Quantifying the cascading effects of passenger delays

Cats, Oded; Hijner, Anne Mijntje

DOI
10.1016/j.ress.2021.107629

Publication date
2021

Document Version
Final published version

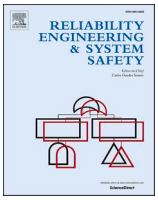
Published in
Reliability Engineering and System Safety

Reliability Engineering and System Safety 212 (2021) 107629

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress



Quantifying the cascading effects of passenger delays

Oded Cats *, Anne Mijntje Hijner

Department of Transport & Planning, Delft University of Technology, the Netherlands

ARTICLE INFO

Keywords:
Network reliability
Bayesian network
Passenger delay
Informativity Indicators
Cascading

ABSTRACT

Delays in transport networks has adverse implications for infrastructure and service managers as well as travellers. While it is widely acknowledged that delays occurring across the transport networks may be related, there are is lack of knowledge on the underlying properties of these relations and means for quantifying them. To this end, we develop a network-wide data-driven delay analysis method. First, we construct a Bayesian network to represent the relations between delays associated with different transport network elements and assess the reliability of critical infrastructure elements. Second, we propose a series of original metrics denominated informativity indicators for quantifying the spatial extent of the delays observed based on the Bayesian network obtained. The proposed approach is applied to the Washington DC metro network. Time-dependent passenger waiting and transferring delays inferred over more than a year from smartcard data are utilized as input to the Bayesian network. Our findings indicate that passenger delays at few selected stations are directly informative of delays occurring at many other stations. We also examine the relation between the proposed informativity metrics and the topological properties of metro stations, concluding that the latter have a limited value in approximating network-wide delay correlations.

1. Introduction

While delays in transport networks are experienced locally they have ramifications for overall network reliability. Infrastructure managers and service providers are interested in encapsulating delays and devising effective mitigation measures. To effectively improve transport reliability, it is essential to understand how the inter-relations between delays across the network. This is especially paramount in public transport networks, where delays may cascade not only to neighbouring segments but also further downstream of upstream services as well as across services due to inter-dependencies of assets and rolling stock as well as passenger interchange patterns.

Past work on delay propagation has approached this problem from the perspective of the operator of the network, rather than the passenger [2, 17]. Manitz et al. [26] proposed a method for identifying the source of train vehicle delays inspired by the analysis of epidemiological processes. Kafle and Zou [16] proposed an analytical model for modeling flight delay propagation in relation to the flight schedule of a given airplane. A time-series analysis was conducted by Du et al. [7] to analyze delay propagation at the airport level. Harrod et al. [13] formulated an analytical model to estimate vehicle delay propagation along a rail line. Daniele Marra and Corman [27] identified clusters of delays in Automated Vehicle Location (AVL) data using a variant of the DBSCAN algorithm. A time-series analysis and a stochastic railway operations simulation model were used by Buchel et al. [3] to analyze delay propagation for the Swiss railway network. In summary, previous studies have focused on the analysis of vehicle punctuality and its propagation along an individual line, over trips of a given vehicle, or across the network. These studies deployed an array of analytical, statistical and simulation techniques to analyze delays in passenger transport systems, yet little is known about the properties of network-wide delays as experienced by passengers.

Passenger journeys consist of several successive travel segments. Journey times, and therefore also delays, are comprised of waiting times, in-vehicle times and possibly transfer times. Inter-vehicle headway, vehicle capacity constraints and transfer synchronization are therefore among the determinants of passenger delays. The literature on public transport network robustness assessment often takes generalized passenger travel time losses as an indicator of network vulnerability and as a mean for identifying the most critical links in the network [30, 4]. Dynamic assignment models enable the assessment of the spatial extent of delay propagation in public transport networks by analysing disruption scenarios [25]. All of the aforementioned robustness analysis were performed based on either an analytical or a simulation-based public

* Corresponding author.
E-mail address: o.cats@tudelft.nl (O. Cats).

https://doi.org/10.1016/j.ress.2021.107629
Received 11 August 2020; Received in revised form 7 December 2020; Accepted 11 March 2021
Available online 13 March 2021

transport assignment model. Moreover, the scenario-based analysis was based on full breakdowns while passenger delays stem from a variety of small-scale disturbances. Furthermore, the relations between delays occurring at different locations were not examined. It remains therefore unknown how passenger delays in different locations throughout the network are co-related in reality. This arguably stems from the lack of empirical knowledge on passenger trajectories as well as the need to develop methods and techniques for inferring passenger delays and systematically quantify their inter-dependencies. The analysis of such relations will reveal the spatial patterns of delays occurring simultaneously across the network and contribute to the analysis of system reliability.

Measuring passenger delays is becoming increasingly possible by means of processing passively collected data. Several studies have analysed passenger journey delays using smart card records [34, 14]. Notwithstanding, there is a lack of empirical knowledge on network-wide passenger delay properties, and in particular the relations between delays occurring across the network. To this end, input on passenger delay associated with individual journey segments, i.e. network elements, is needed. Krishnakumari et al. [20] developed an estimation method for inferring the delay associated with each journey component, i.e. waiting time, in-vehicle time, transfer time from fusing passively collected data concerning both vehicle and passenger movements (and without relying on any disruption log files). This opens opportunities for examining the inter-relations between these delays and uncover potential patterns that can be used by modellers and planners. Based on passenger delays inferred from smart card data, Yap and Cats [33] perform a supervised learning approach for predicting passenger delays for individual stations and time periods and identify the most critical stations in terms of exposure, impact and overall criticality (measured in pass-hour losses). The results were then used for clustering stops based on their contribution to overall network vulnerability and a feature analysis identified the most relevant variables in predicting disruption impact.

Bayesian networks have been previously used to assessing the risk and reliability of critical infrastructure with applications ranging from nuclear waste disposal [22] and chemical capacities under intentional attacks [1] to maritime accident analysis [9], road networks [11] or water and power supplies [21] under seismic hazards, and inland waterway port and surrounding supply chain network [15]. In the context of train operations, Bayesian networks have been used by Zilko et al. [35] and Lessan et al. [23] to predict disruption length and train delays, respectively. Ghaemi et al. [12] demonstrated how the outputs of a Bayesian network prediction can be integrated in a disruption management model to forecast the consequent of disruptions as well as rescheduling measures on passengers delays. Ulak et al. [32] proposed using two network metrics to distinguish between stations inducing delays and stations susceptible to delays and applied it analysing commuter train vehicle delays. Following past studies, we adopt a data-driven approach which employs principles from machine learning and network science. This approach allows establishing statistical relations between system elements while alleviating the need to make many assumptions of the underlying relations between system components. This enables unraveling unexpected relationships and consider all possible dependencies involved in the delay spreading mechanism.

None of the abovementioned studies has analysed the network-wide patterns of passenger delays and their spatial relations. To the best of our knowledge, this is the first study to empirically examine these network-wide relations among passenger delays. In this study, we develop a method for analysing network-wide passenger delays and their cascading patterns. The contributions of this study are as follows: (i) estimate a Bayesian Network for passenger delays associated with individual network elements; (ii) develop a series of original metrics denominated *informativity indicators* for quantifying the spatial extent of the delays observed based on the topological characteristics of the Bayesian network obtained; (iii) apply the proposed method for an extensive empirical passenger delay data from the Washington DC metro network; (iv) examine the relation between the proposed informativity metrics and the topological properties of metro stations.

