Edge Evaluation in Bayesian Network Structures

Saaid Baraty[1]   **Dan A. Simovici**[1]

[1] University of Massachusetts Boston
Department of Mathematics and Computer Science,
100 Morrissey Blvd, Boston, Massachusetts 02125, USA
Email: {sbaraty,dsim}@cs.umb.edu

Abstract

We propose a measure for assessing the degree of influence of a set of edges of a Bayesian network on the overall fitness of the network, starting with probability distributions extracted from a data set. Standard fitness measures such as the Cooper-Herskowitz score or the score based on the minimum description length are computationally expensive and do not focus on local modifications of networks. Our approach can be used for simplifying the Bayesian network structures without significant loss of fitness. Experimental work confirms the validity of our approach.

Keywords: Bayesian belief network, Kullbach-Leibler divergence, entropy, edge pruning

1 Introduction

The construction of a Bayesian Network Structure from a data set that captures the probabilistic dependencies among the attributes of the data set has been one of the prominent problems among community of uncertainty researchers since early 90s. The problem is particularly challenging due to enormity of number of possible structures for a given collection of data.

Formally, a *Bayesian Belief Network* is a pair $(\mathcal{B}_s, \mathcal{B}_p)$, where $\mathcal{B}_s$ is a DAG (directed acyclic graph) which is commonly referred to as a *Bayesian Network Structure* (BNS), and $\mathcal{B}_p$ is a collection of distributions which quantifies the probabilistic dependencies present in the structure, as we discuss in detail below.

Each node of the BNS corresponds to a random variable; edges represent probabilistic dependencies among these random variables. BNS captures the split of the joint probability of a set of random variables, presented by its nodes, into a product of probabilities of its nodes conditioned upon a set of other nodes, namely the set of its *predecessors* or *parents*.

The set of values (or states) of a random variable Z is referred to as the *domain of* Z, borrowing a term from relational databases. This set is denoted by $\mathrm{Dom}(Z)$.

If a random variable X is a node of $\mathcal{B}_s$ with $\mathrm{Dom}(X) = \{1, \cdots, R_X\}$ and set of random variables $Pa_X = \{Y_1, Y_2, \cdots, Y_k\}$ as its set of parents, and if we agree upon some enumeration of set $\mathrm{Dom}(Pa_X) = \prod_{i=1}^{k} \mathrm{Dom}(Y_i)$, then we denote by θ_{lj}^X the conditional probability $P(X = l | Y_1 = y_1, Y_2 = y_2, \cdots, Y_k = y_k)$, where l is some state of X and $(y_1, \cdots, y_k)$ is the jth element of the enumeration. Also, we denote with $\theta_{\cdot j}^X$, the

probability distribution of X conditioned on its set of parents taking on the jth assignment of its domain. $\mathcal{B}_p$ is collection of distributions $\theta_{\cdot j}^X$ for all nodes X of $\mathcal{B}_s$ and $1 \le j \le |\mathrm{Dom}(Pa_X)|$.

Several scoring solutions have been proposed for evaluating the fitness of a BNS for representing probabilistic dependencies among attributes of a data set. The are two major approaches: scores based on maximization of the posterior probability of the network structure conditioned upon data, and scores based on MDL (Minimum Description Length) principle.

The first approach was initially introduced in Cooper & Herskovits (1993), where the scoring formula was derived based on a number of assumptions such as assuming that the distribution of tuples $(\theta_{1j}^X, \cdots, \theta_{R_X j}^X)$ is uniform for all X and j, or is a Dirichlet distribution. In Heckerman et al. (1995) the Dirichlet distribution assumption was replaced by the *likelihood equivalence* assumption and it was shown that under this new assumption tuples $(\theta_{1j}^X, \cdots, \theta_{R_X j}^X)$ obey a Dirichlet distribution.

The second approach is based on the minimum description length principle, introduced in Rissanen (1978), which stipulates that the best model for data is the one that minimizes the combined description length of the model and data. Later, in Lam & Bacchus (1994) this principle was applied to learning a BNS from data. The close relationship between these two approaches was shown in Suzuki (1999).

The application of these methods on learning the local structure in the conditional probability distributions with variable number of parameters that quantify these networks as opposed to attempting to learn the global structure at once was studied in Friedman & Goldszmidt (1998).

Both approaches are expensive to compute and do not focus on local modifications of networks. Our scoring scheme is much cheaper to compute when local modifications are desired and allows the assessment of the importance of individual edges on the global fitness of the network.

We examined this problem from the perspective of conditional entropy Simovici & Baraty (2008) by seeking a set of parents for a node that reduces the conditional entropy of that node in presence of its parents as much as possible. Our main interest in this paper is to evaluate the "importance" of a set of edges of a BNS in the presence of data by measuring the fitness loss of the BNS due to pruning the set of edges. The evaluation is obtained starting from the Kullbach-Leibler divergence between two probability distributions.

Later, we examine the relationship of the fitness loss measure introduced in this article and the conditional entropy, in particular, with the measure introduced in Simovici & Baraty (2008). Finally, we combine the two measures to get a new formula and justify its use to simplify a BNS that represents expert's prior knowledge of the domain without considering the data.

Let $\mathcal{D}$ be a data set and let us denote by $\mathbf{Attr}(\mathcal{D})$ its set of attributes. For $K = \{A_{i_1}, \ldots, A_{i_k}\} \subseteq \mathbf{Attr}(\mathcal{D})$ let $I_K = \{i_1, \ldots, i_k\}$ be its index set.

If $\mathbf{A} = (A_1, A_2, \ldots, A_n)$ is a permutation of $\mathbf{Attr}(\mathcal{D})$, let $\mathbf{A}_{I_K}$ be the sequence of attributes of set K ordered according to $\mathbf{A}$. Denote by $\mathrm{Dom}(\mathbf{A}_{I_K})$ the Cartesian product of the domains of the attributes in the sequence $\mathbf{A}_{I_K}$, that is,

$$\mathrm{Dom}(\mathbf{A}_{I_K}) = \mathrm{Dom}(A_{i_1}) \times \cdots \times \mathrm{Dom}(A_{i_k}).$$

For $\mathbf{a} = (a_1, \ldots, a_k) \in \mathrm{Dom}(\mathbf{A}_{I_K})$ we denote by $\mathbf{A}_{I_K} = \mathbf{a}$, the event

$$A_{i_1} = a_1, \ldots, A_{i_k} = a_k.$$

A BNS for data set $\mathcal{D}$ is a structure $\mathcal{B}_s$ with set of nodes $\mathcal{V}_s = \mathbf{Attr}(\mathcal{D})$ and set of edges $\mathcal{E}_s \subseteq \mathbf{Attr}(\mathcal{D}) \times \mathbf{Attr}(\mathcal{D})$. The attributes of the data set are treated as random variables. The BNS represents probabilistic dependencies among these attributes.

