Objective Bayesian Nets From Consistent Datasets

Jürgen Landes[1,a)] and Jon Williamson[1,b)]

[1] *Department of Philosophy, University of Kent, Canterbury, United Kingdom*

[a)] Corresponding author: juergen_landes@yahoo.de
[b)] j.williamson@kent.ac.uk

Abstract. This paper addresses the problem of finding a Bayesian net representation of the probability function that agrees with the distributions of multiple consistent datasets and otherwise has maximum entropy. We give a general algorithm which is significantly more efficient than the standard brute-force approach. Furthermore, we show that in a wide range of cases such a Bayesian net can be obtained without solving any optimisation problem.

INTRODUCTION

These days, several datasets involving hundreds of variables and thousands of observations are routinely collected in many applications. Different datasets tend to measure different variables, even when the datasets are collected with the same application in mind. For instance, it is common in systems medicine to have datasets measuring proteomics, transcriptomics, metabolomics, clinical data, and patient-reported outcomes, and for these datasets to have very few variables in common. How do we integrate all this data?

One approach to data integration is motivated by objective Bayesian epistemology (OBE), which holds that a rational agent ought to adopt as a representation of her degrees of belief the probability function with maximum entropy, $P^\dagger$, from all those calibrated to her evidence [1]. In this paper we shall assume that the agent's body of evidence consists of a collection of datasets and nothing else. Furthermore, we assume that the datasets are large and reliable enough that each dataset distribution provides an accurate estimate of the frequency distribution of the measured variables, and that they are consistent in the sense that these marginal frequency distributions are satisfiable by some joint probability function defined on the set V of all the variables measured by the datasets. The agent's credence function $P^\dagger$ will be defined on this larger set V of variables. OBE holds that $P^\dagger$ should be calibrated to each marginal distribution of observed frequencies, i.e., $P^\dagger$ should agree with each dataset distribution.

In general, finding the function on a convex set of probability functions which has maximum entropy is a computationally hard optimisation problem [2, Chapter 10]. In this paper we show how, in a wide range of cases, one can compute $P^\dagger$ without optimising at all, via a Bayesian net representation of $P^\dagger$. A Bayesian net representation of the credence function $P^\dagger$ which is motivated by OBE is called an *objective Bayesian net* (OBN). The goals of this paper are to show that an OBN can be constructed efficiently in typical cases and that, in a wide range of cases, an OBN can even be found without solving any optimisation problem at all.

In the following section we show how to efficiently compute an OBN in typical cases. Later, we will present algorithms which run much faster but are only applicable in particular situations.

A GENERAL ALGORITHM

Notation is aligned with [3]. Variables are denoted by (subscripted) upper case letters A, B, C, the set of all variables is $V \neq \emptyset$ and has size $n \geq 1$. Variables have arbitrary arity. An assignment of values to a set of variables $U \subseteq V$ is written as $u@U$ and the value of A under u is a^u. A probability function P maps each assignment $v@V$ to $[0, 1]$ such that $\sum_{v@V} P(v) = 1$. Let $\#V$ denote the number of states of V.

The machine learning community has developed efficient algorithms to learn a Bayesian network from a dataset [4]. Each variable in the dataset corresponds to a vertex in the net and we will use "variable" and "vertex" interchangeably. We shall take such an algorithm as given, and apply it to each dataset—i.e., use it to learn a Bayesian net $\mathcal{B}_i$ which represents the marginal frequency distribution P_i^* determined by dataset DS_i over its set V_i of variables. Each such Bayesian net consists in a directed acyclic graph (DAG) on the set V_i of vertices together with the dataset distribution of each variable conditional on its parents in the DAG. These are related by the Markov Condition, which holds that for all $A \in V_i$, A is probabilistically independent of its non-descendants conditional on its parents.

While a Bayesian network represents conditional probabilistic independencies in the dataset distribution by a DAG, it is also standard to use an undirected graph to represent such independencies. This forms the basis of a Markov net representation of the distribution. This graph can be constructed by 'marrying' the parents of any variable in the DAG by linking them with an edge and dropping the orientations of the remaining arrows in the DAG. We thus obtain an undirected graph $\mathcal{G}_i$ representing the independence structure of each dataset distribution: if Z separates X from Y in $\mathcal{G}_i$, for sets of vertices $X, Y, Z \subseteq V_i$, then X and Y are probabilistically independent conditional on Z, $X \perp\!\!\!\perp_{P_i^*} Y|Z$ for the dataset distribution P_i^*.

