On filling-in missing conditional probabilities in causal networks

Paris, J. B.

2008

MIMS EPrint: **2008.110**

Manchester Institute for Mathematical Sciences
School of Mathematics

The University of Manchester

Reports available from: http://eprints.maths.manchester.ac.uk/
And by contacting: The MIMS Secretary
School of Mathematics
The University of Manchester
Manchester, M13 9PL, UK

ISSN 1749-9097

On filling-in missing conditional probabilities in causal networks

J. B. Paris
Department of Mathematics
University of Manchester
Manchester M13 9PL
UK
jeff.paris@manchester.ac.uk

August 15, 2008

Abstract

This paper considers the problem and appropriateness of filling-in missing conditional probabilities in causal networks by the use of maximum entropy. Results generalizing earlier work of Rhodes, Garside & Holmes are proved straightforwardly by the direct application of principles satisfied by the maximum entropy inference process under the assumed uniqueness of the maximum entropy solution. It is however demonstrated that the implicit assumption of uniqueness in the Rhodes, Garside & Holmes papers may fail even in the case of inverted trees. An alternative approach to filling in missing values using the limiting centre of mass inference process is then described which does not suffer this shortcoming, is trivially computationally feasible and arguably enjoys more justification in the context when the probabilities are objective (for example derived from frequencies) than by taking maximum entropy values.

Keywords: Missing Information, Causal Networks, Maximum Entropy, Centre of Mass.

1 Introduction

In papers [1],[3],[5],[6],[21], Rhodes, Garside and Holmes described efficient algorithms for filling in missing conditional probabilities in various classes of causal

networks by the maximum entropy method. Whilst their main interest, apparently, was the formulation of algorithms a key step in their methods is to first isolate comparatively simple, computationally manageable, subsets of the set of probabilistic independence constraints associate with such causal networks whose maximum entropy solution (hereafter shortened to *maxent* solution) is the unique solution of the full set of constraints. Once this has been achieved missing conditional probabilities can be 'filled-in' by computing the maxent solution of the existing conditional probability constraints plus the above mentioned computationally manageable subsets of the independence constraints, now a computationally relatively tractable task, and using the conditional probabilities provided by this solution to fill-in any omissions. The point here being that this maxent solution remains the maxent solution even after this filling-in, and will, by the choice of these manageable subsets, still satisfy the necessary full set of independence conditions. In other words these filled-in values are exactly the ones that would have been obtained by using the maxent solution of the original set of constraints, with the full set of independencies, in the first place.

The initial aim of this paper is to give straightforward proofs, *under the assumption of a unique maxent solution*, of some of the Rhodes-Garside-Holmes results by arguing directly from well known general principles that the maximum entropy inference process satisfies. We shall then show that unfortunately there need not be a unique maxent solution, even in the case of inverted trees[1]. In the final section we shall show however that there is an alternative to maxent, the so called limiting centre of mass inference process (see [12]), CM^{∞}, which does enjoy uniqueness, is trivially computationally tractible and is, in the context where the existing conditional probabilities are objective (for example when obtained from frequencies), arguably more justified than maxent.

2 Background and Notation

To fit in best with the formulation of the maxent paradigm (also referred to as the *maximum entropy inference process*) as given in [16], [12], [13], [18] we shall adopt the notation of [12], limiting ourselves to networks whose vertices take just two values, 0 and 1, the generalization to more values being straightforward. So let $L = L(p_1, p_2, ..., p_n)$ be a finite propositional language with propositional variables $p_1, p_2, ..., p_n$ and let SL be the set of sentences formed from L, using say the connectives $\neg$, $\vee$, $\wedge$. For a propositional variable p let p^1, p^0 stand for p, $\neg p$ respectively. As usual a probability function on SL is a function $w : SL \rightarrow [0, 1]$ such that for all all $\theta, \phi \in SL$:

(P1) If $\models \theta$ then $w(\theta) = 1$.

[1]In this case Rhodes et al incorrectly assumed that the maxent solution was unique on the basis of results of Garside in [2]

(P2) If $\models \neg(\theta \wedge \phi)$ then $w(\theta \vee \phi) = w(\theta) + w(\phi)$.

Such a function is uniquely determined by its values on the *atoms* of L, that is the sentences of L of the form

$$p_1^{\epsilon_1} \wedge p_2^{\epsilon_2} \wedge ... \wedge p_n^{\epsilon_n}$$

where $\epsilon_1, \epsilon_2, ..., \epsilon_n \in \{0, 1\}$. For $w(\theta) \neq 0$ the conditional probability function $w(.|\theta)$ is defined as usual by

$$w(\phi|\theta) = \frac{w(\theta \wedge \phi)}{w(\theta)}.$$

Since it will also be convenient to use this notation even when possibly $w(\theta) = 0$ we shall adopt the convention that expressions such as $w(\phi|\theta) = X$ are shorthand for

$$w(\theta \wedge \phi) = Xw(\theta). \tag{1}$$

A *causal network* (on L) is (or can be taken to be) a set of probabilistic constraints on a probability function w on SL of the form

$$\begin{aligned} w(p_i|p_1^{\epsilon_1} \wedge p_2^{\epsilon_2} \wedge ... \wedge p_{i-1}^{\epsilon_{i-1}}) &= w(p_i|p_{i1}^{\epsilon_{i1}} \wedge p_{i2}^{\epsilon_{i2}} \wedge ... \wedge p_{im_i}^{\epsilon_{im_i}}) && (2) \\ w(p_i|p_{i1}^{\epsilon_{i1}} \wedge p_{i2}^{\epsilon_{i2}} \wedge ... \wedge p_{im_i}^{\epsilon_{im_i}}) &= b(i; \epsilon_{i1}, \epsilon_{i2}, ..., \epsilon_{im_i}) && (3) \end{aligned}$$

where $i = 1, 2, ..., n$, the $\epsilon_{i1}, \epsilon_{i2}, ..., \epsilon_{im_i}$ come from $\{\epsilon_1, \epsilon_2, ..., \epsilon_{i-1}\} \in \{0, 1\}$ and the $b(i; \epsilon_{i1}, \epsilon_{i2}, ..., \epsilon_{im_i}) \in [0, 1]$.

Such a set of constraints (2), (3) has a unique solution given by

$$\begin{aligned} & w(p_1^{\epsilon_1} \wedge p_2^{\epsilon_2} \wedge ... \wedge p_n^{\epsilon_n}) \\ &= \prod_{i=1}^{n} w(p_i^{\epsilon_i}|p_{i1}^{\epsilon_{i1}} \wedge p_{i2}^{\epsilon_{i2}} \wedge ... \wedge p_{im_i}^{\epsilon_{im_i}}) && (4) \\ &= \prod_{i=1}^{n} \left(\epsilon_i b(i; \epsilon_{i1}, \epsilon_{i2}, ..., \epsilon_{im_i}) + (1 - \epsilon_i)(1 - b(i; \epsilon_{i1}, \epsilon_{i2}, ..., \epsilon_{im_i}))\right). \end{aligned}$$

It is usual to think of the p_i as vertices of a directed graph with an edge from p_j to p_i if $p_j = p_{ir}$ for some $1 \leq r \leq m_i$ and to classify the network in terms of this graph. As far as this paper is concerned we shall limit ourselves to networks where the graph is *acyclic*, that is there are no cycles $p_{j_1}, p_{j_2}, ..., p_{j_s}$ with edges from p_{j_s} to p_{j_1} and from p_{j_r} to $p_{j_{r+1}}$ for $r = 1, 2, ..., s - 1$. In particular we shall be interested in *trees*, which are acyclic graphs in which no vertex has more than one edge directed to it, *inverted trees*, which are acyclic graphs in which no vertex

has more than one edge directed from it, and *singly connected* graphs where there are no undirected cycles.