We denote by $\mathsf{BNS}(\mathcal{D})$ the set of all possible structures for $\mathcal{D}$ and by $\mathsf{BNS_A}(\mathcal{D})$ the set of all structures of $\mathsf{BNS}(\mathcal{D})$ which only contain edges (A, A') such that A precedes A' in the permutation $\mathbf{A}$. If $\mathcal{B}_s \in \mathsf{BNS_A}(\mathcal{D})$, then $\mathsf{Par}^{\mathbf{A}}_{\mathcal{B}_s}(i)$ is the index set of the set of parents of $A_i \in \mathbf{Attr}(\mathcal{D})$ in $\mathcal{B}_s$ according to $\mathbf{A}$. This notation is extended to sets of nodes as follows. If $V \subseteq \mathcal{V}_s$ and I_V is the corresponding index set of V, then $\mathsf{Par}^{\mathbf{A}}_{\mathcal{B}_s}(I_V)$ is $\cup_{i \in I_V} \mathsf{Par}^{\mathbf{A}}_{\mathcal{B}_s}(i)$. The $\mathcal{B}_s$ subscript and $\mathbf{A}$ superscript are omitted when it is clear from context.

The BNS in $\mathsf{BNS_A}(\mathcal{D})$ that contains the maximum number of edges is called *the complete BNS* for sequence $\mathbf{A}$, denoted by $\mathcal{B}^{\mathbf{A}}_{cs}$, and is depicted in Figure 1.
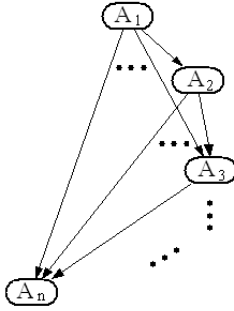

Figure 1: The complete BNS for ordering $\mathbf{A}$.

We make two basic assumptions:

1. The joint probability on the attributes of $\mathcal{D}$ can accurately be represented by a BNS $\mathcal{B}^{\max}_s \in \mathsf{BNS_A}(\mathcal{D})$ and such a structure maximizes the posterior probability of the structure conditioned upon the data set.

2. An uniform prior probability distribution exists on all possible Bayesian network structures for $\mathcal{D}$.

Under these assumptions, $\mathcal{B}^{\mathbf{A}}_{cs}$ has maximum posterior probability in the presence of data.

For any $\mathcal{B}_s \in \mathsf{BNS}(\mathcal{D})$ we have

$$P(\mathcal{B}_s|\mathcal{D}) \cdot P(\mathcal{D}) = P(\mathcal{D}|\mathcal{B}_s) \cdot P(\mathcal{B}_s)$$

by Bayes' Theorem. Since $\mathcal{D}$ is fixed, and we assume $P(\mathcal{B}_s)$ is uniform, $P(\mathcal{B}_s|\mathcal{D})$ is proportional to $P(\mathcal{D}|\mathcal{B}_s)$. Thus, it suffices to show that

$$\frac{P(\mathcal{D}|\mathcal{B}^{\mathbf{A}}_{cs})}{P(\mathcal{D}|\mathcal{B}^{\max}_s)} = 1.$$

Now, if we assume $\mathcal{D} = \{t_1, \cdots, t_d\}$, by independence assumption of tuples of data set we have,

$$\frac{P(\mathcal{D}|\mathcal{B}^{\mathbf{A}}_{cs})}{P(\mathcal{D}|\mathcal{B}^{\max}_s)} = \frac{\prod_{i=1}^{d} P(t_i|\mathcal{B}^{\mathbf{A}}_{cs})}{\prod_{i=1}^{d} P(t_i|\mathcal{B}^{\max}_s)}.$$

and by Bayesian network split of joint probability we have,

$$\frac{\prod_{i=1}^{d} P(t_i|\mathcal{B}^{\mathbf{A}}_{cs})}{\prod_{i=1}^{d} P(t_i|\mathcal{B}^{\max}_s)} =$$

$$\frac{\prod_{i=1}^{d} \prod_{j=1}^{n} P(A_j = t_i[j] | \mathbf{A}_{\mathsf{Par}_{\mathcal{B}^{\mathbf{A}}_{cs}}(j)} = t_i[\mathsf{Par}_{\mathcal{B}^{\mathbf{A}}_{cs}}(j)])}{\prod_{i=1}^{d} \prod_{j=1}^{n} P(A_j = t_i[j] | \mathbf{A}_{\mathsf{Par}_{\mathcal{B}^{\max}_s}(j)} = t_i[\mathsf{Par}_{\mathcal{B}^{\max}_s}(j)])}$$

where if t is a tuple in $\mathcal{D}$ and L is a set of attributes, then we denote the restriction of the tuple t to L be $t[L]$; we refer to $t[L]$ as the *projection* of t on L. Occasionally, we use $t[I_L]$ instead of $t[L]$, where I_L is the index set of L.

A Bayesian network structure for a data set incorporates a collection of conditional independence properties among attributes of that data set. This is captured by the *directed Markov property* of Bayesian networks Cowell (1998). This property stipulates that for any node X we have:

$$P(X|\mathsf{nd}(X), \mathsf{Par}(X)) = P(X|\mathsf{Par}(X)), \qquad (1)$$

which is denoted with $X \perp \mathsf{nd}(X) \mid \mathsf{Par}(X)$, where $\mathsf{nd}(X)$ is the set of non-descendent nodes of X. Note that $\mathsf{Par}_{\mathcal{B}^{\max}_s}(j) \subseteq \mathsf{Par}_{\mathcal{B}^{\mathbf{A}}_{cs}}(j)$ for $1 \leq j \leq n$. Then, since we assumed $\mathcal{B}^{\max}_s$ accurately represents the distribution over $\mathbf{Attr}(\mathcal{D})$ and by directed Markov property we have,

$$P(A_j = t_i[j] | \mathbf{A}_{\mathsf{Par}_{\mathcal{B}^{\mathbf{A}}_{cs}}(j)} = t_i[\mathsf{Par}_{\mathcal{B}^{\mathbf{A}}_{cs}}(j)])$$
$$= P(A_j = t_i[j] | \mathbf{A}_{\mathsf{Par}_{\mathcal{B}^{\max}_s}(j)} = t_i[\mathsf{Par}_{\mathcal{B}^{\max}_s}(j)], \mathbf{A}_{\mathcal{C}} = t_i[\mathcal{C}])$$
$$= P(A_j = t_i[j] | \mathbf{A}_{\mathsf{Par}_{\mathcal{B}^{\max}_s}(j)} = t_i[\mathsf{Par}_{\mathcal{B}^{\max}_s}(j)])$$
$$\text{(by Markov property)}$$

for all i and j where $\mathcal{C} = \mathsf{Par}_{\mathcal{B}^{\mathbf{A}}_{cs}}(j) - \mathsf{Par}_{\mathcal{B}^{\max}_s}(j)$. This justifies our proposition.

