

Article

Calculating the Prior Probability Distribution for a Causal Network Using Maximum Entropy: Alternative Approaches

Michael J. Markham

Ex-Bradford University, 22 Wrenbury Crescent, Leeds LS16 7EG, UK;

E-Mail: mjmarkham@btinternet.com; Tel.: 0113-2675235

Received: 20 April 2011; in revised form: 23 May 2011 / Accepted: 23 June 2011 /

Published: 18 July 2011

Abstract: Some problems occurring in Expert Systems can be resolved by employing a causal (Bayesian) network and methodologies exist for this purpose. These require data in a specific form and make assumptions about the independence relationships involved. Methodologies using Maximum Entropy (ME) are free from these conditions and have the potential to be used in a wider context including systems consisting of given sets of linear and independence constraints, subject to consistency and convergence. ME can also be used to validate results from the causal network methodologies. Three ME methods for determining the prior probability distribution of causal network systems are considered. The first method is Sequential Maximum Entropy in which the computation of a progression of local distributions leads to the over-all distribution. This is followed by development of the Method of Tribus. The development takes the form of an algorithm that includes the handling of explicit independence constraints. These fall into two groups those relating parents of vertices, and those deduced from triangulation of the remaining graph. The third method involves a variation in the part of that algorithm which handles independence constraints. Evidence is presented that this adaptation only requires the linear constraints and the parental independence constraints to emulate the second method in a substantial class of examples.

Keywords: maximum entropy; sequential maximum entropy; causal networks; Method of Tribus; independence

1. Introduction

In Expert Systems the Maximum Entropy (ME) methodology [1–3] can be employed to augment that of Causal Networks (CNs), e.g., HUGIN [4] and Lauritzen-Spiegelhalter [5]; also see Neapolitan [3]. The philosophy and motivation for using ME has already been examined in earlier work, see [6–10] and [11]. ME provides a parallel methodology which can be used to validate results from the CNs methodologies and there is also the potential for application beyond the strict limits of CNs. The work here will be confined to binary variables (*true* or *false*), the aim being to deduce prior probability distributions (*solutions*) which exhibit maximal entropy and match outcomes from CNs when appropriate. Three methods to this effect will be presented.

A causal network relates propositional variables and has the following properties: a prior probability is required for each root (a source node with no edges directed into it) and a set of prior conditional probabilities must be available at every other vertex giving the probability of the associated variable for all assignments to the variables representing its parents. These probabilities constitute the linear constraints and such networks will be described as *complete*. The system is represented by a *directed acyclic graph (DAG)* and a set of independence constraints is implied. The discussion in the paper will focus on complete causal networks (Note that *incomplete* systems may exhibit multiple solutions, see Paris [12]. Some confusion exists regarding the appropriate terminology, quoting Neapolitan: “*Causal networks are also called Bayesian networks, belief networks and sometimes influence diagrams*” [3]. The first of these terms will be used throughout.

Independence constraints will be an important element of this paper (discussion of the Independence concept can be found in [13,14]) and the terminology to be used will now be clarified:

Absolute and conditional independence constraints between propositional variables will be represented, respectively, by forms such as:

$$b \text{ ind } c \text{ and } d \text{ cid } e \mid fg$$

The term *independency* will be used to identify the set of individual independence constraints generated by the possible assignments to the conditioning set.

For example, the generic independency $d \text{ cid } e \mid fg$ will consist of four individual constraints:

$$d \text{ cid } e \mid fg, d \text{ cid } e \mid f\bar{g}, d \text{ cid } e \mid \bar{f}g \text{ and } d \text{ cid } e \mid \bar{f}\bar{g}$$

Independencies between parents are formally termed *moral* independencies (*moral* graphs and also *triangulation* are discussed in [3], see also Figures 1,2,3,4, and associated text. Moral independencies are conditioned by a set of variables blocking all ancestral paths between the parents. This set is empty in the case of absolute independence.

With regard to the first method, Lukasiewicz has proved that the application of ME to a causal network produces a unique solution using *Sequential ME* [15]. This process will be exemplified in Section 1. An understanding of *d-Separation* is required and an explanation will be provided there.

ME methodologies based on the Method of Tribus [16] need an appropriate set of explicit independence constraints (void if only linear constraints are to be considered), and a technique for finding such a set is described in [17]. For the second method, both moral and triangulation independencies are required, in general. When these are in place the *solution* will match that from

CNs [11]. The algorithm advanced handles consistent sets of linears and explicit independencies presented in a standard format.

The third method was suggested by a series of experimental results using a variation of the algorithm which requires just the linear constraints and the moral independencies. The handling of the independencies is accomplished by a technique based, initially, on the method developed for linears. This variant method will be applied to an elementary network, and then the technique developed will be extended to a more general system. Further generalisation of this method remains to be explored. This paper relates to an aspect of the work covered in my Ph.D. thesis successfully submitted to the University of Bradford in 2005 [11], under the supervision of P.C. Rhodes.

2. Method 1: Sequential Maximum Entropy (SME)

Quoting from the abstract to his paper, Lukasiewicz says:

“We thus present a new kind of maximum entropy model, which is computed sequentially. We then show that for all general Bayesian networks, the sequential maximum entropy model coincides with the unique joint distribution.”

The method uses *d-Separation* (*d-sep*) to determine local probabilities in descending order.

2.1. Definition of *d-Separation*

This is a mechanism for determining the implicit independencies of a causal network [3,18]. It is, however, first necessary to understand the principle of *blocking*.

Following Neapolitan, suppose that a DAG, G , is constructed from a set of vertices, V , and a set of edges, E , consisting of directed edges $\langle v_i, v_j \rangle$ where v_i, v_j belong to V . Suppose also that W is a subset of V and that u, v are vertices outside W . A path, p , between u and v is blocked by W if one of the following is true:

- (1) *d-sep1*: There is a vertex w , in W , on p such that the edges, which determine that w is on p , meet tail-to-tail at w . I.e. There are edges $\langle w, u \rangle$ and $\langle w, v \rangle$, where u and v are adjacent to w on the path p .
- (2) *d-sep2*: There is a vertex w , in W , on p such that the edges, which determine that w is on p , meet head-to-tail at w . I.e. There are edges $\langle u, w \rangle$ and $\langle w, v \rangle$, where u and v are adjacent to w on the path p .
- (3) *d-sep3*: There is a vertex x on p , for which neither x nor any of its descendents are in W , such that the edges which determine that x is on p meet head-to-head at x . I.e. There are edges $\langle u, x \rangle$ and $\langle v, x \rangle$, where u and v are adjacent to x on the path p , but u and v are not members of W . Note also that W can be empty.

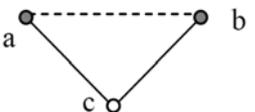
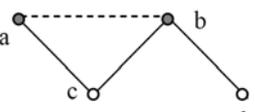
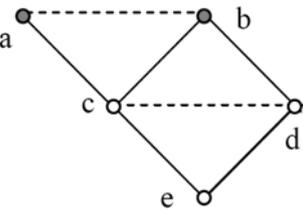
Sets of vertices U and V are said to be *d-separated* by the set W if every path between U and V is blocked by W . Furthermore, a theorem has been proved which asserts that, if all the paths are blocked, then the variables in U are independent of those in V given the outcomes for those in W [19]. The converse is also true, and furthermore a DAG, together with the explicit constraints and the definition

of d -Separation, *i.e.*, d -sep1, d -sep2 and d -sep3, is sufficient to define a causal network and to identify all the implicit independencies [20]. Much of the work on d -Separation was developed by J. Pearl and his associates.

2.2. The SME Technique

The technique requires the establishment of an ancestral ordering [3], of the variables in the network, and then proceeds to calculate the maximised local entropy for each set of variables, using the linear constraints and local d -seps1&2. (It is already known that ME support d -seps1&2 but not d -sep3, see [11,21]. The process starts with the first variable, and then the set is incremented by each variable in turn. Crucially the result of every maximisation is carried forward to the next step as a set of probabilities. Such steps may result in the rapid onset of exponential complexity [22,23], but the d -sep3 independencies (or equivalent alternatives) implied by CNs are automatically built in. An illustration of the technique follows in Table 1.

Table 1. Steps in the SME Analysis of Vee-Loz5.