In this paper we are interested in the case where the graphical structure is fully known but some of the conditional probabilities $b(i; \epsilon_{i1}, \epsilon_{i2}, ..., \epsilon_{im_i})$ are missing. Filling-in missing values is necessary if one wishes to access the computationally efficient algorithms for solving for probabilities which such a complete causal network permits, see for example the seminal works, [10],[20].

Filling-in using maxent

Of course 'filling-in' values is just another way of saying 'guessing', or even 'inventing' values. Nevertheless one might hope to adopt some procedure with a vestige of justification. One general procedure for supplying missing probabilities with such pretentions is the maxent paradigm, that is, take the (assumed unique) probability function satisfying the remaining constraints with the maximum entropy, where in our context the entropy is given by the sum over the atoms α of L

$$E(w) = -\sum_{\alpha} w(\alpha) \log w(\alpha), \tag{5}$$

and use the values given by this function. The popular justification here being that from Shannon's arguments, see [24], entropy is a measure of lack of information, so that by choosing the maxent value one is choosing the least informative possibility, the value which goes as little as possible beyond what is actually known. We shall return to this point later, but for the moment we should point out that, presumably in consequence, there are already in the literature a number of papers on filling in missing values using various applications of maxent, in particular [7], [11], [22], [23], [25] in addition to the already cited works of Rhodes-Garside-Holmes.

From a theoretical standpoint the only possible problem with filling-in the missing values using the maxent solution of the constraints of type (2) and those of type (3) that we do have, is that there may not be a unique such solution (a real possibility as we shall see later). Even assuming uniqueness however, the number of constraints of type (2) will in practice tend to be prohibitively large, rendering actual calculation of the missing values infeasible. A possible alternative might be to drop the constraints of type (2) and simply take the maximum entropy solution of the existing constraints of type (3). [In real, tractable, cases the m_i would not be so large as cause a repeat of these same computational headaches.] A possible problem with this approach is that the resulting maximum entropy solution (now unique alright because the solution space is convex – see, for example, [12] p66) may not satisfy the constraints of type (2), see for example [20] p463-464, [7] (and also section 6 of [26] for further illumination). What seems to be needed then is

some canonical set of constraints which follows from those of types (2) and (3), is computationally acceptable and whose maximum entropy solution satisfies all the existing constraints of types (2) and (3). What Rhodes, Garside and Holmes show (under the running assumption of uniqueness) in a series of papers is that for certain families tree-like graphs such a set exists.

Specifically, in the case of trees they show that this canonical set can be taken to be empty whilst in the case of singly connected graphs the set of constraints

$$w(p_{i_1}^{\epsilon_{i_1}} \wedge p_{i_2}^{\epsilon_{i_2}} \wedge ... \wedge p_{i_{m_i}}^{\epsilon_{i_{m_i}}}) = \prod_{j=1}^{m_i} w(p_{i_j}^{\epsilon_{i_j}}) \tag{6}$$

suffices (and for such networks is directly derivable from (4) by marginalization).

Their results of this form (on which they ground their main objective of formulating feasible algorithms) will follow from Theorems 1 and 3 and Corollary 2 below. Before proving this result however we need to introduce some notation and a key property (Separation) that maxent satisfies..

Hybridizing somewhat the notation given in [12] and [18] let CL be the collection of finite, satisfiable, sets of constraints, on a probability function $w : SL \to [0, 1]$, of the form

$$f_j(w(\theta_1), w(\theta_2), ..., w(\theta_m)) = 0, \quad j = 1, 2, ..., r,$$

where the f_j are continuous functions over the reals, the $\theta_{ij} \in SL$ and by *satisfiable* we mean that there is a probability function w on SL satisfying these constraints. For $K \in CL$ let $V^L(K)$ denote the set of probability functions on SL satisfying K and let $ME^L(K)$ denote the *set* of probability function on SL satisfying K whose entropy is maximal amongst all such functions in $V^L(K)$. In general there will not be a unique such probability function. However when the f_j are linear $V^L(K)$ is convex and $ME^L(K)$ is unique. In this case the *inference process* ME which picks out this unique solution (thought of as a function both of L and $K \in CL$) has been studied extensively (see for example [12]) and is well known to be uniquely characterized by a number of 'common sense principles of uncertain reasoning'[2].

This characterization was extended to the case of (continuous) non-linear constraints in [18]. For the purposes of this paper however it will be enough to point out a handful of key properties, or principles, that carry over even to the case where there is not necessarily a unique maxent solution.

The first such property is *Language Invariance*, namely that if $L_1 \subseteq L_2$ and $K \in CL_1 (\subseteq CL_2)$ and w $\in ME^{L_2}(K)$ then w restricted to SL_1 is in $ME^{L_1}(K)$[3].

[2]Indeed we have repeatedly argued that in this context it exactly coincides with what we mean by common sense uncertain reasoning.

[3]We shall endeavor to use w for the variable, or 'unknown', probability function appearing

The second key property is *Obstinacy*; if $K_1, K_2 \in CL$ and $\mathrm{w} \in ME^L(K_1)$ satisfies K_2 then $\mathrm{w} \in ME^L(K_1 \cup K_2) \subseteq ME^L(K_1)$.

The third property we shall need is not explicitly stated in the above mentioned papers so we state it as a lemma. [This, and related properties of maxent, are folklore in the subject, see for example also the conditional independence properties in [27], sections 4 and 5 of [26], [28] and the atomicity principle in [12].]

The Separation Property Suppose that $L_0, L_1, L_2, ..., L_m$ are finite, pairwise disjoint, propositional languages, $L = L_0 \cup L_1 \cup ... \cup L_m$, $K_i \in C(L_0 \cup L_i)$ for $i = 1, 2, ..., m$. Let $\mathrm{w} \in ME^L(\bigcup_{i=1}^m K_i)$, let K_0 be the set of constraints

$$w(\alpha_0) = \mathrm{w}(\alpha_0),$$

as α_0 ranges over the atoms of L_0, and for $i = 1, 2, ..., m$ let w_i be the restriction of w to $S(L_i \cup L_0)$. Then $\mathrm{w}_i \in ME^{L_i \cup L_0}(K_i \cup K_0)$ and for α_i atoms of L_i, $i = 0, 1, ..., m$,

$$\mathrm{w}(\alpha_0 \wedge ... \wedge \alpha_m) \cdot \mathrm{w}(\alpha_0)^{m-1} = \prod_{i=1}^m \mathrm{w}(\alpha_0 \wedge \alpha_i) = \prod_{i=1}^m \mathrm{w}_i(\alpha_0 \wedge \alpha_i). \tag{7}$$

Conversely if $\mathrm{v}_i \in ME^{L_i \cup L_0}(K_i)$ for $i = 1, 2, ..., m$ and

$$\mathrm{v}_1(\alpha_0) = \mathrm{v}_2(\alpha_0) = ... = \mathrm{v}_m(\alpha_0)$$

for all atoms α_0 of L_0 then the probability function v on SL defined by

$$\mathrm{v}(\alpha_0 \wedge ... \wedge \alpha_m) \cdot \mathrm{v}_1(\alpha_0)^{m-1} = \prod_{i=1}^m \mathrm{v}_i(\alpha_0 \wedge \alpha_i)$$

is in $ME^L(\bigcup_{i=1}^m K_i)$.