Yet, the complexity of a complete structure for a given sequence makes any computation prohibitively expensive. This demands the introduction of a measure which allows the simplification of the structure without incurring a significant loss of fitness. Such a measure may also be used to incrementally modify a BNS as new data becomes available.

2 Entropy and Partitions

A *partition* of a set S is non-empty collection of non-empty subsets of S, $\pi = \{B_i | i \in I\}$, such that $\bigcup_{i \in I} B_i = S$ and $B_i \cap B_j = \emptyset$ for all $i, j \in I$ where $i \neq j$. The set of partitions of a set S is denoted by $\mathsf{PART}(S)$.

A partial order relation on $\mathsf{PART}(S)$ is defined by $\pi \leq \sigma$ for $\pi, \sigma \in \mathsf{PART}(S)$ where $\sigma = \{C_1, C_2, \ldots, C_n\}$, if every block B_i of π is included in a block C_j of σ. The partially ordered set $(\mathsf{PART}(S), \leq)$ is actually a bounded lattice. The infimum of two partitions π and $\pi' = \{B_j | j \in J\}$ on S, denoted with $\pi \wedge \pi'$, is the partition $\{B_i \cap B_j | i \in I, j \in J, B_i \cap B_j \neq \emptyset\}$ on S . The least element of this lattice is the partition $\alpha_S = \{\{s\} \mid s \in S\}$; the largest is the partition $\omega_S = \{S\}$.

The notion of entropy for partitions of finite sets was and axiomatized in Simovici & Jaroszewicz (2002). If S

is a finite set and $\pi = \{B_1, \ldots, B_m\}$ is a partition of S, the entropy of π is the number

$$\mathcal{H}(\pi) = -\sum_{i=1}^{m} \frac{|B_i|}{|S|} \log_2 \frac{|B_i|}{|S|}.$$

Clearly, this is the Shannon entropy of a probability distribution $(p_1, \ldots, p_m)$, where $p_i = \frac{|B_i|}{|S|}$ for $1 \le i \le m$. The main advantage of using partitions rather than probability distributions is the possibility of using the partial order defined on $\mathsf{PART}(S)$. The following statement, proven in Simovici & Jaroszewicz (2002) is used in the sequel.

Theorem 2.1 *The entropy* $\mathcal{H} : \mathsf{PART}(S) \longrightarrow \mathbb{R}_{\ge 0}$ *is anti-monotonic; in other words, if* $\pi \le \pi'$, *then* $\mathcal{H}(\pi) \ge \mathcal{H}(\pi')$ *for every* $\pi, \pi' \in \mathsf{PART}(S)$.

The *trace of a partition* π on a subset T of S is the partition $\pi_T = \{T \cap B_i \mid i \in I \text{ and } T \cap B_i \ne \emptyset\}$ of T. Let $\pi, \sigma \in \mathsf{PART}(S)$ be two partitions, where $\pi = \{B_1, \ldots, B_m\}$ and $\sigma = \{C_1, \ldots, C_n\}$. The entropy of π conditioned on σ is the number:

$$\mathcal{H}(\pi|\sigma) = \sum_{j=1}^{n} \frac{|C_j|}{|S|} \mathcal{H}(\pi_{C_j}).$$

It is immediate that $\mathcal{H}_\beta(\pi|\omega_S) = \mathcal{H}_\beta(\pi)$ and that $\mathcal{H}(\pi|\alpha_S) = 0$. Also, in Simovici & Jaroszewicz (2006) it is shown that $\mathcal{H}_\beta(\pi|\sigma) = \mathcal{H}_\beta(\pi \wedge \sigma) - \mathcal{H}_\beta(\sigma)$, a property that extends the similar property of Shannon entropy.

The next theorem proven in Simovici & Jaroszewicz (2002), Simovici (2007) states that conditional entropy is anti-monotonic with respect to its first argument and is monotonic with respect to its second argument.

Theorem 2.2 *Let* $\pi, \sigma, \sigma' \in \mathsf{PART}(S)$, *where* S *is a finite set. If* $\sigma \le \sigma'$, *then* $\mathcal{H}(\sigma|\pi) \ge \mathcal{H}(\sigma'|\pi)$ *and* $\mathcal{H}(\pi|\sigma) \le \mathcal{H}(\pi|\sigma')$.

Finally, we mention the following corollary, also proven in Simovici & Jaroszewicz (2002).

Corollary 2.3 *Let* S *be a finite set. For every* $\pi, \sigma \in \mathsf{PART}(S)$ *we have* $\mathcal{H}(\pi|\sigma) \le \mathcal{H}(\pi)$.

Definition 2.4 The equivalence relation "$\sim^{\mathbf{A}_I}$" defined by the sequence of attributes $\mathbf{A}_I$ on $\mathcal{D}$, consists of those pairs $(t, t') \in \mathcal{D}^2$ such that $t[\mathbf{A}_I] = t'[\mathbf{A}_I]$.

The corresponding partition $\pi^{\mathbf{A}_I} \in \mathsf{PART}(\mathcal{D})$ is the *partition generated by* $\mathbf{A}_I$. □

It is clear that if $I' \subseteq I$ then $\pi^{\mathbf{A}_I} \le \pi^{\mathbf{A}_{I'}}$.

3 A Distribution Distortion Measure

Let us denote by $\mathbf{p}_{I_V}^{\mathsf{Par}(I_V)}(\mathbf{a})$ the conditional probability distribution,

$$(P(\mathbf{A}_{I_V} = \mathbf{b}_1 | \mathbf{A}_{\mathsf{Par}(I_V)} = \mathbf{a}), \ldots,$$
$$P(\mathbf{A}_{I_V} = \mathbf{b}_m | \mathbf{A}_{\mathsf{Par}(I_V)} = \mathbf{a})),$$

where $\mathrm{Dom}(\mathbf{A}_{I_V}) = \{\mathbf{b}_1, \ldots, \mathbf{b}_m\}$ and $\mathbf{a} \in \mathrm{Dom}(\mathbf{A}_{\mathsf{Par}(I_V)})$.

To avoid unnecessary complications we assume $I_V \cap \mathsf{Par}(I_V) = \emptyset$ in what follows, although the results that we have hold without this condition.

Let $E = \{(A_{s_1}, A_{d_1}), \ldots, (A_{s_r}, A_{d_r})\}$ be a subset of the set $\mathcal{E}$ of the edges of $\mathcal{B}_s$ and let $S_E = \{s_1, \ldots, s_r\}$ be the set of source nodes of edges of E and $D_E = \{d_1, \ldots, d_r\}$ be the set of destination nodes for E. We assume that $\mathsf{Par}(D_E) \cap D_E = \emptyset$. Note also that $S_E \subseteq \mathsf{Par}(D_E)$.

