Bayesian Networks, Total Variation and Robustness

Sophia K. Wright & Jim Q. Smith

Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK

Abstract

Now that Bayesian Networks (BNs) have become widely used, an appreciation is developing of just how critical an awareness of the sensitivity and robustness of certain target variables are to changes in the model. When time resources are limited, such issues impact directly on the chosen level of complexity of the BN as well as the quantity of missing probabilities we are able to elicit. Currently most such analyses are performed once the whole BN has been elicited and are based on Kullback-Leibler information measures. In this paper we argue that robustness methods based instead on the familiar total variation distance provide simple and more useful bounds on robustness to misspecification which are both formally justifiable and transparent. We demonstrate how such formal robustness considerations can be embedded within the process of building a BN. Here we focus on two particular choices a modeller needs to make: the choice of the parents of each node and the number of levels to choose for each variable within the system. Our analyses are illustrated throughout using two BNs drawn from the recent literature.

KEYWORDS: Bayesian Networks, Robustness, Total Variation Distance, Elicitation, Decision Support Systems.

The first author would like to acknowledge their EPSRC grant EP/L016710/1, awarded through the OxWaSP CDT programme. The second author contributed to this work within their EPSRC grant EP/N510129/1.

1 Introduction

Bayesian Networks (BNs) are now a widely used probabilistic modelling tool, particularly in the field of decision support. It is now acknowledged as best practice [Cowell et al., 1999, Laskey and Mahoney, 2000, Smith, 2010] that these models are set up in two distinct stages. Firstly the structure of the BN, as expressed by its Directed Acyclic Graph (DAG), is either directly elicited from domain experts or when sufficient supporting data exists, learned from the data using a model search algorithm with default priors on the hyperparameters, see Boneh [2010] and Korb and Nicholson [2010]. Once this graphical framework has been discovered, the graph is embellished into a full probabilistic model. In the case of a discrete BN, this second stage involves eliciting or estimating, using priors on probabilities informed by expert judgements, the entries of its
conditional probability tables (CPTs). These CPTs provide the numerical pre-
specification of all the conditional probabilities needed to generate the full joint
probability mass function and hence a fully specified probability model.

When engaging in this two stage process the analyst needs to be fully aware
of precisely which inputs might be critical to the inferences made through the
BN, see Albrecht et al. [2014b]. One critical element in an elicitation, or statisti-
cal estimation of the graph is to ensure these critical features are specified as
accurately as possible. This is especially important when elicitation or estima-
tion is resource limited, as is usually the case in practice. The modeller can then
optimize their allocation of resources to concentrate on eliciting those elements
of the model whose misspecification might most influence the required outputs.

To this end, the practitioner, prompted by the functionality of various soft-
ware, is currently encouraged to develop awareness of the robustness of a chosen
model to its inputs by performing a one-at-a-time numerical sensitivity analysis
of the preliminary BN. Here various different forms of numerical contaminations
of the model are investigated, where effects are usually measured in terms of mu-
tual information/Kullback-Leibler divergence [Albrecht et al., 2014a, Friedman
et al., 1997, Nicholson and Jitnah, 1998, Zaragoza et al., 2011]. This type of
study is obviously extremely useful. On the other hand it has drawbacks. First,
it relies on the chosen enacted perturbations covering the entire space which
becomes more challenging as models become increasingly large. Furthermore,
even if such a search is performed systematically, the impacts (most currently
measured by mutual information), are not directly relevant to the impact on en-
suing decisions, see below for further clarification. Additionally, such an analysis
must perforce be performed after the model has been fully specified. This means
that the whole probability model is needed before the sensitivity analysis can
be performed. One interesting recent attempt to provide such assessments after
the structural elicitation phase, but before the probabilistic embellishment is
through the use of distance weighted sensitivity measures [see Albrecht et al.,
2014a]. However, these do not dovetail with the mutual information measures
described above and have a level of arbitrariness in the choice of weight function
needed to use this method.

Over recent years more formal and systematic robustness analyses have ap-
peared. Robustness of probability models has been studied by statisticians for
many decades, and specific methodology for Bayesian Networks has also been re-
cently developed: Coupé and van der Gaag [2002], Gómez-Villegas et al. [2013],
Laskey [1995], O’Neill [2009], Renooij [2010]. These fall into two main streams:
local robustness studies and global studies. In the former, a chosen probability
model is perturbed using a finite parametrised modification. The latter, termed
global analyses, does not rely on perturbation lying within a given parametric
family [O’Neill, 2009, Smith and Daneshkhah, 2010]. Instead, an appropriate di-
vergence measure is applied to first specify a neighbourhood system around each
model. Bounds are then calculated for the maximum deviation in the inference
that could be achieved by a model in this neighbourhood. If this deviation is
small then the model is deemed to be robust [Gustafson and Wasserman, 1995,
Smith and Rigat, 2012]. Both types of robustness analysis have been applied
to BNs in work such as Smith and Daneshkhah [2010]. In this paper we focus
solely on global robustness studies as applied to finite discrete BNs.

Thus far, global robustness studies for BNs have mainly centred around
the analysis of how robust a model might be to perturbations, with respect to
Kullback-Leibler (KL) or Chan-Darwiche divergences [see Chan and Darwiche, 2005, Gómez-Villegas et al., 2013, Leonelli et al., 2017]. Both of these divergence measures benefit from some helpful technical properties which allow various measures of dependence to have explicit formulae. These measures are specified in terms of log probabilities in the KL case or equivalently ratios of probabilities in the Chan-Darwiche instance. Therefore, both have the disadvantage that they depend very heavily on the accurate specification of small probabilities. However, it is well documented that it is precisely these small probabilities that typically exhibit the largest elicitation error [see O’Hagan et al., 2006, Smith, 2010]. Furthermore, when BNs are learned from data, any associated small probabilities are difficult to reliably estimate from data, because almost by definition we will see very few of these events in any training set we use to estimate a model.

In many circumstances (especially in decision analysis), the misspecification of improbable event probabilities has only a small impact on the required outputs of a decision analysis: see below. For the purposes of the two stage process described above, the Kullback-Leibler and Chan-Darwiche divergence measures are hardly ideal as they can be highly sensitive to misspecification which may have little effect on any supported decision analysis.

In this paper we demonstrate that an alternative robustness study based on a more conventional divergence measure (widely used in probability theory and stochastic analysis), which is the total variation distance, has some serious practical and theoretical advantages over its main competitors. Although it is often difficult to derive explicit formulae for the impacts of deviation in variation, it is nevertheless straightforward to tightly bound such deviations in variation distance. Deviation in variation corresponds much more closely to the types of error we would envisage experiencing within either an elicitation exercise or through misestimation. Perhaps most important, the expectation of a fixed bounded utility function $U$, under various decisions (induced by an approximation) are simply bounded by linear functions of the total variation in the probability distributions of the attributes of $U$ [see e.g. Smith, 2010]. Note that in a BN these attributes will typically constitute a small subset of the totality of its variables. Hence small variation distances (between probability mass functions) on these small subsets translate directly into small effects in the pertinent expected utilities. Conversely, large deviations translate into large effects that might have a greater impact on some specification of a utility.

In the following section we review the BN framework and introduce our examples. Then in Section 3 we review some simple properties of the total variation distance and show that the effect in variation distance of the misspecification of the probability mass function of one random variable in a BN to another diminishes exponentially. We then discover explicit bounds for this error both when the BN is decomposable and more generally. We demonstrate that this impact can be bounded explicitly in terms of a simple function of the extreme entries of the CPTs within the BN. These results have the useful spin-off that CPTs do not necessarily need to be fully elicited before the robustness analysis can take place. In Section 4 we show how these explicit measures of robustness can be applied to determine the effect of approximating simplifications on the topology of the BN and additionally, to decide the number of levels into which to categorise each variable. We demonstrate how by using total variation, robustness analyses can be performed in a harmonious composite way that
2 Hypotheses of a Bayesian Network

We begin by giving a short review of BNs and some of its properties we use later in the paper. A discrete Bayesian Network (BN) $G$, or DAG, on a random vector $X \triangleq (X_1, X_2, \ldots, X_m)$ represents a family of models which respect a set of conditional independence hypotheses so that for $i = 2, 3, \ldots, m$

$$X_i \perp \!\!\!\!\perp X_{R(i)} | X_{Pa(i)}$$

where $Pa(i) \subseteq \{1, 2, \ldots, i-1\}$ are the parents of $X_i$, i.e. those indices of the previously listed variables on which $X_i$ depends, and $R(i) \triangleq \{1, 2, \ldots, i-1\} \setminus Pa(i)$.

An equivalent expression is that the joint probability mass function $p(x)$ of $X$ factorises as

$$p(x) = p(x_1)p(x_2|x_{Pa(2)}) \cdots p(x_i|x_{Pa(i)}) \cdots p(x_m|x_{Pa(m)}).$$

(1)

In either formulation the directed graph of the BN has vertex set $\{X_1, X_2, \ldots, X_m\}$ and a directed edge from $X_j$ to $X_i$ iff $j \in Pa(i)$.

An important subclass of BNs whose properties we discuss later, are those which are called decomposable. A decomposable BN $G$ is a BN in which every parent set of each node $X_i$ in the graph forms a complete subgraph of $G$. It is simple to show that any BN can (albeit inefficiently) be re-expressed in a decomposable BN containing it [Lauritzen, 1996, Korb and Nicholson, 2010, Smith, 2010]. This property, widely used for propagation algorithms, can also be used for robustness analyses.

When a BN is decomposable it can be shown [see Lauritzen, 1996, Smith, 2010] that the joint density factors in the following way. The cliques $\{X_{C_1}, X_{C_2}, X_{C_3}, \ldots, X_{C_m}\}$ i.e. the maximal connected subsets of the decomposable graph $G$ can be totally ordered starting with any clique, label this $X_{C_1}$. We call $X_{S(i)}$ where $S_i = C_i \cap \cup_{k=1}^{i-1} C_k$ the separator of $X_{C_i}$ from $\cup_{k=1}^{i-1} C_k$. An indexing is said to satisfy the running intersection property if for all $i = 1, 2, \ldots, m$ there exists some index $j < i$ such that $S_i = C_i \cap \cup_{k=1}^{j-1} C_k \subseteq C_j$. This implies that the result of intersecting a clique with all previous cliques is contained within one or more earlier cliques Lauritzen [1996], Smith [2010]. Note the choice of $C_i$ may not be unique.