Proof. For $i = 1, 2, ..., m$ let $\mathrm{u}_i \in ME^{L_i \cup L_0}(K_i \cup K_0)$ and define the probability function u on $S(\bigcup_{i=0}^m L_i)$ by

$$\mathrm{u}(\bigwedge_{i=0}^m \alpha_i) = \mathrm{w}(\alpha_0)^{1-m} \cdot \prod_{i=1}^m \mathrm{u}_i(\alpha_0 \wedge \alpha_i).$$

Then since u_i satisfies K_0, $\mathrm{u}_i(\alpha_0) = \mathrm{w}(\alpha_0)$ and

$$\begin{aligned} E(\mathrm{u}) &= -\sum_{i=1}^m \sum_{\alpha_0, \alpha_i} \mathrm{u}_i(\alpha_0 \wedge \alpha_i) \log \mathrm{u}_i(\alpha_0 \wedge \alpha_i) \\ &\quad -(m-1) \sum_{\alpha_0} \mathrm{w}(\alpha_0) \log \mathrm{w}(\alpha_0) \end{aligned} \tag{8}$$

in the sets of constraints and to use an w for actual probability function, though at times the distinction may unavoidably become rather blurred.

whereas

$$\begin{aligned} E(\mathrm{w}) &= -\sum_{\alpha_0,\ldots,\alpha_m} \mathrm{w}(\bigwedge_{i=0}^{m}\alpha_i)\log \mathrm{w}(\bigwedge_{i=0}^{m}\alpha_i) \\ &= -\sum_{\alpha_0,\ldots,\alpha_m} \mathrm{w}(\bigwedge_{i=0}^{m}\alpha_i)\log\left\{\frac{\mathrm{w}(\bigwedge_{i=0}^{m}\alpha_i)\cdot \mathrm{w}(\alpha_0)^{m-1}}{\prod_{i=1}^{m}\mathrm{w}(\alpha_0\wedge\alpha_i)}\cdot\frac{\prod_{i=1}^{m}\mathrm{w}(\alpha_0\wedge\alpha_i)}{\mathrm{w}(\alpha_0)^{m-1}}\right\} \\ &= -\sum_{i=1}^{m}\sum_{\alpha_0,\alpha_i}\mathrm{w}(\alpha_0\wedge\alpha_i)\left\{\log\mathrm{w}(\alpha_0\wedge\alpha_i)-(m-1)\log\mathrm{w}(\alpha_0)\right\} \\ &\quad -\sum_{\alpha_0,\ldots,\alpha_m}\mathrm{w}(\bigwedge_{i=0}^{m}\alpha_i)\log\left\{\frac{\mathrm{w}(\bigwedge_{i=0}^{m}\alpha_i)\cdot \mathrm{w}(\alpha_0)^{m-1}}{\prod_{i=1}^{m}\mathrm{w}(\alpha_0\wedge\alpha_i)}\right\} \qquad (9) \end{aligned}$$

where summands for which $\mathrm{w}(\alpha_0)=0$ are taken to be zero. The last term in (9) is actually a cross entropy, so non-negative (see for example [12], p119). Also the $\mathrm{w}_i \in V^{L_i\cup L_0}(K_i\cup K_0)$ so by comparing (8) and (9), $E(\mathrm{u})$ will strictly exceed $E(\mathrm{w})$ if any $E(\mathrm{u}_i) > E(\mathrm{w}_i)$. Since u satisfies $\bigcup_{i=1}^{m} K_i$ it follows then from the choice of w that $E(\mathrm{u}_i) = E(\mathrm{w}_i)$, and hence $\mathrm{w}_i \in ME^{L_0\cup L_i}(K_i\cup K_0)$, for each $i = 1, 2, ..., m$. Similarly the above mentioned cross-entropy term cannot be strictly positive, from which we conclude (see again, for example, [12], p119) that

$$\mathrm{w}(\bigwedge_{i=0}^{m}\alpha_i)\cdot\mathrm{w}(\alpha_0)^{m-1} = \prod_{i=1}^{m}\mathrm{w}(\alpha_0\wedge\alpha_i).$$

The last part now follows by reversing the above arguments, noticing that as defined $\mathrm{v}\in V^{L}(\bigcup_{i=1}^{m}K_i)$. ■

Notice that the same proof goes through with $L_0=\emptyset$ to give in this case that

$$\mathrm{w}(\alpha_1\wedge\ldots\wedge\alpha_m) = \prod_{i=1}^{m}\mathrm{w}(\alpha_i) = \prod_{i=1}^{m}\mathrm{w}_i(\alpha_i). \qquad (10)$$

We now prove a result for general acyclic graphs. The results of Rhodes-Garside-Holmes will follow as corollaries under the assumption of uniqueness. So assume that our graph is acyclic and for $1\le i\le n$ let C_i be the set of $j<i$ such that there is an edge in the graph from p_j to some p_r with $r\ge i$. In particular then $p_{i1}, p_{i2}, ..., p_{im_i}\in C_i$. Let $q_{i1}, q_{i2}, ..., q_{ig_i}$ list the remaining elements of C_i and let $L_i = \{p_j \mid j\in C_i\}$.

Theorem 1 *In the case where the graph of the constraint set given by (2), (3) is acyclic, if* w *is a maxent solution to the set K of constraints of type (3) and of type*

$$w(p_i|\bigwedge_{j=1}^{m_i}p_{i_j}^{\epsilon_{i_j}}\wedge\bigwedge_{j=1}^{g_i}q_{i_j}^{\delta_{i_j}}) = w(p_i|\bigwedge_{j=1}^{m_i}p_{i_j}^{\epsilon_{i_j}}), \qquad (11)$$

for $\epsilon_{i_1}, ..., \epsilon_{i_{m_i}}, \delta_{i_1}, ..., \delta_{i_{g_i}} \in \{0,1\}$*, then* w *already satisfies the constraints of type (2).*

[Notice that the constraints of type (11) are derivable from the constraints of type (2).]

