DIAGNOSIS AND PREDICTION OF TRAFFIC CONGESTION ON URBAN ROAD NETWORKS USING BAYESIAN NETWORKS

Jiwon Kim, Corresponding Author
The University of Queensland
Brisbane St Lucia, QLD 4072, Australia
Tel: +61 7 3346 3008; Fax: +61 7 3365 4599; Email: jiwon.kim@uq.edu.au

Guangxing Wang
The University of Queensland
Brisbane St Lucia, QLD 4072, Australia
Email: guangxing.wang@uq.net.au

Submitted for presentation and publication to the
95th Annual Meeting of the Transportation Research Board, January, 2016
Diagnosis and Prediction of Traffic Congestion on Urban Road Networks
Using Bayesian Networks

ABSTRACT

This paper proposes a Bayesian Network (BN) analysis approach to modelling the probabilistic dependency structure among causes of congestion on a particular road segment and analyzing the probability of traffic congestion given various roadway condition scenarios. A BN is used to encode the joint probability distribution over a set of random variables that describe scenario variables, which represent factors affecting the congestion level of a target segment such as time-of-day, incident, weather, and traffic states on adjacent links, and output variables, which represent traffic performance measures of the target segment such as flow, density, and speed. This study develops a methodology to build a BN model based on historical traffic and event data, and demonstrates the BN-based traffic analysis using a study network in Queensland, Australia. The study discusses applications of the proposed BN model in urban traffic congestion management, focusing on identifying leading causes for congestion diagnosis and identifying critical scenarios for congestion prediction.

Keywords: Bayesian networks, congestion diagnosis and prediction, scenario analysis
1. INTRODUCTION

In dealing with traffic congestion problems, three main questions asked by traffic managers are: “Why is this congestion occurring?”, “Can we anticipate when it will occur?”, and “How can we prevent or mitigate them?” The first question leads to addressing a congestion diagnosis problem, which aims to understand the causes of a particular congestion situation and identify the most critical causes, and the second question leads to addressing a congestion prediction problem, which aims to predict the occurrence of traffic congestion based on the observations of the causes. These two problems are closely connected and together provide important grounds for answering the third question; that is, the ability to accurately diagnose and predict traffic congestion allows traffic managers to identify and closely monitor the root causes of congestion and therefore proactively detect and manage congestion hotspots.

The general objective of this study is to apply a probabilistic graphical modelling approach and machine learning techniques to solve congestion diagnosis and prediction problems in urban road networks. To this end, this paper proposes a Bayesian Network (BN) model that is capable of capturing the probabilistic dependency structure among causes of congestion on a particular road segment and assessing the probability of traffic congestion given various roadway condition scenarios.

2. BAYESIAN NETWORKS

A Bayesian Network (BN) is a probabilistic graphical model that represents probabilistic relationships among a set of variables via a directed acyclic graph (DAG) (1–3). A BN consists of a set of nodes and a set of arcs, where nodes represent random variables and arcs connecting pairs of nodes represent direct dependencies between variables. In general, constructing a BN model requires the following three steps: (i) defining variables (nodes), (ii) specifying structure (arcs), and (iii) specifying parameters (conditional probability distribution for each node). The second step is to determine a qualitative property of the BN, which is causality or dependence relationships between variables, and the third step is to determine its quantitative part, which are probability distributions that quantify these relationships. Once a set of nodes of a BN are defined, specifying its structure and parameters can be done in two ways: manual specification based on domain expert knowledge or automatic specification using machine learning techniques. In this study, we take an approach to manually specifying the network structure while learning parameters from data. Once built, a BN provides a compact representation of the full joint probability distribution over its variables, which allows us to compute the probability of each state of a node conditioned on any subset of other variables. This process is called probabilistic inference—computing the posterior distribution of variables $X$ given evidence $e$, $P(X|e)$—and there are a number of efficient exact and approximate inference algorithms for performing complex probabilistic reasoning tasks in BN. Reasoning can be performed in two different directions, i.e., from known causes to unknown effects (predictive reasoning) and from known effects to unknown causes (diagnostic reasoning). These features make BN a powerful tool for diagnosing and predicting traffic congestion under uncertainty. Examples of questions that can be answered by a BN model include: what are the probability of having a severe congestion at a particular link when there is rain (congestion prediction) and what are the probability that there is rain if we observe a severe congestion (congestion diagnosis).