We can depict one of these choices of order and containment by a junction tree $J(G)$. This is an undirected tree whose vertices are $\{X_{C_1}, X_{C_2}, X_{C_3}, \ldots, X_{C_m}\}$ and whose $m-1$ undirected edges simply connect $X_{C_i}$ to $X_{C_{j(i)}}$, $i = 2, 3, \ldots, m$. Note that these edges can be labelled by a corresponding separator of $G$. Here we will for simplicity assume that the entries of the joint mass function $p(x)$ are all strictly positive, although this is not strictly necessary [see Lauritzen and Spiegelhalter, 1988]. In fact this is advised from a practical point of view by a number of authors e.g. Korb and Nicholson [2010] when dealing with no known functional relations. It can then be proved [e.g. see Cowell et al., 2007, Smith, 2010] that $p(x)$ of any such decomposable BN $G$ respects the following formula:

$$p(x) = \frac{p(x_{C_1})p(x_{C_2})p(x_{C_3}) \cdots p(x_{C_m})}{p(x_{S_1})p(x_{S_2}) \cdots p(x_{S_m})}.$$
One straightforward but important consequence of this decomposition used later is that given any BN \( G \) and an associated junction tree \( J(G) \), then for any two cliques \( X_{C_1}, X_{C_k} \) there is a unique sequence of cliques \( (X_{C_1}, X_{C_2}, X_{C_3}, \ldots, X_{C_k}) \) with no repeats, and separators \( (X_{S_2}, X_{S_3}, \ldots, X_{S_k}) \) between \( X_{C_1} \) and \( X_{C_k} \) within \( J(G) \), called a simple path. If we write \( C_k \triangleq \bigcup_{j=1}^{k} C_j \), then since \( S_i \subseteq C_i \) we know that \( x_{S_i} \) is a subvector of \( x_{C_i} \) giving

\[
p(x_{C_1}) = p(x_{C_1})p(x_{C_2})p(x_{C_3})\ldots p(x_{C_k}) = p(x_{C_1})p(x_{C_2}|x_{S_2})p(x_{C_3}|x_{S_3})\ldots p(x_{C_k}|x_{S_{k-1}}).
\]

Lemma 2.1. It follows from Equation 2 and the conditional independence in \( G \) that if \( T_{1:k} \triangleq \bigcup_{i=3}^{k-1} S_i \) then

\[
p(x_{C_1} \cup C_k) = \sum_{x_{T_{1:k}}} p(x_{C_1})p(x_{S_1}|x_{S_2})p(x_{S_2}|x_{S_3})\ldots p(x_{S_{k-1}}|x_{S_k})p(x_{C_k}|x_{S_k}).
\]

Thus we have a formula for the joint mass function of a “donating” clique \( C_1 \) and a “target” clique \( C_k \) depending on \( C_1 \), expressed in terms of a sequence of transitions in a non-homogeneous Markov Chain. Although this property derives directly from the elementary properties of trees it is important, and an often overlooked property. It means that standard results from non-homogeneous Markov Chain theory can be used to measure the extent of the diminishing effect of information as it passes along this simple path. In particular it is well-known that variation distance in an ergodic, acyclic Markov Chain contracts as information is propagated through the system. The observation in Lemma 2.1 is therefore critical to the development of some of the robustness bounds we develop here.

2.1 Applying a Bayesian Network in Practice

A BN is generally selected in one of two ways. Occasionally we may have access to a complete training data set from which we can select the most promising explanatory BN whose associated \( p(x) \) respecting Equation 1 appears to best fit the data. There are many ways to do this, including using software packages such as ‘bnlearn‘ in R [see Scutari and Denis, 2014]. However, when applying such a model selection method in practice, even for low dimensional BNs, it is common to find many models score similarly well. A BN may not adequately describe all features in the data set. Even if we know this model to be true, as in a simulation exercise or even a moderately sized problem, it has been demonstrated that the best model is only close to the generating process, unless the training data set is absolutely enormous Cussens [2011]. There are also the obvious statistical errors associated with the representativeness of the data set used, even if sampling is performed at random. Hence it is rare for a single data generating model to be unequivocally identified. Considering the robustness of the critical outputs of the fitted model is therefore a critical element of any ensuing statistical analysis.

The second way to create a BN is by performing a direct elicitation from an expert. Here, having listed the variables in an order which might be compatible with the sequence in which those measurements may occur, the expert is asked
for each $X_i$ ($i = 2, 3, \ldots, n$) of the previously mentioned variables which might be relevant to forecasting it. Building on this qualitative framework, hopefully faithful to the expert’s actual judgements, we then proceed to embellish the graph by supplementing the structure with the specification of the corresponding CPTs $\{p(x_i|x_{P(i)}) : i = 1, 2, 3, \ldots, m\}$. These probabilities will be subject to elicitation error, although the preceding structural elicitation process aims to mitigate this specification error [Korb and Nicholson, 2010, EFSA, 2014, Smith, 2010]. Again an understanding of the robustness to perturbations of the hypothesized graphical framework and also the entries in the CPTs of any inferential assumptions we make here, will clearly be critical to a good statistical analysis.

2.2 Applications

2.2.1 Food Security System

To illustrate the uses and practicalities of our results we shall be using the Food Security Integrated Decision Support System (IDSS) described in greater depth in Barons et al. [2018b] and Smith et al. [2015]. The aim of this massive multi-layered dynamic BN is to ascertain which local government policies influence or improve the level of food poverty within the UK. However, the targeted user here is primarily interested in three specific classes of outputs: Health (of its local constituents), Educational Attainment of children and measures of Social Cohesion [Smith et al., 2015], measures of which, in the terminology of this paper, will form our final vector of target variables.

The overarching IDSS model is a DBN model as shown in Figure 1 in which the target nodes are classified as Level 1. Each component of this model can be broken down into detailed subnetworks. For example, the Level 2 ‘UK Food Costs’ depends on the availability of food, production costs and so on. Specifically we may be interested in access to healthy food necessities such as fruit and vegetables which rely heavily on pollinator abundance. A sub-subnetwork to determine the factors which affect the pollinator abundance is therefore required, a fragment of which is shown in Figure 2. A subset of this BN has been elicited from experts and the results can be found in Barons et al. [2018a].
2.2.2 An Ecological Demonstration

To illustrate our approach we also use a well known ecological BN called the “Native Fish” example as introduced in Nicholson et al. [2010] and discussed further in Nicholson and Flores [2011]. This BN was designed specifically for demonstration purposes, notably introducing non-statisticians to BNs, and is therefore simplified version of a much more complicated process. However, because the meaning of its variables are transparent and its topology

Figure 1: Food Security IDSS, red arcs indicating dynamic relationships, from Barons et al. [2018b].

Figure 2: Fragment of the pollinator abundance BN sub-subnetwork, from Barons et al. [2018b].
of this model) is just large enough to demonstrate our arguments, this DSS is ideal for illustrating some of our methods.

This ecological BN is used to model the impact on native fish abundance of pesticide usage on surrounding fields as well as levels of rainfall. The structure of the BN is given in Figure 3. Our target node is ‘Native Fish Abundance’.

![BN Structure](image)

Figure 3: “Native Fish” Version 2 BN structure, from Nicholson et al. [2010].

3 Properties of Total Variation Distance for BNs

We begin by outlining the total variation distance, highlighting some of its useful properties which we can directly apply to this robustness analyses.

Assume \( \mathbf{X} \triangleq (X_1, X_2, \ldots, X_m) \) is a vector of finite discrete random variables taking values \( \mathbf{x} = (x_1, x_2, \ldots, x_m) \in X_1 \times X_2 \times \cdots \times X_m \). Let \( X_A \), taking values \( X_A \subseteq \prod_{i=1}^{r(A)} X_{i_j} \), denote the subvector of \( \mathbf{X} \) comprising those components with indices \( i \in A \), where \( A = \{i_1, i_2, \ldots, i_{r(A)}\} \) denotes a subset of \( \{1, 2, \ldots, m\} \). Let \( p_A, q_A \) denote a hypothesised and an alternative joint mass function on \( X_A \) and \( P_A(E), Q_A(E) \) denote the probability with respect to the mass functions \( p_A, q_A \) respectively of the set \( E = E_{i_1} \times E_{i_2} \times \cdots \times E_{i_{r(A)}} \) where \( E_{i_j} \subseteq X_{i_j}, j = 1, 2, \ldots, r(A) \). Nearly all inferential methodology and certainly all robustness analyses focus on properties of such events [Smith, 2010].

**Definition 3.1.** The (Total) Variation distance, \( d_V(p_A, q_A) \), is defined in the discrete case by

\[
\frac{1}{2} \sum_{\mathbf{x}_A \in X_A} |p_A - q_A|
\]
3.1 Variation under Marginalisation and Conditioning

Measures of variation distance can be applied directly to CPTs. In this section we define some new objects which will be especially useful in our later development.

Let $P$ and $Q$, with rows $p_i, q_i$ respectively for $i = 1, 2, \ldots, n$, be two $n \times n'$ CPT matrices of a random vector $Y$, taking $n'$ levels given another random vector $X$, taking $n$ levels. For a BN, $Y$ will typically be a random variable whilst $X$ will be the vector of its parents; nevertheless when studying junction trees it is also helpful to consider cases when $Y$ is a vector.