Proof. Let K_i^- be the set of constraints

$$w(p_k| \bigwedge_{j=1}^{m_k} p_{kj}^{\epsilon_{kj}} \wedge \bigwedge_{j=1}^{g_k} q_{kj}^{\delta_{kj}}) = w(p_k| \bigwedge_{j=1}^{m_k} p_{kj}^{\epsilon_{kj}}),$$
$$w(p_k|p_{k1}^{\epsilon_{k1}} \wedge p_{k2}^{\epsilon_{k2}} \wedge ... \wedge p_{km_k}^{\epsilon_{km_k}}) = b(k; \epsilon_{k1}, \epsilon_{k2}, ..., \epsilon_{km_k}),$$

of types (11), (3) for $k < i$. Similarly let K_i^+ be the set of such constraints for $k \geq i$.

Let w be a maxent solution to K and let K_i^0 be the set of constraints

$$w(\bigwedge_{j=1}^{m_i} p_{ij}^{\epsilon_{ij}} \wedge \bigwedge_{j=1}^{g_i} q_{ij}^{\delta_{ij}}) = \mathrm{w}(\bigwedge_{j=1}^{m_i} p_{ij}^{\epsilon_{ij}} \wedge \bigwedge_{j=1}^{g_i} q_{ij}^{\delta_{ij}}), \tag{12}$$

for $\epsilon_{i1}, ..., \epsilon_{im_i}, \delta_{i1}, ..., \delta_{ig_i} \in \{0,1\}$.

Let $L_i^- = \{p_j | j < i, j \notin C_i\}$, $L_i^+ = \{p_j | i \leq j \leq n\}$. It is easy to check that $K_i^- \in C(L_i^- \cup L_i)$, $K_i^+ \in C(L_i^+ \cup L_i)$. Then by Separation, $\mathrm{w}^-, \mathrm{w}^+$, the restrictions of w to $S(L_i^- \cup L_i), S(L_i^+ \cup L_i)$ respectively, are in $ME^{L_i^-}(K_i^- \cup K_i^0), ME^{L_i^+}(K_i^+ \cup K_i^0)$ respectively and

$$\mathrm{w}(\alpha^+ \wedge \alpha_0 \wedge \alpha^-) \cdot \mathrm{w}(\alpha_0) = \mathrm{w}(\alpha^+ \wedge \alpha_0) \cdot \mathrm{w}(\alpha^- \wedge \alpha_0) \tag{13}$$

for atoms $\alpha^+, \alpha_0, \alpha^-$ of L_i^+, L_i, L_i^- respectively.

Summing (13) over literals $\pm p_k$ with $k > i$ now gives

$$\mathrm{w}(p_i^{\epsilon_i} \wedge \alpha_0 \wedge \alpha^-) \cdot \mathrm{w}(\alpha_0) = \mathrm{w}(p_i^{\epsilon_i} \wedge \alpha_0) \cdot \mathrm{w}(\alpha_0 \wedge \alpha^-), \tag{14}$$

and hence

$$\mathrm{w}(p_i^{\epsilon_i} | \alpha_0 \wedge \alpha^-) = \mathrm{w}(p_i^{\epsilon_i} | \alpha_0). \tag{15}$$

Combining this with the constraint (11), which w of course satisfies, gives (2), as required. ■

Notice that the particular conclusion of this proof for p_i, that

$$w(p_i|p_1^{\epsilon_1} \wedge p_2^{\epsilon_2} \wedge ... \wedge p_{i-1}^{\epsilon_{i-1}}) = w(p_i|p_{i1}^{\epsilon_{i1}} \wedge p_{i2}^{\epsilon_{i2}} \wedge ... \wedge p_{im_i}^{\epsilon_{im_i}}),$$

did not require that *all* constraints of type (11) were actually present in K, only that the particular constraint of that form for p_i, that is,

$$w(p_i| \bigwedge_{j=1}^{m_i} p_{i_j}^{\epsilon_{i_j}} \wedge \bigwedge_{j=1}^{g_i} q_{i_j}^{\delta_{i_j}}) = w(p_i| \bigwedge_{j=1}^{m_i} p_{i_j}^{\epsilon_{i_j}}),$$

was present. This observation leads to the following corollary which appears in the work of Rhodes-Garside-Holmes (see specifically [3], [5]) under the implicit assumption of uniqueness.

Corollary 2 *In the case where the graph is a tree any maxent solution to the constraints of type (3) also satisfies all the constraints of type (2).*

Proof. Take a particular vertex p_i and renumber the vertices so that in this new numbering the only vertices with numbers higher than p_i are those that can be reached by a directed path from p_i. Then, since our graph is a tree, for this numbering the constraint of type (11) for p_i is trivial (there are no q_{ij}) so by the above observation the constraint of type (2) for p_i and this new numbering holds. But then since any vertex with a lower number than i retains this property in the new numbering the constraint of type (2) for p_i in the old numbering must also hold. ■

The next theorem (under an implicit uniqueness assumption) appears, with a different proof, in the paper [6] of Holmes.

Theorem 3 *In the case where the graph is singly connected any maxent solution to the constraints of type (3) together with the constraints*

$$w(p_{i_1}^{\epsilon_{i_1}} \wedge p_{i_2}^{\epsilon_{i_2}} \wedge ... \wedge p_{i_{m_i}}^{\epsilon_{i_{m_i}}}) = \prod_{j=1}^{m_i} w(p_{i_j}^{\epsilon_{i_j}}) \tag{16}$$

of type (6) also satisfies the constraints of type (2).

Notice that the constraints of type (6) do follow from those of type (2).

Proof. Let w be a maxent solution to these constraints of types (3) and (6). In view of Theorem 1 it would be enough to show that w satisfies the constraints of type (11). However it turns out to be just as easy (or hard) to prove the result directly.

Towards this end, we first introduce a little notation. For vertices p_i, p_j such that there is an edge from p_j to p_i let $L[j,i]$ be those verticess p_k for which there is a path (not necessarily directed) from p_j to p_k which does not pass through p_i. Notice then that $p_{ij} \in L[ij,i]$ since the graph is singly connected the $L[ij,i]$ are all disjoint.

By Separation with $L_0 = \{p_{i1}\}$ we have

$$\mathrm{w}(\bigwedge_{p_k \in L[i1,i]} p_k^{\epsilon_k} \wedge \bigwedge_{p_k \notin L[i1,i]} p_k^{\epsilon_k}) \cdot \mathrm{w}(p_{i1}^{\epsilon_{i1}}) = \mathrm{w}(\bigwedge_{p_k \in L[i1,i]} p_k^{\epsilon_k}) \cdot \mathrm{w}(p_{i1}^{\epsilon_{i1}} \wedge \bigwedge_{p_k \notin L[i1,i]} p_k^{\epsilon_k}), \tag{17}$$

which, by summing over a suitable set of literals, gives

$$\mathrm{w}(\bigwedge_{p_k \in L[i1,i]} p_k^{\epsilon_k} \wedge \bigwedge_{k=2}^{m_i} p_k^{\epsilon_k}) \cdot \mathrm{w}(p_{i1}^{\epsilon_{i1}}) = \mathrm{w}(\bigwedge_{p_k \in L[i1,i]} p_k^{\epsilon_k}) \cdot \mathrm{w}(\bigwedge_{k=1}^{m_i} p_{ik}^{\epsilon_{ik}}). \tag{18}$$

Similar identities hold of course for the other p_{ij}.

