APPLICABILITY OF BAYESIAN NETWORK IN TRANSPORTATION ENGINEERING FOR SAFETY*

By Moinul HOSSAIN** and Yasunori MUROMACHI***

1. Introduction

Over the last few decades, the main focus in the field of transportation research has been on moving people and/or vehicles from one place to another on time at an acceptable cost, unfortunately, without adequate consideration on safety issues. Road traffic crash is a highly complex, yet, little understood phenomena in transportation engineering which is associated with substantial socio-economic impact. Some of the major obstacles for the advancement of this research field have been the lack of crash data, difficulty in field experimental design, absence of quality simulation methods, and the high uncertainty associated with the phenomena itself. Thus, modeling techniques that are capable of reasoning under situations where causality plays a role but we do not have a clear understanding of the phenomena, can play a significant role in reducing the knowledge gap in road safety research. Such a method widely known as Bayesian Network was introduced by the Artificial Intelligence community during the mid 1980s. However, despite its importance in modeling highly uncertain scenarios, the method became popular until recently with the introduction of new algorithms, software packages and the increased capability of computers. Several researchers from various disciplines have since then taken initiatives to introduce Bayesian Network to their research community by publishing papers on step by step explanation and implementation procedures through examples. In this paper, we have also attempted to make Bayesian Network more accessible to the road safety researchers by providing the basic ideas through a formulated example

2. Formulation of the example

In Bayesian Network (BN), we create a model of certain problem domain in such way that it can support experts in performing their tasks rather than substituting them[1)]. In this section, we present a simple Bayesian Network to predict the probability of road crash to introduce its concept, mechanism, properties and use. We have assumed the values of the variables as it is for illustration purpose and is not intended to present a valid model based on real data.

Road crash is a complex phenomena and is caused by multiple interrelated factors, such as, speed of vehicle, traffic flow on the road, weather condition, time of day, road geometry, etc. So far researchers have been successful in finding correlations of these causalities with crash but research regarding the understanding of crash phenomena is still in its infancy. However, from previous research findings and expert opinion, we can identify that time of day and weather can affect traffic flow; weather and road geometry can influence the speed of vehicles and crash is related to speed and flow of vehicles in the stream. We have prepared a simple directed acyclic causal graph with this information where the variables are presented as nodes and their inter-relationship is presented with arcs (Figure 1).

Now, when an arc is drawn from one variable to the other, the former is called the parent (denoted as 'pa') and the later as child when we mention their relationship. In our example, 'Flow' has two parent nodes – 'Weather' and 'Time of day'. The variables represented with nodes in Figure 1 have finite number of discrete values. Let us assume that each variable has two categories as presented in Table 1. The state of the variables are represented with probabilities as rainy weather or non-straight road sections will not always cause low speed, in the same way, high speed will not always cause crashes. We can obtain these probabilities from frequency based data analysis, from previous studies in a similar site or even from expert opinions.

*Keywords: Bayesian Network, Transportation Engineering, Safety

**Student Member of JSCE, M. Engg.., Dept. of Built Environment, Tokyo Institute of Technology (Nagatsuta-machi 4259, Midori-ku, Yokohama-shi, Kanagawa,226-8502 Japan, Telefax: +81-(0)45-924-5524, E-mail address: moinul048i@yahoo.com)

***Member of JSCE, Ph.D., Dept. of Built Environment, Tokyo Institute of Technology (Nagatsuta-machi 4259, Midori-ku, Yokohama-shi, Kanagawa, 226-8502 Japan, Telefax: +81-(0)45-924-5524)

Thus, when complete information is available, we can use such a causal diagram to predict events (e.g., predict crash probability) or infer causes from observed effects[2] (e.g., was the crash caused due to over speeding?). However, it is difficult to obtain information about all the variables at a time. Moreover, it becomes difficult to make inferences when different evidences are suggesting contradictory conclusions. Bayesian Network allows us to calculate such probabilities by considering only a small set of probabilities, relating only conditionally dependent neighboring nodes[2].

