Root Cause Analysis

Introduction

Congestion modeling is a critical tool for finding congestion propagation patterns in a road network. This can give us insight into understanding the sources of congestion that aren't immediately obvious or distinguishable. This information can be leveraged in traffic signal control systems for more robust traffic management.

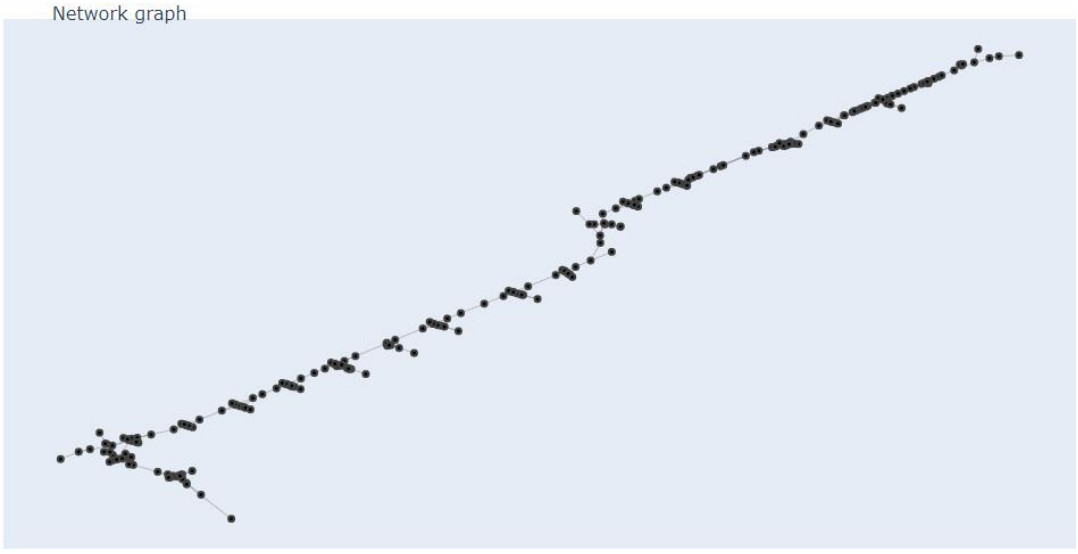
We propose a Dynamic Bayesian Network (DBN) for congestion modeling. A DBN allows for extremely customizable, and useful queries that go beyond vanilla "prediction" models. What sets the DBN apart from traditional regression models is the ability to determine probabilities of congestion in the past, given future congestion. This is formally referred to as "root cause analysis".

Network Building

Only using immediate neighbours in congestion analysis is quite trivial, and something we wouldn't need a DBN to analyze. It is much more interesting to look at a larger number of parents to find patterns that aren't immediately obvious to someone observing the data.

It's important to cast a wider net and take into account the effects of surrounding links at further timesteps and farther distances.

To do this, a road network graph is constructed. Using a path algorithm, all of the downstream neighbours are found for each link within a given distance. From here, the parents of each link is determined in order to start building the DBN.