Clearly, if we remove the set of edges E from $\mathcal{B}_s$, the effect of this pruning on the joint probability distribution of the data set represented by $\mathcal{B}_s$ will only be through conditional distributions attached to nodes of D_E. Thus, to assess the effect of pruning the edges of E from $\mathcal{B}_s$ on the joint probability distribution of attributes of data set, consider the conditional probability distribution $\mathbf{p}_{D_E}^{(\mathsf{Par}(D_E)-S_E)}(\mathbf{a}')$,

$$(P(\mathbf{A}_{D_E} = \mathbf{b}_1 | \mathbf{A}_{(\mathsf{Par}(D_E)-S_E)} = \mathbf{a}'), \ldots,$$
$$P(\mathbf{A}_{D_E} = \mathbf{b}_m | \mathbf{A}_{(\mathsf{Par}(D_E)-S_E)} = \mathbf{a}')),$$

where $\mathrm{Dom}(\mathbf{A}_{D_E}) = \{\mathbf{b}_1, \ldots, \mathbf{b}_m\}$ and $\mathbf{a}' \in \mathrm{Dom}(\mathbf{A}_{(\mathsf{Par}(D_E)-S_E)})$.

Note that $\mathbf{p}_{D_E}^{(\mathsf{Par}(D_E)-S_E)}(\mathbf{a}')$ is the probability distribution of tuple $\mathbf{A}_{D_E}$ conditioned on its set of parents after the removal of the set of edges E instantiated with $\mathbf{a}'$. To see how much the distribution is distorted as we prune the set of edges E from $\mathcal{B}_s$, we compare the probability distributions of $\mathbf{A}_{D_E}$ conditioned on its set of parents in $\mathcal{B}_s$ before and after removal of the set of edges for all possible instantiations of the sequence of parents.

For each instantiation $\mathbf{a}' \in \mathrm{Dom}(\mathbf{A}_{(\mathsf{Par}(D_E)-S_E)})$ of the sequence of parents of $\mathbf{A}_{D_E}$ after pruning, there are several instantiations of the sequence of parents of $\mathbf{A}_{D_E}$ before pruning, $\mathbf{a}_1, \ldots \mathbf{a}_z$, where $\mathbf{a}_i \in \mathrm{Dom}(\mathbf{A}_{\mathsf{Par}(D_E)})$ such that $\mathbf{a}_i[\mathsf{Par}(D_E) - S_E] = \mathbf{a}'$ for $1 \le i \le z$. Thus, we need to compare the probability distributions

$$\mathbf{p}_{D_E}^{\mathsf{Par}(D_E)}(\mathbf{a}) \text{ and } \mathbf{p}_{D_E}^{(\mathsf{Par}(D_E)-S_E)}(\mathbf{a}'),$$

for $\mathbf{a} \in \mathrm{Dom}(\mathbf{A}_{\mathsf{Par}(D_E)})$ and $\mathbf{a}' \in \mathrm{Dom}(\mathbf{A}_{(\mathsf{Par}(D_E)-S_E)})$ such that $\mathbf{a}' = \mathbf{a}[\mathsf{Par}(D_E) - S_E]$. Then, we can linearly combine the divergence between the pairs, weighted by the probability of occurrence of $\mathbf{a}$ derived from data set $\mathcal{D}$.

To compare two finite probability distributions

$$\mathbf{p} = (p_1, p_2, \cdots, p_n) \text{ and } \mathbf{q} = (q_1, q_2, \cdots, q_n)$$

we use the *Kullbach-Leibler* divergence measure given by

$$\mathsf{KL}(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} p_i \cdot \log_2 \frac{p_i}{q_i}.$$

KL has well-known properties:

1. $\mathsf{KL}(\mathbf{p}, \mathbf{q}) \ge 0$ for all finite probability distributions $\mathbf{p}$ and $\mathbf{q}$.

2. $\mathsf{KL}(\mathbf{p}, \mathbf{q}) = 0$ if and only if $\mathbf{p} = \mathbf{q}$ element-wise.

However, $\mathsf{KL}(\mathbf{p}, \mathbf{q})$ has no upper bound which makes comparisons for realizing the level of differences among a set of distributions difficult. We overcome this problem by dividing our linearly weighted measure of differences among the conditional probability distributions before and after edge pruning by the same weighted measure, but this time among the conditional probability distributions before the edge removal and the non-informative uniform probability distribution $\mathbf{u}_m \in [0, 1]^m$:

$$\mathbf{u}_m = \left(\frac{1}{m}, \frac{1}{m}, \cdots, \frac{1}{m} \right),$$

where $m = |\mathrm{Dom}(\mathbf{A}_{D_E})|$. If $\mathbf{p}_{D_E}^{\mathsf{Par}(D_E)}(\mathbf{a}) = \mathbf{u}_m$ for all $\mathbf{a} \in \mathrm{Dom}(\mathsf{Par}(\mathbf{A}_{D_E}))$, then knowing the assignment of values to $\mathbf{A}_{\mathsf{Par}(\mathbf{A}_{D_E})}$ is completely non-informative in predicting the value of $\mathbf{A}_{D_E}$. Also, assuming $|\mathcal{D}|$ is a multiple

of m, the m-block partition on $\mathcal{D}$ which corresponds to the finite probability distribution $\mathbf{u}_m$, $\pi^{\mathbf{u}_m} = \{B_1, \cdots B_m\}$ where $|B_1| = \cdots = |B_m| = \frac{|\mathcal{D}|}{m}$ is referred to as m-*block uniform partition of* $\mathcal{D}$ and it has the maximum entropy, $\mathcal{H}(\pi^{\mathbf{u}_m}) = \log_2(m)$, over all possible partitions of $\mathcal{D}$ with m blocks.

Definition 3.1 The *distribution distortion* caused by removing the set of edges E from the BNS $\mathcal{B}_s$ denoted by $\mathsf{DD}_{\mathcal{B}_s}(E)$ where $D_E \cap \mathsf{Par}(D_E) = \emptyset$, is defined as

$$\frac{\sum_{\mathbf{a}} P(\mathbf{A}_{\mathsf{Par}(D_E)} = \mathbf{a}) \cdot \mathsf{KL}(\mathbf{p}_{D_E}^{\mathsf{Par}(D_E)}(\mathbf{a}), \mathbf{p}_{D_E}^{Q_E}(\mathbf{a}[Q_E]))}{\sum_{\mathbf{a}} P(\mathbf{A}_{\mathsf{Par}(D_E)} = \mathbf{a}) \cdot \mathsf{KL}(\mathbf{p}_{D_E}^{\mathsf{Par}(D_E)}(\mathbf{a}), \mathbf{u}_m)},$$

where D_E and S_E are defined as before and $Q_E = \mathsf{Par}(D_E) - S_E$. Also the sums are over all $\mathbf{a} \in \mathrm{Dom}(\mathbf{A}_{\mathsf{Par}(D_E)})$. $\square$

Theorem 3.2 *We have:*

$$\mathsf{DD}_{\mathcal{B}_s}(E) = \frac{\mathcal{H}(\pi^{\mathbf{A}_{D_E}}|\pi^{\mathbf{A}_{Q_E}}) - \mathcal{H}(\pi^{\mathbf{A}_{D_E}}|\pi^{\mathbf{A}_{\mathsf{Par}(D_E)}})}{\mathcal{H}(\pi^{\mathbf{u}_m}) - \mathcal{H}(\pi^{\mathbf{A}_{D_E}}|\pi^{\mathbf{A}_{\mathsf{Par}(D_E)}})},$$

where $m = |\mathrm{Dom}(\mathbf{A}_{D_E})|$.

