Bayesian Network Project Report

Parvaneh Soleimany

This work implements the BN introduced in the paper [1].

1. Introduction

The dangerous goods belong to an important class of chemical raw materials and industrial materials. The dangerous goods can be classified into nine categories including explosives, gas, flammable liquids, flammable solid, oxidants and organic peroxides, toxic and infectious materials, radioactive material, corrosive substance and others. These goods have some unique physical and chemical properties, such as explosive, corrosive, toxic, radioactive etc. Hence, the transportation and storage of dangerous goods must be handled under strict rules and regulations. As one of the most important dangerous goods transport modes in China - the case study of the selected paper- railway transportation mode moves more than 36% of the total dangerous goods volume each year, and also has a higher accident rate. In order to prevent the leakage, fire, explosion, poisoning and burn accidents in the railway dangerous goods transportation system (RDGTS) as much as possible, it is necessary to analyze the past typical accident events, investigate the relationships among these risk factors that caused the accidents in the system and take relevant measures to control risks and prevent accidents. This work aimed at analyzing the relationships and interaction strengths among the risk factors or accident causes of RDGTS using Bayesian network.

2. Data

Transportation Committee of China Railway Enterprise Management Association analyzed the past typical events and selected some basic risk factors that influence the safety of the system based on the accidents statistical data. In general, the risk factors can be classified into five categories, include risk factors of Human (H), risk factors of Machine (M1), risk factors of Materials (M2), risk factors of Environment (E) and risk factors of Management. Furthermore, the five risk factors can be divided into various sub-indicators after analyzing the statistical historical data of railway dangerous goods transportation accidents (the historical data of Chinese railway dangerous goods transportation accidents from 1986 to 2017). We use ui to represent each *i* sub-indicator of the risk factor, the description of the five classifications as well as the sub-indicators are shown in Table1.

Table. 1

The risk factors (RF), sub-indicators (SI) and their incidents frequencies (F) and percentages (P) from 1986 to 2017

RF	SI	F	P of SI	P of RF
H	*u*1: physical discomfort and poor working environment of the staffs	175	6.4%	44.5%
	*u*2: inaccurate work attitude and operation of the staffs	190	7.0%	
	*u*3: staffs are lack of technical and knowledge during transportation processes	386	14.1%	
	*u*4: staffs of goods manufacturer illegally overload or entrain the goods	269	9.8%	
	*u*5: non-railway personnel's illegal dangerous goods stealing	223	8.2%	
	*u*6: failure of transportation equipment	192	7.0%	
M1	*u*7: failure of dangerous goods storage equipment	137	5.0%	17.4%
	*u*8: failure of loading and unloading equipment in the handling stations	146	5.3%	
	*u*9: dangerous nature of the loaded and transported goods	181	6.6%	
M2	*u*10: packaging of dangerous goods	117	4.3%	16.4%
	*u*11: volume of the dangerous goods	149	5.5%	
	*u*12: extreme weather condition	75	2.7%	
E	*u*13: railway lines condition	100	3.7%	9.4%
	*u*14: sudden natural disaster	83	3.0%	
	*u*15: failure of transportation laws and safety management	86	3.1%	
M3	*u*16: failure of safety education management	71	2.6%	11.3%
	*u*17: failure of Corporate Qualification Management	151	5.5%	

The prior probabilities and CPTs for the sub-indicators in the BN can be determined by using data analysis approach or by the expert scoring. In this study, the probabilities have been developed by experts. The *u*17, *u*14 and *u*2 need to score the prior probability, other sub-indicators need to score the CPTs. If one sub-indicator occurs, then it will be marked as O; otherwise, it will be marked as N.

In the following we show prior probabilities of u2 and CPT of u3 (which has one parent u17) as two examples. Other CPTs are available in the notebook.

```
+-------+-----+
| u2(N) | 0.4 |
+-------+-----+
| u2(O) | 0.6 |
+-------+-----+
```

Prior probabilities of u2

```
+--------+--------+--------+
| u17    | u17(N) | u17(O) |
+--------+--------+--------+
| u3(N)  | 0.3    | 0.15   |
+--------+--------+--------+
| u3(O)  | 0.7    | 0.85   |
+--------+--------+--------+
```

CPT of u3

3. BN structure

The Bayesian Network structure of the paper is constructed using ISM method.