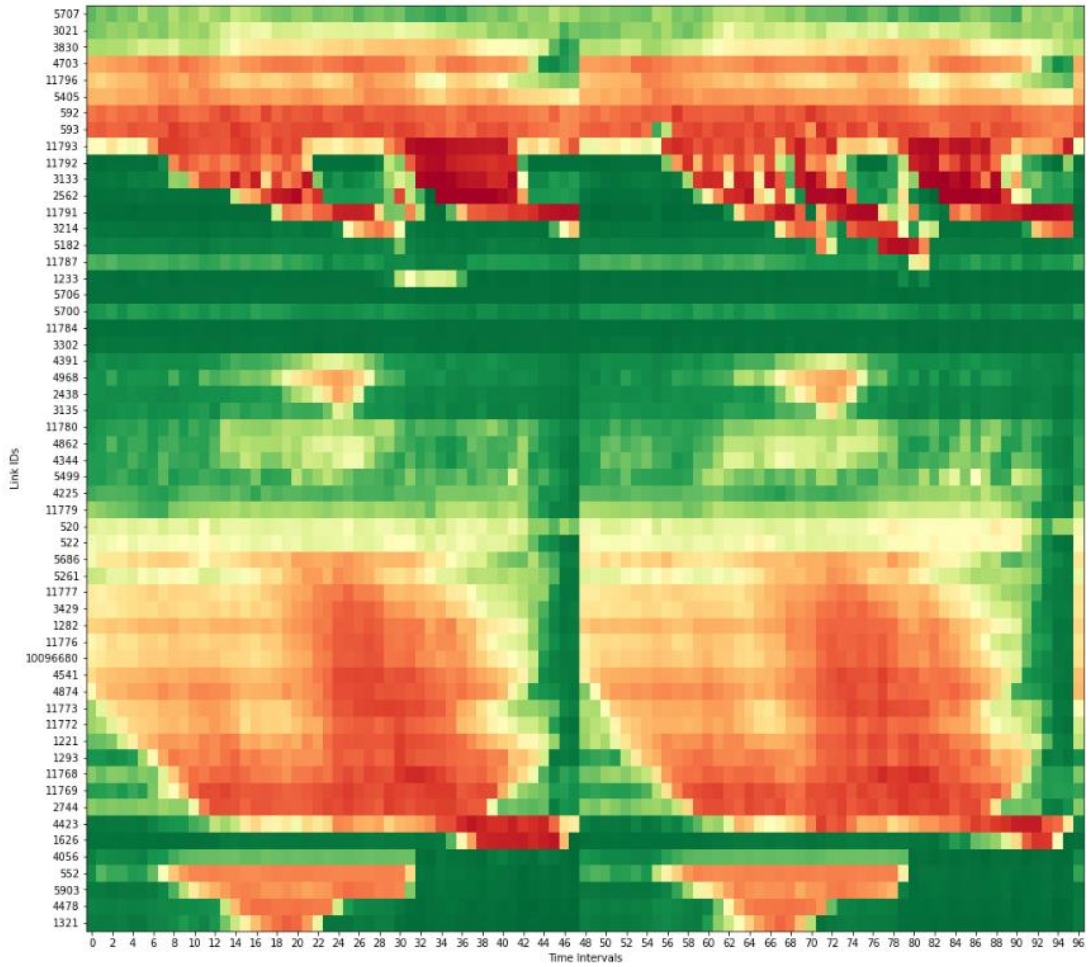
Data Processing

Once the speed data is collected, we have to convert it so something more manageable in our model. For both simulation and collected data, traffic speed is a continuous variable, however we would like discrete variables for simplification in the DBN model. Using the traffic data and speed limits of each road, we find the relative speed of each data point. If the relative speed falls below 0.5, the link segment is considered

congested, otherwise uncongested. Different congestion categories could be defined in future research to go beyond this binary model.

TrafficData is the base class for processing and storing the speed data, and converting it to binary congestion values. In this example, the data is sourced from simulations in Aimsun, therefore AimsunData class is used.

Data Exploration



Using the heatmap, some links of interest are chosen to perform root cause analysis on. These are links which are congested during the simulation that we would like to analyze which links are causing this congestion.