There is a natural variation distance we can now define between $P$ and $Q$:

**Definition 3.2.** Let the variation distance $d^+_V(P, Q)$ between conditional probability tables $P$ and $Q$ be defined by

$$d^+_V(P, Q) \triangleq \max_{1 \leq i \leq n} d_V(p_i, q_i).$$

**Example 3.1.** Assume that the CPT in Nicholson et al. [2010], represented by the transition matrix $P$ below, gives the elicited combined matrix of a panel of experts using a standard protocol [see EFSA, 2014, for example]. Suppose expert A’s individually elicited elicited probabilities are given by the matrix $Q$:

| Drought Conditions | Annual Rainfall | $P(\text{Tree Condition} | \text{Drought, Rainfall})$ |
|--------------------|----------------|----------------------------------|
| yes                | below average  | Good: 0.2, Damaged: 0.6, Dead: 0.2 |
| yes                | average        | Good: 0.25, Damaged: 0.6, Dead: 0.15 |
| yes                | above average  | Good: 0.3, Damaged: 0.6, Dead: 0.1 |
| no                 | below average  | Good: 0.7, Damaged: 0.25, Dead: 0.05 |
| no                 | average        | Good: 0.8, Damaged: 0.18, Dead: 0.02 |
| no                 | above average  | Good: 0.9, Damaged: 0.09, Dead: 0.01 |

We can simplify this to matrix form, denoted by $P$. Let us assume that this CPT was elicited from experts who disagree on a couple of probabilities resulting in an alternate CPT, $Q$:

$$P = \begin{pmatrix} 0.20 & 0.60 & 0.20 \\ 0.25 & 0.60 & 0.15 \\ 0.30 & 0.60 & 0.10 \\ 0.70 & 0.25 & 0.05 \\ 0.80 & 0.18 & 0.02 \\ 0.90 & 0.09 & 0.01 \end{pmatrix}, \quad Q = \begin{pmatrix} 0.20 & 0.60 & 0.20 \\ 0.30 & 0.50 & 0.20 \\ 0.30 & 0.60 & 0.10 \\ 0.65 & 0.25 & 0.10 \\ 0.80 & 0.18 & 0.02 \\ 0.90 & 0.10 & 0.00 \end{pmatrix}.$$

We can now compute $d^+_V(P, Q) = \max\{0, 0.10, 0, 0.05, 0, 0.01\} = 0.1$. Expert A will often be concerned that the substitution of $P$ for $Q$ will not effect significantly the conclusions about the target variables of this panel. We show below how we can use the diameter calculation to directly measure, in a formal sense, the extent of this substitution and thus allay Expert A’s fears that the panel’s judgement might be substantially at variance with their own.

Note that if $\rho(P)$ and $\rho(Q)$ are the vectors of marginal mass functions of $Y$ and $\pi$ is a margin on $X$ then for all possible margins $\pi$

$$d_V(\rho(P), \rho(Q)) \leq d^+_V(P, Q),$$
where $d_V(\rho(P), \rho(Q)) = d^+_V(P, Q)$ whenever $\pi$ puts all its mass on atoms indexed by $i^+$ where

$$i^+ \triangleq \arg \max_{1 \leq i \leq n} d_V(p_i, q_i).$$

Thus we have that for all possible margins $\pi$

$$d_V(\rho(P), \rho(Q)) \leq d^+_V(P, Q).$$

This therefore gives rather coarse, but quick bounds which require only comparisons of the pairs of individual rows of the perturbed CPT.

Earlier we highlighted that when eliciting a BN we first elicit hypotheses of conditional independence. Only then do we expand this with a full probability specification through the numerical values in its CPTs. So we next consider robustness measures associated with small deviations from conditional independence. The definition we present below is, to our knowledge, a new construction using variation distance on CPTs to determine the measure of dependence between variables.

**Definition 3.3.** The diameter, $d^+(P)$, and the I-local diameter $d^{I+}(P)$ of a stochastic matrix $P \triangleq \{p_{ij}\}$ are respectively defined as

$$d^+(P) = \frac{1}{2} \max_{1 \leq i, i' \leq n} \left\{ \sum_{j=1}^{n'} |p_{ij} - p_{i'j}| \right\},$$

$$d^{I+}(P) = \frac{1}{2} \max_{i, i' \in I} \left\{ \sum_{j=1}^{n'} |p_{ij} - p_{i'j}| \right\}.$$

**Example 3.2.** The values of diameters (typical of those found in many exercises) for each of the CPTs of the Native Fish BN from Nicholson and Flores [2011], together with those obtained in an elicitation exercise associated with the pollinator example Barons et al. [2018a] are given in the tables below. Discrepancies passing through CPTS with diameters close to 1 might be retained as different target distributions. However, once discrepancies pass through more than two CPTs with diameters less than 0.7 these usually quickly dissolve, for reasons we discuss below.

| Node                  | Diameter | Node                  | Diameter |
|-----------------------|----------|-----------------------|----------|
| Annual Rainfall       | 0.65     | Crop Yield            | 0.98     |
| River Flow            | 0.98     | Irrigation            | 0.94     |
| Pesticide in River    | 0.7      | Tree Condition        | 0.7      |
| Native Fish Abundance | 0.84     |                       |          |

Figure 4: Diameters of each CPT in the Native Fish BN.

| Node                          | Diameter |
|-------------------------------|----------|
| Honey Bee Abundance           | 0.66     |
| Other Bee Abundance           | 0.55     |
| Other Pollinator Abundance    | 0.54     |

Figure 5: Diameters of each CPT in Pollinator sub-network BN.
The size of the diameter of a conditional probability table $P$ is a measure of the dependence of $Y$ on $X$. This is because whenever $Y \perp X$ all rows of $P$ will be equal and so $d^+(P) = 0$. It is easy to check that whenever some non-trivial function $\tau(Y)$ of $Y$ can be written as a deterministic function of $X$ then $d^+(P) = 1$, its maximum value. So when there is only a weak relationship between $Y$ and $X$, in the sense that changing the different levels of $X$ impacts only slightly on the conditional mass function of $Y$, then $d^+(P) \approx 0$. Note that unless $P$ is symmetric, the diameter of $Y$ on $X$ is not the same as the diameter of $X$ on $Y$, in fact the difference between these can be arbitrarily close to 1 [see Wright, 2018].

The $I$-local diameter has the same property, where this time it is conditional on $X$ taking values only in the set of levels $I$. This is useful when comparing the efficacy of deleting a parent in a BN or when combining a collection of rows of the CPT/levels of $X$ into a single entry: see below.

### 3.2 Variation and Mixtures

#### 3.2.1 Approximations associated with mixing

A useful and well-known property of total variation is its convexity under mixing in the following sense. Let $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$, $\pi' = (\pi'_1, \pi'_2, \ldots, \pi'_{n'})$, $q_i = (q_{i1}, q_{i2}, \ldots, q_{im})$, $p_i = (p_{i1}, p_{i2}, \ldots, p_{im'})$ and define

$$q_\pi \triangleq \sum_{i=1}^{n} \pi_i q_i, \quad p_{\pi'} \triangleq \sum_{i=1}^{n'} \pi'_i p_i,$$

then

**Lemma 3.1.**

$$d_V(p_{\pi'}, q_\pi) \leq \sum_{i=1}^{n} \sum_{i' = 1}^{n'} \pi_i \pi'_{i'} d_V(p_{i'}, q_i).$$

**Proof.** See Appendix A.1.

In particular, if we know extremal distributions are small then so are convex linear combinations of these. Such processes occur for example in the calculation of a margin: here of a target variable. This enables us to prove a number of useful results concerning the contraction of error under learning in a BN: see below.

Combining our new definitions of diameter with variation distance we prove the following result that enables us to track this distance through a given BN:

**Theorem 3.2.** Let $\pi_1$ and $\pi_2$ be two possible margins of vectors $X$ and $Y$ of random variables and suppose that $P(Y|X)$ is the (shared) CPT of the concatenated levels of the conditional $Y|X$ and that $\rho_1$ and $\rho_2$ are the margins of $Y$. Then

$$d_V(\rho_1, \rho_2) \leq d^+(P(Y|X)) d_V(\pi_1, \pi_2).$$

**Proof.** See Appendix A.2.
This property will be exploited below in the study of BNs. Note for example that if \( P(Y|X) \) has been specified accurately, but that the margin \( \pi_1 \) is uncertain, then our marginal beliefs about \( Y \) are no more uncertain than those about \( X \), because by definition \( d^+(P(Y|X)) \leq 1 \). More importantly we have a bound on how much our uncertainty, quantified in terms of total variation, reduces in terms of \( d^+(P(Y|X)) \) – a measure of how far away \( Y \) is from independence of \( X \).

**Example 3.3.** Let us once again look at the CPT of ‘Tree Condition’, \( P \), which had a binary parent ‘Drought’ and a three-state parent ‘Rainfall’. The joint distribution can be calculated from CPTs as 

\[
\pi(X,Y) = \sum_X \pi(X) \cdot P(Y|X)
\]

However, is we knew only the extreme entries of \( P \) then we could still calculate our bound which is of the right order of magnitude: a property we have found to be typical of the types of CPTs we habitually elicit.

### 3.2.2 A Global Bound Approximation

There is another bound which applies when not only a margin \( \pi_1 \) of \( X \) is perturbed to \( \pi_2 \), but also the conditional mass functions of \( Y|X \) is simultaneously perturbed. Occasionally we need variation bounds on the consequent perturbation on the margins \( \rho_1, \rho_2 \) of \( Y \):

**Definition 3.4.** Let the superbound, \( d_V^*(P,Q) \), between stochastic matrices \( P \) and \( Q \) be defined by

\[
d_V^*(P,Q) \triangleq \max_{1 \leq i, i' \leq n} d_V(P_{i,i'}, Q_{i,i'}) \leq 1.
\]

So here we compare variation distances between each row of \( P \) and possibly different rows of \( Q \) before selecting the largest difference. Note by definition and the triangle inequality that

\[
d_V^*(P,Q) \leq d_V^+(P,Q) \leq \max \{ d_V^+(P,Q) + \max \{ d(P), d(Q) \} , 1 \}.
\]

**Example 3.4.** Let us compare the two alternative CPTs, \( P \) and \( Q \), for the ‘Tree Condition’ node as introduced in Example 3.1. The value of \( d_V^*(P,Q) \) can be calculated directly from the total variation distance between every possible pairwise combination of rows in \( P \) and \( Q \). For this example \( d_V^*(P,Q) = 0.7 \) corresponding to \( d_V(p_1, q_0) = d_V(p_0, q_1) \).