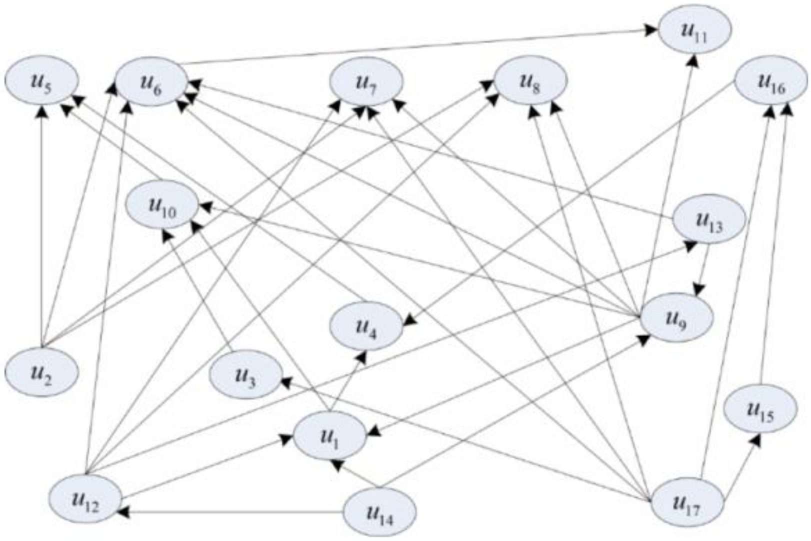


Fig. 1. **The BN based on the directed graph**

4. Flow of probabilistic influence

We investigate six types of probabilistic influence introduced in the class for our BN. Here is the representation of a sample of each type:

4.1. Direct cause

We can see from the network *extreme weather condition* (u12) directly influences on *railway lines condition* (u13) that means whenever some changes in weather condition occurs, it will have effect on lines condition.

4.2. Direct effect

On the other hand, we can see that *extreme weather condition* (u12) can be an effect of *sudden natural disaster* (u14).

4.3. Causal trail

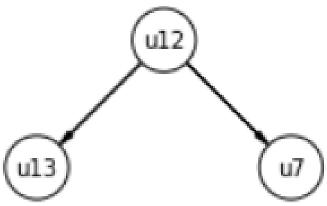
If we consider more than two connected nodes, we can find some trails in the network which clarify indirect cause between nodes. For example, any changes in the *extreme weather condition* (u12), will have effect on *railway lines condition* (u13) and it will change the values in *failure of transportation equipment* (u6). Now if *railway lines condition* (u13) is observed i.e., we know its value, so in this case any change in *extreme weather condition* (u12) won't affect *railway lines condition* (u13) since we already know the value. Hence there won't be any change in *failure of transportation equipment* (u6) as it depends only on its parent.

4.4. Evidential trail

If we investigate the example of causal trail in reverse order, we get evidential trail which says evidence *failure of transportation equipment* (u6) can influence *extreme weather condition* (u12) via *railway lines condition* (u13) only if u13 is unobserved. Observed u13 blocks influence.

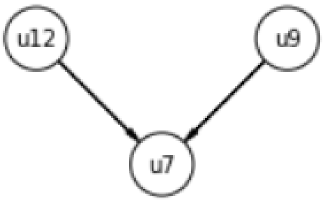
4.5. Common cause

A common cause sample in the network is as follows. It refers to the influence flows from *railway lines condition* (u13) to *failure of dangerous goods storage equipment* (u7) when *extreme weather condition* (u12) is not observed. But when *extreme weather condition* is observed, change in *railway lines condition* doesn't affect *failure of dangerous goods storage equipment* since it's only dependent on extreme weather condition.



4.6. Common effect

This structure is commonly known as V structure. When *failure of dangerous goods storage equipment* (u7) is not observed, any change in *extreme weather condition* (u12) reflects some changes in u7 but not in *dangerous nature of the loaded and transported goods (u9)*. But when *failure of dangerous goods storage equipment* (u7) is observed, if *extreme weather condition* (u12) is also observed, this will increase the probability of occurring *dangerous nature of the loaded and transported goods (u9)* since we already know that *dangerous goods storage equipment has failed.*



5. Independency

5.1. Conditional Independence

A variable is conditionally independent of non-descendant variables given its parents. As an example, we considered independencies of *extreme weather condition* (u12). The result is as follows:

(u12 ⊥ u16, u2, u3, u15, u17 | u14) ==> (Extreme weather condition ⊥ failure of safety education management, inaccurate work attitude and operation of the staffs, staffs are lack of technical and knowledge during transportation processes, failure of transportation laws and safety management, failure of Corporate Qualification Management | sudden natural disaster)

5.2. Markov blanket

A variable is conditionally independent of all other variables given its Markov blanket which is its parents, children, and children's parents. Markov blanket nodes of *extreme weather condition* is as follows:
[u6, u2, u7, u13, u14, u1, u9, u8, u17]

5.3. Active trails

When influence can flow from X to Y via Z then trail X—Z—Y is active. We investigated active trails for all the nodes in the network. Node *inaccurate work attitude and operation of the staffs* (u2) has the least active trail with 6 other nodes and the following nodes are those which have active trail to all the nodes in the network: *non-railway personnel's illegal dangerous goods stealing* (u5), *failure of transportation equipment* (u6), *failure of dangerous goods storage equipment* (u7), *failure of loading and unloading equipment in the handling stations* (u8), *packaging of dangerous goods* (u10), *volume of the dangerous goods* (u11).