The proposed informativity indicators allow quantifying stations' capability to provide information on the delay state of the rest of the network. This allows us to establish the extent to which a given station state in terms of the amount of passenger delay observed there can be indicative of the delay experienced elsewhere across the network. We demonstrate how these indicators contain valuable information that can potentially be used to determine critical locations for deploying delay mitigation measures, and to provide information to both operators and passengers. In order to shed light on the underlying processes related the found passenger delay correlations, we analyze the relation between the calculated informativity indicators and a range of topological indicators of the physical and service networks which reflect various system functions. This also allows identifying whether topological indicators could potentially be used as proxies for the proposed informativity indicators in the absence of suitable data.

The rest of this paper is structured as follows: the method - including the proposed informativity indicators - is described in Section 2; then the application of this study, including a description of the case study, is briefly presented in Section 3; Results are presented and discussed in Section 4 and conclusions are drawn in Section 5.

2. Method

The input to our method includes an estimated passenger delay for each network element (station, track segment) and several graph representations of the respective network. We first identify relations between delays experienced at different locations by means of constructing a Bayesian Network (Section 2.1). Thereafter, we quantify the extent to which delays occurring at one network element coincide with delays occurring elsewhere by calculating a series of original metrics that incorporate information related to the topology of the yielded Bayesian network into what we denominate informatively indicators (Section 2.2). Finally, we cross-check the relation between the obtained informatively indicators and topological indicators of the public transport network itself (Section 2.3).

2.1. Determining delay relations: creating the Bayesian network

First, based on the data available, a Bayesian Network (BN) is constructed. A BN is by definition a Directed Acyclic Graph (DAG) - $G(N, A)$ where N is the set of nodes and A is the set of arcs - which represents the inter-dependencies of its variables. Here, a node that has an arc pointing to another node is called a parent, and the node the arc is pointed towards is called a child. A BN also contains information on the dependency, 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 respective parents. For more information on BN, the interested reader is referred to Kjaerulff and Madsen [18] and Koller et al. [19].

In this study, the set of potential nodes in the BN corresponds to the set of stations where each line-direction is denoted as a separate station node. The arcs of the graph represent co-relations between the nodes regardless of the physical or service connections between stations in the transport network. The mean passenger waiting time delay per station is used as input and is obtained from an estimation algorithm applied to individual passenger trajectories as detailed in the description of input preparation in Section 3.1.

Using the conditional probability tables, the arcs of the BN can be labelled, according to the strength of the correlation between each pair of connected nodes. The labeling is performed by using the Link Strength method introduced by Nicholson and Jitnah [28]. Only arcs with labels that reflect an inter-dependence – measured by means of a correlation coefficient - that exceeds a certain significance are maintained in the

final presentation of the BN. The most common approach for labeling the BN arcs is based on mutual information (MI) [29, 28, 18, 8], which quantifies how much information can be gained concerning a certain node, once information concerning another node becomes available. This concept can be extended to include the possibility of a node having multiple parents, so that inter-dependencies among the parent nodes are taken into account when determining the arc strength [8, 28]. The label, link strength r_{ij}, of the link connecting nodes i and j, both members of the set of nodes V and can be calculated as follows:

$$r_{ij} = \sum_{x,z} p(x,z) \sum_{y} p(y|x,z) \log_2 \frac{p(y|x,z)}{p(y|z)} \tag{1}$$

Where x and y are vectors representing all possible states of nodes i and j, respectively. Similarly, z is a single vector containing all the possible states of all parents of node j, except for i. r_{ij} conveys thus the strength of the dependency between two nodes in the BN graph, independently of the value of the other parents of the downstream node. The probability terms in Eq. (1) refer to variable states, for example $p(y|x,z)$ is the (conditional) probability of finding the values associated with node j at any particular state given a certain combination of states of the variables at node i and all other parenting nodes.

2.2. Informativity indicators

Next, we leverage on the structure of the BN and the arc labels to calculate the informativity indicators for each node. In the following we introduce a set of original indicators. Different indicators are proposed as they encompass different notions of informativity, and might lead to different observations. They are defined and shortly discussed below.

The first proposed informativity indicator is the *outgoing node degree*. In this context, this indicator provides information on how many nodes a certain node can provide direct information on, and is calculated as follows

$$d_i^+ = \sum_{j \in V} a_{i,j} \quad \forall i \in V \tag{2}$$

where d_i^+ is the outgoing node degree of node $i \in V$; N is the set of nodes in the BN; and $a_{i,j} \in A$ is the entrance in the adjacency matrix A of the BN graph that equals 1 if there is a link connecting node n and node i and 0 otherwise.

While the outgoing node degree indicates how many nodes a certain node provides information on, it does not reflect the extent of this information. We therefore propose as a second indicator, the *expected direct informativity*. It is intended to describe how informative a certain node is expected to be in regard to any node that it is connected to. The expected direct informativity, e_i, is calculated using a weighted average term:

$$e_i = \frac{1}{d_i^+} \sum_{j \in V} \left[a_{i,j} \cdot r_{ij} \right] \quad \forall i \in V \tag{3}$$

The two above mentioned indicators only provide an indication of the amount of information that can be provided through the direct downstream neighbours (also known as 'children' in the BN terminology) of a node. However, a node can also indirectly provide information on its further descendants through higher order relations. This can be accounted for by multiplying the link weights along the path connecting to a descendant. When multiple possible paths exist, we consider two options: (i) the most informative path can be taken into account, or (ii) all paths can be taken into account since different paths can provide novel information. In the former case the informativity might be underestimated because not all informative paths are taken into account. Conversely, in the latter case the informativity metric may be overestimated as some paths may overlap, i.e. are not independent, correlated. Therefore, we undertake both approaches and use them to establish the lower and upper bounds of the informativity of a given node for all of its descendants.

In order to determine the *total informativity*, meaning the total information a node can provide on the delay state of the rest of the network, the informativity value for all node's descendants across the network should be summed.

We let $(k,m) \in p$ mean that there is a link with an upstream node k and a downstream node m (i.e. $a_{k,m} = 1;\ k,m \in V$) which belongs to path p. The upper bound of this indicator, t_i^{max}, can then be expressed as

$$t_i^{max} = \sum_{j \in V} \sum_{p \in P_{i,j}} \prod_{(k,m) \in p} \left[a_{k,m} \cdot r_{k,m} \right] \quad \forall i \in V \tag{4}$$

where $P_{i,j}$ is the set of all paths connecting nodes i and j.

The lower bound of this indicator, t_i^{min}, is formulated as

$$t_i^{min} = \sum_{j \in V} \max_{p \in P_{i,j}} \prod_{(k,m) \in p} \left[a_{k,m} \cdot r_{k,m} \right] \quad \forall i \in V \tag{5}$$

2.3. Relating centrality indicators to informativity indicators

The construction of the BN based on which the informativity indicators are calculated requires a large amount of empirical data which is often not readily available. Moreover, we are interested in investigating the relation between the found passenger delay correlations and system functionality characteristics. The third step of the methodology aims therefore at testing the potential usage of topological indicators – which do not require any data other than static network and service configurations - as proxies for capturing the informativity of nodes. Luo et al. [24] provide evidence that strictly topological indicators, i.e. station centrality indicators, allow predicting the distribution of passenger flows across the network, especially once considering the service layer – including the number of transfers and service frequencies implied by the service network design - in addition to the infrastructure layer.