While the use of BN in transportation research areas is relatively new, there are a number of areas where the BN approach has been applied. Two main areas of application are traffic estimation and forecasting and congestion prediction (4–9) and accident detection and crash
prediction (10–13).

2. METHODOLOGY

2.2 Building a BN model for Traffic Congestion Diagnosis and Prediction

This paper considers a BN model that represents the dependency structure of link-level measures. For a given link, we design a BN model that describes relationships between link performance measures (e.g., flow, density, and speed) and external factors that affect the target link (e.g., time-of-day, weather, and incident). The goal of this model is to assess the impacts of external factors on traffic conditions on a target link. Thus, the impacts of upstream or downstream traffic conditions are not considered in this model. The BN model in this study is considered to be static in that the model represents a time-independent knowledge of dependency relationships among variables (i.e., long-term average patterns). To take into account the temporal dimension, one may use Dynamic Bayesian Networks (DBN). This paper uses a static BN and, for more information about DBN, readers are referred to (14).

2.2.1 Variables

Variables used in the proposed BN model are presented in TABLE 1. A total of 12 variables were selected. In this study, we will consider only discrete variables, i.e., nodes that take discrete values. The variables are categorized into three groups: network environment, external event, and traffic condition variables:

- **Network environment variables** represent network-wide environmental factors such as link direction, day-of-week, time-of-day, and weather conditions. Direction (DR) specifies the direction of a link, which in this paper takes two values: \{Southbound, Northbound\} as will be described in the case study below. Day-of-week (D) takes two discrete states: \{Weekday, Weekend\}; Time-of-day (H) takes five states: \{Morning, AM peak, Off-peak, PM peak, Night\}; and Weather (W) takes three states based on rain intensity: \{Clear, Light rain, Moderate rain, Heavy rain\}. The detailed descriptions for the state definitions are presented in in TABLE 1.

- **External event variables** represent events or activities that are external to the traffic stream itself and cause interruptions to traffic flow such as incidents, work-zone, and traffic control signals. In this study, we only include incident factors, defined as three different variables: Incident on a target link (I_O), Incident on upstream links (I_U), and Incident on downstream links (I_D). The incident variables take two states: \{No incident, Incident\}, where the “Incident” state indicates that there is at least one incident occurrence detected during a measurement interval of link traffic parameters.

- **Traffic condition variables** represent link performance measures describing traffic states on a target link. In this study, we include five variables consisting of three basic traffic stream parameters: Flow (F), Occupancy (O), and Speed (S) and two indicators: Level-of-service (L) and Congestion indicator (C). Flow (F), Occupancy (O), and Speed (S) take four discrete states \{Very low, Low, High, Very high\}. These states are defined based on the respective value range of each variable as depicted in FIGURE 1. For each link, we
first plot the diagrams of flow-occupancy-speed relationships and identify the maximum value for each parameter. By dividing the parameter values by the associated maximum value, we obtain normalized flow, occupancy, and speed values, which range between 0 and 1, as shown in FIGURE 1. Then, we divide the range [0, 1] into four equal parts and assign the states {Very low, Low, High, Very high} accordingly. *Level-of-service* ($L$) takes five states {A, B, C, D, E-F}, where level E and level F are merged into one state because the frequency of Level F was too low to be categorized as a separate state. *Congestion indicator* ($C$) is a binary variable that indicates whether the given link is congested or not, which takes two states {Uncongested, Congested}. We use occupancy and flow values to determine the value of $C$: “Uncongested” if occupancy $< \text{Occ}_{\text{crit}}$ and “Congested” if occupancy $\geq \text{Occ}_{\text{crit}}$, where $\text{Occ}_{\text{crit}}$ represents the critical occupancy at which the link flow becomes maximum, as shown in FIGURE 1 (a). We will use this binary congestion indicator $C$ as a main target variable in our BN in performing the congestion diagnosis and prediction analysis.
TABLE 1 Variables and State Definitions for the proposed BN Model