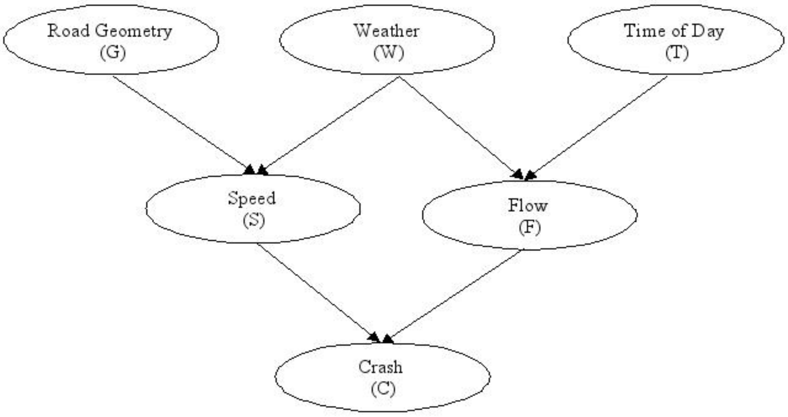


Figure 1. A causal graph representing a simplified model for road crash

Table 1. States of different variables

	Road Geometry	Weather	Time of Day	Speed	Flow	Crash
States	Straight/Not straight	Rainy/Not rainy	Peak/Off-peak	High/Low	High/Low	Yes/No

3. The formal definition and independent assumptions

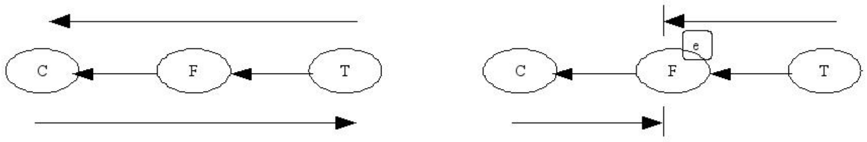
Bayesian Network can be defined as an acyclic directed graph (DAG) which defines a factorization of a joint probability distribution over the variables that are presented by the nodes of the DAG, where the factorization is given by the directed links of the DAG[1]. Thus, if a BN contains 'n' number of variables, then we can represent the complete problem domain as Equation 1.

$$P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | pa(x_i)) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(1)}$$

In order to specify a BN, we need to provide the probability (or conditional probability) distributions of all the nodes. Thus, as we have six variables each with two states in our example, we are expected to need 2^6-1, or, 63 joint probabilities. However, due to the conditional independence assumptions (explained later) conditional probability of each child node can be calculated only with respect to its direct parent nodes. Thus, Equation 1 can be written as Equation 2 for our example.

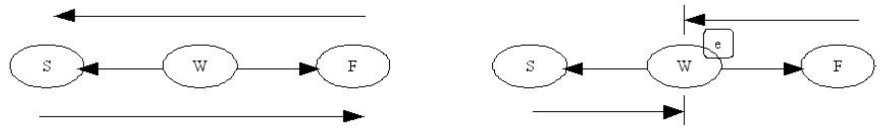
$$P(G,W,T,S,F,C) = P(G)P(W)P(T)P(S|G,W)P(F|W,T)P(C|S,F) \quad \ldots\ \ldots\ \ldots\ \ldots\ \ldots \text{(2)}$$

The conditional independence assumption of BN can be explained by explaining how information flows through its connections. A BN can have three kinds of connections – serial, diverging and converging. One of the serial connections in our example is from 'Time of Day' to 'Crash' through 'Flow' (Figure 2(a)). Now, if we know the time of day, we can revise our belief about the flow condition at that time and thereby update our belief regarding the crash probability. However, if we already know about the flow condition, any extra information regarding the time of the day will not alter our belief about crash, i.e., flow of information will be blocked as shown in Figure 2(b). The connection among 'Weather', 'Speed' and 'Flow' is diverging with 'Weather' as a parent to both 'Speed' and 'Flow' (Figure 3a). Like the serial connection, evidence on 'Flow' can modify our belief regarding the weather and thus influence our belief about 'Speed' and vice versa. If we already have evidence related to 'Weather' then new information related to 'Speed' (or 'Flow') does not change our belief about 'Flow' (or 'Speed') (Figure 3(b)). The connection among 'Weather', 'Flow' and 'Time of Day' is converging (Figure 4(a)). Here, if we assume that both rainy weather and off-peak hour cause the traffic flow to be low, then if we know that the traffic flow is low and if we also observe that the weather is not rainy, our belief regarding the time of day to be off-peak hour will increase. In the same way, if we observe high flow on a rainy day and do not know about the period of time when the data was collected, our belief about the probability that the time of day is peak hour will increase. Thus, in case of converging connection, we need evidence regarding the child node and one of the parent nodes for the propagation of information (Figure 4(b)).