We compute the correlations of the three proposed informativity indicators using a series of node centrality indicators, i.e. node degree, closeness and betweenness centrality for three different public transport network graph variants based on what is sometimes referred to in the literature as L-, P- and B-spaces [6, 10]. We hereby denote them as *Infrastructure, Service-* and *Transfer-* spaces, respectively. These spaces represent the physical and logical layers of the public transport system. The *Infrastructure*-space represents all stations as nodes and any pair of stations that are served successively by a given line are connected by a link. While the same set of nodes is used in the *Service*-space, links are introduced between all pairs of stops that are served by at least one common line, regardless of their position along the line. The *Transfer*-space contains two sets of nodes: in addition to the station nodes, each line is represented by a node. Links are introduced only between nodes belonging to different subsets – between a station and a line – denoting that a station is served by a given line.

The different graph representations allow identifying the characteristics of stations with high informativity value. The same node centrality indicator carries different meanings when applied in different spaces, all of which can be relevant for informativity. For example, node degree centrality corresponds to the number of direct neighbours in the graph. It carries a different interpretation depending on the space in question: in the *Infrastructure*-space the node degree reflects the local role as an intersection/interchange; in the *Service*-space it corresponds to the number of other stations that are affected by the same line, and; in the *Transfer*-space it is the number of lines intersecting at a given station (or the number of stations served by a line in case of a line node). Similarly, node betweenness centrality corresponds to the share of shortest paths traversing through a node. While in the *Infrastructure*-space it can be used a proxy of passenger flows traversing a station, in the *Service*-space it reflects the likelihood of a station been used as an interchange location.

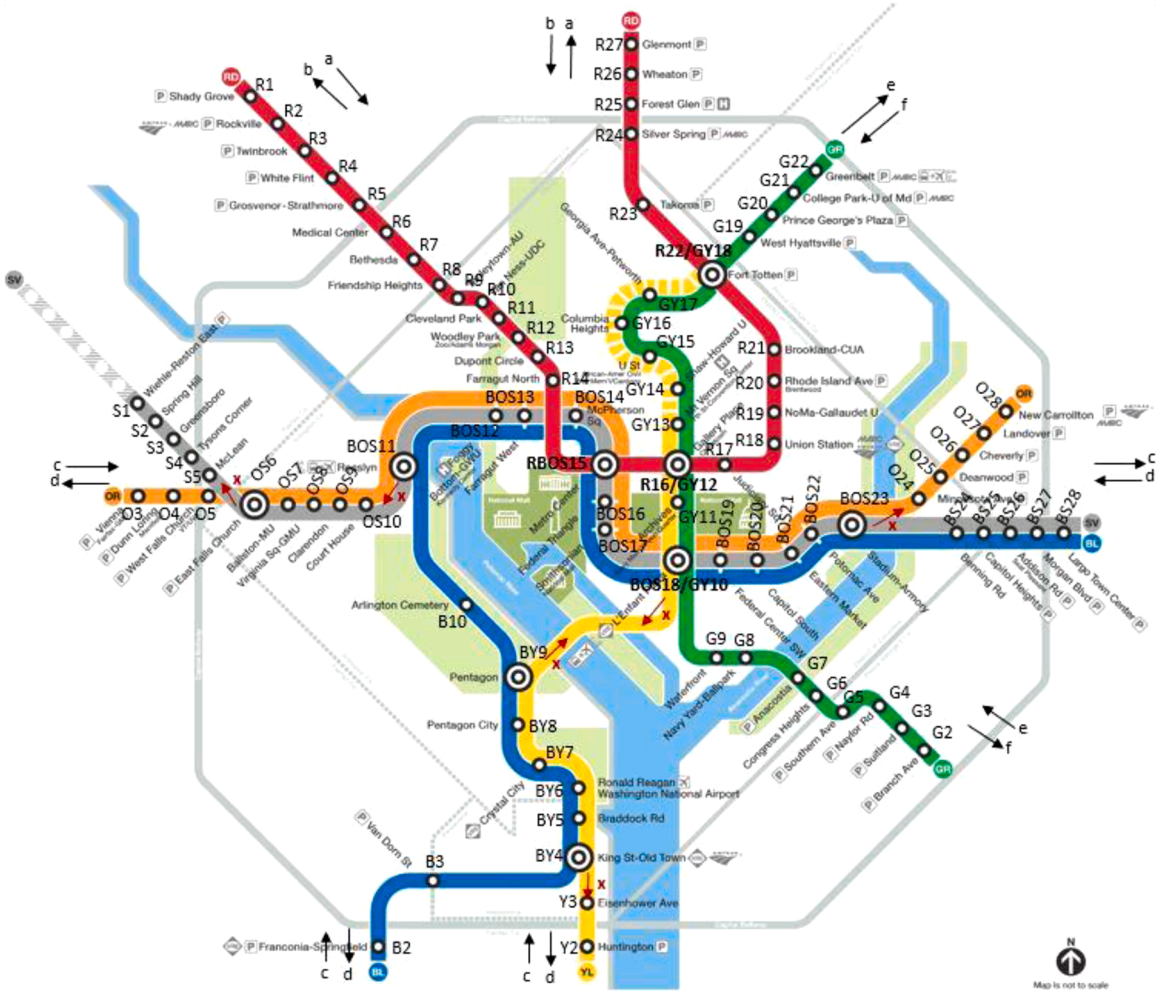


Fig. 1. Map of the Washington DC metro network with our encoding of all stations and transfer directions.

3. Application

The method described in the previous section is applied to a case study public transport network for which individual passenger trajectories and passenger delays are available (Section 3.1). After reporting the passenger delay characteristics (Section 3.2), we discuss the results of the three key methodological steps, namely the BN results (Section 3.3), the values of the informativity indicators (Section 3.4) and their correlations with the topological indicators for the case study network (Section 3.5).

3.1. Washington DC metro case study and implementation

For this study, a year's worth of train movement and passenger-train assignment output data of the so-called ODX method described in Sanchez-Martinez [31] from the Washington DC metro system (Fig. 1) managed by the Washington DC Area Metro Transit Authority (WMATA) is available. The data is composed of one year of smart card data from 19 August 2017 to 28 August 2018 for the entire metro system which is comprised of six lines serving 91 stations.

WMATA defines for each origin-destination pair of stations in their metro network the *maximum expected travel time*. The latter corresponds to the case where a passenger has just missed a train under the assumption that the next train arrives within a full planned service headway, the time spent on-board follows the scheduled time, and walking times correspond to the 80th percentile of the walking speed distribution. To define it mathematically, let us denote by $\widetilde{t}_{s_0,s_d,\tau}$ the maximum expected travel time for journey r_{s_o,s_d} which occurs during time period τ, where $s_o \in S$ and $s_d \in S$ are the origin (tap-in) and destination (tap-out) stations of this journey, respectively. S is the set of all stations. The maximum expected travel time comprises of three components as follows:

$$\widetilde{t}_{s_0,s_d,\tau} = \sum_{l \in r_{s_o,s_d}} \sum_{s_i \in l} \widetilde{t}^{veh}_{s_i,s_{i+1},\tau} + \sum_{l \in r_{s_o,s_d}} h_{l,\tau} + \sum_{s_i \in r_{s_o,s_d}} \sum_{s_i=s_1}^{s_{m-1}} t^{walk}_{s_i} \tag{6}$$

The first term sums over all scheduled in-vehicle travel times, $\widetilde{t}^{veh}_{s_i,s_{i+1},\tau}$, between successive stations (i.e. $s_i, s_{i+1} \in S$) traversed by line l which is boarded along the respective journey. The second term is the planned service headway of each of the lines boarded along the respective journey. The third term corresponds to all the walking segments along the journey within the metro system (i.e. between tap-in location and boarding platform, between alighting and boarding platforms in case of transfers, and between alighting platform and the tap-out location).