Proof. See Appendix A. ∎

Corollary 3.3 *We have* $0 \le \mathsf{DD}_{\mathcal{B}_s}(E) \le 1$.

Proof. Since $Q_E \subseteq \mathsf{Par}(D_E)$, we have $\pi^{\mathbf{A}_{\mathsf{Par}(D_E)}} \le \pi^{\mathbf{A}_{Q_E}}$. By the monotonicity property of conditional entropy with respect to its second argument we have:

$$\mathcal{H}(\pi^{\mathbf{A}_{D_E}}|\pi^{\mathbf{A}_{\mathsf{Par}(D_E)}}) \le \mathcal{H}(\pi^{\mathbf{A}_{D_E}}|\pi^{\mathbf{A}_{Q_E}}).$$

Also,

$$\mathcal{H}(\pi^{\mathbf{A}_{D_E}}|\pi^{\mathbf{A}_{Q_E}}) \le \mathcal{H}(\pi^{\mathbf{A}_{D_E}}|\{\mathcal{D}\}) = \mathcal{H}(\pi^{\mathbf{A}_{D_E}}).$$

But we know,

$$\mathcal{H}(\pi^{\mathbf{A}_{D_E}}) \le \mathcal{H}(\pi^{\mathbf{u}_m}).$$

The result follows immediately. ∎

Theorem 3.4 *We have* $\mathsf{DD}_{\mathcal{B}_s}(E) = 0$ *if and only if*

$$P(\mathbf{A}_{D_E} = b_i \mid \mathbf{A}_{\mathsf{Par}(D_E)} = \mathbf{a}) =$$
$$P\left(\mathbf{A}_{D_E} = b_i \mid \mathbf{A}_{(\mathsf{Par}(D_E)-S_E)} = \mathbf{a}[\mathsf{Par}(D_E) - S_E]\right)$$

for all i, $1 \le i \le m$, *and* $\mathbf{a} \in \mathrm{Dom}(\mathbf{A}_{\mathsf{Par}(T)})$.

Proof. Note that we implicitly assume that $P(\mathbf{A}_{\mathsf{Par}(D_E)} = \mathbf{a}) \ne 0$ because, otherwise, $P(\mathbf{A}_{D_E} = b_i \mid \mathbf{A}_{\mathsf{Par}(D_E)} = \mathbf{a})$ is undefined. The statement follows from the second property of the KL measure. ∎

Theorem 3.5 *Let* E *and* E' *be two sets of edges of BNS* $\mathcal{B}_s$. *Then, if* $\mathsf{Par}(D_E) \cap D_E = \emptyset$, $D_E = D_{E'}$ *and* $S_E \subseteq S_{E'}$, *we have* $\mathsf{DD}_{\mathcal{B}_s}(E) \le \mathsf{DD}_{\mathcal{B}_s}(E')$.

Proof. Since $S_E \subseteq S_{E'}$, we have $Q_{E'} \subseteq Q_E$. Then, by the monotonicity property of conditional entropy with respect to second argument we have,

$$\mathcal{H}(\pi^{\mathbf{A}_{D_E}}|\pi^{A_{Q_E}}) \le \mathcal{H}(\pi^{\mathbf{A}_{D_E}}|\pi^{A_{Q_{E'}}}).$$

The result follows immediately. ∎

The global search for a set of edges to be pruned to simplify a BNS $\mathcal{B}_s$ can be very expensive. Alternatively, we can examine local structures defined by a node

and its set of parent nodes along with the set of parent-child edges. That is, we can set D_E to be a set of single node, and consider different subsets of its set of parents as S_E. Then, we seek a subset such that the $\mathsf{DD}_{\mathcal{B}_s}$ is close to zero. Since if we prune a set E of incoming edges at X such that $S_E \subseteq \mathsf{Par}(X)$ and $\mathsf{DD}_{\mathcal{B}_s}(E)$ is close to zero, then $\mathbf{p}_X^{\mathsf{Par}(X)}(\mathbf{a}) \approx \mathbf{p}_X^{Q_E}(\mathbf{a}[Q_E])$ for all $\mathbf{a} \in \mathrm{Dom}(\mathsf{Par}(X))$ by Theorem 3.4. This, in turn implies $P(X|\mathbf{A}_{\mathsf{Par}(X)}) \approx P(X|\mathbf{A}_{Q_E})$.

But, the directed Markov property implies

$$P(X|\mathbf{A}_{\mathsf{Par}(X)}, \mathbf{A}_{\mathsf{nd}(X)}) = P(X|\mathbf{A}_{\mathsf{Par}(X)}).$$

Then, we have

$$P(X|\mathbf{A}_{Q_E})$$
$$\approx P(X|\mathbf{A}_{\mathsf{Par}(X)}, \mathbf{A}_{\mathsf{nd}(X)})$$
$$= P(X|\mathbf{A}_{Q_E}, \mathbf{A}_{S_E}, \mathbf{A}_{\mathsf{nd}(X)})$$

Thus, we have

$$X \perp (\mathbf{A}_{S_E}, \mathbf{A}_{\mathsf{nd}(X)}) \mid \mathbf{A}_{Q_E}.$$

Finally, by symmetry and decomposition properties of conditional independence Pearl (1988) we have

$$X \perp \mathbf{A}_{\mathsf{nd}(X)} \mid \mathbf{A}_{Q_E}.$$

Thus, the conditional independence property of a node of a structure is preserved if we prune a set of incoming edges of the node with distribution distortion measure close to zero.

In fact the parent-child fitness measure,

$$0 \le \frac{\mathcal{H}(\pi^X|\pi^{\mathbf{A}_{\mathsf{Par}(X)}})}{\mathcal{H}(\pi^X)} \le 1 \qquad (2)$$

introduced in Simovici & Baraty (2008) has some similarity with $\mathsf{DD}_{\mathcal{B}_s}$.

