technologies*, 68:350–368.