Total passenger journey delay is then calculated as the travel time in excess of the maximum expected travel time for the respective tap-in and tap-out station pair. The delay associated with passenger journey n that involves travelling between origin station s_0 and destination station s_d which took place during time period τ is decomposed as follows:

$$d_{s_o,s_d,\tau,n} = d^{wait}_{s_o,\tau} + \sum_{l \in r_{s_o,s_d,n}} \sum_{s_i \in l} d^{on-board}_{s_i,s_{i+1},\tau} + \sum_{s_i \in r_{s_o,s_d,n}} \sum_{s_i=s_1}^{s_{m-1}} d^{trans}_{s_i,\tau} + \varepsilon_n \tag{7}$$

where $d^{wait}_{s_o,\tau}$ is the delay associated with waiting at the origin station,

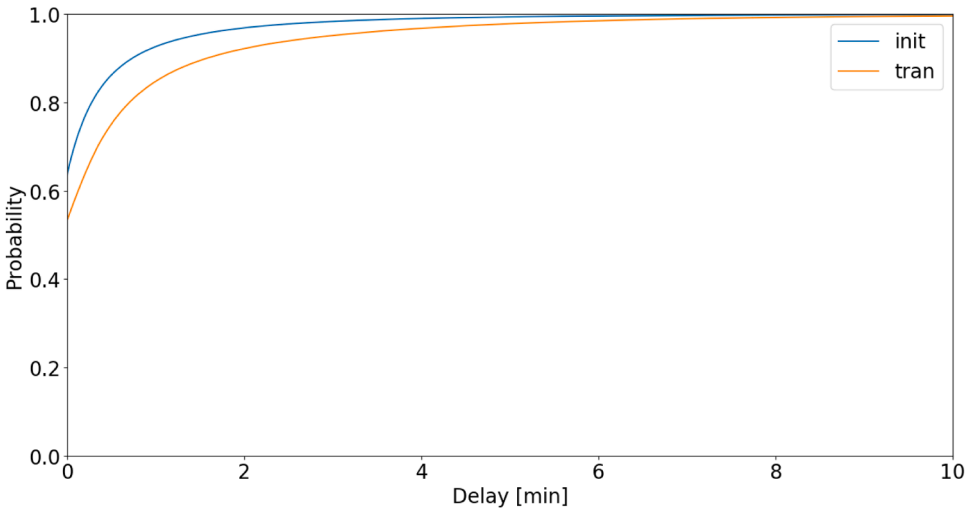


Fig. 2. Cumulative distribution function of initial and transfer passenger delay.

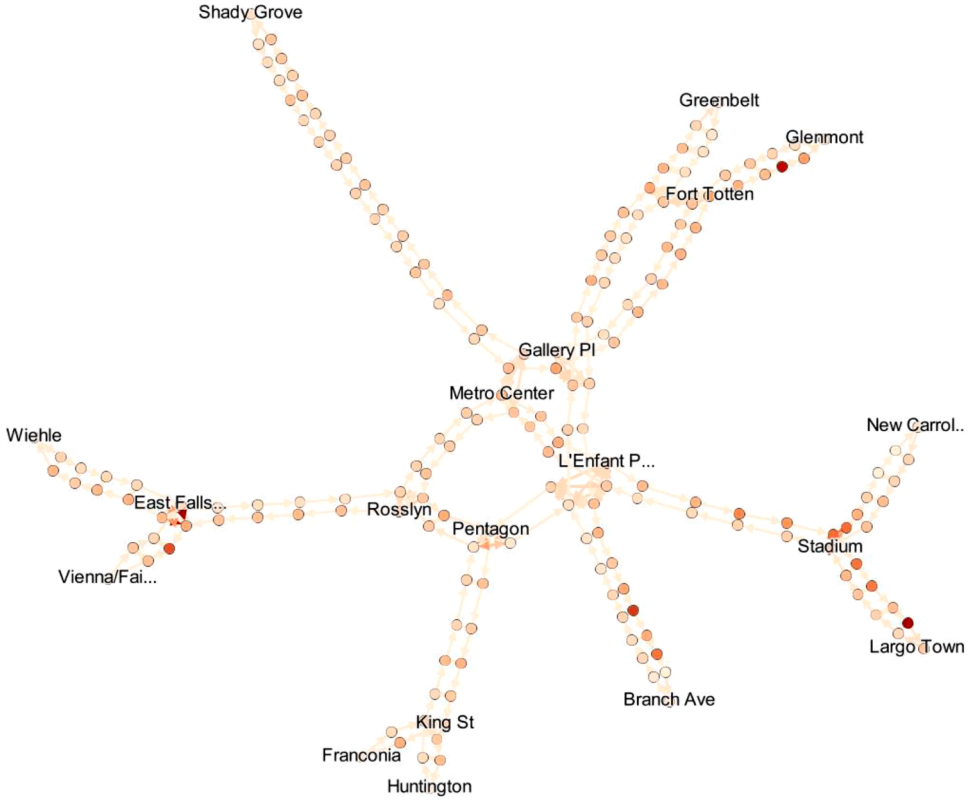


Fig. 3. Average passenger delay per directed initial station, transfer station and link. Darker color indicates longer delays.

$d^{on-board}_{s_i,s_{i+1},\tau}$ is the delay occurring when traversing a certain train segment, $d^{trans}_{s_i,\tau}$ is the delay attributed to a transfer location and ε_n is the remaining journey-specific delay which is not attributed to any of the network elements. Krishnakumari et al. [20] formulated a set of equations to map the passenger journey delays to the passenger delay associated with each network element. The output thereof is available as input for this study.

The input available to this study consists therefore of the estimated average initial and transfer passenger delay per station and transfer direction resulting with 186 nodes in our graph representation. Note that the passenger delay components associated with an interchange station that allows transferring between l lines is represented using $2l$ initial waiting time nodes (one per line-direction) and $2l(2l-1)-2l$ or $4l(l-1)$ transfer time nodes (one per combination of line-directions, excluding opposite directions of the same line).

Average passenger delays were estimated for 30 min time window throughout the analysis period. The passenger delay data is differentiated per direction in which the passengers are travelling. In the following, only regular weekdays (i.e. excluding weekends, holidays and weekdays with large-scale maintenance works) were retained. This was done as it is expected that due to the different demand and supply availability properties on the excluded days, the delay relations might be different and may thus muddle the results. The p-value parameter was set to the standard 0.05 value when deciding whether the significance of a dependence is to be included in the BN.