Let \( P_{A|B}, Q_{A|B} \) represent respectively the conditional probability mass functions of \( X_A | X_B \) under the hypothesis and alternative given \( X_B \), where without loss we can assume that \( A \) and \( B \) are disjoint. Notice that these can be seen as CPTs whose rows correspond to the different values of the vector \( x_B \). Then under our definitions of transition matrices above whenever \( X_A \perp\!\!\!\perp X_B \)

\[
d_V^*(P_{A|B}, Q_{A|B}) = d_V^*(P_{A|B}, Q_{A|B}) = d_V(p_A, q_A).
\]
This arises simply because $X_A \equiv X_B$ implies that all rows in the CPT matrix are equal to each other and so equal to the corresponding margin on $X_A$. Thus we see that standard analyses that elicit irrelevances or independences translate here into equations on variation distance. We will see later that this enables us to study the implications of models where the embedded conditional independences are only approximately true.

**Definition 3.5.** The stochastic variation matrix $D^+(P) = \{D+(i, j)\}_{1 \leq i, j \leq n}$ is the $n \times n$ symmetric matrix whose entries are the variation distances between the different rows of the matrix $P$.

We will later use this construction to draw out useful functions of the explanatory variables associated with a particular variable of focus.

Now note that we can write
\[
\pi_1 = (1 - \beta)\pi_1^* + \beta\pi_{1\&2}, \quad \pi_2 = (1 - \beta)\pi_2^* + \beta\pi_{1\&2},
\]
where $(1 - \beta) = d_V(\pi_1, \pi_2)$ and where without loss we can assume the mixing process is shared by the two mass functions, so points are drawn either from $\pi_{1\&2}$ or alternatively something drawn from either $\pi_1^*$ or $\pi_2^*$ (see Supplementary Material for more a more detailed construction). Using the same argument as for when $P_1 = P_2$
\[
d_V(\rho_1, \rho_2) = d_V(\pi_1 P_1, \pi_2 P_2)
\]
\[
= d_V((1 - \beta)\pi_1^* + \beta\pi_{1\&2}) P_1, ((1 - \beta)\pi_2^* + \beta\pi_{1\&2}) P_2)
\]
\[
\leq \beta d_V(\pi_{1\&2} P_1, \pi_{1\&2} P_2) + (1 - \beta) d_V(\pi_1^* P_1, \pi_2^* P_2)
\]
\[
\leq \beta d_V^*(P_1, P_2) + (1 - \beta) d_V^*(P_1, P_2).
\]

We can then show
\[
d_V(\rho_1, \rho_2) \leq d_V^*(P_1, P_2) + d_V(\pi_1, \pi_2) d_V^*(P_1, P_2).
\]

Using Equation 3 in particular we have that
\[
d_V(\rho_1, \rho_2) \leq \{1 + d_V(\pi_1, \pi_2)\} d_V^*(P_1, P_2) + d_V(\pi_1, \pi_2) \max\{d(P_1), d(P_2)\}.
\]

## 4 Approximations of the CPTs in a known BN

Suppose all clients are content that the conditional independences in a given BN are valid. Without changing the random variables in the system we are now interested in finding ways of approximating the graphical model and refining initial probability estimates within this given BN.

### 4.1 Diameter Bounds when Marginalising or Conditioning

We now present some basic results about diameters of the transition matrices between two vectors of random variables under various marginalisations and conditioning of the subvectors. These bounds are particularly helpful when moving from a BN to a junction tree.

Let $X = (X_1, X_2)$, $Y = (Y_1, Y_2)$ and $P_{Y|X}$ ($P_{Y|X_1}$) be, respectively, the transition matrix associated with the conditional distribution of $Y|X$ (the same conditional distribution $Y|X_1$ but now with $X_2$ marginalised out). Let $d^+(P_{Y|X})$, $(d^+(P_{Y|X_1}))$ denote their respective diameters.
Lemma 4.1.
\[ d^+(P_{Y|X_1}) \leq d^+(P_{Y|X}). \]

Proof. This is immediate since each of the rows of \( P_{Y|X_1} \) is a weighted average (the weights on row labelled \( x_1 \) corresponding to the masses on \( X_2|X_1 = x_1 \)).

Note that this bound is tight in the sense that it is attained for a particular distribution on \( X_2|X_1 = x_1 \). Suppose \( d^+(P_{Y|X}) \) is attained when we compare the row \((x_1, x_2)\) with \((x_1', x_2')\) and \( P(X_2 = x_2|X_1 = x_1) = 1 \) and \( P(X_2 = x_2'|X_1 = x_1') = 1 \), then it is easy to check that \( d^+(P_{Y|X}) \leq d^+(P_{Y|X}). \)

Lemma 4.2. Using the obvious notation, for any two joint probability mass functions \( p_{X,Y}(x, y), p'_{X,Y}(x, y) \) over \( X, Y \)
\[ d_V(p_{X,Y}(x, y), p'_{X,Y}(x, y)) \leq \inf \left\{ d_V(p_X(x), p'_X(x)) + \sup_x d_V(p_{Y|X}(y|x), p'_{Y|X}(y|x)), 1 \right\}. \]

Proof. See Appendix A.3.

Finally, we can determine a bound on the diameter of a CPT in which many variables are dependent on the same set. This will often be the case when we are looking at a simple path of a junction tree in which a separator contains more than one variable:

Lemma 4.3. \( d^+(P_{Y|X}) \leq \inf \{ d^+(P_{Y_1|X}) + d^+(P_{Y_2|X,Y_1}), 1 \}. \)

Proof. See Appendix A.4.

These results may seem trivial, however they enable us to bound the diameters of CPTs in our junction tree path, using the diameters already calculated from the original CPTs in the BN. This enables us to study the robustness to misspecification without calculating any new information.

4.2 Diminishing tree propagated approximation error

The following result explains why when using standard propagation algorithms on updating one of the clique margins \( C_1 \), the knock on effect on the other clique margins becomes weaker and weaker as the updated cliques become progressively more remote from \( C_1 \) - a property Albrecht et al. [2014a] exploit in their work. Furthermore the extent of the deviation can be measured, in the sense that it can be bounded above. This enables us to bound the potential extent of error in the distributions of focus variables induced from the misspecification of structure or various CPTs in the BN. This is particularly useful when we elicit a large BN and want to know how far away from target nodes we need to elicit the corresponding CPTs accurately.

Theorem 4.4. Let \((C_1, C_2, C_3, \ldots, C_k)\), from \( C_1 \) to \( C_k \), be the minimal sequence of cliques with associated separators \((S_2, S_3, \ldots, S_k)\). Let each undirected edge of the marginalised junction tree be denoted by \( \delta_i \) for \( i = 1, 2, \ldots, k \); the diameter of the conditional probability table between the two sequential nodes,
Define the diminishing effect of errors downstream in the causal chain. For example $\delta_1 = d^+(P(S_2|C_1)), \delta_2 = d^+(P(S_2|S_2)), \ldots, \delta_k = d^+(P(C_k|S_k))$. Then

$$d_V(p_{C_i}(x_{C_i}), q_{C_i}(x_{C_i})) \leq d_V(p_{C_1}(x_{C_1}), q_{C_1}(x_{C_1})) \prod_{i=1}^{k} \delta_i.$$

**Proof.** By Lemma 2.1 we can rewrite our junction tree to marginalise over internal cliques leaving us with the graphical structure:

![Graphical Structure](image)

Let each undirected edge be denoted by $\delta_i$ for $i = 1, 2, \ldots, k$; the diameter of the conditional probability table between the two sequential nodes. Giving $\delta_1 = d^+(P(S_2|C_1)), \delta_2 = d^+(P(S_3|S_2)), \ldots, \delta_k = d^+(P(C_k|S_k))$. By successive application of Theorem 3.7:

$$d_V(p_{C_k}(x_{C_k}), q_{C_k}(x_{C_k})) \leq d^+(P(C_k|S_k))d_V(p_{S_k}(x_{S_k}), q_{S_k}(x_{S_k}))$$

$$\leq d^+(P(C_k|S_k))d^+(P(S_k|S_{k-1}))d_V(p_{S_{k-1}}(x_{S_{k-1}}), q_{S_{k-1}}(x_{S_{k-1}}))$$

$$\leq d^+(P(C_k|S_k))d^+(P(S_k|S_{k-1})) \ldots d^+(P(S_3|S_2))d^+(P(S_2|C_1))d_V(p_{C_1}(x_{C_1}), q_{C_1}(x_{C_1}))$$

$$= \left( \prod_{i=1}^{k} \delta_i \right) d_V(p_{C_1}(x_{C_1}), q_{C_1}(x_{C_1}))$$

Next we define the impact of one clique upon another in order to ascertain the diminishing effect of errors downstream in the causal chain.

**Definition 4.1.** Define the impact $I(C_k|C_1)$ of $C_1$ on $C_k$ to be $\prod_{i=1}^{k} \delta_i$.

The impact of one clique on another is a simple measure of the maximum possible influence the misspecification of one set of clique probabilities could have on another as measured by a bound on the variation distance. Note that in general we can label the edges of a junction tree (which are also labelled by a separator between adjacent cliques) $C_i$ and $C_j$ by two diameters $\delta_{i \rightarrow j}$ and $\delta_{j \rightarrow i}$, one measuring the impact of $i$ on $j$ and the other the impact of $j$ on $i$. Note that these two impacts are not necessarily equal, and are often very different. However, in the contexts we consider here (where our primary interest concerns the robustness of the margins of an identified subset of attributes) we usually need to focus on propagation in a single direction. Furthermore, if the BN is constructed consistently with a conjectured causal directionality in mind, then this directionality often tends to have the attributes at the end of the causal chain. This means that the diameters we need can often be calculated directly from the diameter of the elicited CPTs of the BN.