Configuration	Steps
	Starting with variable a , $P(a)$ is given.
	Going on to b and adding b <i>ind</i> a (given by local d -sep2 in SME), $P(b)$ is given so $P(ab) = P(a).P(b)$.
	The independence of a and b determines the distribution of ab and this, together with $P(c ab)$, allows $P(abc)$ to be calculated, viz: $P(c ab) = k_{c ab} \text{ gives } P(abc) = k_{c ab}.P(ab).$
	Local d -sep2 gives d <i>cid</i> $c b$, which implies that d <i>cid</i> $ac b$, therefore $P(abcd).P(b) = P(abc).P(bd),$
	so determining $P(abcd)$. Finally, e <i>cid</i> $ab cd$ (d -sep2) plus $P(e cd)$ provides $P(abcde)$. The probability distribution for the variables of <i>Vee-Loz5</i> has now been found in 5 steps, and the moral indeps b <i>ind</i> a and d <i>cid</i> $c b$ have been generated.

3. Method 2: Development of the Method of Tribus

This method aims to calculate the over-all probability distribution directly. It demands the provision of a set of explicit independencies because ME does not support d -sep3. The version presented below is based on a method first described by Tribus [16], for linear constraints, but it has been re-worked to give potential for the accommodation of independence constraints. This method only finds stationary

points, additional work (Hessians, Hill-Climbing, Probing) would then be needed to determine the nature of any such point, in particular whether a global maximum has been found.

In outline, each constraint function is multiplied by its own, as yet unknown, Lagrange multiplier. These new functions are summed and then subtracted from the expression for entropy (H), thus forming a new function (F). An attempt is then made to find a stationary point of this new function in order to express the state probabilities in terms of the Lagrange multipliers. These generalised state probabilities are substituted into the constraints to produce a set of simultaneous linear equations which may be solvable for the Lagrange multipliers, thus leading to the determination of the state probabilities.

3.1. Analysis Using Lagrange Multipliers

Consider a knowledge domain which contains n variables v_1, \dots, v_n . The system representing this domain can be assigned any one of a finite set of 2^n states, s_0, \dots, s_{2^n-1} with probabilities p_0, \dots, p_{2^n-1} . It is required that a stationary point for the entropy, H , be discovered, whilst conforming to the system constraints, where:

$$H \equiv - \sum_{i=0}^{2^n-1} p_i \ln p_i \tag{1}$$

after Shannon and Weaver [24]. The state probabilities are assumed to be mutually independent but are constrained by the fact that their sum must be unity, since they are mutually exclusive and exhaustive. This fact constitutes a linear constraint and it will be presumed to be the first (or *normalising*) constraint, C_0 , in all the systems to be considered in this paper, thus:

$$C_0 \equiv \sum_{i=0}^{2^n-1} p_i - 1 = 0 \tag{2}$$

The other m constraints will be represented by the forms:

$$C_j = 0, \text{ for } j = 1..m \tag{3}$$

where C_j is some function of the p_i , e.g., $C_j \equiv 0.3 p_0 - 0.7 p_1$.

The constraints must be mutually unrelated so that each contributes a unique piece of information. Guided by Tribus [16], Griffeth [1] and Williamson [10], the Lagrangian function, F , will be taken as:

$$F \equiv - \sum_{i=0}^{2^n-1} p_i \ln p_i - (\lambda_0 - 1).C_0 - \sum_{j=1}^m \lambda_j C_j \tag{4}$$

where the λ s are the Lagrange multipliers (the term $\lambda_0 - 1$ makes the final formula tidier, without affecting the validity of the argument). It is to be noted at this point that F is concerned with the functions C_j , not with the equations $C_j = 0$. The function F has the property that $F = H$ when the constraints are satisfied, hence stationary points of F will be the same as those of H , at satisfaction.

Now stationary points of F occur when all partial derivatives are zero, *i.e.*,

$$\partial F / \partial p_i = 0 \text{ for } i = 0 .. 2^n - 1 \tag{5}$$

and:

$$\partial F / \partial \lambda_j = 0 \text{ for } j = 0..m \tag{6}$$

The p_i will be supposed independent on the grounds that the whole n -dimensional space is being considered during the search for a point which meets the criteria.

Assuming the independence of λ_j from p_i , λ_j from λ_k , and p_i from p_k , Equations 4 and 5 give:

$$-(\ln p_i + 1) - (\lambda_0 - 1) \cdot \partial C_0 / \partial p_i - \sum_{j=1}^m \lambda_j \cdot \partial C_j / \partial p_i = 0, \text{ for } i = 0..2^n - 1 \tag{7}$$

or

$$-(\ln p_i + 1) - (\lambda_0 - 1) - \sum_{j=1}^m \lambda_j \cdot \partial C_j / \partial p_i = 0$$

and Equations 4 and 6 give:

$$C_j = 0, \text{ for } j = 0..m$$

This is a restatement of the fact that the constraints must be satisfied, but Equation 7 provides the *General Tribus Formula* for the state probabilities, viz:

$$p_i = e^{-\lambda_0} \cdot \prod_{j=1}^m e^{-\lambda_j \cdot \partial C_j / \partial p_i}, \text{ for } i = 0..2^n - 1 \tag{8}$$

This formula, which is computationally recursive with respect to p_i , may be used iteratively to attempt to determine the Lagrange multipliers and hence the initial state probabilities, so effecting a solution, assuming consistency and convergence. The algorithm, to be described later, employs a fixed point iteration scheme which cycles around the constraints, updating the Lagrange multipliers and the state probabilities, using a dynamic data structure.

3.2. Updating the Lagrange Multiplier for a Linear Constraint

Developing Tribus, consider a linear constraint in the form:

$$C_L \equiv P(a | x) - k_L = 0$$

where a is a single propositional variable and x is a set of the same type. This equation is equivalent to:

$$C_L \equiv (1 - k_L) \cdot P(ax) - k_L \cdot P(\bar{a}x) = 0 \tag{9}$$

Suppose further that U is the set of state subscripts such that $\sum_{u \in U} p_u = P(ax)$, and u is any one of those subscripts, with V and v similarly defined for $P(\bar{a}x)$.

Now, if Equation 8 is substituted into Equation 9, $e^{-\lambda_0}$ cancels giving:

$$C_L \equiv (1 - k_L) \cdot \sum_U \prod_{j=1}^m e^{-\lambda_j \cdot \partial C_j / \partial p_u} - k_L \cdot \sum_V \prod_{j=1}^m e^{-\lambda_j \cdot \partial C_j / \partial p_v} = 0 \tag{10}$$

Also from Equation 9, by differentiation:

$$\partial C_L / \partial p_u = 1 - k_L \text{ and } \partial C_L / \partial p_v = -k_L$$

It can be seen that there is a common factor in the U summation of $e^{-\lambda_L \cdot (1-k_L)}$, and a factor of $e^{-\lambda_L \cdot (-k_L)}$ in the V summation. The second factor cancels and Equation 10 may now be re-written as:

$$e^{-\lambda_L} \cdot (1-k_L) \cdot \sum_U \prod_{j \neq L} e^{-\lambda_j \cdot \partial C_j / \partial p_u} - k_L \cdot \sum_V \prod_{j \neq L} e^{-\lambda_j \cdot \partial C_j / \partial p_v} = 0$$

and then expressed in quotient form as:

$$e^{-\lambda_L} = \left[k_L \cdot \sum_V \prod_{j \neq L} e^{-\lambda_j \cdot \partial C_j / \partial p_v} \right] / \left[(1-k_L) \cdot \sum_U \prod_{j \neq L} e^{-\lambda_j \cdot \partial C_j / \partial p_u} \right] \tag{11}$$

This equation can now be used to find a new estimate of $e^{-\lambda_L}$. This involves a great deal of computation because the appropriate products must be formed and then each of these has to be summed for both numerator and denominator. It is better to re-absorb the old value of $e^{-\lambda_L}$ into the right-hand side of Equation 11, the products then revert to state probabilities, thus:

$$[e^{-\lambda_L}]_{\text{new}} = [e^{-\lambda_L}]_{\text{old}} \cdot \left[k_L \cdot \sum_V p_v \right] / \left[(1-k_L) \cdot \sum_U p_u \right]$$

the new estimate of the Lagrange multiplier (λ_L) for the linear constraint.

3.3. Updating the Lagrange Multiplier for an Independence Constraint

Extending the work above to independence constraints constituted an original piece of work by Markham & Rhodes, published in revised form in [17].