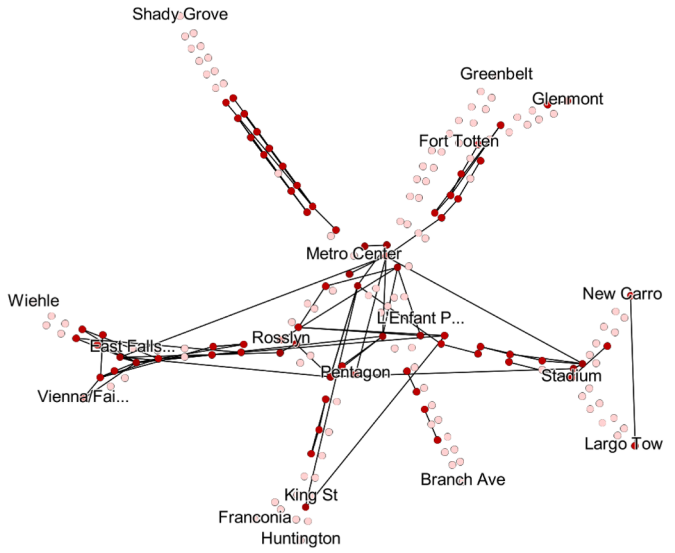


Fig. 4. The recombined BN mapped onto the geographical map of the stations. No distinction is made between transfer and initial stations in the presentation. Nodes that are not connected in the BN are shown in light pink, while nodes that are connected are shown in dark red.

The BN was constructed using the libpgm library for Python 2.7 [5]. Constructing the BN took 15.5 h on a personal HP EliteBook 840 G1 laptop with Intel(R) Core i7–4510U CPU @ 2.00 GHz 2.60 GHz with a RAM of 8GB.

3.2. Descriptive statistics of passenger delay

Passenger delays vary greatly across time slices and network elements. As can be seen in Fig. 2, the distribution of delays is such that most stations are not attributed with any passenger delay during most time slices. Recall that the definition of passenger delay (Eq. (6)) is such that a delay is registered only in case the passenger waits longer than the full service headway. Furthermore, the vast majority of the stations that do induce delays– either initial or upon waiting – are up to 2 min per passenger. Notwithstanding, over the course of the yearlong data there are few instances where individual nodes are associated with 5 and even 10 min delay per passenger.

Next, we turn into examining the spatial distribution of passenger delays amongst network nodes. In Fig. 3 the average (over all 30 min time slices) delay per passenger for each initial station (represented by a node for each direction), link (arcs between nodes of different stations) and transfer movement (an arc connecting the differently direct nodes of the same station), are displayed. Delays are in general uniformly distributed across the network once averaged over all time periods. Notwithstanding, certain stations exhibit recurrent and severe delays with an average passenger delay of several (up to 4.6) minutes when averaged over the entire analysis period. High values are observed at stations located in proximity to terminal stations (i.e. upstream of Glenmont and Largo Town Center) as well as for some transfer movements (e.g. at East Falls Church).

As evident also from Fig. 2, transfer movements induce longer average delays than initial waiting time at stations. However, only 36% of all passenger trips involve a transfer and the average number of trip legs in the Washington metro is 1.43. Even though stations within the inner rings (Gallery Palace-Metro Center-Rosslyn-Pentagon-L'Enfant Plaza) are subject to low-moderate average delays, due to the very high passenger volumes at these stations, they nevertheless account for most of the passenger delays occurring in the network.

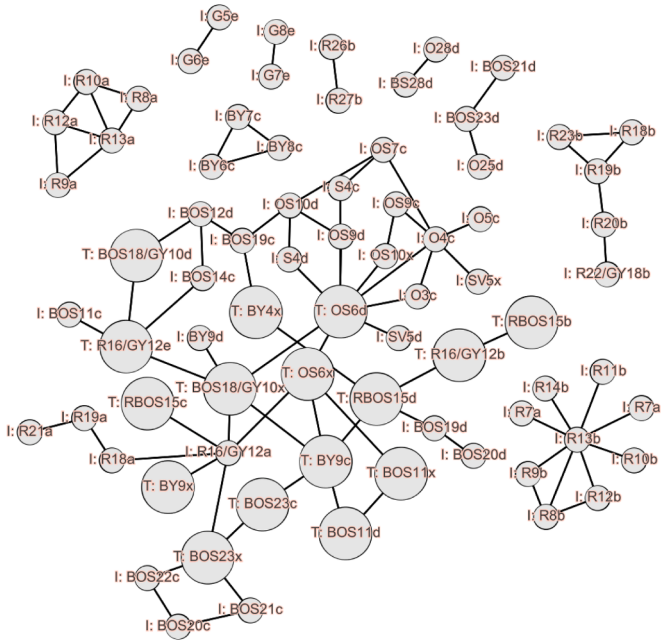


Fig. 5. All connected nodes in the Bayesian Network, transfer nodes are depicted as larger than initial station nodes. The station code (see Fig. 1) is proceeded by an indication of whether the node refers to initial waiting time (I:) or transfer time (T:).

3.3. Bayesian networks results

In the remaining of this study we focus on delays encountered while waiting – either at the initial station or when interchanging. Each entrée in the vector representing all possible states for all nodes (x in Eq. (1)) corresponds of the delay value for a 30-minutes time slice throughout the analysis period. Due to the size of the network, determining the Bayesian Network for all stations at once is computationally prohibitive. Therefore, the network is divided into sectors: one sector for the core of the network, and six sectors for the different sections branching out from the core of the network (e.g. west of Rosslyn, south of Pentagon, see Fig. 1). For each of these sectors, a BN is constructed. By having some overlap between the sectors (namely the first two stations of each radial sector closest to the core, as well as the transfer stations that are present in the radial sectors), it is possible to thereafter recombine the BNs of the sectors into one large BN. This simplification of the BN representation is made, since it is expected that no considerable dependencies are to be found between stations that are further apart; it is much more likely that far apart stations would only be indirectly dependent of each other, through the stations that lie in between.

The results of the Bayesian Network construction, displayed using a geographical map of the metro network, can be seen in Fig. 4. Out of the 186 nodes included in the Washington metro network graph, 75 are connected in the BN yielded using 85 arcs (i.e. the p-value for the respective r_{ij} as defined in Eq. (1) is smaller than 0.05).

Several observations can be made, namely that connections exist predominantly between nearby stations, and between nodes that are heading in the same direction. This corroborates the underlying assumption in constructing BN for sectors. Notwithstanding, there are also a few connections that bind stations that are not in proximity. Furthermore, the results for different folds of the data set (the data was split into several training and test sets for a k-fold approach to calculate the errors of the model) showed that the mean percentage errors for all nodes were well below 10% when compared for different partitioning of the dataset with the highest error of 7.4% observed for the transfer station of Pentagon in the direction of Arlington Cemetery. We also examine the resulted attained for different folds and all cases similar