**Example 4.1.** The two simple BNs we have used in our running example are not deep enough to illustrate the usefulness of this result, whilst the full IDSS is far too complicated. So instead we use here a simplification of a BN used to model radicalisation processes one of the authors has elicited, where the precise
meaning of the nodes is confidential but not relevant to the points we mean to illustrate.

Let us label the cliques to satisfy the running intersection property:

\[ C_1 = \{X_1, X_2\}, \quad C_2 = \{X_2, X_3, X_4\}, \quad C_3 = \{X_4, X_5\}, \]
\[ C_4 = \{X_5, X_6, X_7\}, \quad C_5 = \{X_6, X_7, X_8\}, \quad C_6 = \{X_8, X_{10}\}, \quad C_7 = \{X_7, X_9\} \]

Giving us separators:

\[ S_2 = \{X_2\}, \quad S_3 = \{X_4\}, \quad S_4 = \{X_5\}, \quad S_5 = \{X_6, X_7\} \]
\[ S_6 = C_2 \cap C_6 = \{X_3\}, \quad S_7 = C_5 \cap C_7 = \{X_7\} \]

Suppose we wish to determine the effect on \( X_9 \) if we perturb \( X_1 \). Draw the ancestral graph of nodes \( X_1 \) and \( X_9 \), derive the impact formula (which is simply the product of diameters of each separator conditional on the previous separators):

\[
I(X_9|C_1) = p(X_2|X_1)p(X_4|X_2)p(X_5|X_4)p(X_7|X_5)p(X_9|X_7) 
\leq d^+(X_2)d^+(X_4)d^+(X_5)d^+(X_7)d^+(X_9)
\]

Extending this further, we can determine the impact on cliques \( X_6 \) and \( X_7 \) simultaneously, if we perturb both \( X_1 \) and \( X_2 \). Following the same steps of creating cliques and separators for the ancestral graph of these nodes, the impact is given as:

\[
I(X_6, X_7|X_1, X_2) = p(X_2|X_1)p(X_4|X_2)p(X_5|X_4)p(X_6, X_7|X_5). 
\]

This can be written in terms of the original BN CPTs using Lemma 4.3, as some separators contain more than one node:

\[
I(X_6, X_7|X_1, X_2) \leq d^+(X_2)d^+(X_4)d^+(X_5) \left[ \inf \left\{ d^+(X_6|X_5) + d^+(X_7|X_6, X_5), 1 \right\} \right] 
\leq d^+(X_2)d^+(X_4)d^+(X_5) \left[ \inf \left\{ d^+(X_6|X_5) + d^+(X_7|X_6, X_5), 1 \right\} \right] 
\]

There are various practical corollaries to the simple theorem above:

**Corollary 4.4.1.** If \( G \) is decomposable and \( C_i \) lies on the minimal sequence between \( C_1 \) and \( C_k \) then if all attributes are in \( C_k \) then the probabilities of \( C_i \) have higher influence on \( C_k \) than those of \( C_1 \).
As we indicated above, these bounds can be applied to any BN. We recommend following the construction below to ensure that your BN is in a suitable format to apply Theorem 4.4:

- Begin with a BN $G$, the diameters of whose CPTs have been provisionally elicited.
- Identify a donating variable or complete vector $X_i$ of $G$ and the vector of focus $X_k$.
- Find the ancestral set of $X_i, X_k$ in $G$.
- Construct the ancestral graph, $A$, which has variables $(X_1, X_2, ..., X_n)$ where the order of these vertices are chosen compatible with $G$.
- Create a triangularised version, $A^*$, of $A$ and find its junction tree $J$.
- Denote the clique containing $X_i$ as $C_1$ and the clique containing $X_k$, $C_k$.
- Find the single path $J^*$ starting from clique $C_1$ to $C_k$ labelling the cliques in order $C_1, C_2, ..., C_k$.
- Remove all variables that are not in one of these cliques.

Note that these influences provide a very useful tool for prioritisation of the elicitation in a BN. For example, if we can obtain estimates of influence across a junction tree (either from direct elicitation of $\delta$ or alternatively after having performed a preliminary coarse elicitation of the corresponding CPTs) then we can use these influences to identify which of those CPTs to refine. For example suppose all attributes consisting of the subvectors of variables of interest lie in a single clique. We can then follow the simple guidelines:

- Refine the elicitation of the CPTs whose attributes and parents lie in this clique,
- Elicit the CPTs associated with parents/separators with the most influence,
- Use the influence formula (Theorem 4.4) to guide the refinement of the CPTs associated with other parents or parents of parents.

### 4.3 Approximations associated with a general BN

In a junction tree each vector has just a single parent within a given compatible ordering. Of course in the case of a BN this is no longer necessarily true. We would still like to find the impact bound of one variable on another and so annotate each of its directed edges with a value between zero and one which reflects this. The result below gives us a way of coding this impact in a useful way.

Suppose $Y$, taking values $y \in Y$, is potentially dependent on $k$ vectors $X = (X_1, X_2, ..., X_k)$, taking values $x = (x_1, x_2, ..., x_k) \in X = X_1 \times X_2 \times \cdots \times X_k$. For $j = 1, 2, ..., k$ let $x_j^* \in X_j \triangleq X_{\{1,2,\ldots,k\}\setminus\{j\}}$ be a vector of values of other variables $X_j$. Let the CPT of $Y$ given $X$ be $P$ so that its diameter is given by

$$d^+(P) = \frac{1}{2} \max_{x,x' \in \mathcal{X}} \left\{ \sum_{y \in \mathcal{Y}} |p_{xy} - p_{x'y}| \right\}.$$
Definition 4.2. Let the diameter $d_j^+$ of $Y$ to $X_j$ be defined by

$$d_j^+ = \frac{1}{2} \max_{x_j \in \mathcal{X}_j} \max_{x_j' \in \mathcal{X}_j} \left\{ \sum_{y \in Y} |p_{xy} - p_{x'y}| \right\}.$$ 

So the diameter $d_j^+$ is the maximum extra effect varying the value of $x_j$ can have on the distribution of $Y$ for any fixed value $x_j^* \in \mathcal{X}_j$ of the other variables. Notice in particular that

$$Y \perp X_j | X_j^* \iff d_j^+ = 0.$$ 

Thus in a formal sense, $d_j^+$ is a measure of the extent by which this conditional independence is violated and the merit of knowing the value of $X_j$ once we knew the value of $X_j^*$. We now have the simple but pleasing additive relationship between $d_j^+(P)$ and $d_j^+, j = 1, 2, \ldots, k$:

Theorem 4.5. Under the notation above

$$d_j^+(P) \leq k \sum_{j=1}^k \delta_j.$$ 

Proof. Here we simply use the triangle inequality to bound $d_j^+(P)$ changing the entries of the conditioning variables $X_j$ one at a time. So if $x = (x_1, x_2, \ldots, x_k), x' = (x'_1, x'_2, \ldots, x'_k), x(0) = x, x(1) = (x'_1, x_2, \ldots, x_k), x(2) = (x'_1, x'_2, x_3, \ldots, x_k), \ldots, x(k) = (x'_1, x'_2, \ldots, x'_k)$, then

$$\max_{x,x' \in \mathcal{X}} \left\{ \sum_{j=1}^{n'} |p_{xy} - p_{x'y}| \right\} \leq \sum_{i=1}^k \max_{x(i), x(i-1) \in \mathcal{X}} \left\{ \sum_{j=1}^{n'} |p_{x(i)y} - p_{x(i-1)y}| \right\}$$

$$= \sum_{i=1}^k \max_{x_j \in \mathcal{X}_j} \max_{x_j' \in \mathcal{X}_j} \left\{ \sum_{j=1}^{n'} \sum_{y \in Y} |p_{xy} - p_{x'y}| \right\}.$$ 

since by definition of $x$ and $x'$, the adjacent CPTs appearing in the sum above differ only in the $i^{th}$ entry. \qed

4.4 Robustness to approximation by a sparser BN and edge deletion

Often a BN is chosen to be sparser than it would be were we to have more information or time. This happens for a variety of reasons. For example when eliciting a BN we often ask for the list of the most important variables on which a specific variable $X$ might depend, defining this phrase by asking that variables not included in the list could be expected to have only a small influence on $X$. This restriction is imposed because it is difficult for a client to think clearly about the interrelationships between more than a handful of variables. Increasing the number of different joint levels on the conditioning variables quadratically increases the number of entries in the CPT that need to be elicited.

If data is used to inform the model choice then a severe penalty is often imposed or is implicit to ensure the selection of models with smaller size parameter
spaces – which in this context usually implies sparser associated graphs. Finally, for reasons of implementability, it is quite common for a search of candidate BNs to include only those graphs whose nodes have no more than a fixed number of parents: the limit often set to be two or three, see e.g. Cussens [2011].

It is therefore very important to properly understand the implications of these potential over-simplifications on the robustness of the BN model. We present some corollaries on the use of these results.

4.5 Edge Deletion

When constructing BN systems we seek the model which best describes the underlying physical process. However, in reality we are often limited by restrictions on resources and time which lead us to instead develop a model that is as large as it need be. The size of a BN can grow exponentially by adding more variables, relationships or even states. So eliciting these probabilities becomes problematic. Therefore, a frequent simplification of a model can be to delete unnecessary edges that impact only a little on the outputs of the system.

The methodology we have introduced in earlier sections can be directly applied to this problem in order to quantify the cost of removing a certain variable from the parent set of another variable.

Example 4.2. Return once again to the CPT of ‘Tree Condition’ given its binary parent ‘Drought’ and three-state parent ‘Annual Rainfall’, as in Example 3.1. Denote the rows of $P$ by $p_i$ for $i = 1, 2, \ldots, 6$. If we are interested in the effects of deleting the arc $\text{Drought} \rightarrow \text{Tree Condition}$, then we can use Definition 4.2 directly:

$$\max_{\{yes, no\}} \max_{\{\text{Below}, \text{Avg.}, \text{Above}\}} \{d_V(p_i, p_i')\} = 0.6$$

Alternatively, deleting arc $\text{Annual Rainfall} \rightarrow \text{Tree Condition}$ gives us

$$\max_{\{\text{Below}, \text{Avg.}, \text{Above}\}} \max_{\{yes, no\}} \{d_V(p_i, p_i')\} = 0.2$$

Below we present the full BN with each edge annotated with the quantitative effect on the child if we delete the parent.
The higher the value of $\delta_j$ the larger the effect of deleting the arc $X_j \rightarrow Y$, because it is indicating that the corresponding rows of other parent responses are dissimilar.