Consider an independence constraint in the form:

$$C_I \equiv P(ab | x) - P(a | x) \cdot P(b | x) = 0$$

which is equivalent to:

$$C_I \equiv P(abx) \cdot P(\bar{a}\bar{b}\bar{x}) - P(a\bar{b}x) \cdot P(\bar{a}bx) = 0 \tag{12}$$

Suppose that U is the set of state subscripts such that $\sum_{u \in U} p_u = P(abx)$, and that u is any one of those subscripts. If V and v are similarly defined for $P(a\bar{b}x)$, W and w for $P(\bar{a}bx)$ and also Z and z for $P(\bar{a}\bar{b}\bar{x})$, then Equation 12 can be expressed as:

$$C_I \equiv \sum_U p_u \cdot \sum_Z p_z - \sum_V p_v \cdot \sum_W p_w = 0 \tag{13}$$

It is further evident that:

$$P(x) = \sum_U p_u + \sum_V p_v + \sum_W p_w + \sum_Z p_z \tag{14}$$

Proceeding as in Equations 9 to 10 before, Equation 13 becomes:

$$\sum_U \prod_{j=1}^m e^{-\lambda_j \cdot \partial C_j / \partial p_u} \cdot \sum_Z \prod_{j=1}^m e^{-\lambda_j \cdot \partial C_j / \partial p_z} - \sum_V \prod_{j=1}^m e^{-\lambda_j \cdot \partial C_j / \partial p_v} \cdot \sum_W \prod_{j=1}^m e^{-\lambda_j \cdot \partial C_j / \partial p_w} = 0 \tag{15}$$

However, differentiating Equation 13 gives:

$$\partial C_I / \partial p_u = \sum_Z p_z, \quad \partial C_I / \partial p_v = -\sum_W p_w$$

$$\partial C_I / \partial p_w = -\sum_V p_v \text{ and } \partial C_I / \partial p_z = \sum_U p_u$$

therefore Equation 15 contains $e^{-\lambda_I}$ raised to the power of $\sum_Z p_z$, $-\sum_W p_w$, $-\sum_V p_v$ and $\sum_U p_u$ in its first, second, third and fourth terms respectively.

Following the argument used between Equations 10 and 11, Equation 15 becomes:

$$e^{-\lambda_I \cdot (\sum_U p_u + \sum_V p_v + \sum_W p_w + \sum_Z p_z)} = \left[\prod_{V, j \neq I} e^{-\lambda_j \cdot \partial C_j / \partial p_v} \cdot \prod_{W, j \neq I} e^{-\lambda_j \cdot \partial C_j / \partial p_w} \right] / \left[\prod_{U, j \neq I} e^{-\lambda_j \cdot \partial C_j / \partial p_u} \cdot \prod_{Z, j \neq I} e^{-\lambda_j \cdot \partial C_j / \partial p_z} \right] \tag{16}$$

Using Equation 14 and re-absorbing $e^{-\lambda_I}$, Equation 16 can be re-written as:

$$[e^{-\lambda_I}]_{\text{new}} = [e^{-\lambda_I}]_{\text{old}} \cdot \left\{ \frac{[\sum_V p_v \cdot \sum_W p_w]}{[\sum_U p_u \cdot \sum_Z p_z]} \right\}^{1/P(x)}$$

the update formula for the Lagrange multiplier (λ_I) of the independency.

3.4. Algorithm and Data Structures

The theory presented above has two properties that suggest a particular approach to the design of an algorithm. The first is that each constraint provides a means by which its associated Lagrange multiplier term can be updated. This suggests that a fixed point iteration scheme would be suitable for solving the set of non-linear simultaneous equations which arise from the constraints. The second property is the fact that the set of states which appear in each sum are disjoint and that these same sets need to be revisited to update the state probabilities. This points to a multi-list structure of state records which enables the appropriate state probabilities to be accessed fluently for each constraint.

This structure facilitates two operations:

- Using a standard procedure to normalise the state probabilities.
- Updating the estimate of state probabilities using the constraints.

The linear constraints have a constant associated with them and two, mutually exclusive, lists of state records. The variables associated with the constraint determine which state records are included in each list. The lists are static for a given application and hence can be constructed during initialisation of the data structure. At execution time, each list is traversed and the associated state probabilities summed. Each sum and the constant are then used to update the appropriate Lagrange multiplier. The independence constraints require four mutually exclusive lists but there is no associated constant. Those state records which are members of these lists can again be determined from the variables associated with the particular constraint and the lists can be created during initialisation of the data structure. These lists are used to update the Lagrange multiplier.

The algorithm cycles through the operations above, repeatedly updating individual multipliers. At the end of each cycle the state probabilities are calculated and re-normalised. It is essential that the state probabilities and Lagrange multipliers continue to be refreshed at each step until the constraints are satisfied, to a set standard of accuracy, and convergent entropy is achieved, if possible. The

algorithm requires that each state has an associated record which can store the latest estimate of the probability along with the constants and the pointers which traverse the sets.

Method 2 has the ability to handle any consistent set of linears and independencies constraints, assuming convergence of the iteration. There remains the potential to develop algorithms for other types of constraint.

4. Method 3: A Variation on Method 2

An alternative method of updating the Lagrange multiplier for an independence constraint will now be advanced. The narrative from this point forward describes original work devised by the author as part of his Thesis [11].

Consider again an independence constraint in the form:

$$C_I \equiv P(ab | x) - P(a | x).P(b | x) = 0 \tag{17}$$

A typical linear appears as:

$$C_L \equiv P(a | x) - k_L = 0, \tag{18}$$

or:

$$C_L \equiv (1 - k_L).P(ax) - k_L.P(\bar{a}x) = 0$$

Equation 17 can be reformulated as:

$$P(abx) = P(ax).P(bx)/P(x) = q, \text{ say} \tag{19}$$

Treating $C_I \equiv P(abx) - q = 0$ as a pseudo-linear, allows the independency to assume a form similar to Equation 18, viz:

$$C_I \equiv (1 - q).P(abx) - q.P(\bar{a}bx) = 0 \tag{20}$$

Assuming that q is constant, the technique already referenced can be applied to this equation to find a procedure for updating the Lagrange multipliers for Method 3, as follows:

Suppose that U is the set of state subscripts such that $\sum_{u \in U} p_u = P(abx)$, and u is any one of those subscripts, with V and v similarly defined for $P(\bar{a}bx)$.

Now, if Equation 8 is substituted into Equation 20, $e^{-\lambda_0}$ cancels giving:

$$C_I \equiv (1 - q) \cdot \sum_U \prod_{j=1}^m e^{-\lambda_j \cdot \partial C_j / \partial p_u} - q \cdot \sum_V \prod_{j=1}^m e^{-\lambda_j \cdot \partial C_j / \partial p_v} = 0 \tag{21}$$

Also from Equation 20, by differentiation:

$$\partial C_I / \partial p_u = 1 - q \text{ and } \partial C_I / \partial p_v = -q$$

It can be seen that there is a common factor in the U summation of $e^{-\lambda_l(1-q)}$, and a factor of $e^{-\lambda_l(-q)}$ in the V summation. The second factor cancels and Equation 21 may now be re-written as:

$$e^{-\lambda_l} \cdot (1 - q) \cdot \sum_U \prod_{j \neq l} e^{-\lambda_j \cdot \partial C_j / \partial p_u} - q \cdot \sum_V \prod_{j \neq l} e^{-\lambda_j \cdot \partial C_j / \partial p_v} = 0$$

and then expressed in quotient form as:

$$e^{-\lambda_l} = \left[q \cdot \sum_V \prod_{j \neq l} e^{-\lambda_j \cdot \partial C_j / \partial p_v} \right] / \left[(1 - q) \cdot \sum_U \prod_{j \neq l} e^{-\lambda_j \cdot \partial C_j / \partial p_u} \right] \tag{22}$$

Re-absorbing the old value of $e^{-\lambda_i}$ into the right-hand side of this equation, the products then revert to state probabilities, thus:

$$[e^{-\lambda_i}]_{\text{new}} = [e^{-\lambda_i}]_{\text{old}} \cdot [q \cdot \sum_V p_v] / [(1-q) \cdot \sum_U p_u]$$

giving a new estimate of the Lagrange multiplier (λ_i) for the independence constraint using the variant algorithm.

5. μ -Notation

The General Tribus Formula (Equation 8) contains a product of terms which represents a joint probability. This formula will now be expressed in a more compact form designed to facilitate further work, viz.:

$$p_i = \mu_0 \cdot \prod_{j=1}^m \mu'_{ji}$$

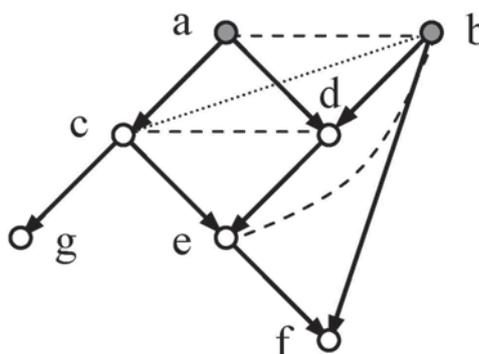
where $\mu_0 \equiv e^{-\lambda_0}$, $\mu'_{ji} \equiv e^{-\lambda_j \cdot \partial C_j / \partial p_i}$ and the index j ranges over the relevant constraints.