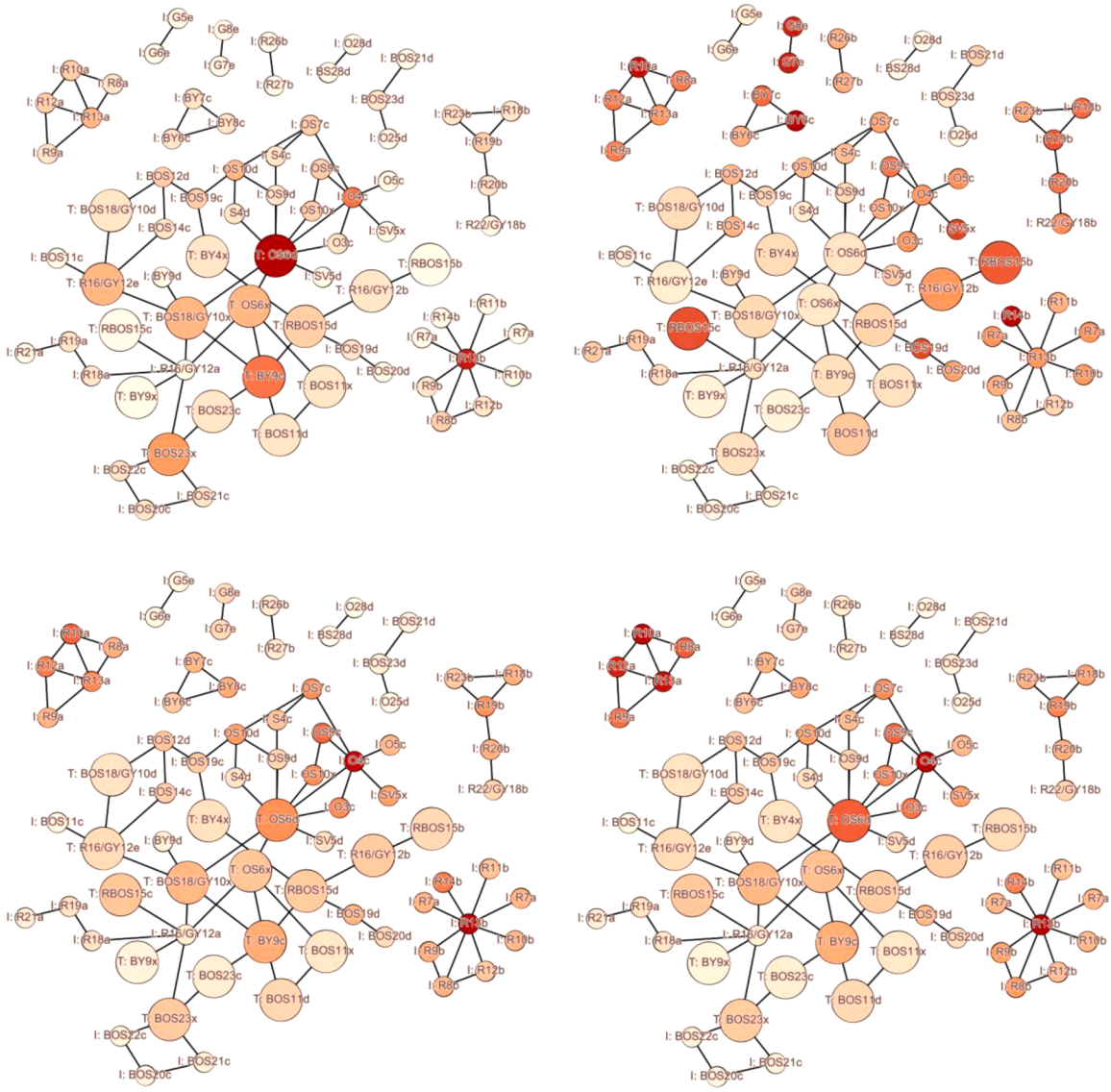


Fig. 6. All connected nodes in the Bayesian Network, transfer nodes are depicted as larger than initial station nodes. The color scale reflects the relative value of the respective informativity indicator shown in clock-wise starting from top-left as follows: Outgoing node degree, Expected direct Informativity, Lower bound of Total informativity and Upper bound of Total informativity.

observations are made regarding the types of dependencies found, namely that most dependencies occur between nearby stations and nodes corresponding to the same service direction. Furthermore, we calculate the Root Mean Squared Error (RMSE) of the obtained BN using the 5-fold method. The test set resulted in an average RMSE of 0.0092 min, with no significant differences between the folds. The RMSE is very low given that delays can range from 0 to over 5 min.

In order to better understand the properties of the obtained BN, we display its configuration in Fig. 5. As is clearly visible in Fig. 5, the obtained BN is a disconnected graph that consists of one large sub-component (42 nodes), one considerably smaller component (9 nodes) and eight additional components ranging in size between 2 and 5 nodes.

The large component clearly visible in Fig. 5 consists of both initial waiting time and transfer nodes, each of which having a higher likelihood to be connected with another node of the same type (initial-initial/transfer-transfer) than with the other type (initial-transfer). Moreover, all other components are exclusively comprised of initial waiting nodes. A closer inspection also shows that in the case of the eight smallest components, with one exception, they all consist of a set of stations along a single service line direction. The exception pertains to the pair of the two eastern terminal stations (visible also in Fig. 4) of the Orange (i.e. New Carrollton) and the Blue/Silver line (i.e. Largo Town Center) for which a large portion of the lines runs along a common corridor and therefore their dispatching regime is likely to be inter-dependent.

Metro lines tend to form cliques within the BN graph, indicating that waiting time delays mostly spread amongst stations served by the same line. All nodes of the mid-size component in the BN (Fig. 4) are located along the Red line, in either direction of its northwest section. Even in the large component, while it consists of a plurality of lines, lines tend to form cliques. The correlations of passenger delays across stations of the same service which are not in close sequence may be attributed to line headways. Note that certain transfer delays connected in the BN are also successive stations along a service line (e.g. R16/GY12b and RBOS15b).

Unlike vehicle delays, passenger delays may very well be correlated also across lines due to transferring flows, facilities used by multiple lines (i.e. terminals, tracks, stations) and system features that extend beyond a single line (e.g. overall demand level, weather conditions). Indeed, while most passenger delays are correlated at the line-level, they are by no means limited to a single line as is clearly visible also in Fig. 4.

Interestingly, some relations do not exhibit transitivity. For instance,

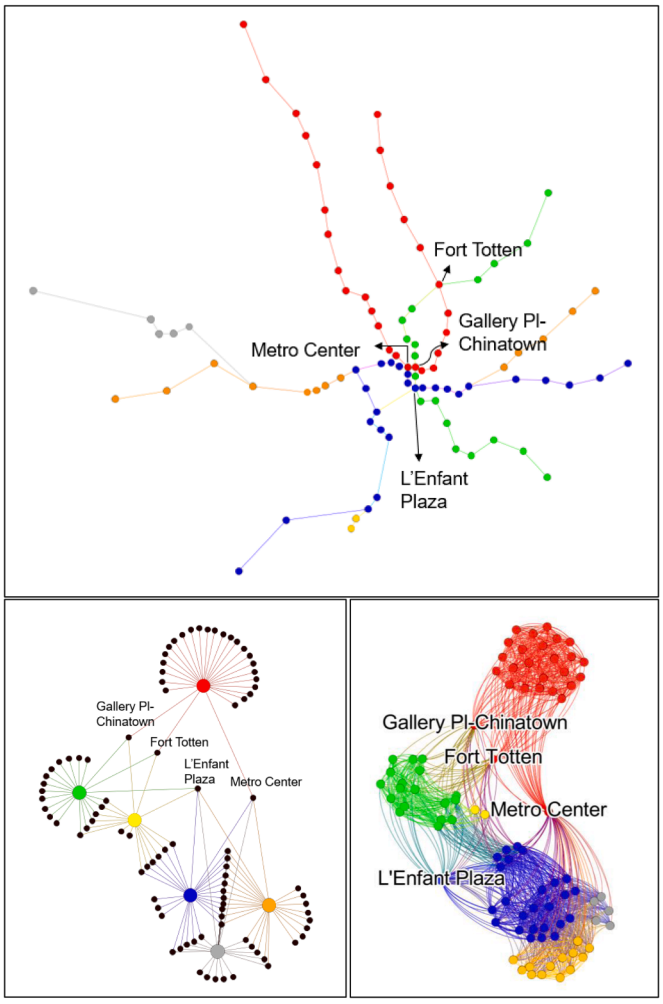


Fig. 7. Washington DC metro network representation in Infra-space (top), Transfer-space (bottom left) and Service-space (bottom right).. Node and link colours correspond to the designated line color (see Fig. 1).

while transfer delay occurring when transferring to the northbound direction of the Green line at L'Enfant Plaza is correlated with transfer delays at the same station for those transferring to the southbound direction of the Yellow line as well as to the transfer delay of those transferring there to the westbound direction of the Orange line, the latter two are not mutually correlated. While this may seem counterintuitive at first, this is reasonable once considering that there is no origin-destination relation for which passengers will interchange between these services, unlike the aforementioned correlated delays.