Now of course this model is already well designed and so we find that most edges need to be there. However, the possible exception if the edge from Irrigation to River Flow. In candidate models devised early in an analysis, or in large BNs, we often find that many edges can be omitted without strong effects on the outputs of the system.

Note that this process scales up easily to handle parent sets of discrete BNs with numerous nodes and large numbers of states, due to the variation distance being a simple and transparent piece of arithmetic.

### 4.6 Level amalgamation

One practical issue found by discrete BN modellers is the number of levels each random variable within the system should be assigned. Obviously there is a trade-off here. The finer the division of levels, the more nuanced the BN can be. On the other hand, the fewer the number of levels, the easier it will be to faithfully elicit or efficiently estimate the probabilities within a BN. One advantage of using the variation approach for robustness is that such considerations can be taken under the same technical umbrella as other necessary approximations, such as edge deletion. We simply evaluate the potential that such simplifications might have on the distribution of the attributes of the problem, just as when considering whether or not to keep a weak edge in the system.

When considering amalgamating levels within a specified random variable we must ensure that the interpretation of the states can still be understood and quantified by experts. When those variables are ordinal and have a monotonic relationship with its neighbours we would therefore recommend that you only consider amalgamating consecutive node levels. For example, the node of Annual Rainfall in Example 3.1 had levels \{BelowAverage, Average, AboveAverage\} and if we were to combine Below Average with Above Average we would not
have a cohesive state for the experts to quantify.

The second step in level amalgamation is deciding how to combine the probabilities associated with those levels that are to be amalgamated. We recommend taking a simple row average, because the convexity of variation distance tells us that this will enable us to avoid increasing the diameter of the original CPT and more importantly this method does not require additional information (which would otherwise be needed to perform a weighted row average).

Occasionally the modeller or expert will have an intuitive feel for which states should be combined, possibly from past experience or relevant data. However, sometimes this may not be obvious. In the latter case, we can calculate the variation distance between the considered states and combine the closest states first, then find the next closest state and add to the amalgamation iteratively until the combination appears to induce significant variation distance from the original.

Example 4.3. For example, let us once again consider the ‘Tree Condition’ CPT from 3.1. To reduce the three-state node Annual Rainfall to a binary node we could either combine Below Average with Average (case i) or Above Average with Average (case ii). To decide we compare the variation distance between rows $p_i$:

Case i: $\max\{d_V(p_1, p_2), d_V(p_4, p_5)\} = 0.1$

Case ii: $\max\{d_V(p_2, p_3), d_V(p_5, p_6)\} = 0.1$

In this instance we can arbitrarily choose between the two, so we shall opt for case (i) to form the amalgamated state ‘Average or Below’ using a simple average of relevant rows to obtain $P'$:

$$P = \begin{pmatrix}
0.20 & 0.60 & 0.20 \\
0.25 & 0.60 & 0.15 \\
0.30 & 0.60 & 0.10 \\
0.70 & 0.25 & 0.05 \\
0.80 & 0.18 & 0.02 \\
0.90 & 0.09 & 0.01
\end{pmatrix}, \quad P' = \begin{pmatrix}
0.225 & 0.60 & 0.175 \\
0.30 & 0.60 & 0.10 \\
0.725 & 0.215 & 0.06 \\
0.90 & 0.09 & 0.01
\end{pmatrix}.$$

Calculating that $d^+_V(P, P') = 0.075$ in results such as Theorem 3.2 shows that the effect of using this amalgamated CPT rather than the original is small.

5 Some principled strategies for BN creation

Obviously the evaluation criteria we indicate here can be embedded into a formal protocol. However, there are many considerations that a user has to consider before undertaking model construction: transparency of the model, computational issues, elicitation constraints and so forth, which vary in importance depending on the context of the model building. So setting a bound on any effects or perturbations against differing approaches is often best undertaken informally. However, we acknowledge that the framework we have presented here is sufficiently formal to admit generalisation and this is work that we plan to undertake next.

To implement our techniques as efficiently as possible we recommend two differing approaches tailored to the specific circumstances of the modeller. Firstly,
there are occasions in which we have obtained provisional information from one expert who can recommend nodes, levels, interactions and provisional CPTs before undertaking a more formal elicitation conference with multiple experts. Such was the case in the pollinator example discussed in Barons et al. [2018a] and Barons et al. [2018b]. In this particular scenario we can begin to design the analysis by using the bounds discussed earlier on the preliminary values stated by the expert. We recommend starting by eliciting attributes and nodes of interest before working systematically backwards along the chain of inference to discover parent nodes and conditional independences, performing variation measures on preliminary CPT values to determine the efficacy of including variables in the model. Of course after the full elicitation has taken place the robustness analyses suggested above can be repeated for a final sensitivity analysis.

In situations when we are starting the model with no such preliminary information it may be wisest to attempt to elicit the value of the diameter of the CPT directly, before eliciting the full matrix, so that full elicitation is not undertaken before we can derive concrete bounds on the usefulness of this data harvesting exercise. This can then be bounded and decisions undertaken on whether to include certain variables in the chain or not. To elicit the diameter directly we need to ascertain the largest differences between rows of a CPT, which corresponds to requesting the “best case scenario” probabilities and the “worst case scenario” probabilities before calculating the variation distance between the two.

By following this procedure we therefore continuously appraise and compare each possible simplification against the potential accuracy of an analysis, weighted against the issues provided by a simpler representation of a model. We have demonstrated above that it can be proved that in many cases the effects of various simplifications are often very small, and approximations based on these simplifications are justified from a pragmatic point of view. We also note that some of the best approximations to use are often not the ones currently undertaken in practice. For example we often find that using an approximation which deletes an edge can cause significant changes, whilst allowing dependence only on subsets of levels performs much better.

6 Conclusions

We have demonstrated here how the properties of variation distance can be harnessed to study the robustness of a discrete BN, if certain target variables are known a priori to be those of primary interest. Although all our illustrative examples in this paper have perforce been of moderate size, our methods become ever more useful as the number of nodes in the BN increases. Even when this number is huge, we can show it is possible to identify a priori which features of the full joint distribution will have the strongest impact on the target variables of interest, and therefore employ effective and expedient approximations to make inferences which are both provably accurate and feasible for the task at hand. In such models, since the simple paths between learned variables and attribute variables is typically much longer, it is possible to formally demonstrate that some remote variables are just not worth eliciting directly, but should be marginalised over.

The approach we have introduced in this paper relies on the well-studied
variation distance which naturally embeds conditional independence relationships between variables. We have therefore devised a seamless way of looking at perturbed versions of a BN in a manner which enables us to apply the same devices to generic effects, be these perturbations associated with edge deletion, the knock-on effect of learning certain variables, misquoting probabilities within CPTs or changing the number of levels for nodes. All of these different alterations can be compared on an equal footing whereas previous work usually depended on model selection of BNs, using methodology such as Bayes Factors, which focused on one particular perturbation at a time. We note that many of the techniques to communicate visually the bounding effect of one node on a target node can be straightforwardly adapted and reapplied to this domain, such as the use of heatmaps demonstrated in Albrecht et al. [2014b].

There is of course much work to still be undertaken in this field, starting with refining the bounds we have developed here. The robustness studies we have introduced can also be applied to context-specific BNs where we have a natural trade-off between the number of probabilities to elicit and the robustness of the model. Typically if we elicit less probabilities for a context-specific model we can show that we weaken the robustness of the system due to the constraint of forcing inputs to be the same. Similarly within this paper we have had no space to consider the robustness to the choice of probability distribution on the entries of the CPTs of a BN. In Smith and Daneshkhah [2010], BN robustness associated with the inputs of the distribution in terms of the local DeRobertis distance is studied. Smith and Rigat [2012] provide bounds on posterior variation distances. Therefore a fairly straightforward extension to the variation bounds we have presented here can be developed by carefully combining our results with the local DeRobertis distance to provide a comprehensive robustness analysis when necessary. Essentially we can show that with sufficient data and global independence assumptions the most robust CPTs are the ones whose probabilities are best known.

Finally, recent theoretical advances have suggested that if a Bayesian accepts the their model is only approximate, the M-open scenario [Bernardo and Smith, 1994], then Bayesian learning using Bayes Rule may not be optimal and that other updating rules based on divergences other than the KL divergence should be considered. There are exciting new possibilities of combining this technology with the robustness methods described here when that divergence is defined as the variation distance [Jewson et al., 2018].