This notation was suggested by Garside [25]. The μ s will be described as *contributions* from their respective constraints. At this point, the contributions from the set of linear constraints for a given probability group will be combined into a single *combined contribution*, to give a more *economical* representation. This is illustrated by the following example system which relates seven propositional variables, a..g.

5.1. Example 1

The moral graph for Figure 1 requires that vertices a and b, the parents of d, be joined (dashes), and also c and d, the parents of e, together with b and e the parents of f. If the network is to be triangulated, then b and c have to be joined (dots) as well.

Figure 1. Tree-Loz-Vee-Tri7.



5.1.1. Linear Constraints

For this network to be complete, values for the following set of probabilities must be made available (see Section 1):

$$\begin{array}{c}
 \underline{P(a)}, \underline{P(b)}, \\
 \underline{P(c|a)}, \underline{P(c|\bar{a})}, \\
 \underline{P(d|ab)}, \underline{P(d|a\bar{b})}, \underline{P(d|\bar{a}b)}, \underline{P(d|\bar{a}\bar{b})}, \\
 \underline{P(e|cd)}, \underline{P(e|c\bar{d})}, \underline{P(e|\bar{c}d)}, \underline{P(e|\bar{c}\bar{d})}, \underline{P(f|be)}, \underline{P(f|b\bar{e})}, \underline{P(f|\bar{b}e)}, \underline{P(f|\bar{b}\bar{e})}, \\
 \underline{P(g|c)}, \underline{P(g|\bar{c})}
 \end{array}$$

This would imply eighteen contributions, but economy reduces this to seven combined contributions (underlined), one for each group of probabilities.

5.1.2. Independence Constraints

Method 2 requires both moral (the first two rows below) and triangulation independence constraints for a solution; a suitable set, see [17], is:

$$\begin{array}{c}
 \underline{a \text{ ind } b}, \underline{c \text{ cid } d | a}, \underline{c \text{ cid } d | \bar{a}}, \\
 \underline{b \text{ cid } e | cd}, \underline{b \text{ cid } e | c\bar{d}}, \underline{b \text{ cid } e | \bar{c}d}, \underline{b \text{ cid } e | \bar{c}\bar{d}}, \\
 \underline{b \text{ cid } c | ad}, \underline{b \text{ cid } c | a\bar{d}}, \underline{b \text{ cid } c | \bar{a}d}, \underline{b \text{ cid } c | \bar{a}\bar{d}}
 \end{array}$$

Economy reduces the contributions for the independence constraints to four. Using μ -notation, a typical joint probability, $P(abcdefg)$, can now be expressed as:

$$\mu_0 \cdot \mu_a \cdot \mu_b \cdot \mu_{c|a} \cdot \mu_{d|ab} \cdot \mu_{e|cd} \cdot \mu_{f|be} \cdot \mu_{g|c} \cdot \mu_{I(a,b)} \cdot \mu_{I(c,d|a)} \cdot \mu_{I(b,e|cd)} \cdot \mu_{I(b,c|ad)}$$

where the suffixes locate the combined contributions rather than i and j .

Here, for example:

$$\mu_{e|cd} \text{ represents } \mu'_{e|cd} \cdot \mu'_{e|c\bar{d}} \cdot \mu'_{e|\bar{c}d} \cdot \mu'_{e|\bar{c}\bar{d}}$$

and:
$$\mu_{I(b,e|cd)} \text{ represents } \mu'_{I(b,e|cd)} \cdot \mu'_{I(b,e|c\bar{d})} \cdot \mu'_{I(b,e|\bar{c}d)} \cdot \mu'_{I(b,e|\bar{c}\bar{d})}$$

5.2. Using Method 3 with μ -Notation

Early experiments with this method applied to a series of causal systems suggested that this algorithm, when given a full set of moral independencies, would generate the triangulation independencies required by Method 2. With this possibility in mind, the discussion will proceed by considering independence under Method 3 for a complete version of a system of six variables represented by a lozenge shaped network, *Loz6*. (NB. A lozenge will have at least five sides.)

Example 2: In Figure 2, given the moral independency for the sink vertex f , $d \text{ cid } e | c$, analysis will be employed in an attempt to show that the non-moral independencies $c \text{ cid } d | b$ and $b \text{ cid } c | a$, are implied.

The independence constraints $d \text{ cid } e | c$ and $d \text{ cid } e | \bar{c}$, may be expressed as:

$$P(cde) = P(cd) \cdot P(ce) / P(c) = r$$

and:

$$P(\bar{c}de) = P(\bar{c}d) \cdot P(\bar{c}e) / P(\bar{c}) = s, \text{ say}$$

Treating $P(cde) - r = 0$ and $P(\bar{c}de) - s = 0$ as pseudo-linears gives:

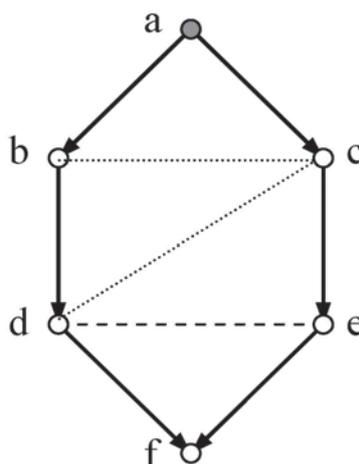
$$(1-r).P(cde) - r.P(\neg(cde)) = 0$$

and:

$$(1-s).P(\bar{c}de) - s.P(\neg(\bar{c}de)) = 0$$

The state probabilities in $P(cde)$ all differentiate to give $(1-r)$, and those in $P(\neg(cde))$ give $(-r)$, with similar results for the second constraint.

Figure 2. Loz6.



According to the General Tribus Formula (Equation 8), and reverting to individual exponentials for the independency, to allow closer examination, the respective contributions for states in cde and $\neg(cde)$ are $e^{-\lambda_i \cdot (1-r)}$ and $e^{-\lambda_i \cdot (-r)}$. These may be written, in terms of a multiplier (g), as g^{1-r} and g^{-r} , say. A similar provision will be made for the states of $\bar{c}de$, using h therefore, following work in Section 3 above, the update formulæ for the Lagrange multipliers of the independency are:

$$[g]_{\text{new}} = [g]_{\text{old}} \cdot [r.P(\neg(cde))] / [(1-r).P(cde)]$$

and:

$$[h]_{\text{new}} = [h]_{\text{old}} \cdot [s.P(\neg(\bar{c}de))] / [(1-s).P(\bar{c}de)]$$

5.2.1. Grg-Elimination

A method of deriving an independence constraint from a ratio due to G.R.Garside (outlined in a seminar) will now be described as it will be needed in ensuing sections:

Given:

$$P(yzx) / P(y\bar{z}x) = P(\bar{y}zx) / P(\bar{y}\bar{z}x)$$

this equals $P(zx) / P(\bar{z}x)$ after addition of the two numerators, and then the addition of the two denominators (this is a valid algebraic manipulation !)

i.e.,
$$P(yzx) / P(y\bar{z}x) = P(zx) / P(\bar{z}x)$$

or, after re-arrangement, $P(yzx) / P(zx) = P(y\bar{z}x) / P(\bar{z}x)$.

Adding numerators and denominators again:

$$P(yzx) / P(zx) = P(yx)/P(x) \text{ or } P(yzx) \cdot P(x) = P(yx) \cdot P(zx)$$

which shows that $y \text{ cid } z | x$, as required. This division and reduction technique will be referred to as *grg-elimination* and can be applied in reverse.