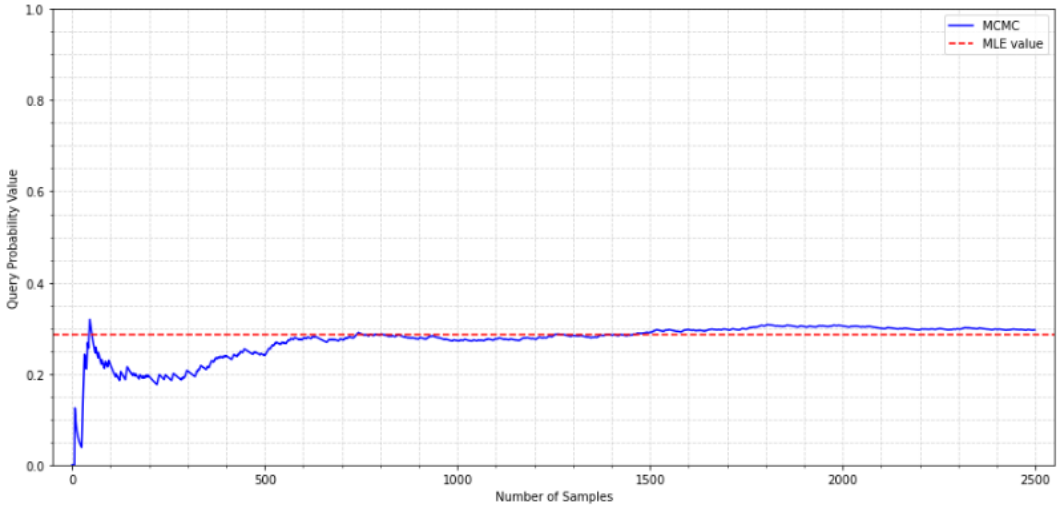
Dynamic Bayesian Network

Conventional PGM libraries (such as pgmpy) cannot handle exact inference algorithms on even our simplest model. As models become more complex such as the GTHA, some nodes could have hundreds of parents, so conditional probability tables cannot be used. A link with 100 parents would have a conditional probability table with 2^100 rows. Instead, we use Logistic regression models to define these conditional probability distributions for each link.

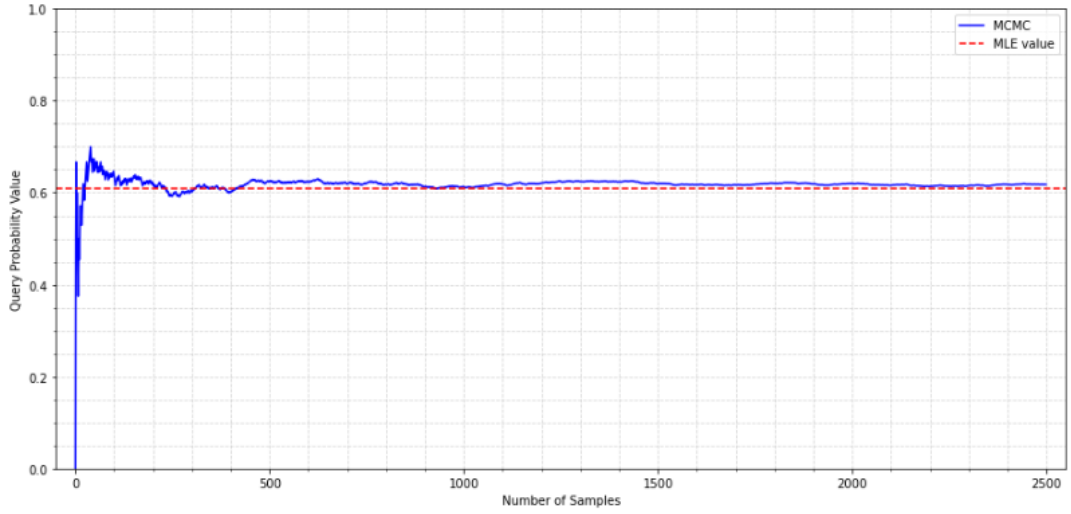
The DynamicBayesianNetwork class is used here to build the DBN using the road network structure and the processed data and define all marginal and conditional probabilities of the DBN.

It is imperative that the DBN method is tested on a variety of test cases to verify the validity of the model. In the following test cases, a simpler DBN is defined with a small number of parents, looking only at two timesteps. The small number of parents is necessary so that the results from each query can be compared to the actual value by calculating the maximum likelihood estimation using cpd tables.

In the graph below, we have the probability that link ‘520’ is congested at t=0, given no evidence



In the graph below, we have the probability that link ‘11768’ is congested at t=1, given that link '11768' is congested at t=0 and link '11769' is uncongested at t=0.