We have shown that if this measure is close to zero, then $\mathsf{Par}(X)$ is a suitable parent set for node X. Finding parent-child relationships among the attributes of $\mathcal{D}$ such that the measure (2) is close to zero, increases the posterior probability of $\mathcal{D}$ conditioned upon the inducing BNS $\mathcal{B}_s$. As stated before, an increase in this probability leads to an increase in the posterior probability of the structure conditioned on data (assuming a uniform prior on possible Bayesian network structures for a data set, as in Cooper & Herskovits (1993)). This happens because if we choose a set of parents, $\mathsf{Par}(X)$ for a node X such that $\mathcal{H}(\pi^X|\pi^{\mathbf{A}_{\mathsf{Par}(X)}})$ is close to zero, then for those $\mathbf{a} \in \mathrm{Dom}(\mathbf{A}_{\mathsf{Par}(X)})$ such that $P(\mathbf{A}_{\mathsf{Par}(X)} = \mathbf{a})$ is nontrivial, the probability $P(X = x \mid \mathbf{A}_{\mathsf{Par}(X)} = \mathbf{a})$ is close to 1 for some $x \in \mathrm{Dom}(X)$ and close to 0 for all other $x' \in \mathrm{Dom}(X)\backslash\{x\}$.

If $P(\mathbf{A}_{\mathsf{Par}(X)} = \mathbf{a})$ is large and $P(X = x \mid \mathbf{A}_{\mathsf{Par}(X)} = \mathbf{a}) \approx 1$, this implies that $P(\mathbf{A}_{\mathsf{Par}(X)} = \mathbf{a}, X = x)$ is large. Thus, for a BNS $\mathcal{B}_s$ that we obtain in this way we have

$$P(\mathcal{D} \mid \mathcal{B}_s)$$
$$= \prod_{i=1}^{|\mathcal{D}|} \prod_{X \in \mathbf{Attr}(\mathcal{D})}$$
$$\qquad P_{\mathcal{B}_s}(X = t_i[X] \mid \mathbf{A}_{\mathsf{Par}(X)} = t_i[\mathsf{Par}(X)])$$
$$= \prod_{X \in \mathbf{Attr}(\mathcal{D})} \prod_{\substack{x \in \mathrm{Dom}(X) \\ \mathbf{a} \in \mathrm{Dom}(\mathbf{A}_{\mathsf{Par}(X)})}}$$
$$\qquad \left(P_{\mathcal{B}_s}(X = x \mid \mathbf{A}_{\mathsf{Par}(X)} = \mathbf{a})\right)^{\mathcal{N}(X, \mathbf{A}_{\mathsf{Par}(X)})(x, \mathbf{a})},$$

where $\mathcal{N}_{(X,\mathbf{A}_{\mathsf{Par}(X)})}(x,\mathbf{a})$ is the number of tuples t in $\mathcal{D}$ with $t[X] = x$ and $t[\mathsf{Par}(X)] = \mathbf{a}$. Having a BNS $\mathcal{B}_s$ such that $P_{\mathcal{B}_s}(X = x \mid \mathbf{A}_{\mathsf{Par}(X)} = \mathbf{a})$ is close to one for those pairs $(x,\mathbf{a})$ with large $\mathcal{N}_{(X,\mathbf{A}_{\mathsf{Par}(X)})}(x,\mathbf{a})$ justifies the increase in posterior probability of $\mathcal{D}$.

4 Constructing a BNS for a Data Set

Recall that if E is a set of edges, then $Q_E = \mathsf{Par}(D_E) - S_E$ is the set of nodes that remain parents of the nodes in D_E after the edges in E are removed.

Definition 4.1 Let $A_i \in \mathbf{Attr}(\mathcal{D})$ for $1 \leq i \leq n$. Then, *the total measure of fitness loss* by pruning the set of converging edges E at node A_i in BNS $\mathcal{B}_s$, is the number,

$$\alpha \cdot \frac{\mathcal{H}(\pi^{A_i}|\pi^{\mathbf{A}_{Q_E}}) - \mathcal{H}(\pi^{A_i}|\pi^{\mathbf{A}_{\mathsf{Par}(A_i)}})}{\log_2 |\mathrm{Dom}(A_i)| - \mathcal{H}(\pi^{A_i}|\pi^{\mathbf{A}_{\mathsf{Par}(A_i)}})}$$
$$+ (1-\alpha) \cdot \frac{\mathcal{H}(\pi^{A_i}|\pi^{\mathbf{A}_{Q_E}}) - \mathcal{H}(\pi^{A_i}|\pi^{\mathbf{A}_{\mathsf{Par}(A_i)}})}{\mathcal{H}(\pi^{A_i})},$$

denoted by $\mathbf{FL}_\alpha(A_i, \mathcal{B}_s, E)$, where $0 \leq \alpha \leq 1$. □

Clearly, this measure is always in the range $[0, 1]$. Note that the left component of the sum is the distribution distortion measure of pruning the set of incoming edges, E. The right component measures the decrease in reduction of entropy of node A_i in presence of its parents after pruning of the set E. Thus, if $\mathbf{FL}_\alpha(A_i, \mathcal{B}_s, E)$ is close to zero, then both components will be close to zero. The left component ensures that the conditional independence is preserved after the pruning of E, while the right component preserves the posterior probability of $\mathcal{D}$ conditioned upon the structure. We can choose $\alpha = \frac{1}{2}$ if we have no preference over any of the two measures. Note that, if E is a set of converging edges at node A_i and $E' \subseteq E$ then, $\mathbf{FL}_\alpha(A_i, \mathcal{B}_s, E') \leq \mathbf{FL}_\alpha(A_i, \mathcal{B}_s, E)$. This enables us to use the heuristic method explained in Simovici & Baraty (2008) to find a structure that fits the given data from a complete network structure. This complete structure can be induced from a total order of attributes that represents the expert's knowledge of the domain.

5 Experimental Results

As our first experiment, we started with a Bayesian network Structure for Neapolitan Cancer data set with 5 attributes and 10000 rows, pruned different subsets of converging edges at a single node and computed the total measure of fitness loss for each pruning. Figure 2 visualizes these pruned structures and their relation with each other as a graph which we refer to as meta-graph to avoid confusion with the Bayesian graphs for the data set. Also, we refer to the edges and nodes of the meta-graph as meta-edges and meta-nodes, respectively. Each meta-node represents a BNS for Neapolitan data set and each edge, a pruning transformation. That is, the destination meta-node of a meta-edge is obtained by removing a subset of converging edges at a single node from the source meta-node of that meta-edge. Each meta-node is labeled with a letter from A to I.

Table 1 represents the scores for each meta-node in figure 2 based on two schemes MDL and C-H score.