3.4. Informativity indicators

We now turn to extracting the informativity indicators proposed in Section 2.2 for the BN obtained from our passenger delay data as reported in the previous section. We calculate the Outgoing node degree, Expected direct Informativity, Lower bound of Total informativity and Upper bound of Total informativity for each node in our BN graph. Fig. 6 shows again the BN configuration where now the node (delay at either initial waiting or transfer station) color corresponds to the relative value of the respective informativity indicator. For ease of reference, the positions of all nodes – which do not correspond here to their geographical locations - is the same as in Fig. 5 and is kept the same in all graphs included in Fig. 6.

The Outgoing node degree indicator directly reflects the local connectivity of nodes in the BN. It has a very skewed distribution with few nodes that constitute hubs in the BN having a high degree and most nodes having a low degree since they are connected to only one or two other nodes (Fig. 6, top left). This implies that delays at only few stations are directly informative for many other stations. This is reasonable given the very radial structure of the case study network and the ability to encapsulate delays without their correlations extending to more than few stations.

The values of the Expected direct informativity account also for the labels of the links connecting to a node in the BN (Eq. (3)), not only the number thereof (Eq. (2)). It is noticeable that certain nodes – both initial and transfer waiting time – which are not connected to many other nodes have a high Expected direct informativity value (Fig. 6, top right). This is driven by their high correlation with the few other nodes they are correlated with. Consequently, this results with a greater diversity in indicator values across the network. A maximum value of 0.53 is found, while the theoretical upper bound value is one – implying that one node is perfectly correlated with all of its neighbours.

The results for the Lower (Eq. (4)) and Upper (Eq. (5)) bounds of Total informativity are overall similar (Fig. 6, bottom left and bottom right, respectively). High values are found for nodes that are globally well connected since they can provide either directly or indirectly (through other nodes)s information that is relevant to many other nodes. The total informativity value for the small component at the top left of the figures - which corresponds to initial waiting time delays at the middle stations along the counter-clockwise direction of the Red line – stands out, due to the clique it forms. The relatively high meshedness for this component which offers several paths in the BN to skim informativity for each node is also arguably the cause for the more pronounced difference between the lower and upper bound values.

3.5. Relations between centrality and informativity indicators

We represent the case study network in the *Infrastructure*-, *Transfer*- and *Service*-spaces (described in Section 2.3) as depicted in Fig. 7. The *Infrastructure*-space closely follows the familiar representation of the public transport network as shown in public maps. The *Transfer*-space consists of two sets of nodes: lines and stations. The latter are connected to all lines that serve them. The *Service*-space connects all stop pairs that can be reached without performing a transfer, i.e. there is at least one line serving both stops, and therefore comprises of clearly visible line-based cliques.

We examine the correlations between the three node centrality indicators – degree, betweenness and closeness - in each of the three graph representations – *Infrastructure*-, *Transfer*- and *Service*-spaces – and the proposed informativity indicators extracted from the BN, i.e. Outgoing

Table 1
Correlations between node informativity and centrality indicators.

	Infrastructure-space			*Transfer*-space			*Service*-space		
	Deg.	Clos.	Betw.	Deg.	Clos.	Betw.	Deg.	Clos.	Betw.
d_n^+	0.343	0.280	0.390	0.217	0.249	0.144	0.273	0.251	0.143
e_n	0.203	0.280	0.334	0.093	0.221	0.236	0.224	0.230	0.244
t_n^{max}	0.108	0.176	0.250	−0.042	0.045	0.053	0.055	0.051	0.054
t_n^{min}	0.154	0.214	0.292	0.020	0.107	0.099	0.117	0.113	0.102

node degree, Expected direct Informativity, Lower bound of Total informativity and Upper bound of Total informativity. The results are reported in Table 1.

It can be observed that the WMATA metro network topology indicators exhibit low to moderate correlations with the informativity indicators of the passenger delay BN. This indicates that the topological properties of a station contain only limited information on the extent to which delays occurring at this station are informative concerning delays occurring within the same time slice at other stations across the network. Notwithstanding, some moderate correlations can potentially be useful in the absence of operations and passenger flows data. In particular, the indicators calculated in the *Infrastructure*-space representation of the network, and the betweenness centrality in specific, can explain almost 40% of the variability observed. The most encompassing informativity indicators that pertain to the total informativity can be less well approximated by strictly topological indicators. This confirms that network-wide delay relations are the outcome of complex interactions that cannot be easily approximated without adequate data sources.

4. Conclusion

The extent to which delays occur simultaneously in transport networks has implications for network reliability and robustness. We quantify the amount of information contained in observed passenger delay in one location on passenger delays observed elsewhere in the network. We propose a series of informativity indicators based on Bayesian Network estimations. Furthermore, we assess how well the proposed indicators can be approximated by the centrality indicators of the respective nodes.

We apply the proposed method to the Washington DC metro network. The analysis is enabled by a large empirical dataset containing passenger and metro train movements for the case study application. The BN obtained yields an average RMSE of 0.0092 min. As can be expected, passenger delays at one station are most informative for delays occurring at nearby stations on the same line direction. In addition, some further apart stations along the same line or corridor exhibits strong correlations. The latter is arguably due to the dependency between dispatching caused by the constrained infrastructure as well as due to passenger interchange movements.

Passenger delays at few selected stations are directly informative of delays occurring at many other stations. This suggests that service providers may use information from a limited set of stations to analyze network states, a possible direction for future research. The results of the correlation analysis between informativity indicators and centrality indicators suggest that the latter can only be used to a limited extent in the event that data on passenger delay is missing.

Correlation does not imply causation, and hence the results do not allow drawing conclusions on the underlying determinants of observed relations among passenger delays at stations. The informativity indicators proposed in this study and the identified correlations should be supplemented with analytical and simulation models for metro operations in order to examine the relations between the (un)reliability of system components. Such models can be instrumental in examining the causes that give result with the network-wide statistical properties and relations reported in this study. Future research may calibrate such models so as to reproduce the observed delay instances and correlations. A transport modeling approach will involve the combination of a rail traffic model and a passenger assignment model and the calibration and validation thereof using empirical data such as utilized in this study. This will pave the way to potential applications such as disruption mitigation both in real time and at the tactical level. For example, disruption mitigation measures can benefit from incorporating information on which delays can be encapsulated and which are likely to occur in conjunction with delays at other locations.