<table>
<thead>
<tr>
<th>Node Group</th>
<th>Variable</th>
<th>Description</th>
<th>States and State Definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>$DR$</td>
<td>Direction</td>
<td>-Southbound</td>
</tr>
<tr>
<td>Network environment</td>
<td>$D$</td>
<td>Day-of-week</td>
<td>-Weekend: Saturday, Sunday</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-Weekday: Monday – Friday</td>
</tr>
<tr>
<td></td>
<td>$H$</td>
<td>Time-of-day</td>
<td>-Morning: 1AM – 6AM (5hrs)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-AM peak: 6AM – 10AM (4hrs)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-Off-peak: 10AM – 4PM (6hrs)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-PM peak: 4PM – 8PM (5hrs)</td>
</tr>
<tr>
<td></td>
<td>$W$</td>
<td>Weather</td>
<td>-Clear: 0mm/h</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-Light rain: &lt; 2.5mm/h</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-Moderate rain: 2.5 – 7.6mm/h</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-Heavy rain: ≥ 7.6mm/h</td>
</tr>
<tr>
<td>Group 2</td>
<td>$I_U$</td>
<td>Incident on upstream links</td>
<td>-No incident</td>
</tr>
<tr>
<td>External event</td>
<td>$I_O$</td>
<td>Incident on a target link</td>
<td>-Incident</td>
</tr>
<tr>
<td></td>
<td>$I_D$</td>
<td>Incident on downstream links</td>
<td>-No incident</td>
</tr>
<tr>
<td>Group 3</td>
<td>$F$</td>
<td>Flow (veh/h/ln)</td>
<td>-Very low: &lt; 0.25</td>
</tr>
<tr>
<td>Traffic condition</td>
<td>$O$</td>
<td>Occupancy (%)</td>
<td>-Low: 0.25 – 0.5</td>
</tr>
<tr>
<td></td>
<td>$S$</td>
<td>Speed (km/h)</td>
<td>-High: 0.5 – 0.75</td>
</tr>
<tr>
<td></td>
<td>$L$</td>
<td>Level-of-service (LOS)</td>
<td>-Very high: ≥ 0.75</td>
</tr>
<tr>
<td></td>
<td>$C$</td>
<td>Congestion indicator</td>
<td>-Uncongested: occupancy &lt; $\hat{\text{Occ}}_{\text{crit}}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-Congested: occupancy ≥ $\hat{\text{Occ}}_{\text{crit}}$</td>
</tr>
</tbody>
</table>

* based on normalized parameter values, which range between 0 and 1
FIGURE 1 Example of data discretization for Flow (F), Occupancy (O), Speed (S), and Congestion indicator (C) variables.

2.2.2 Structure
Once the nodes of the BN are specified, the next step is to specify qualitative relationships between variables. While the structure of BN can be learned from data by searching for the best model using various learning algorithms (1, 15–18), this paper adopts the approach of manually specifying a set of alternative model structures and selecting the best model from them. This allows us to incorporate our knowledge on the variables into the model building process and compare different configuration assumptions more clearly. To develop candidate models systematically, we first determine relationships between node groups and then apply the group-level relationships to the individual nodes. For instance, if we believe that Group 1 affects (or cause) Group 2, then we add arcs from all nodes in Group 1 to all nodes in Group 2. Given the three node groups defined above, two observations can be made as follows:

- Group 1 (Network environment) affects Group 2 (External event) or Group 3 (Traffic condition), but not the other way around (e.g., weather can cause incident but incident cannot cause weather event).
- Group 2 (External event) and Group 3 (Traffic condition) are dependent each other, but the causal relationship can be either direction, i.e., either Group 2 affects Group 3 or Group 3 affects Group 2.

Based on these observations, we can identify a total of seven possible network configurations as shown in FIGURE 2.

FIGURE 2 Seven possible configurations for the proposed BN model.
Among these seven configurations, we consider type \((g)\) in FIGURE 2 as our main model structure as it seems to be reasonable to assume that the environment variables directly affect both incident occurrence and traffic conditions while there will be also a direct influence of incident occurrence on link traffic. To keep the model structure simple, we assume that variables within each node group are conditionally independent given their common parent or common descendant, i.e., no direct arcs between nodes within the group. This is apparent in Group 1 as variables \(DR, D, H,\) and \(W\) do not affect one another. For Groups 2, this indicates that the incident occurrence on a link is influenced by the incident occurrence on its upstream or downstream link only through a third variable such as network environment and traffic condition, and this seems to be a reasonable assumption. The conditional independence assumption is also viewed acceptable for the variables in Group 3 as those are all different measures of the same link traffic condition and it is not assumed that one variable causes another within this group. The final graph representation for the proposed BN model is thus as shown in FIGURE 3. The model validation results are presented in the later section.