5.2.2. Analysis Phase 1: Determination of the Independence Multipliers

Assuming that a solution exists, the state probabilities will be calculated by the product of the combined contributions from the linear constraints, and the individual contributions from the moral independency in the form of powers of their multipliers. Using μ -notation, the joint probability now assumes the form:

$$P(abcdef) = \mu_0 \cdot \mu_a \cdot \mu_{b|a} \cdot \mu_{c|a} \cdot \mu_{d|b} \cdot \mu_{e|c} \cdot \mu_{f|de} \cdot g^{1-r} \cdot h^{-s}$$

also:

$$P(abc\bar{d}\bar{e}\bar{f}) = \mu_0 \cdot \mu_a \cdot \mu_{b|a} \cdot \mu_{c|a} \cdot \mu_{d|b} \cdot \mu_{e|c} \cdot \mu_{\bar{f}|\bar{d}\bar{e}} \cdot g^{1-r} \cdot h^{-s}$$

Adding these equations will eliminate f from the left-hand side, *viz*:

$$\begin{aligned} P(abcd\bar{e}) &= \mu_0 \cdot \mu_a \cdot \mu_{b|a} \cdot \mu_{c|a} \cdot \mu_{d|b} \cdot \mu_{e|c} \cdot \{\mu_{f|de} + \mu_{\bar{f}|\bar{d}\bar{e}}\} \cdot g^{1-r} \cdot h^{-s} \\ &= \mu_0 \cdot \mu_a \cdot \mu_{b|a} \cdot \mu_{c|a} \cdot \mu_{d|b} \cdot \mu_{e|c} \cdot F_{de} \cdot g^{1-r} \cdot h^{-s}, \text{ say} \end{aligned} \tag{23}$$

where:

$$F_{de} \equiv \{\mu_{f|de} + \mu_{\bar{f}|\bar{d}\bar{e}}\}$$

Varying d and e will give:

$$P(abcd\bar{e}) = \mu_0 \cdot \mu_a \cdot \mu_{b|a} \cdot \mu_{c|a} \cdot \mu_{d|b} \cdot \mu_{\bar{e}|c} \cdot F_{d\bar{e}} \cdot g^{-r} \cdot h^{-s}$$

$$P(ab\bar{c}\bar{d}\bar{e}) = \mu_0 \cdot \mu_a \cdot \mu_{b|a} \cdot \mu_{c|a} \cdot \mu_{\bar{d}|\bar{b}} \cdot \mu_{\bar{e}|c} \cdot F_{\bar{d}\bar{e}} \cdot g^{-r} \cdot h^{-s}$$

and:

$$P(ab\bar{c}d\bar{e}) = \mu_0 \cdot \mu_a \cdot \mu_{b|a} \cdot \mu_{c|a} \cdot \mu_{\bar{d}|\bar{b}} \cdot \mu_{\bar{e}|c} \cdot F_{\bar{d}e} \cdot g^{-r} \cdot h^{-s}$$

Now $d \text{ cid } e | c$, which is given, implies that $d \text{ cid } e | abc$. Therefore, the application of *grg-elimination* to the last four equations must result in complete cancellation, *i.e.*,

$$P(abcd\bar{e}) / P(abcd\bar{e}) \text{ must equal } P(ab\bar{c}d\bar{e}) / P(ab\bar{c}\bar{d}\bar{e})$$

giving:

$$g = \{F_{d\bar{e}} \cdot F_{\bar{d}\bar{e}}\} / \{F_{de} \cdot F_{\bar{d}e}\} \tag{24}$$

The right-hand side of this equation does not depend on c and so, by symmetry, h will return the same value, *i.e.*, $g = h$, as confirmed by experiment.

NB. It can be seen that g^{-r} and h^{-s} have cancelled, leaving just g ! This will be important in subsequent analysis.

5.2.3. Analysis Phase 2: Proof that $c \perp d \mid b$

With the multipliers having been determined, the analysis can move on to testing one of the non-moral independencies. Starting with Equation 23:

$$P(abcde) = \mu_0 \cdot \mu_a \cdot \mu_{b|a} \cdot \mu_{c|a} \cdot \mu_{d|b} \cdot \mu_{e|c} \cdot F_{de} \cdot g^{l-r} \cdot h^{-s}$$

and also:

$$P(abc\bar{d}\bar{e}) = \mu_0 \cdot \mu_a \cdot \mu_{b|a} \cdot \mu_{c|a} \cdot \mu_{d|b} \cdot \mu_{\bar{e}|c} \cdot F_{\bar{d}\bar{e}} \cdot g^{-r} \cdot h^{-s}$$

Addition of these two equations leads to:

$$P(abcd) = \mu_0 \cdot \mu_a \cdot \mu_{b|a} \cdot \mu_{c|a} \cdot \mu_{d|b} \cdot \{ \mu_{e|c} \cdot F_{de} \cdot g + \mu_{\bar{e}|c} \cdot F_{\bar{d}\bar{e}} \} \cdot g^{-r} \cdot h^{-s}$$

The following may be reached by a similar process:

$$P(abc\bar{d}) = \mu_0 \cdot \mu_a \cdot \mu_{b|a} \cdot \mu_{c|a} \cdot \mu_{\bar{d}|b} \cdot \{ \mu_{e|c} \cdot F_{\bar{d}\bar{e}} + \mu_{\bar{e}|c} \cdot F_{\bar{d}\bar{e}} \} \cdot g^{-r} \cdot h^{-s}$$

$$P(ab\bar{c}d) = \mu_0 \cdot \mu_a \cdot \mu_{b|a} \cdot \mu_{\bar{c}|a} \cdot \mu_{d|b} \cdot \{ \mu_{e|\bar{c}} \cdot F_{de} \cdot g + \mu_{\bar{e}|\bar{c}} \cdot F_{\bar{d}\bar{e}} \} \cdot g^{-r} \cdot h^{-s}$$

$$P(ab\bar{c}\bar{d}) = \mu_0 \cdot \mu_a \cdot \mu_{b|a} \cdot \mu_{\bar{c}|a} \cdot \mu_{\bar{d}|b} \cdot \{ \mu_{e|\bar{c}} \cdot F_{\bar{d}\bar{e}} + \mu_{\bar{e}|\bar{c}} \cdot F_{\bar{d}\bar{e}} \} \cdot g^{-r} \cdot h^{-s}$$

The condition required by *grg*-elimination is that:

$$P(abcd) / P(abc\bar{d}) - P(ab\bar{c}d) / P(ab\bar{c}\bar{d}) \equiv 0$$

or:

$$P(abcd) \cdot P(ab\bar{c}\bar{d}) - P(abc\bar{d}) \cdot P(ab\bar{c}d) \equiv 0$$

Representing the left-hand side of this identity by D, and noting that all the contribution terms cancel except for the four braces:

$$D \propto \{ \mu_{e|c} \cdot F_{de} \cdot g + \mu_{\bar{e}|c} \cdot F_{\bar{d}\bar{e}} \} \cdot \{ \mu_{e|\bar{c}} \cdot F_{\bar{d}\bar{e}} + \mu_{\bar{e}|\bar{c}} \cdot F_{\bar{d}\bar{e}} \} - \{ \mu_{e|c} \cdot F_{\bar{d}\bar{e}} + \mu_{\bar{e}|c} \cdot F_{\bar{d}\bar{e}} \} \cdot \{ \mu_{e|\bar{c}} \cdot F_{de} \cdot g + \mu_{\bar{e}|\bar{c}} \cdot F_{\bar{d}\bar{e}} \} \tag{25}$$

The result after multiplication is:

$$D \propto \{ \mu_{e|c} \cdot \mu_{e|\bar{c}} \cdot F_{de} \cdot F_{\bar{d}\bar{e}} \cdot g + \mu_{e|c} \cdot \mu_{\bar{e}|\bar{c}} \cdot F_{de} \cdot F_{\bar{d}\bar{e}} \cdot g + \mu_{\bar{e}|c} \cdot \mu_{e|\bar{c}} \cdot F_{\bar{d}\bar{e}} \cdot F_{\bar{d}\bar{e}} + \mu_{\bar{e}|c} \cdot \mu_{\bar{e}|\bar{c}} \cdot F_{\bar{d}\bar{e}} \cdot F_{\bar{d}\bar{e}} \} - \{ \mu_{e|c} \cdot \mu_{e|\bar{c}} \cdot F_{\bar{d}\bar{e}} \cdot F_{de} + \mu_{e|c} \cdot \mu_{\bar{e}|\bar{c}} \cdot F_{\bar{d}\bar{e}} \cdot F_{\bar{d}\bar{e}} + \mu_{\bar{e}|c} \cdot \mu_{e|\bar{c}} \cdot F_{\bar{d}\bar{e}} \cdot F_{de} \cdot g + \mu_{\bar{e}|c} \cdot \mu_{\bar{e}|\bar{c}} \cdot F_{\bar{d}\bar{e}} \cdot F_{\bar{d}\bar{e}} \}$$

The first and fifth terms of this expression cancel, as do the fourth and eighth, giving:

$$D \propto \mu_{e|c} \cdot \mu_{\bar{e}|\bar{c}} \cdot \{ F_{de} \cdot F_{\bar{d}\bar{e}} \cdot g - F_{\bar{d}\bar{e}} \cdot F_{\bar{d}\bar{e}} \} + \mu_{\bar{e}|c} \cdot \mu_{e|\bar{c}} \cdot \{ F_{\bar{d}\bar{e}} \cdot F_{\bar{d}\bar{e}} - F_{\bar{d}\bar{e}} \cdot F_{de} \cdot g \} \propto \{ F_{de} \cdot F_{\bar{d}\bar{e}} \cdot g - F_{\bar{d}\bar{e}} \cdot F_{\bar{d}\bar{e}} \} \cdot \{ \mu_{e|c} \cdot \mu_{\bar{e}|\bar{c}} - \mu_{\bar{e}|c} \cdot \mu_{e|\bar{c}} \}$$

But Equation 24 insists that $g = \{F_{d\bar{e}} \cdot F_{\bar{d}e}\} / \{F_{de} \cdot F_{\bar{d}\bar{e}}\}$, therefore $D \equiv 0$ and the condition for *grg*-elimination is satisfied, giving the independency:

$$c \text{ cid } d | ab$$

If this independency is compared with one given by *d-Separation* [9,18,21], $a \text{ cid } d | bc$, then:

$$P(abcd) \cdot P(ab) = P(abc) \cdot P(abd)$$

and:

$$P(abcd) \cdot P(bc) = P(abc) \cdot P(bcd)$$

division then gives:

$$P(bcd) \cdot P(ab) = P(abd) \cdot P(bc)$$

Summation over a, \bar{a} now gives the required independency, $c \text{ cid } d | b$.