Construct an OBN as follows. First form an undirected graph $\mathcal{G}$ by joining the $\mathcal{G}_i$: take the variables in $V = \bigcup_i V_i$ as vertices and connect every pair of vertices that are connected in some $\mathcal{G}_i$. This graph represents the independence structure of the maximum entropy function: if Z separates X from Y in $\mathcal{G}$ then $X \perp\!\!\!\perp_{P^\dagger} Y|Z$ [3, Theorem 5.1].

Next transform this into a DAG $\mathcal{H}$ that also represents the independence structure of $P^\dagger$ in the sense that if Z D-separates X from Y in $\mathcal{H}$ then $X \perp\!\!\!\perp_{P^\dagger} Y|Z$. A standard algorithm for achieving this proceeds as follows [3, §5.7]. (i) Triangulate $\mathcal{G}$ (i.e., ensure that each simple cycle of length greater or equal than 4 possesses a chord) to give $\mathcal{G}^T$. (ii) Order the vertices of $\mathcal{G}^T$ with vertex set V according to maximum cardinality search: at each step select a vertex which is adjacent to the largest number of previously numbered vertices. (iii) Let $D_1, \ldots, D_l$ be the cliques of $\mathcal{G}^T$, ordered according to the highest labelled vertex. (iv) Let $E_j := D_j \cap (\bigcup_{i=1}^{j-1} D_i)$ and $F_j := D_j \setminus E_j$. (v) Add an arrow from each vertex in E_j to each vertex in F_j. (vi) Add further arrows to ensure there is an arrow between each pair of vertices in D_j such that the resulting directed graph $\mathcal{H}$ is acyclic. Arbitrarily break ties and arbitrarily make unconstrained choices.

Finally, determine the conditional probabilities in the OBN. These can be found by computing the probability function, from all those that agree with the dataset distributions P_i^*, that has maximum entropy. Let $V = \{A_1, \ldots, A_n\}$. Denoting by Anc_i the ancestors of A_i in $\mathcal{H}$ and $Anc_i' := \{A_i\} \cup Anc_i$, the entropy of a probability function P that satisfies the probabilistic independencies represented by $\mathcal{H}$ is

$$H(P) = -\sum_{i=1}^{n} \sum_{v @ Anc_i'} \Big(\prod_{A_j \in Anc_i'} y_j^v \Big) \log y_i^v. \tag{1}$$

Here each y_i^v is a parameter which denotes $P(a_i^v | Anc_i^v)$.

Computational Complexity. The complexity of learning the $\mathcal{B}_i$ in [5] is polynomial, as long as the maximal degree of a vertex is bounded by a polynomial. One can find a minimal triangulation in polynomial time [6]. Maximum cardinality search can be completed in linear time [6, §3]. Since $\mathcal{G}^T$ is triangulated, it has at most $|V|$-many cliques [7] which can be found in linear time [8]. Orienting all arrows is achievable in polynomial time [9].

In typical cases, $\mathcal{G}^T$ is a sparse graph, and maximising (1) is computationally much simpler than brute-force maximising of entropy $H(P) = -\sum_{v @ V} P(v) \log P(v)$ expressed by in terms of exponentially many states of V [3, p. 95]. Roughly speaking, the sparser the graph, the fewer conditional dependencies there are, and the fewer the y-parameters there are in (1). This leads to a dimension reduction in the optimisation problem. This is reduced further as follows. If a variable A_i and all its parents are measured in the same dataset, DS_j say, then for all $v @ Anc_i'$, $P_j^*(v)$ has been measured in DS_j. We can hence calculate y_i^v from $\mathcal{B}_j$. Since $P^\dagger$ and P_j^* have to agree, such parameters can thus be computed *without* solving an optimisation problem.

At times, we shall later add auxiliary edges to $\mathcal{G}$; this does not invalidate any of the relevant formal properties. At worst, adding further edges may lead to a more complex optimisation problem.

If a constraint graph is not connected, then its maximum entropy function is simply the product of the maximum entropy functions of its connected components. Hence, we shall restrict our attention to connected graphs.

TWO DATASETS

In this section we consider the case of two datasets DS_1 and DS_2, where DS_1 uses variables $A_1^1, \ldots, A_{l_1}^1, C_1, \ldots, C_k$ and DS_2 uses the variables $A_1^2, \ldots, A_{l_2}^2, C_1, \ldots, C_k$. All variables are distinct. Let $C := \{C_1, \ldots, C_k\}$; we refer to C as the *centre* and to $A^1 := \{A^1_1, \dots, A^1_{l_1}\}, A^2 := \{A^2_1, \dots, A^2_{l_1}\}$ as *appendices*. We will call a graph on V *simply-connected* if there is no edge linking different appendices. An edge is simply-connecting if and only if does not link A^1 to A^2.