Again by Separation with $L_0 = \{p_{i1}, p_{i2}, ..., p_{im_i}\}$, $L_j = L[ij, i]$, $j = 1, 2, ..., m_i$, J the remaining vertices, and summing over the literals from J (notice by single-connectedness this application is valid) we obtain

$$\mathrm{w}(\bigwedge_{j=1}^{m_i} \bigwedge_{p_k \in L[ij,i]} p_k^{\epsilon_k}) \cdot \mathrm{w}(\bigwedge_{k=1}^{m_i} p_{ik}^{\epsilon_{ik}})^{m_i} = \prod_{j=1}^{m_i} \mathrm{w}(\bigwedge_{p_k \in L_0 \cup L[ij,i]} p_k^{\epsilon_k}) \cdot \mathrm{w}(\bigwedge_{k=1}^{m_i} p_{ik}^{\epsilon_{ik}}). \tag{19}$$

Using (18), (6) and cancelling now gives

$$\mathrm{w}(\bigwedge_{j=1}^{m_i} \bigwedge_{p_k \in L[ij,i]} p_k^{\epsilon_k}) = \prod_{j=1}^{m_i} \mathrm{w}(\bigwedge_{p_k \in L[ij,i]} p_k^{\epsilon_k}). \tag{20}$$

Having established these identities we now press ahead to show that the constraints of type (2) must hold. Let L_i^* be the set of vertices p_j such that there is a path from p_j to some parent p_{ik} of p_i which does not pass through p_i, together with p_i itself. Let L_i^- be the set of p_k from which there is a directed path to p_i. Our plan is to show that if J is a set of vertices such that

$$L_i^* \cap \bigcup_{p_j \in J} L_j^- = \emptyset \tag{21}$$

then

$$\mathrm{w}(\bigwedge_{p_j \in L_i^*} p_j^{\epsilon_j} \wedge \bigwedge_{p_j \in J} p_j^{\epsilon_j}) = \mathrm{w}(\bigwedge_{p_j \in L_i^*} p_j^{\epsilon_j}) \cdot \mathrm{w}(\bigwedge_{p_j \in J} p_j^{\epsilon_j}). \tag{22}$$

The required result follows from this by taking J to be the set of p_k with $k < i$ and $p_k \notin L_i^*$ and summing over the literals not in $\{\pm p_k \mid k \leq i\}$ to give

$$\mathrm{w}(\bigwedge_{k \leq i} p_k^{\epsilon_k}) = \mathrm{w}(\bigwedge_{\substack{p_k \in L_i^* \\ k \leq i}} p_k^{\epsilon_k}) \cdot \mathrm{w}(\bigwedge_{\substack{p_k \notin L_i^* \\ k \leq i}} p_k^{\epsilon_k}). \tag{23}$$

Summing over $\pm p_i$ and dividing both sides of (23) gives

$$\mathrm{w}(p_i | \bigwedge_{k<i} p_k^{\epsilon_k}) = \mathrm{w}(p_i | \bigwedge_{\substack{p_k \in L_i^* \\ k<i}} p_k^{\epsilon_k}). \tag{24}$$

Again by Separation (with $\{p_{i1}, p_{i2}, ..., p_{im_i}\}$, $L_i^* - (L_0 \cup \{p_i\})$, and the set of remaining vertices corresponding to the L_0, L_1, L_2 of that formulation) we have

$$\mathrm{w}(\bigwedge_{k=1}^{n} p_k^{\epsilon_k}).\mathrm{w}(\bigwedge_{p_k \in L_0} p_k^{\epsilon_k}) = \mathrm{w}(\bigwedge_{p_k \in L_0 \cup L_1} p_k^{\epsilon_k}) \cdot \mathrm{w}(\bigwedge_{p_k \in L_0 \cup L_2} p_k^{\epsilon_k}). \tag{25}$$

Summing over the literals $\pm p_k$ with $k > i$ or $p_k \in L_2$ gives

$$\mathrm{w}(\bigwedge_{\substack{p_k \in L_i^* \\ k \leq i}} p_k^{\epsilon_k}).\mathrm{w}(\bigwedge_{p_k \in L_0} p_k^{\epsilon_k}) = \mathrm{w}(\bigwedge_{\substack{p_k \in L_i^* \\ k<i}} p_k^{\epsilon_k}) \cdot \mathrm{w}(\bigwedge_{p_k \in L_0 \cup \{p_i\}} p_k^{\epsilon_k}), \tag{26}$$

and hence, with (24),

$$\mathrm{w}(p_i|\bigwedge_{k<i} p_k^{\epsilon_k}) = \mathrm{w}(p_i|\bigwedge_{p_k\in L_0} p_k^{\epsilon_k}), \tag{27}$$

as required.

It remains only to show (22). The proof is by induction on $|J|$. Clearly it holds if $J = \emptyset$ so suppose $|J| > 0$ and the result holds for any smaller such set. If the set J' of vertices in J which are in the same component of the graph as p_i is not all of J then by the degenerate version of Separation (and summing over the remaining literals)

$$\mathrm{w}(\bigwedge_{p_j\in L_i^*} p_j^{\epsilon_j} \wedge \bigwedge_{p_j\in J} p_j^{\epsilon_j}) = \mathrm{w}(\bigwedge_{p_j\in L_i^*} p_j^{\epsilon_j} \wedge \bigwedge_{p_j\in J'} p_j^{\epsilon_j}) \cdot \mathrm{w}(\bigwedge_{p_j\in J-J'} p_j^{\epsilon_j}), \tag{28}$$

and the required conclusion follows straightforwardly.

Otherwise let $p_k \in J$. Then there is a non-self intersecting path (not directed) from p_i to p_k and by (21) on this path there must be a vertex p_t with neighbors p_s, p_h on this path, in that order, such that $s, h < t$, say $p_h = p_{t1}$. Now by (20),

$$\mathrm{w}(\bigwedge_{j=1}^{m_t} \bigwedge_{p_k\in L[tj,t]} p_k^{\epsilon_k}) = \prod_{j=1}^{m_t} \mathrm{w}(\bigwedge_{p_k\in L[tj,t]} p_k^{\epsilon_k}),$$

from which it follows that

$$\mathrm{w}(\bigwedge_{j=1}^{m_t} \bigwedge_{p_k\in L[tj,t]} p_k^{\epsilon_k}) = \mathrm{w}(\bigwedge_{j=2}^{m_t} \bigwedge_{p_k\in L[tj,t]} p_k^{\epsilon_k}) \cdot \mathrm{w}(\bigwedge_{p_k\in L[t1,t]} p_k^{\epsilon_k}). \tag{29}$$

Hence if $J' = J \cap L[t1, t]$ then $J' \neq \emptyset$ and by summing over suitable literals in (29),

$$\mathrm{w}(\bigwedge_{p_j\in L_i^*} p_j^{\epsilon_j} \wedge \bigwedge_{p_j\in J} p_j^{\epsilon_j}) = \mathrm{w}(\bigwedge_{p_j\in L_i^*} p_j^{\epsilon_j} \wedge \bigwedge_{p_j\in J-J'} p_j^{\epsilon_j}) \cdot \mathrm{w}(\bigwedge_{p_j\in J'} p_j^{\epsilon_j}),$$

from which, by induction, the required result follows.