5.2.4. Tokenised Algebra

Consider Equation 25 again:

$$D \propto \{ \mu_{e|c} \cdot F_{de} \cdot g + \mu_{\bar{e}|c} \cdot F_{d\bar{e}} \} \cdot \{ \mu_{e|\bar{c}} \cdot F_{\bar{d}e} + \mu_{\bar{e}|\bar{c}} \cdot F_{\bar{d}\bar{e}} \} \\ - \{ \mu_{e|c} \cdot F_{\bar{d}e} + \mu_{\bar{e}|c} \cdot F_{d\bar{e}} \} \cdot \{ \mu_{e|\bar{c}} \cdot F_{de} \cdot g + \mu_{\bar{e}|\bar{c}} \cdot F_{d\bar{e}} \}$$

The author has devised a scheme whereby each term in the equation can be represented by its position within the set of possibilities for that term, for example:

$$\mu_{e|c}, \mu_{e|\bar{c}}, \mu_{\bar{e}|c} \text{ and } \mu_{\bar{e}|\bar{c}} \text{ would appear as the tokens 0, 1, 2 and 3}$$

As long as strict ordering is maintained it becomes possible to reduce the mass of algebraic manipulation. The current operation would start with:

$$D \propto \{00g + 21\} \cdot \{12 + 33\} - \{02 + 23\} \cdot \{10g + 31\}$$

Some simple arrays are needed to accommodate the result of the products:

$$D \propto \left\{ \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0 \\ 2 \end{bmatrix} g + \begin{bmatrix} 0 \\ 3 \end{bmatrix} \begin{bmatrix} 0 \\ 3 \end{bmatrix} g + \begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} + \begin{bmatrix} 2 \\ 3 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \end{bmatrix} \right\} \\ - \left\{ \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0 \\ 2 \end{bmatrix} g + \begin{bmatrix} 0 \\ 3 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} + \begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 3 \\ 0 \end{bmatrix} g + \begin{bmatrix} 2 \\ 3 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \end{bmatrix} \right\}$$

where the first array shows the product of $\mu_{e|c}$ and $\mu_{e|\bar{c}}$, and so on (it is to be noted that the elements are interchangeable in any such array).

The first and fifth terms cancel, as do the fourth and eighth to leave:

$$D \propto \left\{ \begin{bmatrix} 0 \\ 3 \end{bmatrix} \begin{bmatrix} 0 \\ 3 \end{bmatrix} g + \begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right\} - \left\{ \begin{bmatrix} 0 \\ 3 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} + \begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 3 \\ 0 \end{bmatrix} g \right\}$$

Considering the second array position:

$$g = \{F_{d\bar{e}} \cdot F_{\bar{d}e}\} / \{F_{de} \cdot F_{\bar{d}\bar{e}}\} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} / \begin{bmatrix} 0 \\ 3 \end{bmatrix},$$

so:

$$D \propto \left\{ \begin{bmatrix} 0 \\ 3 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} + \begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right\} - \left\{ \begin{bmatrix} 0 \\ 3 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} + \begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right\} \equiv 0, \text{ as before.}$$

5.2.5. Analysis Phase 3: Proof that $b \perp c \mid a$

To test this independency, Equation 23 (and similar equations) will be used as a starting point, but this time d and e will be removed:

$$P(abcde) = \mu_0 \cdot \mu_a \cdot \mu_{b|a} \cdot \mu_{c|a} \cdot \mu_{d|b} \cdot \mu_{e|c} \cdot F_{de} \cdot g^{1-r} \cdot h^{-s}$$

$$P(abcd\bar{e}) = \mu_0 \cdot \mu_a \cdot \mu_{b|a} \cdot \mu_{c|a} \cdot \mu_{d|b} \cdot \mu_{\bar{e}|c} \cdot F_{d\bar{e}} \cdot g^{-r} \cdot h^{-s}$$

$$P(abc\bar{d}e) = \mu_0 \cdot \mu_a \cdot \mu_{b|a} \cdot \mu_{c|a} \cdot \mu_{\bar{d}|b} \cdot \mu_{e|c} \cdot F_{\bar{d}e} \cdot g^{-r} \cdot h^{-s}$$

and:

$$P(abc\bar{d}\bar{e}) = \mu_0 \cdot \mu_a \cdot \mu_{b|a} \cdot \mu_{c|a} \cdot \mu_{\bar{d}|b} \cdot \mu_{\bar{e}|c} \cdot F_{\bar{d}\bar{e}} \cdot g^{-r} \cdot h^{-s}$$

Addition applied to these four equations gives:

$$P(abc) = \mu_0 \cdot \mu_a \cdot \mu_{b|a} \cdot \mu_{c|a} \cdot \{ \mu_{d|b} \cdot [\mu_{e|c} \cdot F_{de} \cdot g + \mu_{\bar{e}|c} \cdot F_{d\bar{e}}] + \mu_{\bar{d}|b} \cdot [\mu_{e|c} \cdot F_{\bar{d}e} + \mu_{\bar{e}|c} \cdot F_{\bar{d}\bar{e}}] \} \cdot g^{-r} \cdot h^{-s}$$

Likewise:

$$P(ab\bar{c}) = \mu_0 \cdot \mu_a \cdot \mu_{b|a} \cdot \mu_{\bar{c}|a} \cdot \{ \mu_{d|b} \cdot [\mu_{e|\bar{c}} \cdot F_{de} \cdot g + \mu_{\bar{e}|\bar{c}} \cdot F_{d\bar{e}}] + \mu_{\bar{d}|b} \cdot [\mu_{e|\bar{c}} \cdot F_{\bar{d}e} + \mu_{\bar{e}|\bar{c}} \cdot F_{\bar{d}\bar{e}}] \} \cdot g^{-r} \cdot h^{-s},$$

$$P(a\bar{b}c) = \mu_0 \cdot \mu_a \cdot \mu_{\bar{b}|a} \cdot \mu_{c|a} \cdot \{ \mu_{d|\bar{b}} \cdot [\mu_{e|c} \cdot F_{de} \cdot g + \mu_{\bar{e}|c} \cdot F_{d\bar{e}}] + \mu_{\bar{d}|\bar{b}} \cdot [\mu_{e|c} \cdot F_{\bar{d}e} + \mu_{\bar{e}|c} \cdot F_{\bar{d}\bar{e}}] \} \cdot g^{-r} \cdot h^{-s}$$

and:

$$P(a\bar{b}\bar{c}) = \mu_0 \cdot \mu_a \cdot \mu_{\bar{b}|a} \cdot \mu_{\bar{c}|a} \cdot \{ \mu_{d|\bar{b}} \cdot [\mu_{e|\bar{c}} \cdot F_{de} \cdot g + \mu_{\bar{e}|\bar{c}} \cdot F_{d\bar{e}}] + \mu_{\bar{d}|\bar{b}} \cdot [\mu_{e|\bar{c}} \cdot F_{\bar{d}e} + \mu_{\bar{e}|\bar{c}} \cdot F_{\bar{d}\bar{e}}] \} \cdot g^{-r} \cdot h^{-s}$$