Table 2 shows the total fitness loss of each meta-edge for parameter $\alpha = \frac{1}{2}$ and its two components, distribution distortion and entropy loss for each pruning of set of edges. The fitness loss measure is strongly correlated with both scoring schemes (C-H and MDL), which shows the usefulness of this measure for simplifying a Bayesian network structure. Also, our measure can be used to assess the importance of various edges. For example, the total fitness loss measure suggests that the edge that whose
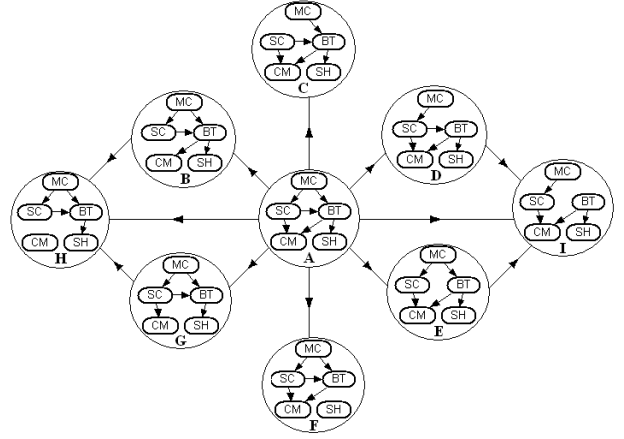

Figure 2: Neapolitan Cancer Bayesian structure was introduced in Cooper (1984). This structure has five nodes: Metastatic Cancer (MC), Serum Calcium (SC), Brain Tumor (BT), Coma (CM) and Severe Headaches (SH). Meta-node E corresponds to this structure.

Table 1: Neapolitan Cancer Scores

Structure	log(C-H Score)	MDL Score
A	-7505	24948
B	-7586	25214
C	-7938	26384
D	-7509	24958
E	-8120	26986
F	-7505	24947
G	-7697	25581
H	-7713	25632
I	-8341	27719

source is *Serum Calcium* and destination is *Brain Tumor* (SC to BT) is important since its removal causes a significant degradation of both the C-H and MDL scores. On the other hand, TFL assess the edge from Brain Tumor to Severe Headaches (SH) as not important. Again CH and MDL scores confirm this assessment.

Table 2: Neapolitan Cancer Pruning Measures with $\alpha = \frac{1}{2}$.

Edge	DD	Ent. Loss	TFL
AB	0.0526	0.092	0.0723
AC	0.684	0.2067	0.4454
AD	0.0068	0.0035	0.005
AE	0.5777	0.2976	0.4377
AF	0.023	0.001	0.012
AG	0.12155	0.2126	0.167
AH	0.1328	0.2322	0.1825
AI	0.786	0.405	0.5957
BH	0.0846	0.1402	0.1124
GH	0.0128	0.0196	0.0162
DI	0.7848	0.4016	0.593
EI	0.4938	0.1074	0.301

We also applied our approach on a Bayesian network structure for the ALARM data set, originally described in Beinlich et al. (1988) as a network for monitoring patients in intensive care. Table 3 contains scores for this structure which is labeled as A. Structures B to H are generated from A by pruning different subsets of parents for three nodes selected at random. Table 4 shows the exact parent pruning specification applied to obtain B to H from A. The child column represents the node we have chosen to prune its incoming edges. The original parent column shows the set of parents in the original structure, namely

A. New parent column represents the set of parents of the child after pruning. The other columns are the same as in previous example. Note that again, there is a close correlation between fitness loss measure and different scores. Since the structures are much larger than in previous experiment, pruning an edge or two has a milder effect on the magnitude of the scores of the global structure than in the Neapolitan case for about the same total fitness loss which is a local measure.

Table 3: Alarm Scores

Structure	log(C-H Score)	MDL Score
A	-159636	530806
B	-164287	546157
C	-162785	541189
D	-161372	536491
E	-161731	537644
F	-161684	537514
G	-159767	531136
H	-159638	530802

Table 4: ALARM pruning measures with parameter $\alpha = \frac{1}{2}$. The nodes of the ALARM network are traditionally numbered from 1 though 37 in the literature. The correspondence between the node numbers mentioned in the table and the real attributes are as follows (8, HREKG), (9, HRSat), (27, Catecholamine), (29, Heart Rate) and (30, Error Cauter).

Struct.	Child	O. Par	N. Par	DD	Ent. Loss	TFL
B	8	30,29	none	0.4388	0.778	0.6083
C	8	30,29	30	0.297	0.5266	0.4118
D	8	30,29	29	0.1655	0.293	0.2294
E	9	30,29	none	0.2847	0.319	0.3018
F	9	30,29	30	0.2767	0.31	0.2933
G	9	30,29	29	0.0212	0.0237	0.0225
H	29	27	none	0.0006	0.001	0.0008

Finally, Table 5 shows the correlations between TFL and changes in logarithm of CH score and also between TFL and changes in MDL score for Neapolitan Cancer and ALARM structures as a result of edge removals explained in Tables 2 and 4. Interestingly, although TFL is a local measure, it has very close correlations with MDL and CH scores which are global measures.

Also note that while the correlations between TFL and MDL are positive, the correlations between TFL and CH score are negative, since as TFL increases, the probability of the structure conditioned upon the data set decreases and as a result the CH score decreases.

Table 5: Correlations

Data Set	TFL/ log(CH)	TFL/MDL
Neapolitan	-0.97324	0.9733857
ALARM	-0.9983168	0.9980918

6 Conclusions and Future Work

We proposed a method for assessing the degree of influence of a set of edges of a Bayesian network structure on local conditional probability distributions. In particular, for the purpose of constructing a BNS from data, we concentrate on pruning a set of converging edges at a single node. This local pruning has a direct effect on the global fitness of the Bayesian network structure, measured by scoring schemes such as MDL or CH, which appear to be strongly correlated to the distribution distortion proposed by us. Thus, pruning is useful for adjusting a Bayesian network structure obtained from an expert's prior knowledge of the domain to a data set.

The distribution distortion could be used as measure of importance and interestingness of the edges of the Bayesian network structure and we intend to further pursue this issue. Another open technical problem is to explore whether by pruning a complete Bayesian network structure in the presence of a data set can lead to a network structure that best fits the data.