**FIGURE 3** Selected graph representation for the proposed BN model (configuration type \(g\)).

2.2.3 Parameter Learning

Once the structure of the BN is determined, the next step is to quantify the relationships between connected nodes. This is done by computing a conditional probability distribution for each node, considering all the possible combinations of values of its parent nodes. For discrete variables, conditional probability distribution is expressed in the form of conditional probability table (CPT), where each element of the table represents the probability that a given variable take a particular value given a particular combination of its parent node values. For instance, the CPT of node \(F\)
represents the probability values of all possible configurations of \( P(F = f | D = d, H = h, W = w, I\_U = iu, I\_O = io, I\_D = id) \). For a node that does not have parents, the CPT becomes the marginal distribution of its own. The size of CPT can get very large if a node has many parents or if the parents can take many states. Thus, in many real-world applications it is not feasible to manually calculate CPTs and instead machine learning techniques are used to automatically learn the CPTs from data.

3. EXPERIMENTAL SETUP AND RESULTS

3.1 Study Site

This section presents a case study implementing the proposed BN approach. The study area is Brisbane, Australia and the study sites are 19 highway links selected from Pacific Motorway in Brisbane as shown in FIGURE 4. TABLE 2 presents the basic information about these two links.

<table>
<thead>
<tr>
<th>Link</th>
<th>Location</th>
<th>Direction</th>
<th>Length (m)</th>
<th>Lanes</th>
<th>Design speed (km/m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Pacific Motorway</td>
<td>Southbound</td>
<td>143</td>
<td>2</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>Pacific Motorway</td>
<td>Southbound</td>
<td>361</td>
<td>3</td>
<td>70</td>
</tr>
<tr>
<td>3</td>
<td>Pacific Motorway</td>
<td>Southbound</td>
<td>334</td>
<td>3</td>
<td>70</td>
</tr>
<tr>
<td>4</td>
<td>Pacific Motorway</td>
<td>Southbound</td>
<td>346</td>
<td>4</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>Pacific Motorway</td>
<td>Southbound</td>
<td>520</td>
<td>4</td>
<td>100</td>
</tr>
<tr>
<td>6</td>
<td>Pacific Motorway</td>
<td>Southbound</td>
<td>642</td>
<td>3</td>
<td>90</td>
</tr>
<tr>
<td>7</td>
<td>Pacific Motorway</td>
<td>Southbound</td>
<td>1116</td>
<td>3</td>
<td>100</td>
</tr>
<tr>
<td>8</td>
<td>Pacific Motorway</td>
<td>Southbound</td>
<td>1177</td>
<td>3</td>
<td>100</td>
</tr>
<tr>
<td>9</td>
<td>Pacific Motorway</td>
<td>Southbound</td>
<td>2500</td>
<td>3</td>
<td>100</td>
</tr>
<tr>
<td>10</td>
<td>Pacific Motorway</td>
<td>Southbound</td>
<td>1925</td>
<td>3</td>
<td>100</td>
</tr>
<tr>
<td>11</td>
<td>Pacific Motorway</td>
<td>Northbound</td>
<td>1988</td>
<td>3</td>
<td>100</td>
</tr>
<tr>
<td>12</td>
<td>Pacific Motorway</td>
<td>Northbound</td>
<td>2047</td>
<td>3</td>
<td>100</td>
</tr>
<tr>
<td>13</td>
<td>Pacific Motorway</td>
<td>Northbound</td>
<td>1333</td>
<td>3</td>
<td>100</td>
</tr>
<tr>
<td>14</td>
<td>Pacific Motorway</td>
<td>Northbound</td>
<td>1049</td>
<td>4</td>
<td>90</td>
</tr>
<tr>
<td>15</td>
<td>Pacific Motorway</td>
<td>Northbound</td>
<td>612</td>
<td>3</td>
<td>90</td>
</tr>
<tr>
<td>16</td>
<td>Pacific Motorway</td>
<td>Northbound</td>
<td>580</td>
<td>4</td>
<td>100</td>
</tr>
<tr>
<td>17</td>
<td>Pacific Motorway</td>
<td>Northbound</td>
<td>283</td>
<td>4</td>
<td>100</td>
</tr>
<tr>
<td>18</td>
<td>Pacific Motorway</td>
<td>Northbound</td>
<td>329</td>
<td>3</td>
<td>70</td>
</tr>
<tr>
<td>19</td>
<td>Pacific Motorway</td>
<td>Northbound</td>
<td>116</td>
<td>3</td>
<td>70</td>
</tr>
</tbody>
</table>
3.2 Data Collection and Discretization