As you can tell from the plots above, the gibb's sampler roughly converges the approximated probability to the actual value. This validates the use of logistic regression and the gibb's sampler to perform root cause analysis.

Root Cause Analysis

At a very broad level, there are two types of queries we are mainly interested in: prediction, and root cause. Using our DBN we can make these probabilistic queries to determine conditional probabilities.

Prediction Queries

Probability of a link i being congested at time t, given that link j was congested at time t-5:

$$P(cong_{i,t} \mid cong_{j,t-5} = true)$$

Root Cause Queries

Probability of a link i being congested at time t, given that link j will be congested at time t+5:

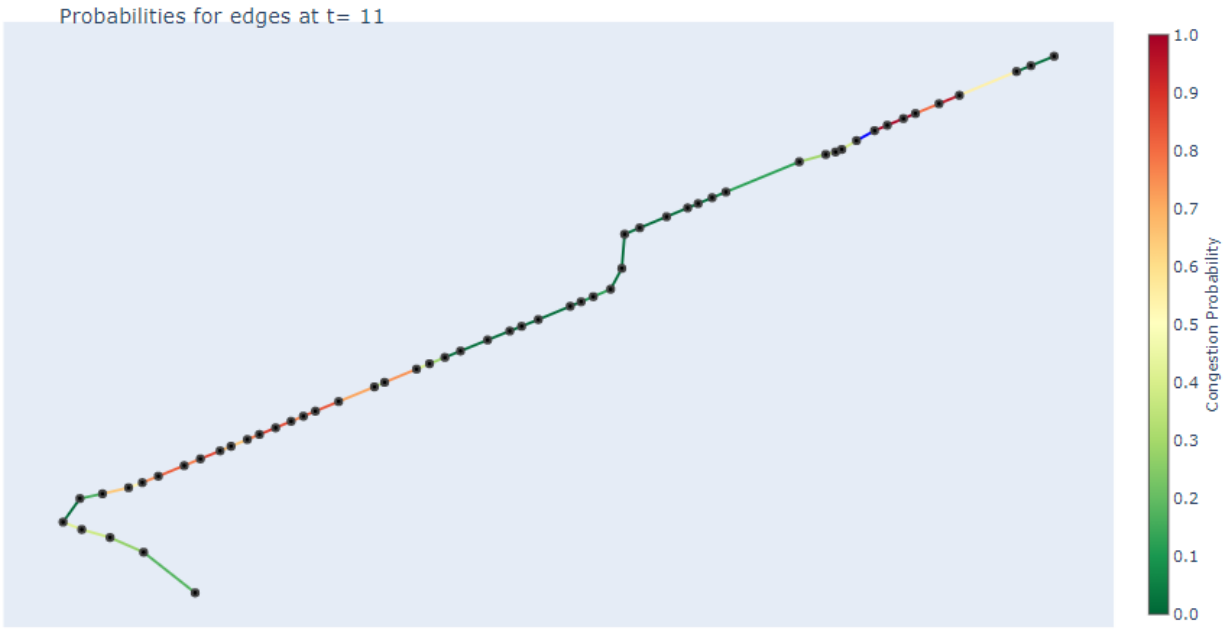
$$P(cong_{i,t} \mid cong_{j,t+5} = true)$$

Once we determine these conditional probabilities, there is information we can gather:

1. If we find a probability of 0.5, this means there's a 50-50 chance the link is congested given the congestion state of another link. Meaning, there is no correlation between these two links
2. As the probability approaches 1, there is a significantly strong positive correlation
3. As the probability approaches 0, there is a significantly negative correlation

Gibb's sampling is used to approximate the joint distribution when each of the links of interest are congested at t=12. We then make queries to find the probability of congestion for each link at all timesteps before t=12.

The plot below shows the probability of congestion for each link given the evidence. As you can see, there are several links that have a very high probability, however this does not tell us much in regards to root cause of our evidence link shown in blue. This is because, if a link already is congested most of the time as seen in the data, then it will have a high probability given the evidence. This does not necessarily mean the high probability of the link is correlated to the evidence, but that the link has an intrinsic property of being congested most of the time.