The condition required by *grg*-elimination to establish the required independency is:

$$P(abc) / P(ab\bar{c}) = P(a\bar{b}c) / P(a\bar{b}\bar{c})$$

Let:

$$D \equiv P(abc) \cdot P(a\bar{b}\bar{c}) - P(ab\bar{c}) \cdot P(a\bar{b}c)$$

then, after cancelling, it follows that:

$$D \propto \{ \mu_{d|b} \cdot [\mu_{e|c} \cdot F_{de} \cdot g + \mu_{\bar{e}|c} \cdot F_{d\bar{e}}] + \mu_{\bar{d}|b} \cdot [\mu_{e|c} \cdot F_{\bar{d}e} + \mu_{\bar{e}|c} \cdot F_{\bar{d}\bar{e}}] \} \\ \cdot \{ \mu_{d|\bar{b}} \cdot [\mu_{e|\bar{c}} \cdot F_{de} \cdot g + \mu_{\bar{e}|\bar{c}} \cdot F_{d\bar{e}}] + \mu_{\bar{d}|\bar{b}} \cdot [\mu_{e|\bar{c}} \cdot F_{\bar{d}e} + \mu_{\bar{e}|\bar{c}} \cdot F_{\bar{d}\bar{e}}] \} \\ - \{ \mu_{d|b} \cdot [\mu_{e|\bar{c}} \cdot F_{de} \cdot g + \mu_{\bar{e}|\bar{c}} \cdot F_{d\bar{e}}] + \mu_{\bar{d}|b} \cdot [\mu_{e|\bar{c}} \cdot F_{\bar{d}e} + \mu_{\bar{e}|\bar{c}} \cdot F_{\bar{d}\bar{e}}] \} \\ \cdot \{ \mu_{d|\bar{b}} \cdot [\mu_{e|c} \cdot F_{de} \cdot g + \mu_{\bar{e}|c} \cdot F_{d\bar{e}}] + \mu_{\bar{d}|\bar{b}} \cdot [\mu_{e|c} \cdot F_{\bar{d}e} + \mu_{\bar{e}|c} \cdot F_{\bar{d}\bar{e}}] \}$$

Applying tokens, where for example:

$$\mu_{\bar{d}|b} \cdot [\mu_{e|c} \cdot F_{de}^- + \mu_{\bar{e}|c} \cdot F_{d\bar{e}}^-] \text{ becomes } 3[02 + 23]$$

$$\begin{aligned} D &\propto \{0[00g + 21] + 2[02 + 23]\} \cdot \{1[10g + 31] + 3[12 + 33]\} \\ &- \{0[10g + 31] + 2[12 + 33]\} \cdot \{1[00g + 21] + 3[02 + 23]\} \\ &\propto \{000g + 021 + 202 + 223\} \cdot \{110g + 131 + 312 + 333\} \\ &- \{010g + 031 + 212 + 233\} \cdot \{100g + 121 + 302 + 323\} \end{aligned}$$

Multiplying out gives the difference of two sixteen term array sums. If they cancel out then the independency is established. The first term in the product is given by:

{000g } . {110g } to which is added the second term {000g} . {131}, etc.

$$\begin{array}{l} \left(\begin{array}{cccc} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} g^2 & \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0 \\ 3 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} g & \begin{bmatrix} 0 \\ 3 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0 \\ 2 \end{bmatrix} g & \begin{bmatrix} 0 \\ 3 \end{bmatrix} \begin{bmatrix} 0 \\ 3 \end{bmatrix} \begin{bmatrix} 0 \\ 3 \end{bmatrix} g \\ \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} g & \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} & \begin{bmatrix} 0 \\ 3 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} & \begin{bmatrix} 0 \\ 3 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \end{bmatrix} \\ \begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \end{bmatrix} g & \begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 0 \\ 3 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} & \begin{bmatrix} 2 \\ 3 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 2 \\ 2 \end{bmatrix} & \begin{bmatrix} 2 \\ 3 \end{bmatrix} \begin{bmatrix} 0 \\ 3 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix} \\ \begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 3 \\ 0 \end{bmatrix} g & \begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \end{bmatrix} & \begin{bmatrix} 2 \\ 3 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \end{bmatrix} & \begin{bmatrix} 2 \\ 3 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix} \begin{bmatrix} 3 \\ 3 \end{bmatrix} \end{array} \right) \\ \text{minus} & \\ \left(\begin{array}{cccc} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} g^2 & \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} g & \begin{bmatrix} 0 \\ 3 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 2 \end{bmatrix} g & \begin{bmatrix} 0 \\ 3 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} \begin{bmatrix} 0 \\ 3 \end{bmatrix} g \\ \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 3 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} g & \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} & \begin{bmatrix} 0 \\ 3 \end{bmatrix} \begin{bmatrix} 3 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} & \begin{bmatrix} 0 \\ 3 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \end{bmatrix} \\ \begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \end{bmatrix} g & \begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} & \begin{bmatrix} 2 \\ 3 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 2 \\ 2 \end{bmatrix} & \begin{bmatrix} 2 \\ 3 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix} \\ \begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 3 \\ 0 \end{bmatrix} \begin{bmatrix} 3 \\ 0 \end{bmatrix} g & \begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \end{bmatrix} & \begin{bmatrix} 2 \\ 3 \end{bmatrix} \begin{bmatrix} 3 \\ 0 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \end{bmatrix} & \begin{bmatrix} 2 \\ 3 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \end{bmatrix} \begin{bmatrix} 3 \\ 3 \end{bmatrix} \end{array} \right) \end{array}$$

All the terms cancel which are off the trailing diagonals in the two arrays. The remaining terms require the application of:

$$g = \begin{bmatrix} 1 \\ 2 \end{bmatrix} / \begin{bmatrix} 0 \\ 3 \end{bmatrix}$$

to the third array in each element for cancellation. It follows that $D \equiv 0$ and so the independence constraint $b \text{ cid } c \mid a$ has been shown to be implied and, by a parallel process, $b \text{ cid } c \mid \bar{a}$ can be inferred.

The deduction can now be made that, given the explicit moral independencies, the triangulation independencies are induced in *Loz6* under Method 3.

6. The Embedded Lozenge

Method 3 required the linear constraints and just the moral independency to determine a solution for *Loz6*, but it remained to be discovered whether this property would persist with more general

To engineer results of complete generality the graph, or an algebraic equivalent, would have to be comprehensive. It might then be possible to deduce a result by either mathematical induction or by graph reduction.

The system in this section, *Emb-Loz N* (Figure 3, with $N + 1$ vertices, is intended to mark a step towards such generality. An extended lozenge (double circles) has new network elements added so that that every vertex of the lozenge has the following two properties:

- (1) It is directly linked to both ancestors and descendants from among the new elements.
- (2) There are paths which bypass such a vertex.

Ideally the new elements should include vertices representing sets of variables but the work done was restricted to single variables because of the inherent algebraic complexity. The graph displays an ancestral ordering, of single variables. The configuration at the top of the graph will depend on the number of sides to left and right.

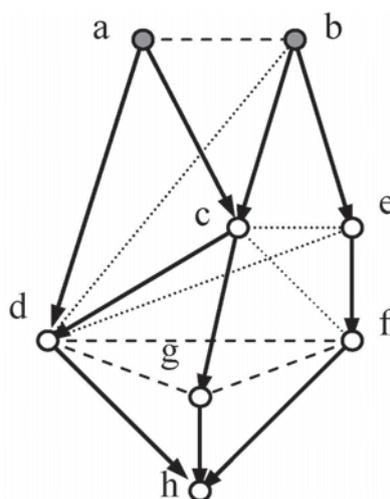
The moral independency $3 \text{ cid } 5 \mid 4,7$ is given together with the set typified by:

$$2 \text{ cid } 5 \mid 3,4 \text{ and } 4 \text{ cid } 7 \mid 5,6$$

6.2. Further Testing

Another system, *Random-Test8* (Figure 4), was subjected to examination to determine whether Methods 2, 3 and CNs gave coincident solutions.

Figure 4. *Random-Test8.*



Suitable Independence Sets for Figure 4:

For Method 2:

$$a \text{ ind } b, d \text{ cid } f \mid e, g \text{ cid } df \mid c$$

and:

$$b \text{ cid } d \mid ac, c \text{ cid } f \mid de, e \text{ cid } cd \mid b.$$

For Method 3: The first three only.

The augmented system was tested using the following linear constraints:

$$\begin{aligned}
 P(a) &= 0.56, P(b) = 0.06, \\
 P(c \mid ab, a\bar{b}, \bar{a}b, \bar{a}\bar{b}) &= 0.71, 0.50, 0.34, 0.39, \\
 P(d \mid ac) &= 0.58, 0.53, 0.32, 0.49, \\
 P(e \mid b) &= 0.05, 0.49, P(f \mid e) = 0.88, 0.92, P(g \mid c) = 0.51, 0.61, \\
 P(h \mid dfg) &= 0.19, 0.40, 0.24, 0.45, 0.75, 0.58, 0.56, 0.70
 \end{aligned}$$

The two algorithms and CNs produced identical solutions with entropy of:

$$4.526\ 672\ 923$$

Two lozenges embedded in Random-Test8 (see Figure 5,6), sections of Figure 4, will now be used to illustrate the outcome of the analysis above:

Figure 5. Loz1.