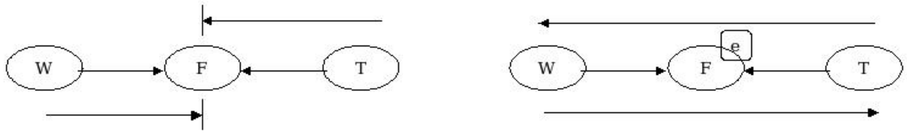
(a) (b)

Figure 2. A serial connection (upper and lower arrows indicating propagation of evidence)



(a) (b)

Figure 3. A diverging connection (upper and lower arrows indicating propagation of evidence)



(a) (b)

Figure 4. A converging connection (upper and lower arrows indicating propagation of evidence)

As these three cases cover all the ways evidence can be propagated through a BN, based on these concepts, we can identify if two variables are conditionally independent. This is formulated as a rule called d-separation rule. According to Jensen and Nielsen[3)], two distinct variables A and B in a causal network are d-separated if for all paths between A and B, there is an intermediate variable V (distinct from A and B) such that either

– the connection is serial or diverging and V received evidence, or,

– the connection is converging, and neither V nor any of V's descendants have received evidence.

If A and B are d-separated then changes in the certainty of A have no impact on the certainty of B. If they are not d-separated then they are called d-connected. In case of our example, any propagation of evidence to 'Flow' can be either through 'Weather' or 'Time of Day', or through 'Crash'. As 'Crash' is the child node of a converging connection, it will block the evidence propagation when we do not have any information regarding crash. In the same way, evidence regarding 'Weather' and 'Time of Day' will block the propagation to 'Flow' (imagine that the network was larger and 'Weather' and/or 'Time of Day' had parent nodes). Thus, any child node in a BN can be conditioned only with their direct parent nodes as presented in Equation 2.

4. Calculating the probabilities in a Bayesian Network

At this point, we need to specify the prior probability of nodes which have no parents as well as the conditional probabilities of nodes with parents to specify the probability distribution of our example Bayesian Network. Now, let us assume that we have obtained the probability of road section being straight, weather being rainy and time of day being peak period to be 0.8, 0.15 and 0.25. Similarly, let us also assume that $P(S_{High} \mid G_{Straight}, W_{Rain})$ =0.25, $P(S_{High} \mid G_{Straight} \overline{W}_{Rain})$ =0.85, $P(S_{High} \mid \overline{G}_{Straight}, W_{Rain})$ =0.05, $P(S_{High} \mid \overline{G}_{Straight}, \overline{W}_{Rain})$ =0.35, $P(F_{High} \mid T_{Peak}, W_{Rain})$ =0.70, $P(F_{High} \mid T_{Peak}, \overline{W}_{Rain})$ =0.99, $P(F_{High} \mid \overline{T}_{Peak}, W_{Rain})$ =0.10, $P(F_{High} \mid \overline{T}_{Peak}, \overline{W}_{Rain})$ =0.20, $P(C_{Yes} \mid S_{High}, F_{High})$ =0.01, $P(C_{Yes} \mid S_{High}, \overline{F}_{High})$ =0.008, $P(C_{Yes} \mid \overline{S}_{High}, F_{High})$ =0.006, $P(C_{Yes} \mid \overline{S}_{High}, F_{High})$ =0.006, $P(C_{Yes} \mid \overline{S}_{High}, \overline{F}_{H})$ =0.0001. Thus, the joint probability of speed being high, flow being high and crash can be calculated as respectively 0.669, 0.375 and 0.0066 as presented in Figure 5 (we used free version of Netica, a BN software, to calculate the probabilities).