Therefore, an alternate measure is important to analyze. It is much more interesting to find the relative probability for each link. Comparing a link's probability of congestion given the evidence, to some baseline will give us more information to how the link and the evidence are correlated.

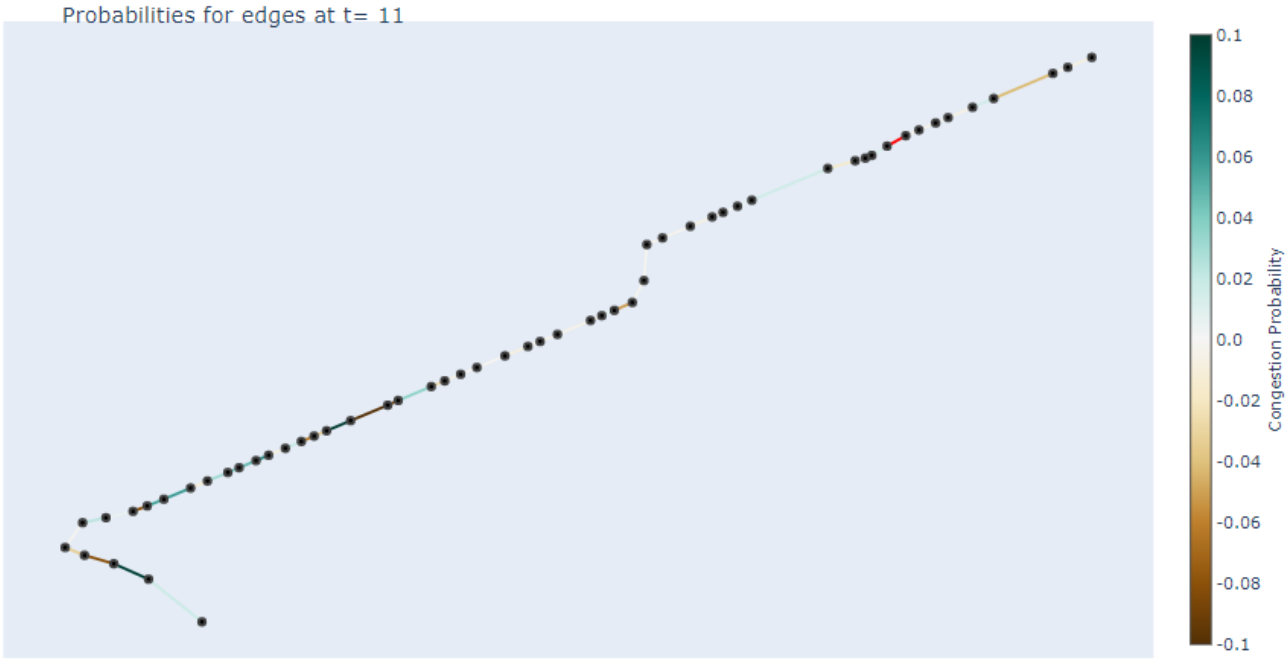
The plot below shows how each link in the network has an effect on the evidence link, shown in red. The probabilities are calculated by finding the joint distribution when the evidence is congested (p) and when it's uncongested (p2). We then find the difference and plot.

p = P(congi,11 | congj,12 = true)

p2 = P(congi,11 | congj,12 = false)