Traffic and incident data were obtained from the Queensland Department of Transport and Main Road (DTMR) through Public Traffic Data system (PTDS), which provide link measures including flow, occupancy, speed and level-of-service from available loop detectors around South East Queensland (SEQ) network and the list of incidents reported DTMR’s incident management system. Link measures in PTDS traffic data are recorded every 3 minutes. Weather data were received from the Bureau of Meteorology station located within the study area. Weather data reports weather parameters including precipitation, visibility, temperature, and humidity, which are recorded every 30 minutes. For the case study, traffic, incident, and weather data were collected from 608 days between 2011-01-01 and 2013-09-30. These data are used to learn the parameters of the BN model in FIGURE 3. In order to match the data with the model, the data should be organized in a matrix form, where columns represent 12 variables defined in TABLE 1 and rows represent observations for these variables for the entire study period covering all the study links. To do this, we map weather and incident data to traffic data such that each 3-min traffic observation (link measures) has its associated weather and incident information attached. After
fusing data in this way, we obtained a 4,787,932-by-12 matrix for all 19 links, where each row represent a 3-min traffic, incident, and weather observation. Since the data are originally numerical and continuous, the data were discretized on the categorical states defined in TABLE 1. After the discretization, we obtain the matrices of discrete (categorical) data converted from the original numeric matrices. An example of the processed data matrix is provided in FIGURE 5.

<table>
<thead>
<tr>
<th>DR</th>
<th>D</th>
<th>H</th>
<th>W</th>
<th>I_U</th>
<th>I_O</th>
<th>I_D</th>
<th>F</th>
<th>O</th>
<th>S</th>
<th>L</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Southbound</td>
<td>Weekday</td>
<td>PM_peak</td>
<td>Clear</td>
<td>No_incident</td>
<td>No_incident</td>
<td>No_incident</td>
<td>High</td>
<td>Low</td>
<td>High</td>
<td>C</td>
<td>Uncongested</td>
</tr>
<tr>
<td>Southbound</td>
<td>Weekday</td>
<td>PM_peak</td>
<td>Clear</td>
<td>No_incident</td>
<td>No_incident</td>
<td>No_incident</td>
<td>High</td>
<td>Low</td>
<td>High</td>
<td>D</td>
<td>Uncongested</td>
</tr>
<tr>
<td>Southbound</td>
<td>Weekday</td>
<td>PM_peak</td>
<td>Clear</td>
<td>No_incident</td>
<td>No_incident</td>
<td>No_incident</td>
<td>High</td>
<td>Low</td>
<td>High</td>
<td>C</td>
<td>Uncongested</td>
</tr>
<tr>
<td>Southbound</td>
<td>Weekday</td>
<td>PM_peak</td>
<td>Clear</td>
<td>No_incident</td>
<td>No_incident</td>
<td>No_incident</td>
<td>High</td>
<td>Low</td>
<td>E_F</td>
<td>Congested</td>
<td></td>
</tr>
<tr>
<td>Southbound</td>
<td>Weekday</td>
<td>PM_peak</td>
<td>Clear</td>
<td>No_incident</td>
<td>No_incident</td>
<td>No_incident</td>
<td>High</td>
<td>Low</td>
<td>E_F</td>
<td>Congested</td>
<td></td>
</tr>
</tbody>
</table>

FIGURE 5 Example of processed data.

3.3 Model Validation

Once the model is built, it is important to assess how good the selected BN model is (i.e., how well the model describes the underlying data), compared to the other alternatives. We perform two different tests for assessing the goodness-of-fit of a BN: (i) measuring a network-level scoring function from the fitted BN model and (ii) measuring a variable-specific classification error by performing cross-validation. There are a number of scoring functions available for assessing a BN structure (19). We use four well-known scores: Log-likelihood, Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC), and K2 (20). For measuring classification error, we perform k-fold cross-validation, where the original dataset is randomly partitioned into k-equal sized subsamples and the model is trained using k-1 subsamples (called training set) and then validated on the remaining one sample (called testing set). This procedure is repeated k times. The classification error for a particular variable is measured as the percentage that the BN model predicts a wrong state for this variable during testing, averaged over the k folds. We used k=5 in this study. FIGURE 6 shows the network scores and classification errors obtained from the seven BN structures in FIGURE 2 including the one selected for this study (type g). FIGURE 6 (a) presents the network scores, where the higher the score, the better the model fits the data. Overall, the scores are low in a and b and high in c, d, e, f, and g. By closely looking at the numbers, we find that the score of g is the highest among the seven in AIC and K2, the second highest in Log-likelihood, and the third highest in BIC. FIGURE 6 (b) presents the classification errors tested for five variables F, S, L, O, and C, where the lower the error, the better a learned model predicts the state of a variable. The classification error for F shows the highest variability across models, where the error is the highest in a and b and the lowest in f and g. By comparing the numbers, we find that the error rate produced by g is the lowest among the seven models when predicting S, L, and, C and the second lowest when predicting F and O. From these results, we conclude that the model structure selected for this study, g, describes the underlying data best, among the tested seven structures.
FIGURE 6 Comparing the goodness-of-fits of seven different BN structures based on (a) network scores (the higher, the better) and (b) classification errors for target variables F, O, S, L, and C (the lower, the better).