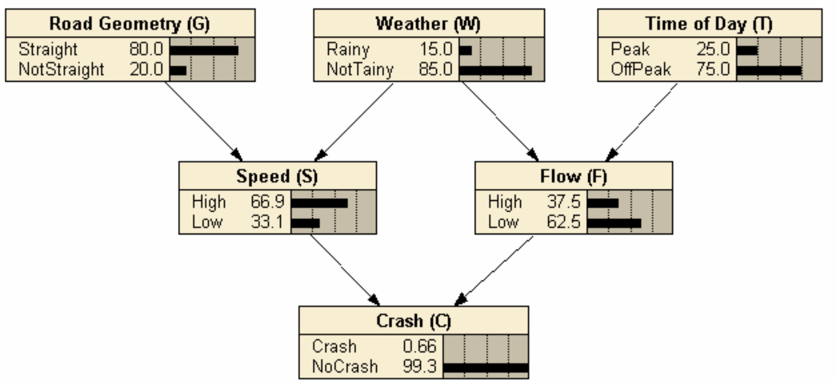


Figure 5. Probability of the example Bayesian Network

Now, this model can be used for inference purpose when we have information regarding the state of any/several variable(s). For illustration purpose, let us assume that we would like to know the probability of crash on a straight road section during peak hour. Thus, when G = 'Straight' and T = 'Peak' probability of P(C = 'Crash') can be calculated as 0.88:

$$P(C=Crash|G=Straight, T=Peak)=\frac{\sum_W \sum_S \sum_F P(G=Straight, W, T=Peak, S, F, C=Crash)}{\sum_W \sum_S \sum_F \sum_C P(G=Straight, W, T=Peak, S, F, C)}$$

Thus, when built with real data, our example model can be used to predict crash probability depending on various road, weather and traffic conditions, unveil the relationship between speed and flow during crash, identify specific combination of situations resulting in high crash potential, etc.

5. Key Advantages of Bayesian Network, its application and conclusion

One of the key features in BN is the status of its variables in the model. Unlike classical modeling approaches, BN does not have the concept of dependent and independent variable and treats each one equally. Thus, once we build the model, state of any variable can be predicted with available information about any/some variable(s), which eliminates the necessity to build separate models for each variable. BN's ability to update the belief about any variable in presence of new evidences resemblance with human rational way of thinking. Moreover, when we have more information regarding some parts of the network, we can update the corresponding probabilities without needing to re-model the problem domain. Bayesian Network has provided researchers with an essential tool to address an array of problems in which one is willing to draw conclusions based on a probabilistic approach. At present, the use of BN in transportation engineering has been limited. As many problems in transportation engineering, specifically in road safety, deal with high uncertainty, lack of data availability and require probabilistic approach, we can expect a rapid growth in use of Bayesian Network in this research field. However, all these benefits are associated with cost, too. Preparing the Bayesian Network and setting up the causal directions in it may require experience as wrong causal directions can represent erroneous conditional independence. For example, in Figure 1, if we alter the direction of arrows from 'Weather' and 'Time of Day' to 'Flow', i.e., create a diverging connection with 'Flow' as the parent node, it will suggest a dependency between 'Weather' and 'Time of Day' – which is wrong. Moreover, the calculations associated with Bayesian Network is considered as NP-hard and require a substantial amount of resources to model large networks. The available software packages to apply Bayesian Network are expensive and limited in number. However, as a large number of research communities are now using Bayesian Network, we can expect that soon more efficient and cost effective tools will be available, which will facilitate the use of Bayesian Network for researchers outside the information science and artificial intelligence community.

References

1) Charniak, E (1991). Bayesian Network Without Tears, AI Magazine, vol.. 12 issue 4, pp. 50-63.

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

3) Jensen, F. V. and Nielsen, T. D. (2007). Bayesian Network and Decision Graphs, 2nd Edition, Springer.