diff = p - p2

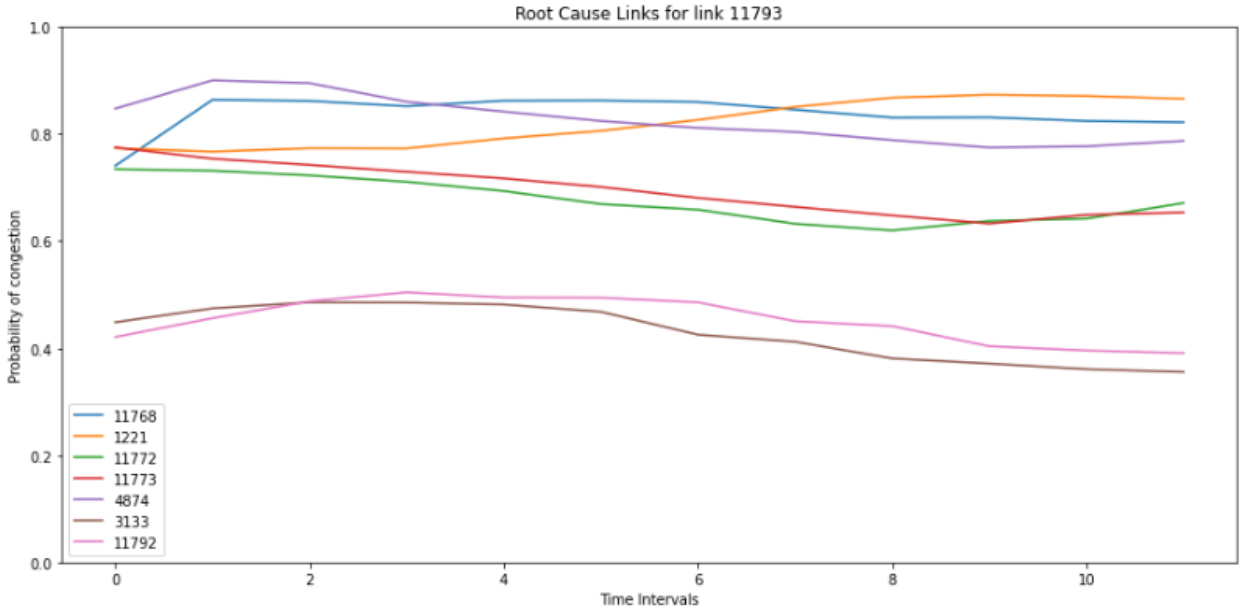
A negative value means the link has a negative correlation with the evidence and a positive value means the link has a positive correlation with the evidence. In this plot, the darker the colour of the link, the stronger the correlation is.



The probability differential tells us that links 4478 and 3429 have a strong positive correlation with the evidence at t=11 and may be a root cause for congestion. Alternately, the links 5903, 2744, 11777, and 5261 have a strong negative correlation with the evidence at t=11.

Another approach to analyzing root cause is to see how the probability of congestion for each link changes over time, given the evidence. The probability for congestion for most links does not change with each timestep. This makes sense, as the conditional probabilities of every link are stationary. Meaning, conditional probabilities for a link at t=20 is the same at t=2. Therefore, if a link's probability of congestion does have a significant change over time, this may be the result of the evidence link being congested.

The links with a change in probability of 10% or more were found. The probability of congestion at each time step was then plotted as shown below:



None of the links found seem to agree with the previous analysis, suggesting that more information may be needed to do an accurate temporal analysis.

Conclusions and recommendations

The results of the root cause analysis seem to be inconclusive, however there are no real metrics to benchmark these results against. There are several limitations which may introduce error into the model.

Limitations in the data.

There is not enough variety in congested/uncongested states for each link. A link that is either congested or uncongested 99% of the time will skew the results and are much more difficult to analyze. Running simulations in a wider variety of traffic scenarios may assist this.

Definition of congestion

Congestion is modelled as binary, given an arbitrary threshold value. A binary model is not very representative of the real world state of traffic. It's recommended to model congestion with several traffic states, and eventually use a continuous model.

Stationary

The model is assumed to be stationary, meaning the conditional probabilities of each link do not change with time. This is a massive oversimplification of the model, as time of day has a large effect on whether a link will be congested. It's suggested that a time component, such as a clock variable be introduced to the model so that queries can be made in both the spatial and temporal dimensions of the data.

Road Network Simplification

The QEW road network structure was simplified, as this model only considers highway sections. No speed data was collected for on/off ramps which can be a major source of congestion. We may get better results by introducing more links to the model.