3.4 Parameter estimation results
The parameters of the BN, i.e., the CPT for each node, were learned from the data that contain a total of 4,787,932 observations from all 19 links. In this paper, we used the R software package bnlearn (21) for parameter learning and probabilistic inference. FIGURE 7 presents the estimation result for the marginal probability distribution of each node. The distributions of direction (DR), day-of-week (D), and time-of-day (H) are consistent with what can be expected from the state definitions. The probability that the weather (W) is clear is 93.3% and rain events occur with the probabilities of 5.2% for light rain, 1.0% for moderate rain, and 0.4% for heavy rain, respectively. For any given link, the probability of incident occurrence (I_O) is 0.44%. The probabilities of having an incident on its upstream link (I_U) and downstream link (I_D) are 0.43% and 0.44%, respectively. Since all target links are upstream or downstream links of one another, the marginal distributions of these three incident variables should be very similar. From the distributions of flow (F), occupancy (O), and speed (S), we observe that O is heavily skewed compared F and S, showing that 90.9% of the time the occupancy is very low and the percentage of high or very high is 1.7%. Based on Congestion indicator (C), the probability that any given link within the study corridor is congested is 4% and the probability of uncongested is 96%.
FIGURE 7 Marginal probability distributions for 12 variables in the proposed BN model.

4. ANALYSIS
In this section, we present a number of analysis methods to identify factors that affect traffic congestion using the BN model constructed above. The main variable of interest is congestion indicator $C$ and its parent nodes $DR, D, W, H, I_U, I_O$, and $I_D$ are potential causes of congestion or congestion factors, which will be called scenario variables. The focus of our analysis is on understanding the relationships between the target variable and the scenario variables by performing various diagnostic and predictive reasoning tasks.

4.1 Identifying Leading Causes for Congestion Diagnosis
Using the BN model, we can perform diagnostic reasoning, i.e., reasoning from effect (congestion occurrence) to cause (scenario variables). FIGURE 8 shows the posterior probability distribution of each scenario variable $S$ given that congestion has been observed, i.e., $P(S|C=$ congested$)$. Notice that the probability of $C=$ congested is 100% in FIGURE 8, meaning that we have updated our belief about the congestion state, e.g., we have observed the traffic congestion, and there is no uncertainty in variable $C$. Compared to the prior distributions $P(S)$ in FIGURE 7, the following changes have been made on the distributions of scenario variables:

- [DR] The probability that a link is northbound when congestion has been observed is 56.3%, increased from 47.5% when there is no information about the congestion state.
- [D] The probability that it is weekday has increased from 71.1% to 97.1%.
• [H] The probability of being PM-peak has increased from 17% to 49% and that of AM-peak has increased from 17% to 38.1%.
• [W] The probability that it is raining (regardless of the rain intensity) has increased by 3.1%.
• [I_U, I_O, I_D] The probabilities that there exists an incident on the downstream link, the current link, and the upstream link have increased from the level of 0.44% to 1.85%, 1.41%, and 1.00%, respectively.
• [F] The probability that the flow rate is very high has increased from 5.6% to 26.8%.
• [O] The probability that the occupancy is high or very high has increased from 1.7% to 7.9%.
• [S] The probability that the speed is very low or low has increased from 4.1% to 82.6%.
• [L] The probability that the level-of-service is E-F has increased from 3.6% to 15.6%.

![Graph showing posterior distributions of scenario variables S given that congestion has been observed](image)

**FIGURE 8** The posterior distributions of scenario variables $S$ given that congestion has been observed, i.e., $P(S|C=\text{congested})$.