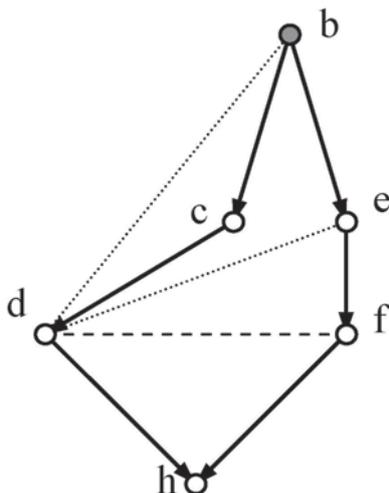
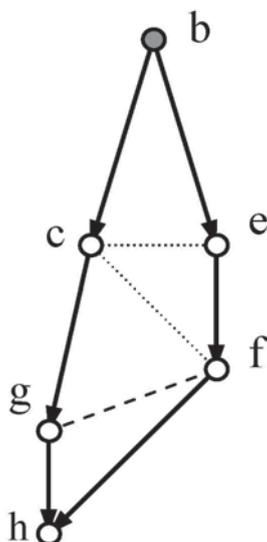


Figure 6. Loz2.



Loz1 and *Loz2*, from Figure 4, exhibit their explicit moral independencies (dashes) and implied triangulation independencies (dots).

7. Discussion

The determination of the prior distribution of a causal network, using ME, by the three algorithms described, reveals their respective properties (NB. Some additional work was performed on incomplete networks).

7.1. Properties

Sequential ME

- Operates on complete networks,
- Requires *d*-seps1&2 to propagate probabilities,
- Suffers from rapid exponential complexity.

Development of Tribus

- Operates on complete and incomplete networks,
- Requires both triangulation and moral independencies,
- Can be used to validate CNs methodologies,
- Can be used to detect the stationary points for an incomplete network by varying the initial conditions,
- Potential for application beyond conventional networks.

Variation on Tribus

- Operates on complete networks only,
- Requires moral independencies alone,
- Potential for wider application,
- Further generalisation of proof needed.

7.2. Further Testing

A total of thirty systems were tested, including *Loz6*, *Tree-Loz-Vee-Tri7*, and *Random-Test8*. All the tests gave matching distributions under Methods 1,2,3 and CNs methodologies.

Method 1 relied on *d*-Separation alone and Method 3 required fewer explicit independencies than Method 2, as shown in Table 2.

Table 2. Independence Averages.

Average Number of Explicit Independence Constraints Rqd.		
METHOD 1 & CNS: NIL,	METHOD 2: 8.8,	METHOD 3: 5.4

8. Conclusions

Three methods for calculating the *a priori* probability distribution of a causal network, using ME, were discussed.

The first method used sequential determination of local probabilities, following a paper by Lukasiewicz, and this was illustrated with an example. Only the first two clauses of *d*-Separation were required in this process.

Method 2 started from the General Tribus Formula and aimed to find the over-all distribution directly. An algorithm was designed which required a set of explicit moral and triangulation independencies, in addition to the linear constraints. The technique for handling linear constraints served as a starting point for Method 3 where only the moral independencies and the linear constraints were required.

A novel representation of the General Tribus Formula, μ -Notation, was then used to represent the contributions from each constraint and was further developed by grouping similar terms. This divergence between Methods 2 and 3 was investigated by analysis applied to an elementary example. The investigation was then extended to a system of greater generality where Bottom-up analysis uncovered a method of finding implied independencies whilst climbing the graph. More work is needed to produce a result of complete generality.

A series of tests was undertaken to compare the independence requirements of the three methodologies. The results showed that Method 3 provides a more efficient method of solution than Method 2 over a significant range of complete causal network systems. Methods 2 and 3 have the potential to solve a wider range of problems than the CNs methodologies, therefore they constitute a useful augmentation of the technology available for solving problems in Expert Systems.

Acknowledgements

Suggestions and corrections were provided by R. S. Roberts (Durham).

References

1. Griffeath, D.S. Computer solution of the discrete maximum entropy problem. *Technometrics* **1972**, *14*, 891–897.
2. Rhodes, P.C; Garside, G.R. Use of maximum entropy method as a methodology for probabilistic reasoning. *Knowl. Based Syst.* **1995**, *8*, 249–258.
3. Neapolitan, R.E. *Probabilistic Reasoning in Expert Systems: Theory and Algorithms*; John Wiley & Sons: Hoboken, NJ, USA, 1990.
4. Anderson, S.K.; Oleson, K.G.; Jenson F.V.; Jenson, F. HUGIN-A Shell for Building Bayesian Belief Universes for Expert Systems. *Proc. Int. Jt. Conf. Artif. Intell.* **1989**, *11*, 1080–1085.
5. Lauritzen, S.L.; Spiegelhalter, D.J. Local computations with probabilities on graphical structures and their applications to expert systems. *J. Roy. Stat. Soc. B* **1988**, *50*, 157–224.
6. Jaynes, E.T. Where do we stand on maximum entropy? In *The ME Formalism*; The MIT Press: Cambridge, MA, USA, 1979.

7. Paris, J.B; Vencovská, A. A Note on the Inevitability of Maximum Entropy. *Int. J. Approximate Reasoning* **1990**, *4*, 183–223.
8. Shore, J.E.; W.Johnson, R,W. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans. Inform. Theory* **1980**, *IT-26*, 26–37.
9. Williamson, J.; Corfield, A.; Williamson, J. Foundations for bayesian networks. In *Foundations of Bayesianism (Applied Logic Series)*; Kluwer: Dordrecht, The Netherlands, 2001; pp. 75–115.
10. Williamson, J. Maximising entropy efficiently. *Linköping Electron. Artic. Comput. Inform. Sci.* **2002**, *7*, 1–32.
11. Markham, M.J. *Probabilistic independence and its impact on the use of maximum entropy in causal networks*. Ph.D. Thesis, School of Informatics, Department of Computing, University of Bradford, Bradford, England, 2005.
12. Paris, J.B. On filling-in missing information in causal networks. *Int. J. Uncertainty Fuzziness Knowl. Based Syst.* **2005**, *13*, 263–280.
13. De Campos, L.; Moral, S. Independence concepts for convex sets of probabilities. In Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence, Montreal, Canada, 18–20 August 1995; Morgan Kaufmann: San Francisco, CA, USA, 1995; pp. 108–115.
14. Wally, P. *Statistical Reasoning with Imprecise Probabilities*; Chapman and Hall/CRC: Boca Raton, FL, USA, 1991; pp. 443–444.
15. Lukasiewicz, T. Credal Networks under Maximum Entropy. In Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence, Stanford, California, USA, July 2000; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2000; pp. 63–70.
16. Tribus, M. *Rational Descriptions, Decisions and Designs*; Pergamon Press: Elmsford, NY, USA, 1969, 120–123.
17. Markham, M.J. Independence in complete and incomplete causal networks under maximum entropy. *Int. J. Uncertainty Fuzziness Knowl. Based Syst.* **2008**, *16*, 699–713.
18. Geiger, D.; Verma, T.; Pearl, J. *d*-Separation: From theorems to algorithms. In Proceedings of the UAI'89 Fifth Annual Conference on Uncertainty in Artificial Intelligence; North-Holland Publishing Co.: Amsterdam, The Netherlands, 1990; pp. 139–148.
19. Verma, T.S.; Pearl, J. Causal networks: Semantics and expressiveness. In Proceedings of the 4th AAAI Workshop on Uncertainty in AI; North-Holland Publishing Co.: Amsterdam, The Netherlands, 1988; pp. 352–359.
20. Geiger, D.; Pearl, J. On the logic of causal models. In Proceedings of 4th AAAI Workshop on Uncertainty in AI; North-Holland Publishing Co.: Amsterdam, The Netherlands, 1988; pp. 136–147.
21. Holmes, D.E. Independence relationships implied by D-separation in the bayesian model of a causal tree are preserved by the maximum entropy model. In Proceedings of the 19th International Workshop on Bayesian Interference and Maximum Entropy Methods, Boise, ID, USA, 2–6 August 1999; pp. 296–307.
22. Grove, A.J.; Halpern; J.Y.; Koller, D. Random worlds and maximum entropy. *J. Artif. Intell. Res.* **1994**, *2*, 33–88.
23. Hunter, D. Uncertain reasoning using maximum entropy inference. In Proceedings of the UAI-Uncertainty in Artificial Intelligence; Elsevier Science: New York, NY, USA, 1986; Volume 4, pp. 203–209.

24. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.
25. Garside, G.R.; Rhodes, P.C. Computing marginal probabilities in causal multiway trees given incomplete information. *Knowl. Based Syst.* **1996**, *9*, 315–327.

© 2011 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).