A Proof of Theorem 3.2

We substitute the Kullbach-Leibler measure in the numerator,

$$
\sum_{\mathbf{a} \in \mathrm{Dom}(\mathbf{A}_{\mathsf{Par}(D_E)})} P(\mathbf{A}_{\mathsf{Par}(D_E)} = \mathbf{a}) \cdot \sum_{i=1}^{m} \left[P(\mathbf{A}_{D_E} = b_i | \mathbf{A}_{\mathsf{Par}(D_E)} = \mathbf{a}) \right.
$$
$$
\left. \cdot \log_2 \frac{P(\mathbf{A}_{D_E} = b_i | \mathbf{A}_{\mathsf{Par}(D_E)} = \mathbf{a})}{P(\mathbf{A}_{D_E} = b_i | \mathbf{A}_{Q_E} = \mathbf{a}[Q_E])} \right]
$$
$$
= \sum_{\mathbf{a} \in \mathrm{Dom}(\mathbf{A}_{\mathsf{Par}(D_E)})} \sum_{i=1}^{m} \left[P(\mathbf{A}_{D_E} = b_i, \mathbf{A}_{\mathsf{Par}(D_E)} = \mathbf{a}) \right.
$$
$$
\left. \cdot \log_2 P(\mathbf{A}_{D_E} = b_i | \mathbf{A}_{\mathsf{Par}(D_E)} = \mathbf{a}) \right]
$$
$$
- \sum_{\mathbf{a} \in \mathrm{Dom}(\mathbf{A}_{\mathsf{Par}(D_E)})} \sum_{i=1}^{m} \left[P(\mathbf{A}_{D_E} = b_i, \mathbf{A}_{\mathsf{Par}(D_E)} = \mathbf{a}) \right.
$$
$$
\left. \cdot \log_2 P(\mathbf{A}_{D_E} = b_i | \mathbf{A}_{Q_E} = \mathbf{a}[Q_E]) \right]
$$
$$
= - \mathcal{H}(\pi^{\mathbf{A}_{D_E}} | \pi^{\mathbf{A}_{\mathsf{Par}(D_E)}})
$$
$$
- \sum_{\substack{\mathbf{a}' \in \mathrm{Dom}(\mathbf{A}_{Q_E}) \\ \mathbf{a}'' \in \mathrm{Dom}(\mathbf{A}_{S_E})}} \sum_{i=1}^{m} \left[P(\mathbf{A}_{D_E} = b_i, \mathbf{A}_{Q_E} = \mathbf{a}', \mathbf{A}_{S_E} = \mathbf{a}'') \right.
$$
$$
\left. \cdot \log_2 P(\mathbf{A}_{D_E} = b_i | \mathbf{A}_{Q_E} = \mathbf{a}') \right]
$$
$$
= - \mathcal{H}(\pi^{\mathbf{A}_{D_E}} | \pi^{\mathbf{A}_{\mathsf{Par}(D_E)}})
$$
$$
- \sum_{\mathbf{a}' \in \mathrm{Dom}(\mathbf{A}_{Q_E})} \sum_{i=1}^{m} \left[\log_2 P(\mathbf{A}_{D_E} = b_i | \mathbf{A}_{Q_E} = \mathbf{a}') \right.
$$
$$
\left. \cdot \sum_{\mathbf{a}'' \in \mathrm{Dom}(\mathbf{A}_{S_E})} P(\mathbf{A}_{D_E} = b_i, \mathbf{A}_{Q_E} = \mathbf{a}', \mathbf{A}_{S_E} = \mathbf{a}'') \right]
$$
$$
= - \mathcal{H}(\pi^{\mathbf{A}_{D_E}} | \pi^{\mathbf{A}_{\mathsf{Par}(D_E)}}) - \sum_{\substack{\mathbf{a}' \in \mathrm{Dom}(\mathbf{A}_{Q_E}) \\ i \in [1..m]}} \left[\log_2 P(\mathbf{A}_{D_E} = b_i | \mathbf{A}_{Q_E} = \mathbf{a}') \right.
$$
$$
\left. \cdot P(\mathbf{A}_{D_E} = b_i, \mathbf{A}_{Q_E} = \mathbf{a}') \right]
$$
$$
= \mathcal{H}(\pi^{\mathbf{A}_{D_E}} | \pi^{\mathbf{A}_{Q_E}}) - \mathcal{H}(\pi^{\mathbf{A}_{D_E}} | \pi^{\mathbf{A}_{\mathsf{Par}(D_E)}}).
$$

In the same way we can show,

$$
\sum_{\mathbf{a} \in \mathrm{Dom}(\mathbf{A}_{\mathsf{Par}(D_E)})} P(\mathbf{A}_{\mathsf{Par}(D_E)} = \mathbf{a}) \cdot \mathsf{KL}(\mathbf{p}_{D_E}^{\mathsf{Par}(D_E)}(\mathbf{a}), \mathbf{u}_m)
$$
$$
= \sum_{\substack{\mathbf{a} \in \mathrm{Dom}(\mathbf{A}_{\mathsf{Par}(D_E)}) \\ i \in [1..m]}} \left[P(\mathbf{A}_{D_E} = b_i, \mathbf{A}_{\mathsf{Par}(D_E)} = \mathbf{a}) \right.
$$
$$
\left. \cdot \log_2 P(\mathbf{A}_{D_E} = b_i | \mathbf{A}_{\mathsf{Par}(D_E)} = \mathbf{a}) \right] + \log_2 m
$$
$$
= \mathcal{H}(\pi^{\mathbf{u}_m}) - \mathcal{H}(\pi^{\mathbf{A}_{D_E}} | \pi^{\mathbf{A}_{\mathsf{Par}(D_E)}}).
$$

References

Beinlich, I., Suermondt, J., Chavez, M. & Cooper, G. (1988), The alarm monitoring system: A case study with two probablistic inference techniques for belief networks, *in* 'Second European Conference on Artificial Intelligence in Medicine', London.

Cooper, G. F. (1984), NESTOR: A computer-based medical diagnosis aid that integrates casual and probabilistic knowledge, PhD thesis, Stanford University.

Cooper, G. F. & Herskovits, E. (1993), A Bayesian method for the induction of probabilistic networks from data, Technical Report KSL-91-02, Stanford University, Knowledge System Laboratory.

Cowell, R. (1998), *Learning in Graphical Models*, MIT Press, Cambridge, MA, USA, pp. 9–26.

Friedman, N. & Goldszmidt, M. (1998), *Learning in Graphical Models*, MIT Press, Cambridge, MA, USA, pp. 421–459.

Heckerman, D., Geiger, D. & Chickering, D. M. (1995), Learning Bayesian networks: The combination of knowledge and statistical data, *in* 'Machine Learning', pp. 197–243.

Lam, W. & Bacchus, F. (1994), 'Learning Bayesian belief networks: An approach based on the MDL principle', *Computational Intelligence* **10**, 269–293.

Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers, Inc, San Mateo, CA.

Rissanen, J. (1978), 'Modeling by shortest data description', *Automatica* **14**, 456–471.

Simovici, D. A. (2007), 'On generalized entropies and entropy metrics', *Journal of Multiple-valued Logic and Soft Computing* **13**, 295–320.

Simovici, D. A. & Baraty, S. (2008), Structure inference of Bayesian networks from data: A new approach based on generalized conditional entropy, *in* 'EGC', pp. 337–342.

Simovici, D. A. & Jaroszewicz, S. (2002), 'An axiomatization of partition entropy', *Transactions on Information Theory* **48**, 2138–2142.

Simovici, D. A. & Jaroszewicz, S. (2006), 'A new metric splitting criterion for decision trees', *International Journal of Parallel, Emergent and Distributed Systems* **21**, 239–256.

Suzuki, J. (1999), 'Learning Bayesian belief networks based on the MDL principle: An efficient algorithm using the branch and bound technique', *IEICE Trans. Information and Systems* pp. 356–367.