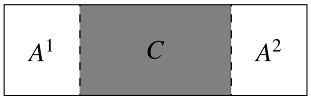

FIGURE 1: Schematic representation of variables for two datasets with non-trivial intersection of variables.

As outlined above, learn Bayesian networks $\mathcal{B}_1$ and $\mathcal{B}_2$ representing P^*_1 and P^*_2 respectively and construct the graph $\mathcal{G}$ on V. Next turn C into a clique by adding further edges and call the esulting graph $\mathcal{G}_*$. $\mathcal{G}_*$ is simply-connected.

Proposition 1. *There exists a triangulation $\mathcal{G}^T_*$ of $\mathcal{G}_*$ which is simply-connected.*

Proof: Add further simply-connecting edges to $\mathcal{G}_*$ so that the restrictions of $\mathcal{G}_*$ to $A^1 \cup C$ and $A^2 \cup C$ are triangulated. Call this new graph $\mathcal{G}^T_*$. We show that $\mathcal{G}^T_*$ is a triangulation of $\mathcal{G}_*$.

If a simple cycle in $\mathcal{G}^T_*$ of length four or greater is contained in $A^i \cup C$, then it contains a chord, since the restriction of $\mathcal{G}^T_*$ to $A^i \cup C$ is triangulated.

If a simple cycle in $\mathcal{G}^T_*$ of length four or greater contains at least one vertex in A^1 and one in A^2, then this cycle contains at least two variables in the centre C, since $\mathcal{G}'$ is simply-connected. Since the cycle is simple two of these vertices are not adjacent on the circle. C is a clique, hence there is an edge between these two vertices, i.e., the cycle possesses a chord. ■

If there was a non-simply-connecting edge in $\mathcal{G}^T_*$, then there would exist variables $A \in A^1, A' \in A^2$ connected by an edge. Every Bayesian net $\mathcal{B}$ with underlying graph $\mathcal{G}^T_*$ would have to specify $P(A|A')$ or $P(A'|A)$. No such conditional probability is given by P^*_1 or P^*_2. Proposition 1 ensures that all edges of $\mathcal{G}^T$ are simply-connecting.

Corollary 2. *There exists an OBN $\mathcal{B}$ with underlying graph $\mathcal{G}^T_*$ such that the conditional probabilities of $\mathcal{B}$ can be obtained directly from $\mathcal{B}_1$ and $\mathcal{B}_2$.*

Proof: Choose any standard enumeration of $V = A^1 \cup A^2 \cup C$ which first enumerates the k vertices in the centre C. Orient all the arrows between the centre and an appendix such that they originate from the centre and point towards the appendix.

The set of parents of a vertex in the centre is a (possibly empty) subset of C. Since $\mathcal{G}^T_*$ is simply-connected, the set of parents of a vertex in A^i is a subset of $C \cup A^i$. Hence, the conditional probability of a variable $A \in A^i$ given its parents only depends on probabilities determined by P^*_i. The centre screens off one appendix from the other.

The so-obtained Bayesian net $\mathcal{B}$ represents $P^\dagger$. Furthermore, all probabilities required to specify $\mathcal{B}$ are conditional probabilities of a variable conditional on its parents, where the variable and all its parents are used in the same dataset DS_i. These conditional probabilities can be obtained from $\mathcal{B}_i$. ■

The above constructed $\mathcal{B}$ is one possible OBN. However, there is a more efficient representation of $P^\dagger$. Prune any arrow from C_i to C_j from the DAG in $\mathcal{B}$ if the parents of C_i separate C_i from C_j in $\mathcal{G}$. In each such case the original arrow is redundant because, as noted above, separation in $\mathcal{G}$ implies conditional independence in $P^\dagger$.

Computational Complexity. Given the $\mathcal{B}_i$, computing $\mathcal{G}^T_*$ is as complex as the finding triangulations which is achievable in polynomial time. Pruning the arrows in the DAG requires at most $k = |C|$-many tests of separation in $\mathcal{G}$ and each such test can be run in polynomial time.

CENTRED DATASETS

In this section we show how the results for a collection of two consistent datasets can be generalised to a larger class of collections of datasets. A collection of $h \geq 2$ datasets is *centred* if and only if there exists a dataset DS_m such that every variable which is measured in more than one dataset is also measured in DS_m.

The variables in DS_m are $A^m_1, \dots, A^m_{l_m}$ and $C_1, \dots, C_k$. The variables in dataset DS_i with $i \neq m$ are the $A^i_1, \dots, A^i_{l_i}$, which are unique to DS_i, together with some variables from $C_1, \dots, C_k$. All variables are distinct. We extend the notions of the centre C and an appendix A^i in the obvious way. We say that a graph $\mathcal{G}$ on V is simply-connected if and only if for all pairs of linked vertices there exists a dataset which uses both these vertices.