Our ideas also apply directly to the Dynamic BN where the robustness of the system can be even more important because the dynamic nature of the problem makes the model much more complex. Throughout this paper, for simplicity, we have considered only robustness as it applies to finite discrete BNs. However the whole technology we describe here of course translates seamlessly into tools for examining continuous and mixed Bayesian Networks. Using the variation distance on these highly structured and complex Markov Processes using the approach highlighted here can help us to determine the robustness of DBNs to dynamic effects. Work in this more general setting has already begun. We hope we have demonstrated that this is actually a very fruitful way of addressing robustness within this family of graphical models.
References

D. Albrecht, A.E. Nicholson, and C. Whittle. Structural sensitivity for the knowledge engineering of bayesian networks. In Probabilistic Graphical Models: 7th European Workshop, chapter 1. Springer, 2014a. 2, 14

D. W. Albrecht, A. E. Nicholson, and C. Whittle. Structural sensitivity for the knowledge engineering of bayesian networks. In L. C. van der Gaag and A. J. Feelders, editors, Proceedings of the Seventh European Workshop on Probabilistic Graphical Models (PGM 2014), volume 8754, pages 1–16. Springer, Cham, 2014b. 2, 23

M. J. Barons, A. M. Hanea, S. K. Wright, K. C. R. Baldock, L. Wilfert, D. Chandler, S. Datta, J. Fannon, C. Hartfield, A. Lucas, J. Ollerton, S. G. Potts, and N. L. Carreck. Assessment of the response of pollinator abundance to environmental pressures using structured expert elicitation. Journal of Apicultural Research, 57(5):593–604, 2018a. 6, 10, 22

M. J. Barons, S. K. Wright, and J. Q. Smith. Eliciting probabilistic judgements for integrating decision support systems. In Dias L. C., A. Morton, and J. Quigley, editors, Elicitation: The science and art of structuring judgement, chapter 17, pages 445–478. Springer, 2018b. 6, 7, 22

J. M. Bernardo and A. F. M. Smith. Bayesian Theory. Wiley, Chichester, 1994. 23

T. Boneh. Ontology and Bayesian decision networks for supporting the meteorological forecasting process. PhD thesis, Clayton School for Information Technology, Monash University, 2010. 1

H. Chan and A. Darwiche. A distance measure for bounding probabilistic belief change. International Journal of Approximate Reasoning, 38(2):149–174, 2005. 3

V. M. Coupé and L. C. van der Gaag. Properties of sensitivity analysis of bayesian belief networks. Annals of Mathematics and Artificial Intelligence, 36(4):323–356, 2002. 2

R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. Probabilistic Networks and Expert Systems. Springer-Verlag, New York, 1999. 1

R. G. Cowell, R. J. Verrall, and Y. K. Yoon. Modeling operational risk with bayesian networks. Journal of Risk and Insurance, 74(4):795–827, 2007. 4

J. Cussens. Bayesian network learning with cutting planes. In F. Cozman and A. Pfeffer, editors, Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence, pages 153–160, 2011. 5, 19

EFSA. Guidance on expert knowledge elicitation in food and feed safety risk assessment. EFSA Journal, 12(6):3734, 2014. 6, 9

N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. Machine Learning, 29(2-3):131–163, 1997. 2
Gómez-Villegas, M. A., P. Main, and R. Susi. The effect of block parameter perturbations in Gaussian Bayesian networks: sensitivity and robustness. *Information Sciences*, 222:439–458, 2013. 2, 3

P. Gustafson and L. Wasserman. Local sensitivity diagnostics for Bayesian inference. *The Annals of Statistics*, 23(6):2153–2167, 1995. 2

J. Jewson, J. Q. Smith, and C. Holmes. Principles of Bayesian inference using general divergence criteria. *Entropy*, 20(6):442, 2018. 23

K. B. Korb and A. E. Nicholson. *Bayesian artificial intelligence, second edition*. CRC press, 2010. 1, 4, 6

K. B. Laskey. Sensitivity analysis for probability assessments in Bayesian networks. In *IEEE Transactions on Systems, Man, and Cybernetics*, volume 25, pages 901–909, 1995. 2

K. B. Laskey and S. M. Mahoney. Network engineering for agile belief network models. *IEEE Transactions on Knowledge and Data Engineering*, 12(4):487–498, 2000. 1

S. L. Lauritzen. *Graphical models*. Clarendon Press, Oxford, 1996. 4

S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 157–224, 1988. 4

M. Leonelli, C. Görgen, and J. Q. Smith. Sensitivity analysis in multilinear probabilistic models. *Information Sciences*, 411:84–97, 2017. 3

A. Nicholson, O. Woodberry, and C. Twardy. The “native fish” Bayesian networks. Technical report, Technical Report 2010/3, Bayesian Intelligence., 2010. 7, 8, 9

A. E. Nicholson and M. J. Flores. Combining state and transition models with dynamic Bayesian networks. *Ecological Modelling*, 222(3):555–566, 2011. 7, 10

A. E. Nicholson and N. Jitnah. Using mutual information to determine relevance in Bayesian networks. In *Pacific Rim International Conference on Artificial Intelligence*, pages 399–410. Springer, Berlin, Heidelberg, 1998. 2

A. O’Hagan, C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow. *Uncertain judgements: eliciting experts’ probabilities*. John Wiley & Sons, 2006. 3

B. O’Neill. Importance sampling for Bayesian sensitivity analysis. *International Journal of Approximate Reasoning*, 50(2):270–278, 2009. 2

S. Renooij. Bayesian network sensitivity to arc-removal. In *Proceedings of the Fifth European Workshop on Probabilistic Graphical Models, Helsinki.*, page 233, 2010. 2
M. Scutari and J. B. Denis. *Bayesian networks: with examples in R.* CRC press, 2014. 5

J. Q. Smith. *Bayesian decision analysis: principles and practice.* Cambridge University Press, 2010. 1, 3, 4, 6, 8

J. Q. Smith and A. Daneshkhah. On the robustness of bayesian networks to learning from non-conjugate sampling. *International Journal of Approximate Reasoning,* 51(5):558–572, 2010. 2, 23

J. Q. Smith and F. Rigat. Ioseparation and robustness in finite parameter bayesian inference. *Annals of the Institute of Statistical Mathematics,* 64(3): 495–519, 2012. 2, 23

J. Q. Smith, M. J. Barons, and M. Leonelli. Coherent frameworks for statistical inference serving integrating decision support systems. *arXiv preprint arXiv:1507.07394,* 2015. 6

S. K. Wright. *Robustness in Bayesian Networks.* PhD thesis, Department of Statistics, University of Warwick, 2018. 11

J. C. Zaragoza, L. E. Sucar, and E. F. Morales. A two-step method to learn multidimensional bayesian network classifiers based on mutual information measures. In *Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference (FLAIRS),* pages 644–649, 2011. 2
Appendix A  Proofs

A.1 Proof of Lemma 3.1

To prove Lemma 3.1 we shall first need some simple intermediary results:

Let \( p, q_0, q_1 \) be three vectors of mass functions and define

\[ q_\alpha \triangleq (1 - \alpha)q_0 + \alpha q_1, \]

for \( 0 \leq \alpha \leq 1 \).

**Lemma A.1.**

\[ d_V(p, q_\alpha) \leq (1 - \alpha)d_V(p, q_0) + \alpha d_V(p, q_1) \]

**Proof.** Note that for \( 0 \leq \alpha \leq 1 \),

\[ p = (1 - \alpha)p + \alpha p \]

so that

\[ 2d_V(p, q_\alpha) = \sum_{j=1}^{n} |p_j - (1 - \alpha)q_{0j} - \alpha q_{1j}| \]

\[ = \sum_{j=1}^{n} |(1 - \alpha)p_j + \alpha p_j - (1 - \alpha)q_{0j} - \alpha q_{1j}| \]

\[ = \sum_{j=1}^{n} |(1 - \alpha)(p_j - q_{0j}) + \alpha(p_j - q_{1j})| \]

Next note that if \((p_j - q_{0j}), (p_j - q_{1j})\) are the same sign then

\[ |(1 - \alpha)(p_j - q_{0j}) + \alpha(p_j - q_{1j})| = (1 - \alpha)|p_j - q_{0j}| + \alpha|p_j - q_{1j}| \]

whilst if \((p_j - q_{0j}), (p_j - q_{1j})\) are of different sign then

\[ |(1 - \alpha)(p_j - q_{0j}) + \alpha(p_j - q_{1j})| \leq \max\{(1 - \alpha)|p_j - q_{0j}|, \alpha|p_j - q_{1j}|\} \]

\[ \leq (1 - \alpha)|p_j - q_{0j}| + \alpha|p_j - q_{1j}|. \]

It follows that

\[ 2d_V(p, q_\alpha) \leq \sum_{j=1}^{n} (1 - \alpha)|p_j - q_{0j}| + \alpha|p_j - q_{1j}| \]

\[ = (1 - \alpha)\sum_{j=1}^{n} |p_j - q_{0j}| + \alpha\sum_{j=1}^{n} |p_j - q_{1j}| \]

\[ = 2\{(1 - \alpha)d_V(p, q_0) + \alpha d_V(p, q_1)\} \]

proving the result. \( \square \)
This results proves that the variation distance between a mass function $p$ and the mixture of two others is less than the mixture of the mass function $p$ and the two extremal distributions $q_0, q_1$. We can now extend this result so that it applies to any finite mixtures: let $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$ with $\sum_{i=1}^n \pi_i = 1$, $q_i = (q_{i1}, q_{i2}, \ldots, q_{in})$ and define

$$q_\pi \triangleq \sum_{i=1}^n \pi_i q_i.$$

Lemma A.2.

$$\text{d}_V(p, q_\pi) \leq \sum_{i=1}^n \pi_i \text{d}_V(p, q_i).$$

Proof. We shall proceed using induction. For the case $n = 2$ we have $q_\pi = \pi_1 q_1 + \pi_2 q_2$ with $\pi_1 + \pi_2 = 1$ which means $q_\pi = (1 - \pi_2)q_1 + \pi_2 q_2$. By applying the previous lemma we know that

$$\text{d}_V(p, q_\pi) \leq (1 - \pi_2)\text{d}_V(p, q_1) + \pi_2 \text{d}_V(p, q_2) = \sum_{i=1}^2 \pi_i \text{d}_V(p, q_i).$$

So this statement certainly holds true for $n = 2$. Assume that it is true for $n - 1$. Then note that for the case $n = n$:

$$q_\pi = \sum_{i=1}^n \pi_i q_i = \pi_1 q_1 + \pi_2 q_2 + \ldots + \pi_{n-1} q_{n-1} + \pi_n q_n$$

$$= (\pi_1 + \pi_2 + \ldots + \pi_{n-1})q_0 + \pi_n q_n$$

$$= (1 - \pi_n)q_0 + \pi_n q_n$$

$$= (1 - \alpha)q_0 + \alpha q_n$$

where $\alpha \triangleq \pi_n$ and for $i = 1, 2, \ldots, n - 1$, $\pi_i' \triangleq \pi_i(1 - \pi_n)^{-1}$, giving

$$q_0 \triangleq \sum_{i=1}^{n-1} \pi_i' q_i = \frac{1}{1 - \pi_n} \sum_{i=1}^{n-1} \pi_i q_i.$$