■

All of this has been proved for a particular maxent solution w. Unfortunately the assumption that there is always a *unique* maxent solution (made implicitly in some of the Rhodes-Garside-Holmes algorithms) turns out to be false, even in the case where the graph is simply an inverted tree, as we now show.

A counter example to the uniqueness assumption

Fix large k to be a large natural number. Consider the following inverted tree. On the top we have p_1, p_2. They have a child, p_3, who in turn is the single

parent of a child p_4, who in turn is the single parent of a child p_5, and so on down to p_{m+3} for some large (compared even with k) m.

The $w(p_1), w(p_2)$ are unknown but apart from that

$w(p_3|p_1 \wedge p_2) = w(p_3|\neg p_1 \wedge \neg p_2) = 0,$
$w(p_3|p_1 \wedge \neg p_2) = w(p_3|\neg p_1 \wedge p_2) = 1,$
$w(p_{i+1}|p_i) = k/(k+1),\ i = 3, 4, 5, ..., m+2,$
$w(p_{i+1}|\neg p_i) = 0,\ i = 3, 4, 5, ..., m+2.$

For this network (assuming the full set of independence conditions (2)) the entropy, $E(w(p_1), w(p_2))$, comes out to be

$$
\begin{aligned}
&-w(p_1)\log w(p_1) - (1-w(p_1))\log(1-w(p_1)) \\
&-w(p_2)\log w(p_2) - (1-w(p_2))\log(1-w(p_2)) \qquad (30)\\
&-(w(p_1)(1-w(p_2)) + w(p_2)(1-w(p_1)))\cdot\{ \\
&\sum_{i=4}^{m+3}\prod_{4\leq j<i} w(p_j|p_{j-1})((w(p_i|p_{i-1})\log w(p_i|p_{i-1}) + \\
&(1-w(p_i|p_{i-1}))\log(1-w(p_i|p_{i-1})))\}\,. \qquad (31)
\end{aligned}
$$

Expanding the term in (31) between the braces $\{,\}$ we see that this equals

$$
\begin{aligned}
&\sum_{i=0}^{m-1}(k/(k+1))^i((k/(k+1))\log(k/(k+1)) + (1/(k+1))\log(1/(k+1)) \\
&= ((1-(k/(k+1))^m)(1-(k/(k+1)))^{-1}((k/(k+1))\log(k) - \log(k+1)) \\
&= ((1-(k/(k+1))^m)(k\log(k) - (k+1)\log(k+1)). \qquad (32)
\end{aligned}
$$

Let $\epsilon > 0$ be small and k so large that

$$\epsilon^2 \log(k+1) > 1. \qquad (33)$$

Then from (31) and (32),

$$E(1/2, 1/2) = 2\log 2 + 1/2((1-(k/(k+1))^m)((k+1)\log(k+1) - k\log(k)) \qquad (34)$$

whilst

$$
\begin{aligned}
E(1/2+\epsilon, 1/2-\epsilon) &= -(1-2\epsilon)\log(1/2-\epsilon) - (1+2\epsilon)\log(1/2+\epsilon) \\
&\quad +((1/2+\epsilon)^2 + (1/2-\epsilon)^2)\cdot\{ \\
&\quad ((1-(k/(k+1))^m)((k+1)\log(k+1) - k\log(k))\} \\
&= 2\log 2 + O(\epsilon) + (1/2+2\epsilon^2)\cdot\{ \\
&\quad ((1-(k/(k+1))^m)((k+1)\log(k+1) - k\log(k))\}.
\end{aligned}
$$

But clearly since m is large compared with k the component involving ϵ^2 in the last term of this expression is at least

$$2\epsilon^2 \cdot (1/2) \cdot \log(k+1)$$

which by (33) is at least 1.

From this, and the fact that ϵ is chosen small, it follows that $E(1/2+\epsilon, 1/2-\epsilon)$ exceeds $E(1/2, 1/2)$ so the maxent solution in this case cannot satisfy $w(p_1) = w(p_2) = 1/2$. However there is clearly complete symmetry here between p_1 and p_2 so if the maxent solution were unique this would be the only possible value. We conclude that in this case there is no unique maxent solution.

This counter-example also deals a blow to the algorithms proposed in [3] and [5] for finding maxent solutions. In short the method is to first guess the unknowns to be 1/2 and then sequentially tune each of them in turn to the value which maximizes the entropy when all the others are fixed at their current values. The problem is that attempting that process here will simply keep you where you started (because if $w(p_2)$ is fixed at 1/2 then the best value for $w(p_1)$ remains at 1/2 etc.). In other words the algorithm will converge, but to a saddle point, not a maximum. Of course in this case an alternative starting point may lead to one of the two points of global maximum entropy. But in general it is not obvious how can we know whether we have found all, or indeed any, global maxima, and even if we have found them all what method for choosing between them can we adopt in this context of (objective) probabilities[4]?

Filling-in using limiting centre of mass

In view of the somewhat disappointing result which concluded the previous section it would seem sensible to briefly reassess the argument for using maxent to fill in missing values in the first place.

It would seem that there are three possible justifications for this choice. The first is that, as shown in [16] (see also [13]), maxent is the only inference process which is consistent with 'common sense', that is satisfies a particular set of 'common sense principles' described in these papers. However these principles, and hence that result, are uncompromisingly set within the context of *subjective probability*, where probabilities correspond to an agent's personal degrees of belief as willingness to bet and the knowledge base *sums up the totality of the agent's knowledge*. As far as causal networks are concerned however that is not at all the situation in general. Typically the data which are known are empirical frequencies and the intention in filling in missing values is to *estimate* values of

[4]See [18] for an analogous, but we would argue, essential unproblematic, problem of assigning *subjective* probabilities.

the underlying, objective, probability function. Furthermore the fact that we are dealing here with a real world probability function would seem to give us background knowledge beyond the specified knowledge base of identities (2), (3) (for more on this point see [19], [14]). In this case it seems hard to manufacture a convincing argument that an objective probability function should in any way be constrained by what we judge to be common sense in an entirely different context.

In a rather similar flavor a second 'justification' for using maxent here is that the resulting probability function has minimum Shannon information content (see [24]) amongst all those probability functions satisfying the knowledge base. In other words that this choice assumes, or goes, as little as possible beyond what is already implicit in the knowledge base. Attractive as this might initially appear it suffers the same criticisms as the first case. Namely, in this case the knowledge base goes beyond simply the constraints (2), (3), and there seems no obvious reason why the 'true' probability function should be in any way obliged to minimize its Shannon information content.