In particular, every collection of two datasets is centred; the centre consists of the variables used in both datasets.

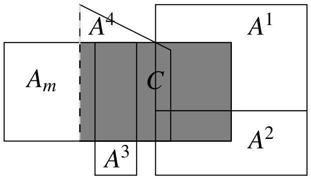

FIGURE 2: DS_m contains the centre; indicated by the dashed line. All other datasets consist of an appendix and a subset of the centre. Every variable measured in two or more datasets is contained in the centre.

An OBN can be found by following the strategy of the previous section. First, learn Bayesian networks $\mathcal{B}_i$ representing P_i^* and obtain the $\mathcal{G}_i$. Next, construct a graph $\mathcal{G}'$ on V by linking two variables if and only if they are linked in at least one of the $\mathcal{G}_i$. Then, add further edges to turn C into a clique to obtain a graph $\mathcal{G}_*$. Proposition 1 holds for the case of a centred collection of N datasets, too. There exists a triangulation $\mathcal{G}_*^T$ of $\mathcal{G}_*$ which is simply-connected.

Corollary 3. *There exists an OBN $\mathcal{B}$ with underlying graph $\mathcal{G}_*^T$ where the conditional probabilities of $\mathcal{B}$ can be obtained from the $\mathcal{B}_i$.*

Proof: Again, we first enumerate the centre and then the vertices in the appendices and ensure that all directed edges between the centre and an appendix originate from the centre. The parents of a vertex in the centre are hence all in the centre and the conditional probability of such a variable given its parents can be obtained from $\mathcal{B}_m$.

Since $\mathcal{G}_*^T$ is simply-connected, the set of parents of a vertex in some appendix A_i is a subset of the variables in dataset DS_i. Hence, the conditional probability of this variable given its parents can be obtained from $\mathcal{B}_i$. ■

As in the previous section, we can drop redundant edges from $\mathcal{B}$ and obtain a sparser OBN.

TRIANGLES

We shall now study, in some detail, a simple case with only binary variables in which we cannot obtain $P^\dagger$ directly from the $\mathcal{B}_i$. Nevertheless, we shall see that one can still obtain $P^\dagger$ without having to solve an optimisation problem. The computational task can be reduced to finding a particular root of a polynomial of degree three. The analysis we give of this simple case turns out to be crucial in more complex situations.

Suppose datasets DS_1, DS_2, DS_3 measure sets of variables $\{A_1, A_2\}$, $\{A_1, A_3\}$ and $\{A_2, A_3\}$, respectively, that these variables are all binary, with each A_i taking $a_i, \neg a_i$ as possible values, and that each $\mathcal{B}_i$ contains an arrow. Hence, there are edges between all three variables in the graph $\mathcal{G}$. Clearly, this collection of datasets is not centred.

Every acyclic orientation of the edges of $\mathcal{G}$ makes one vertex a child of the other two vertices. Two or more edges pointing towards A_3, A_3 is said to be a *collider*, as in Figure 3. Now enumerate the eight assignments $v@\{A_1, A_2, A_3\}$

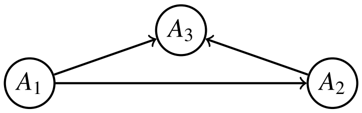

FIGURE 3: Collider at A_3

as follows: assignment v_1 is $a_1a_2a_3$, v_2 is $a_1a_2\neg a_3$, and so on. Since $P^\dagger$ has to match the marginal distributions P_i^*, where they are defined, $P^\dagger$ has to satisfy the following 12 linear constraints and no further constraints:

$$\begin{array}{lll}
P^\dagger(v_1) + P^\dagger(v_2) = P_1^*(a_1a_2) =: a & P^\dagger(v_3) + P^\dagger(v_4) = P_1^*(a_1\neg a_2) =: b & P^\dagger(v_5) + P^\dagger(v_6) = P_1^*(\neg a_1a_2) =: c \\
P^\dagger(v_7) + P^\dagger(v_8) = P_1^*(\neg a_1\neg a_2) =: d & P^\dagger(v_1) + P^\dagger(v_3) = P_2^*(a_1a_3) =: e & P^\dagger(v_2) + P^\dagger(v_4) = P_2^*(a_1\neg a_3) \\
P^\dagger(v_5) + P^\dagger(v_7) = P_2^*(\neg a_1a_3) =: f & P^\dagger(v_6) + P^\dagger(v_8) = P_2^*(\neg a_1\neg a_3) & P^\dagger(v_1) + P^\dagger(v_5) = P_3^*(a_2a_3) =: g \\
P^\dagger(v_2) + P^\dagger(v_6) = P_3^*(a_2\neg a_3) & P^\dagger(v_3) + P^\dagger(v_7) = P_3^*(\neg a_2a_3) & P^\dagger(v_4) + P^\dagger(v_8) = P_3^*(\neg a_2\neg a_3).
\end{array}$$