5.4. Direct separation

Two nodes are d-separated if there is no active trail between them. According to this definition, node *volume of the dangerous goods* (u11) which has active trail to all nodes, is not d-separated from any node in the network. On the other hand, the nodes *physical discomfort and poor working environment of the staffs* (u1) and *inaccurate work attitude and operation of the staffs* (u2) are d-separated.

6. Exact inference

There are some algorithms which can be used for exact inference. These algorithms find the exact probability values for our queries. In this work we used variable elimination for inference.

6.1. Inference by variable elimination

Variable elimination is a dynamic programming algorithm that re-uses computation. For querying we select the risk factor with highest incidents frequency which is *staffs are lack of technical and knowledge during transportation processes* (u3). The probability of u3 is as follows:

u3	phi(u3)
u3(N)	0.2400
u3(O)	0.7600

This variable has just one parent which is *failure of Corporate Qualification Management* (u17). It is crystal clear that in the absence of u17 the probability of occurrence of u3 will decrease and vice versa.

u3	phi(u3)
u3(N)	0.3000
u3(O)	0.7000

P (u3 | u17 Not occurs)

u3	phi(u3)
u3(N)	0.1500
u3(O)	0.8500

P (u3 | u17 Occurs)

The other node we investigated is *inaccurate work attitude and operation of the staffs* (u2) which is a root node. We wanted to see the influence of children nodes on the parent. As we expected if each of the children is declared as “Not occurred”, the probability of occurring the parent tends to decrease. It means if we know *non-railway personnel's illegal dangerous goods stealing* and *failure of transportation equipment* and *failure of dangerous goods storage equipment* and *failure of loading and unloading equipment in the handling stations* have not occurred, we can say u2 occurs with less probability in comparison to the situation that we do not know anything about its children.

u2	phi(u2)
u2(N)	0.4000
u2(O)	0.6000

P (u2)

u2	phi(u2)
u2(N)	0.6501
u2(O)	0.3499

P (u2 | all its children Not Occurs)

The algorithm has exponential time complexity, but could be efficient for low-treewidth graphs, if the proper elimination order is used. We tried some orders to show the time differences between each type in practice. The summary of them is presented in the notebook.

7. Approximate inference

Approximate inference work well for many networks where exact methods are intractable. In this work we tried four sampling algorithms introduced in the course for inference. The results show when the number of samples are low, the predicted value for variable probabilities is very different from the exact probabilities. If we want to reach the results comparable with the exact ones we have to try with higher number of samples.

We investigate variable u6 (failure of transportation equipment). The exact probability of the selected variable is shown here for the purpose of comparison with approximate results.

u6	phi(u6)
u6(N)	0.4054
u6(O)	0.5946

7.1. Sampling from an empty network

Sampling from an empty network generates samples from joint distribution of the Bayesian network.

u6	phi(u6)
u6(N)	0.5500
u6(O)	0.4500

Inference by 20 samples

u6	phi(u6)
u6(N)	0.4040
u6(O)	0.5960

Inference by 1000 samples

7.2. Rejection sampling

Rejection sampling generates samples from joint distribution of the Bayesian network, but reject those that not match the evidences.

u6	phi(u6)
u6(N)	0.1500
u6(O)	0.8500

Inference by 20 samples

u6	phi(u6)
u6(N)	0.4297
u6(O)	0.5710

Inference by 1000 samples

7.3. Likelihood weighting

Likelihood sampling generates weighted samples from joint distribution of the Bayesian network, that comply with the given evidence.

u6	phi(u6)
u6(N)	0.2000
u6(O)	0.8000

Inference by 20 samples

u6	phi(u6)
u6(N)	0.4350
u6(O)	0.5650

Inference by 1000 samples

7.4. Gibbs sampling

Start with an arbitrary instantiation consistent with the evidence. Gibbs algorithm Sample one variable at a time, conditioned on all the rest, but keep evidence fixed.

u6	phi(u6)
u6(N)	0.3500
u6(O)	0.6500

Inference by 20 samples

u6	phi(u6)
u6(N)	0.4265
u6(O)	0.5735

Inference by 1000 samples

8. Conclusion

The final results show that the occurrence probability of *staffs are lack of technical and knowledge during transportation processes* (u3) has the highest impact with probability 0.74, the occurrence probability of *sudden natural disaster* (u14) has the weakest impact with probability 0.05.

9. Reference

[1] Wencheng Huang, Yue Zhang, Xingyi Kou. *Railway dangerous goods transportation system risk analysis: An Interpretive Structural Modeling and Bayesian Network combining approach*. Reliability Engineering and System Safety, 2020. DOI: https://doi.org/10.1016/j.ress.2020.107220