The third possible justification for using maxent here however is altogether more serious and dates back to the origins of maxent in thermodynamics, see in particular the discussion in [8] and [9]. Within the present notation the basic argument is given in [15] and goes as follows. Let us suppose that the constraints we have are actually frequencies derived from some large population, $\mathcal{P}$ say, of N individuals, or more reasonably very close to such frequencies. So, for each propositional variable of the language and each individual from $\mathcal{P}$ that individual either does or does not satisfy the property corresponding to the propositional variable. More formally then we can think of each $x \in \mathcal{P}$ as determining a $\{0,1\}$-valuation V_x on the language and the 'true' probability function $\mathrm{w}_{\mathcal{P}}$ being given by

$$\mathrm{w}_{\mathcal{P}}(\theta) = \frac{|\{\ x \in \mathcal{P} \mid V_x(\theta) = 1\ \}|}{|\mathcal{P}|}.$$

Now, of course, ostensibly all we know about $\mathcal{P}$ is that $\mathrm{w}_{\mathcal{P}}$ satisfies our knowledge base (approximately), and in general there will be many such $\mathcal{P}$. However, it turns out that if N is sufficiently large then for almost all such $\mathcal{P}$ $\mathrm{w}_{\mathcal{P}}$ is correspondingly close to the maxent solution of this knowledge base. In other words, if we accepted that the knowledge base had derived from some such $\mathcal{P}$ in this way and that all such possible $\mathcal{P}$ were equally likely then guessing the missing values to be the maxent values would almost certainly be close to the actual answer.

An immediate criticism of this argument arises once one looks a little more closely at the role of the chosen approximation. For whilst this makes little difference (within some rather loose bounds, see [15]) to the answers obtained it does, in general, make a significant difference to the distribution of the resulting data sets. For apart from some exceptional circumstances almost all the data sets

will actually give frequencies clustering tightly around points which differ from the maxent values. Tightening the approximation (for fixed, large, N) then will, in general, have the effect of simply removing the vast majority of these possible data sets. From the viewpoint then that the constraints are (presumably) being prescribed as accurately as possible this would seem to cast a question mark over how representative the overwhelming majority of the data sets actually are.

Our second criticism of using this justification for applying maxent in the context of causal networks is somewhat less formal and is aimed at what we believe to be the underlying assumption justifying 'indifference' across data sets[5], namely that it is the data set that is primary and determines the distribution.

Certainly there do seem to be situations in which the individuals involved interact together, for example the speeds of the molecules in a container of gas, or trends in fashion wear, and in these cases it may well be argued that the data determines the distribution and that it makes no sense to talk of probabilities applying to individuals in vaccuo. However these probability distributions do not seem to be particularly representative of those currently being modeled by causal networks, indeed it is hard to see that they are even amenable to such modeling. Typically in the distributions modeled by causal networks the individuals involved are influenced by factors existing independently of the particular individuals themselves. For example the ultimate fate of a ball in a pin-ball machine, or the signs and symptoms relevant to diagnosing chest pains. In these cases it surely does seem to make sense to talk of the probability of a pin-ball scoring, say 100 points, simply on the basis of the layout of the pin-ball table itself.

Of course this is a drastic simplification and arguments certainly can be advanced to blur the distinction. Nevertheless we cannot see, in the current realm of application of causal networks, that the assumption that all data sets are equally likely (as described above) is any more justified than the assumption that, say, all probability functions satisfying the known constraints are equally likely[6][7].

This alternative, of treating all *probability functions* satisfying the constraints as equally likely to be the true probability function, is clearly very much in the flavor of Bayesian methods and immediately suggests approximating, or estimating, the true probability by taking the 'average' of all probability function

[5]As a general 'principle' we have little sympathy for 'indifference' in general, unless, as in the Renaming Principle, see [12], it can be justified in terms of invariance under symmetries of the language. In neither this case, nor in the case of CM^{∞} which we shall shortly be considering, are any such supporting arguments apparent.

[6]Of course the assumption of a uniform distribution of probability functions is not the only one we could make here. (For an attempt to derive a justified 'prior' here see [19].)

[7]To quote a (private) comment by Jon Williamson to this effect, 'arguably the distribution determines the the data not vice versa; intuitively one ought to be indifferent over causes rather than effects; in which case one is better to be indifferent over the partition of distributions than the partition of data sets.'

satisfying the constraints, or more precisely the centre of mass of $V^L(K)$, where K is the knowledge base, assuming uniform mass distribution/prior. Alternatively we could think of this as the function minimizing the mean square error, see for example [12], p70. In these ways then taking the centre of mass of $V^L(K)$ could be said to be somewhat justified as an *estimate* for missing probabilities.

Unfortunately however this estimate suffers from a, to our mind serious, fault. Namely the assigned probabilities do not satisfy Language Invariance. That is if $L' \supset L$ then the values given to sentences of L by the centre of mass of $V^L(K)$ may differ from those given by the centre of mass of $V^{L'}(K)$, despite there being no differences in the underlying knowledge K. One solution to this difficulty is to acknowledge that in reality the overlying language, though finite, is indefinitely extendible and replace the values given to the sentences of L by the centre of mass of $V^L(K)$ by the limiting values given to theses sentences by $V^{L'}(K)$, for $L' \supseteq L$, as $|L'|$ tends to infinity. Fortunately this limit probability function on SL, denoted $CM^\infty(K)$ (see [12]), does exist and the corresponding inference process does satisfy Language Invariance. Indeed, as shown in [17],

$$CM^\infty(K) = \text{that } \vec{x} \in V^L(K) \text{ for which } \sum_i \log x_i \text{ is maximal.}$$

where the sum is over those i for which there is an $\vec{x} \in V^L(K)$ with $x_i > 0$[8]. Apart from its being somewhat justified in the envisaged context of practical causal networks it turns out that using CM^∞ in place of maxent avoids the problems on non uniqueness *and* is computationally trivial, just give all missing probabilities value 1/2. [Of course this is a value that one might have chosen of 'indifference' in any case, the difference is that these 1/2's are *justified* !!]

To see this let w be a solution to the constraints of types (2) and the existing constraints of type (3). Using (4) the sum of the logs of the atoms, $\sum_\alpha \log \alpha$, in this case can be expressed as

$$\begin{aligned} &\sum_{\vec{\epsilon}} \log \mathrm{w}(p_1^{\epsilon_1} \wedge p_2^{\epsilon_2} \wedge ... \wedge p_n^{\epsilon_n}) \\ &= \sum_{\vec{\epsilon}} \sum_{i=1}^{n} \log \mathrm{w}(p_i^{\epsilon_i} | p_{i1}^{\epsilon_{i1}} \wedge p_{i2}^{\epsilon_{i2}}. \wedge ... \wedge p_{im_i}^{\epsilon_{im_i}}) \end{aligned} \tag{35}$$

Now clearly the term $\mathrm{w}(p_i | p_{i1}^{\epsilon_{i1}} \wedge p_{i2}^{\epsilon_{i2}})$ only appears in this sum as

$$R\{\log \mathrm{w}(p_i | p_{i1}^{\epsilon_{i1}} \wedge p_{i2}^{\epsilon_{i2}}) + \log(1 - \mathrm{w}(p_i | p_{i1}^{\epsilon_{i1}} \wedge p_{i2}^{\epsilon_{i2}}))\}$$

[8]It is worth remarking that according to Hartley's measure, see [4], $-\log x_i$ is the information contained in the outcome α_i in this context.

for some positive constant factor R (the number of choices for the unspecified ϵ_j). Hence if the constraint

$$w(p_i|p_{i1}^{\epsilon_{i1}} \wedge p_{i2}^{\epsilon_{i2}} \wedge ... \wedge p_{im_i}^{\epsilon_{im_i}}) = b(i; \epsilon_{i1}, \epsilon_{i2}, ..., \epsilon_{im_i})$$

is missing the choice of

$$\mathrm{w}(p_i|p_{i1}^{\epsilon_{i1}} \wedge p_{i2}^{\epsilon_{i2}})$$

which maximizes (35) is independent of what other missing values are assigned and is that $x \in [0, 1]$ which maximizes

$$\log x + \log(1 - x),$$

that is $x = 1/2$.