By the last lemma

$$\text{d}_V(p, q_\pi) \leq (1 - \alpha)\text{d}_V(p, q_0) + \alpha \text{d}_V(p, q_n)$$

where by the inductive hypothesis

$$\text{d}_V(p, q_0) \leq \sum_{i=1}^{n-1} \pi_i' \text{d}_V(p, q_i).$$

Substituting the inductive hypothesis into our equation gives

$$\text{d}_V(p, q_\pi) \leq \sum_{i=1}^{n-1} (1 - \alpha)\pi_i' \text{d}_V(p, q_i) + \alpha \text{d}_V(p, q_n)$$

$$= \sum_{i=1}^n \pi_i \text{d}_V(p, q_i)$$

as required. \qed
Lemma 3.1 stated that
\[ d_V(p_{\pi^*}, q_{\pi^*}) \leq \sum_{i=1}^{n} \sum_{i'=1}^{n'} \pi_i \pi_{i'} d_V(p_{i'}, q_i) \]

Proof. Here we simply use the symmetry of variation distance. From Lemma A.2 we have that
\[ d_V(p_{\pi^*}, q_{\pi^*}) \leq \sum_{i=1}^{n} \pi_i d_V(p_{\pi^*}, q_i) \tag{4} \]
but for \( i = 1, 2, \ldots, n \)
\[ d_V(p_{\pi^*}, q_i) = d_V(q_i, p_{\pi^*}) \leq \sum_{i'=1}^{n'} \pi_{i'} d_V(q_{i'}, p_{\pi^*}) = \sum_{i'=1}^{n'} \pi_{i'} d_V(p_{i'}, q_i) \tag{5} \]
So substituting Inequality 5 into Inequality 4 gives us our result. \( \square \)

A.2 Proof of Theorem 3.2
Theorem 3.2 stated that if we let \( \pi_1 \) and \( \pi_2 \) be two possible margins of \( X \), suppose that \( P \) is the (shared) CPT of \( Y\mid X \) and that \( \rho_1 \) and \( \rho_2 \) are the margins of \( Y \). Then
\[ d_V(\rho_1, \rho_2) \leq d^+(P)d_V(\pi_1, \pi_2) \]
Proof. For \( i = 1, 2, \ldots, n \) let \( \pi_1 \land \pi_2 \triangleq \min_i(\pi_1(i), \pi_2(i)) \) and let us define
\[ \pi_{1 \land 2} = \frac{\pi_{1 \land 2}}{\sum (\pi_1 \land \pi_2)} = \frac{\pi_1 \land \pi_2}{\beta} \]
So by definition \( 1 - \beta = d_V(\pi_1, \pi_2) \).
Let
\[ \pi'_1 = d_V(\pi_1, (\pi_1 \land \pi_2)) \]
\[ \pi'_2 = d_V(\pi_2, (\pi_1 \land \pi_2)) \]
However, we would like \( \pi'_1 \) and \( \pi'_2 \) to be densities, so let:
\[ \pi^*_1 = \frac{\pi'_1}{\sum \pi'_1} = \frac{\pi'_1}{1 - \beta} \]
\[ \pi^*_2 = \frac{\pi'_2}{\sum \pi'_2} = \frac{\pi'_2}{1 - \beta} \]
We know that \( \pi_1 = (\pi_1 \land \pi_2) + \pi'_1 \) and a similar statement holds for \( \pi_2 \), however we can now rewrite \( \pi_1 \) and \( \pi_2 \) as a sum of two densities:
\[ \pi_1 = \beta \pi_{1 \land 2} + (1 - \beta)\pi^*_1 \]
\[ \pi_2 = \beta \pi_{1 \land 2} + (1 - \beta)\pi^*_2 \]
Note that
\[ d_V(\rho_1, \rho_2) = d_V(\pi_1 P, \pi_2 P) \]
\[ = d_V \left( [\beta \pi_{1 \land 2} + (1 - \beta)\pi^*_1] P, [\beta \pi_{1 \land 2} + (1 - \beta)\pi^*_2] P \right) \]
\[ \leq (1 - \beta) d_V(\pi^*_1 P, \pi^*_2 P). \]
Now consider $\pi_1^*P, \pi_2^*P$ as different mixture mass functions on $Y$ given that $X$ takes each of its $n$ values, i.e. the different rows $p_i$ of $P$. Let $\pi_1^* = (\pi_{11}^*, \pi_{12}^*, \ldots, \pi_{1n}^*)$, $\pi_2^* = (\pi_{21}^*, \pi_{22}^*, \ldots, \pi_{2n}^*)$ and the rows of $P$ be $p_i = (p_{i1}, p_{i2}, \ldots, p_{in})$. Note that by simple matrix multiplication on the entries we can now write

$$
\pi_1^*P = ((\pi_{11}^*p_{11} + \pi_{12}^*p_{21} + \ldots + \pi_{1n}^*p_{n1}), (\pi_{11}^*p_{12} + \pi_{12}^*p_{22} + \ldots + \pi_{1n}^*p_{n2}), \ldots, \\
(\pi_{11}^*p_{1n} + \pi_{12}^*p_{2n} + \ldots + \pi_{1n}^*p_{nn}))
$$

$$
= \pi_{11}^* (p_{11}, p_{12}, \ldots, p_{1n}) + \pi_{12}^* (p_{21}, p_{22}, \ldots, p_{2n}) + \ldots + \pi_{1n}^* (p_{n1}, p_{n2}, \ldots, p_{nn})
$$

$$
= \sum_{i=1}^{n} \pi_{1i}^*p_i
$$

Similarly,

$$
\pi_2^*P = \sum_{i=1}^{n} \pi_{2i}^*p_i
$$

which means we can now employ Lemma 3.1

$$
d_V(\rho_1, \rho_2) \leq (1 - \beta)d_V(\pi_1^*P, \pi_2^*P)
$$

$$
\leq (1 - \beta) \sum_{i=1}^{n} \sum_{i'=1}^{n} \pi_{1i}^*\pi_{2i'}^*d_V(p_{i'}, p_i)
$$

$$
= (1 - \beta)\pi_1^TD_0(P)\pi_2^T
$$

So by definition of the diameter $d^+(P)$

$$
d_V(\rho_1, \rho_2) \leq (1 - \beta) \sum_{i=1}^{n} \sum_{i'=1}^{n} \pi_{1i}^*\pi_{2i'}^*d^+(P)
$$

$$
= (1 - \beta)d^+(P)
$$

$$
= d^+(P)d_V(\pi_1, \pi_2)
$$

\[\square\]

### A.3 Proof of Lemma 4.2

Lemma 4.2 stated that for any two joint probability mass functions $p_{X,Y}(x, y), p'_{X,Y}(x, y)$ over $X, Y$

$$
d_V(p_{X,Y}(x, y), p'_{X,Y}(x, y)) \leq \inf \left\{ d_V(p_X(x), p'_X(x)) + \sup_x d_V(p_{Y|X}(y|x), p'_{Y|X}(y|x)), 1 \right\}.
$$

**Proof.** Note

$$
2d_V(p_{X,Y}(x, y), p'_{X,Y}(x, y)) \triangleq \sum_{x,y} |p_{X,Y}(x, y) - p'_{X,Y}(x, y)|
$$

$$
\triangleq \sum_{x,y} |p_{Y|X}(y|x)p_{X}(x) - p'_{Y|X}(y|x)p'_{X}(x)|.
$$
Let \( r(y|x) \triangleq p'_{Y|X}(y|x) - p_{Y|X}(y|x) \). Then
\[
2d_V(p_{X,Y}(x,y), p'_{X,Y}(x,y)) = \sum_{x,y} |p_{Y|X}(y|x) (p_X(x) - p'_X(x)) - r(y|x)p'_X(x)|
\]
\[
\leq \sum_x \left\{ |p_X(x) - p'_X(x)| \left( \sum_y p_{Y|X}(y|x) \right) \right\} + \sum_x p'_X(x) |r(y|x)|
\]
\[
= \sum_x |p_X(x) - p'_X(x)| + \sum_x \left\{ p'_X(x) |r(y|x)| \right\}
\]
\[
\leq \sum_x |p_X(x) - p'_X(x)| + \sup_y |r(y|x)|
\]
\[
\triangleq 2d_V(p_X(x), p'_X(x)) + 2 \sup_x d_V(p_{Y|X}(y|x), p'_{Y|X}(y|x)).
\]

The result follows, since variation distance is, by definition, bounded by 1. This simple result leads to the next useful result: \( \square \)

**A.4 Proof of Lemma 4.3**

Lemma 4.3 stated that \( d^+(P_{Y|X}) \leq \inf \{ d^+(P_{Y_1|X}) + d^+(P_{Y_2|X,Y_1}), 1 \} \).

**Proof.** In Lemma 4.2 substituting \( p = p(y|x) \) and \( p' = p(y|x') \) gives us that for all values of \( x, x' \)
\[
d_V(p_{Y_1|X}(y_1, y_2|x), p_{Y_1|X}(y_1, y_2|x'))
\]
\[
\leq d_V(p_{Y_1|X=x}(y_1|x), p_{Y_1|X=x}(y_1|x')) + \sup_{(y_1, x), (y_1', x')} d_V(p_{Y_2|Y_1, X}(y_2|y_1, x), p_{Y_2|Y_1, X}(y_2|y_1', x')).
\]

But by definition of the diameter
\[
d^+(P_{Y_1|X}) = \sup_{a, a'} \{ d_V(p_{Y_1|X=x}(y_1|x), p_{Y_1|X=x}(y_1|x')) \},
\]
\[
d^+(P_{Y_2|X,Y_1}) = \sup_{(y_1, a), (y_1', a')} \{ d_V(p_{Y_2|Y_1, X}(y_2|y_1, x), p_{Y_2|Y_1, X}(y_2|y_1', x')) \}.
\]

The result follows. \( \square \)