The results reported are valid for the case study network and the mode of operations used during the one-year analysis period. A systematic comparison of informativity indicators for a range of case studies that vary in their topology, infrastructure utilization, service congestion and delay management practices, will allow concluding on the transferability of the results reported in this study.

Further research may further advance the proposed method to include also temporal delay propagation. The BN may be extended to also include delays recorded for different time slices. To reduce the computational effort associated with the increase in the large increased in number of nodes, one may connect each node only to a limited number of directly subsequent time windows, including between delay at the same station at successive time slices to account for auto-correlations. Alternatively, a probabilistic network model that can capture the propagation of disruptions using mechanisms inspired by the spreading of epidemics and rumours may be developed.

CRediT authorship contribution statement

Oded Cats: Conceptualization, Methodology, Project administration, Supervision, Writing – original draft. **Anne Mijntje Hijner:** Data curtion, Formal analysis, Software, Visualization.

Declaration of Competing Interest

The authors have no competing interests to declare.

Acknowledgements

The authors thank WMATA, and Jordan Holt in particular, for the valuable cooperation and data provision. The authors are also grateful for comments made by Marie Schmidt, Yousef Maknoon and Serge Hoogendoorn at an earlier stage of this study.

References

[1] Argenti F, Landucci G, Reniers G, Cozzani V. Vulnerability assessment of chemical facilities to intentional attaches based on Bayesian network. Reliab Eng Syst Saf 2018;169:515–30.
[2] Berger A, Gebhardt A, Müller-Hannemann M, Ostrowski M. Stochastic delay prediction in large train networks. Oasics-open access series in informatics, 20. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik; 2011.
[3] Buchel B, Spanninger T, Corman F. Empirical dynamics of railway delay propagation identified during the large-scale Rastatt disruption. Sci Rep 2020;10:18584.
[4] Cats O, Koppenol G-J, Warnier M. Robustness assessment of link capacity reduction for complex networks: application for public transport systems. Reliab Eng Syst Saf 2017;167:544–53.
[5] Cyberpoint International, LLC (2012). libpgm. Available at: https://pythonhosted.org/libpgm/index.html.
[6] Derrible S, Kennedy C. Applications of graph theory and network science to transit network design. Transport Reviews 2011;31(4):495–519.
[7] Du W-B, Zhang M-Y, Zhang Y, Cao X-B, Zhang J. Delay causality network in air transport systems. Transportation Research Part E 2018;118:466–76.
[8] Ebert-Uphoff, I. (2009). Tutorial on how to measure link strengths in discrete Bayesian networks. Available at: https://pdfs.semanticscholar.org/0cf9/6f545b06d5fd3cc4de4017f833b8eff887c7.pdf.
[9] Fan S, Blanco-Davis E, Yang Z, Zhang J, Yan X. Incorporation of human factors into maritime accident analysis using a data-driven Bayesian network. Reliab Eng Syst Saf 2020:107070.
[10] von Ferber C, Holovatch T, Holovatch Y, Palchykov V. Public transport networks: empirical analysis and modeling. The European Physical Journal B 2009;68(2):261–75.
[11] Gehl P, Cavalieri F, Franchin P. Approximate Bayesian network formulation for the rapid loss assessment of real-world infrastructure systems. Reliab Eng Syst Saf 2018;177:80–93.
[12] Ghaemi N, Zilko AA, Yan F, Cats O, Kurowicka D, Goverde RMP. Impact of railway disruption predictions and rescheduling on passenger delays. Journal of Rail Transport Planning & Management 2018;8(2):103–22.
[13] Harrod S, Cerreto F, Nielsen OA. A closed form railway line delay propagation model. Transportation Research. Part C 2019;102:189–209.
[14] Hendren P, Antos J, Carney Y, Harcum R. Transit travel time reliability: shifting the focus from vehicles to customers. Transp Res Rec 2015;2535:35–44.
[15] Hossain NUI, El Amrani S, Jaradat R, Marufuzzaman M, Buchanan R, Rinaudo C, Hamilton M. Modeling and assessing interdependencies between critical infrastructures using Bayesian network: a case study of inland waterway port and surrounding supply chain network. Reliab Eng Syst Saf 2020:106898.

[16] Kafle N, Zou B. Modeling flight delay propagation: a new analytical-econometric approach. Transportation Research Part B 2016;93:520–42.
[17] Kirchhoff, F. and Kolonko, M. (2015). Modelling delay propagation in railway networks using closed family of distributions. Technical Report.
[18] Kjaerulff UB, Madsen AL. Bayesian networks and influence diagrams: a guide to construction and analysis. Springer; 2008.
[19] Koller D, Friedman N, Bach F. Probabilistic graphical models: principles and techniques. MIT press; 2009.
[20] Krishnakumari P, Cats O, van Lint H. Estimation of network passenger delay form individual trajectories. Transportation Research Part C 2020;117:102704.
[21] Lee S, Choi M, Lee H-S, Park M. Bayesian network-based seismic damage estimation for power and potable water supply systems. Reliab Eng Syst Saf 2020; 197:106796.
[22] Lee C-J, Lee KJ. Application of Bayesian network to the probabilistic risk assessment of nuclear waste disposal. Reliab Eng Syst Saf 2006;91(5):515–32.
[23] Lessan J, Fu L, Wen C. A hybrid Bayesian network model for predicting delays in train operations. Comput Ind Eng 2019;127:1214–22.
[24] Luo D, Cats O, van Lint H. Can passenger flow distribution be estimated solely based on network properties in public transport systems? Transportation 2020;47: 2757–76. https://doi.org/10.1007/s11116-019-09990-w. In press.
[25] Malandri C, Fonzone A, Cats O. Recovery time and propagation effects of passenger transport disruptions. Physica A 2018;505:7–17.
[26] Manitz J, Harberting J, Schmidt M, Kneib T, Schöbel A. Source estimation for propagation processes on complex networks with an application to delays in public transport systems. Appl Stat 2017;66(3):521–36.
[27] Marra AD, Corman F. From delays to disruption: impact of service degradation on public transport networks. Transp Res Rec 2020;2674(10):886–97.
[28] Nicholson AE, Jitnah N. Using mutual information to determine relevance in Bayesian networks. In: In *Pacific Rim International Conference on Artificial Intelligence*. Springer; 1998. p. 399–410. pages.
[29] Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufman Publishers, Inc; 1988.
[30] Rodríguez-Núñez E, García-Palomares JC. Measuring the vulnerability of public transport networks. J Transp Geogr 2014;35:50–63.
[31] Sanchez-Martinez GE. Inference of public transportation trip destinations by using fare transaction and vehicle location data: dynamic programming approach. Transp Res Rec 2017;2652(1):1–7.
[32] Ulak MB, Yazici A, Zhang Y. Analysing network-wide patterns of rail transit delays using Bayesian network learning. Transportation Research Part C 2020;119: 102749.
[33] Yap M, Cats O. Predicting disruptions and their passenger delay impacts for public transport stops. Transportation 2020. https://doi.org/10.1007/s11116-020-10109-9. in press.
[34] Zhao J, Frumin M, Wilson N, Zhao Z. Unified estimator for excess journey time under heterogeneous passenger incidence behavior using smartcard data. Transportation Research Part C 2013;34:70–88.
[35] Zilko AA, Kurowicka D, Goverde RMP. Modeling railway disruption lengths with copula Bayesian networks. Transportation Research Part C 2016;68:350–68.