Conclusion

In this paper we have, under the assumption of uniqueness, given direct proofs using the Separation Principle of earlier results of Rhodes, Garside and Holmes on the extent to which the maxent solution of an incomplete causal network satisfies the full set of associated independences. However we have shown that, contrary to their implicit assumption, uniqueness need not hold even for inverted trees. And indeed that their algorithms (which we do not otherwise consider in this paper) may not even converge to *a* maxent solution.

We then briefly criticized the rationale of using maxent to fill in missing values in a causal network, suggesting instead the limiting centre of mass inference process CM^{∞} which we showed does satisfy uniqueness and is computational trivial, indeed it agrees with, and serves to vindicate, the standard ad hoc choice of simply guessing 1/2 for all missing probabilities!

Acknowledgements

Clearly we would like to thank Paul Rhodes, Gerald Garside and Dawn Holmes for many invaluable discussions, papers and empirical data. We would also like to thank Jon Williamson and Alena Vencovská for their helpful comments and advice on the wider aspects and context of this paper.

References

[1] G.R. Garside, Conditional independence between siblings in a causal binary tree with two valued events, Research Report, Department of Computing, University of Bradford, 1997.

[2] G.R. Garside, Some results supporting the use of maximum entropy in causal inverted binary trees, Research Report of the Department of Computing, Bradford University, No. RS-96/CS/03 (revised version), 1998.

[3] G.R. Garside, P.C. Rhodes & D.E. Holmes, The efficient estimation of missing information in causal inverted multiway trees given incomplete information, Knowledge-Based Systems 12 (1999) 101-111.

[4] R.V.L. Hartley, Transmission of information, Bell System Technical Journal 7 (1928) 535-563. [See also http://www.cs.ucl.ac.uk/staff/S.Bhatti/D51-notes/node28.html.]

[5] D.E. Holmes, P.C. Rhodes & G.R. Garside, Efficient computation of marginal probabilities in multivalued causal inverted multiway trees given incomplete information, International Journal of Intelligent Systems 14 (6) (1998) 535-558.

[6] D.E. Holmes, Efficient estimation of missing information in multivalued singly connected networks using maximum entropy, in: W. van der Linden, V. Dose, R. Fischer & R. Preuss, eds., Maximum Entropy and Bayesian Methods (Kluwer Academic Press, Dordrecht, Netherlands, 1999) 289-300.

[7] D. Hunter, Causality and maximum entropy updating, International Journal of Approximate Reasoning 3 (1) (1989) 87-114.

[8] E.T. Jaynes, Prior Probabilities, IEEE Transactions on Systems Science and Cybernetics SSc-4 (1968) 227-241.

[9] E.T. Jaynes, Concentration of distributions at entropy maxima, in: R.D.Rosenkrantz, ed., E.T.Jaynes: Papers on Probability, Statistics and Statistical Physics (Dordrecht, D.Reidel Publishing, 1983) 317-336.

[10] S.L. Lauritzen & D.J. Spiegelhalter, Local computations with probabilities on graphical structures and their application to expert systems, Journal of the Royal Statistical Society B50 (2) (1988) 154-227.

[11] T. Lukasiewicz, Credal networks under maximum entropy, in: UAI-2000 Proc. 16th Conference on Uncertainty in AI (Morgan Kaufman 2000) 363-370.

[12] J.B. Paris, The Uncertain Reasoner's Companion: A Mathematical Perspective (Cambridge university Press 1994).

[13] J.B. Paris, Common sense and maximum entropy, Synthese 117 (1) (1999) 75-93.

[14] J.B. Paris, On the distribution of probability functions in the natural world, in: V.F.Hendricks, S.A.Pedersen & K.F.Jøgensen, eds., Probability Theory: Philosophy, Recent History and Relations to Science, Synthese Library 27 (2001) 125-145.

[15] J.B. Paris, & A. Vencovská, On the applicability of maximum entropy to inexact reasoning, International Journal of Approximate Reasoning 3 (1) (1989) 1-34.

[16] J.B. Paris, & A. Vencovská, A note on the inevitability of maximum entropy, International Journal of Approximate Reasoning 4 (3) (1990) 183-224.

[17] J.B. Paris, & A. Vencovská, A method of updating justifying minmum cross entropy, International Journal of Approximate Reasoning 7 (1) (1992) 1-18.

[18] J.B. Paris, & A. Vencovská, Common sense and stochastic independence, in: D.Corfield & J.Williamson, eds., Foundations of Bayesianism (Kluwer Academic Press, 2001) 203-240.

[19] J.B. Paris, P.N. Watton & G.M. Wilmers, On the structure of probability functions in the natural world, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 8 (3) (2000) 311-329.

[20] J. Pearl, Probabilistic Reasoning in Intelligent Systems. Networks of Plausible Inference (Morgan Kaufman, 1988).

[21] P.C. Rhodes, & G.R. Garside, Computing marginal probabilities in multiway causal trees given incomplete information, Knowledge-Based Systems 9 (1996) 315-327.

[22] M. Schramm, & B. Fronhöfer, PIT – a system for reasoning with probabilities, Technical Report 287-8/2001, FernUniversität Hagen, Germany, 1996. [See also http://www.pit-systems.de.]

[23] M. Schramm, & B. Fronhöfer, Completing incomplete bayesian networks, in: G.Kern-Isberner & W.Rödder, eds., Proceedings of the Conditionals Information Inference (CII) Workshop (FernUniversität, Hagen, Germany, 2002) 231-244.

[24] C.E. Shannon, & W. Weaver, The Mathematical Theory of Communication (University of Illinois Press, 1948).

[25] J. Williamson, Foundations for Bayesian networks, in: D.Corfield & J.Williamson, eds., Foundations of Bayesianism (Kluwer Academic Press, 2001) 75-115.

[26] J. Williamson, Maximising entropy efficiently, Electronic Transactions of Artificial Intelligence 7 (2002), www.etaij.org.

[27] J. Williamson, Bayesianism and language change, Journal of Logic, Language and Information 12 (1) (2003) 53-97.

[28] J. Williamson, Bayesian Nets and Causality: Philosophical and Computational Foundations (Oxford University Press, 2005).