We can further quantify how strongly a particular scenario state is associated with the congestion occurrence by measuring odds ratio. The odds of an event occurring is the ratio of the probability that the event will happen to the probability that the event will not happen. The odds ratio (OR) is the ratio of the odds of an event occurring in one group to the odds of it occurring in another group, providing a way to measure how strongly the event is associated with the first group compared to
the second group. For instance, if we consider the event of “congestion” and compare its odds
between two groups: “rain” and “no rain”, then the odds ratio is:

\[ OR = \frac{P(\text{congestion}\mid \text{rain})/P(\text{no congestion}\mid \text{rain})}{P(\text{congestion}\mid \text{no rain})/P(\text{no congestion}\mid \text{no rain})}. \]

An odds ratio greater than 1 (less than 1) indicates that the event is more likely (less likely) to
occur in the first group. In the above example, OR < 1 indicates that the “congestion” is more
likely to occur when it is “rain” compared to when there is “no rain”. The fact that the distribution
of OR is highly skewed (the odds ratio is limited to zero at the lower end but unlimited at the upper
end), however, makes it difficult to compare the strength of association across two regimes
0<OR<1 and OR>1. To overcome this, the logarithm of the odds ratio (logOR) is used to
convert the scales of “less-likely” and “more-likely” regimes from (0, 1) and (1, ∞) to (−∞, 0) and
(0, ∞), respectively.

By denoting the event of having congestion and the event of having a particular scenario state
simply by C and S, respectively, the log odds ratio (logOR) for our problem is defined as follows:

\[
\text{logOR} = \log \left( \frac{P(C \mid S)/P(\sim C \mid S)}{P(C \mid \sim S)/P(\sim C \mid \sim S)} \right) = \log \left( \frac{P(C,S) \cdot P(\sim C, \sim S)}{P(C, \sim S) \cdot P(\sim C, S)} \right),
\]

where \( P(S) \) is the probability of scenario event \( S \) occurring (e.g., \( P(W=\text{clear}) \)), \( P(\sim S) \) is the
probability that the scenario event does not occur (e.g., \( P(W=\text{clear}) \)), \( P(C) \) is the probability that
the congestion occurs, i.e., \( P(C=\text{congested}) \), and \( P(\sim C) \) is the probability that the congestion does
not occur, i.e., \( P(C=\text{not congested}) \). The odds ratio can also be defined in terms of the joint
probabilities, as shown in the last term in Eq. (1), where the expression becomes the product of the
probability that both \( C \) and \( S \) occur and the probability that both \( C \) and \( S \) do not occur divided by
the product of the probabilities that only one of them occur. If the logOR is greater than 0, then
having scenario event \( S \) is considered to be "associated" with having congestion \( C \) and events \( S \)
and \( C \) are more likely occur together. If the logOR is less than 0, event \( S \) and \( C \) are less likely occur
together. All the probability values required to compute logOR can be obtained from the BN model
and the results are presented in FIGURE 9. The highest logOR is found in scenario event
“D=Weekday” which is 1.15, indicating that weekday and congestion occurrence is very strongly
associated (highly likely to occur together). The events “H=PM-peak” and “I_D=Incident” is also
strongly associated with congestion occurrence, with the log odds ratios of 0.71 and 0.70,
respectively. The next highest logORs are found in “I_O=Incident” and “H=AM-peak” with the
values of 0.56 and 0.51, respectively. Scenario events “DR=Southbound, D=Weekend,
H=Morning H=Off-peak, H=Night, W=Clear, I_O=No incident, I_D=No incident, I_U=No
incident” are all showing the log odds ratios less than 0, indicating that these events are less likely
to occur together with the congestion event.
FIGURE 9 The log odds ratio between each scenario event and congestion occurrence.

4.2 Identifying Critical Scenarios for Congestion Prediction

So far, we have investigated the relationships between scenario variables and congestion variable focusing on individual scenario variables separately. It is, however, possible to consider all seven scenario variables simultaneously to quantify their impacts on congestion occurrence. Now assume that scenario event $S$ is the combination of all seven scenario variables, e.g., $S = \{ DR=\text{Southbound}, D=\text{Weekday}, H=\text{AM-peak}, W=\text{Clear}, I_U=\text{No incident}, I_O=\text{Incident}, I_D=\text{No incident} \}$. We are interested in identifying the most important scenarios that are highly associated with congestion event $C$, namely, scenarios that have high joint probabilities with the congestion event, $P(S, C)$. While we can compute all possible combinations of $P(S, C)$ and identify the scenarios with highest joint probabilities, the scenarios identified in this manner might include very rare scenarios because $P(C,S) = P(S) \times P(C | S)$ and the combination of very low $P(S)$ and very high $P(C | S)$ can still produce high $P(C,S)$. A better strategy would be to identify scenarios that produce both high $P(S)$ and high $P(C | S)$. This is similar to the concept of identifying risk scenarios in risk analysis. In quantitative risk analysis, risk is often expressed as:

$$\text{Risk} = \text{Probability} \times \text{Impact}.$$ 

The relation $P(C,S) = P(S) \times P(C | S)$ can be interpreted in the similar manner as follows:
where $P(C,S)$ represents the overall importance or risk of scenario $S$, $P(S)$ is the likelihood of the scenario, and $P(C|S)$ is the impact of the scenario expressed in terms of the probability that congestion occurs given this scenario occurring. Based on this framework, we now create the Impact-Probability chart to identify high probability-high impact scenarios (i.e., high $P(S)$ and high $P(C|S)$) as shown in FIGURE 10. This chart allows us to rate potential risks or importance of a scenario on two dimensions and select appropriate cut-off points for $P(S)$ and $P(C|S)$ based on the number of scenarios that we wish to include in the final scenarios set. In this study, we select top 40 important scenarios from the High Probability and High Impact region in the chart and the boundaries for this region were selected as $P(S) \geq 0.0003$ and $P(C|S) \geq 0.0085$, as depicted in FIGURE 10 (b). The selected 40 scenarios are presented in FIGURE 11 in the form of scenario tree. The column chart in the rightmost column of the figure visualizes the relative magnitude of each of $P(S)$, $P(C|S)$, and $P(C,S)$ across 40 scenarios, allowing us to easily find the ranking of a scenario within the selected scenario group. For instance, scenario #5 has relatively high $P(S)$ but low $P(C|S)$ and scenario #15 has relatively low $P(S)$ but high $P(C|S)$ within those 40 scenarios. The resulting $P(C,S)$ are relatively high in both scenarios.

Identifying critical scenarios that affect traffic performance is an important task in many decision-making situations for transportation planning and operations. The proposed approach provides a systematic framework to rank and prioritize important scenarios and can be used in various scenario analysis applications such as scenario-based travel time reliability analysis and simulation modelling (22, 23).

**FIGURE 10** Probability-Impact chart for identifying high probability-high impact scenarios in terms of scenario probability $P(S)$ and scenario impact $P(C|S)$. 

\[
\text{Scenario Risk} = \frac{\text{Probability of Scenario}}{P(S)} \times \text{Impact of Scenario} = \frac{P(C,S)}{P(S)} \times P(C|S) \text{,}
\]
**FIGURE 11** Scenario tree of top 40 important scenarios that have high association with congestion occurrence.
5. CONCLUSIONS
Managing traffic congestion requires accurate knowledge of the causes of congestion on a given road network as well as their relative significance and respective solutions. Previous research has focused on estimating the relative impact of different causes of congestion, which include incidents, weather, work-zone, and special event, using statistical analysis such as linear regression methods. Traditional statistical methods, however, have limitations in capturing complex dependencies and uncertainties in external events and traffic states in urban networks. This study proposes a Bayesian Network (BN) analysis approach to modelling the probabilistic dependency structure among causes of congestion on a particular road segment and analyzing the probability of traffic congestion given various roadway condition scenarios. A BN is used to encode the joint probability distribution over a set of random variables that describe scenario variables, which represent factors affecting the congestion level of a target segment such as time-of-day, incident, weather, and traffic states on adjacent links, and output variables, which represent traffic performance measures of the target segment such as flow, density, and speed. A properly configured BN model can be used to (i) quantify the contribution of each cause or the combination of multiple causes to traffic congestion, thereby allowing the identification of leading causes for the purpose of congestion diagnosis, (ii) predict future congestion levels based on the current network situations, and (iii) analyze the likely scenarios (combinations of causes) that produce worst traffic congestion in a study network and their occurrence probabilities. This study develops a methodology to build a BN model based on historical traffic and event data, and demonstrates the BN-based traffic analysis using a study network in Queensland, Australia. The study discusses applications of the proposed BN model in urban traffic congestion management, focusing on its capability to provide a comprehensive data-driven and probabilistic analysis platform for congestion diagnosis and prediction.

ACKNOWLEDGMENT
This work was funded by the University of Queensland through the New Staff Research Start-up Fund (NSRSF).
REFERENCES