After some linear algebra, we find that this set of equations is equivalent to the following set of systems:

$$\begin{array}{rlrlrl} \mathrm{i}: & P^\dagger(v_1) - P^\dagger(v_7) = g - f & \mathrm{ii}: & P^\dagger(v_2) + P^\dagger(v_7) = a + f - g & & \\ \mathrm{iii}: & P^\dagger(v_3) + P^\dagger(v_7) = e + f - g & \mathrm{iv}: & P^\dagger(v_4) - P^\dagger(v_7) = b - e - f + g & & \\ \mathrm{v}: & P^\dagger(v_5) + P^\dagger(v_7) = f & \mathrm{vi}: & P^\dagger(v_6) - P^\dagger(v_7) = c - f & \mathrm{vii}: & P^\dagger(v_8) + P^\dagger(v_7) = d. \end{array} \tag{2}$$

Seven constraints apply to eight unknowns. Hence, there is one degree of freedom. In particular, there is more than one function consistent with all the P_i^*, in general.

The probability functions on A_1, A_2, A_3 consistent with all the P_i^* lie on a line segment. The entropy of a probability function on this line segment can now be expressed in terms of one unknown, $x := P^\dagger(v_7)$. The $P^\dagger(v_i)$ become dependent on this unknown x by applications of (2). Since Shannon entropy, defined on the set of probability functions, is a strictly concave function, we can use the first derivative to find the unique maximum entropy function along this line segment. With $\varphi_1 := -a - f + g$, $\varphi_2 := -e - f + g$ and $\varphi_3 := -f$ we obtain

$$\frac{d}{dx}H(x) = \frac{d}{dx}\sum_{i=1}^{n} -P(v_i)\log P(v_i) = \log\Big(\frac{x+\varphi_1}{x+\varphi_1+a}\ \frac{x+\varphi_2}{x+\varphi_2+b}\ \frac{x+\varphi_3}{x+\varphi_3+c}\ \frac{x-d}{x}\Big). \tag{3}$$

We thus have to find the unique value $x^* \in [0,1]$ at which $\frac{dH}{dx}|_{x=x^*} = 0$ and which corresponds to a probability function. Substituting $d = 1 - a - b - c$ we find after a lengthy but uneventful calculation that the entropy is maximal if and only if the following polynomial $\mathcal{P}(x)$ has a real root of odd degree and re-substituting $P^\dagger(v_7)$ for x results in a probability function which solves (2):

$$\mathcal{P}(x) = x^3 + \beta x^2 + \gamma x + \delta \qquad \text{where} \qquad \beta := ab + ac + bc + \varphi_1 + \varphi_2 + \varphi_3 - \varphi_1 a - \varphi_2 b - \varphi_3 c$$
$$\gamma := -abc + \varphi_1 bc + \varphi_2 ac + \varphi_3 ab + (1-a-b)\varphi_1\varphi_2 + (1-a-c)\varphi_1\varphi_3 + (1-b-c)\varphi_2\varphi_3 \quad \delta := (1-a-b-c)\varphi_1\varphi_2\varphi_3.$$

We are looking for the unique maximum of $H(x)$ and can hence ignore double roots of $\mathcal{P}(x)$. The discriminant Δ of the polynomial $\mathcal{P}(x)$ and the auxiliary values p, q are defined as usual

$$p := \frac{3\gamma - \beta^2}{3} \qquad q := \frac{2\beta^3 - 9\beta\gamma + 27\delta}{27} \qquad \Delta := \frac{27\delta^2 + 4\beta^3\delta - 18\beta\gamma\delta + 4\gamma^3 - \beta^2\gamma^2}{108} \qquad \Delta = (q/2)^2 + (p/3)^3.$$

If $\underline{\Delta > 0}$, then we find x^* by only considering the real third roots in

$$x^* = \sqrt[3]{-\frac{q}{2} + \sqrt{\Delta}} + \sqrt[3]{-\frac{q}{2} - \sqrt{\Delta}} - \frac{\beta}{3}.$$

If $\underline{\Delta = 0}$, then the roots of $\mathcal{P}(x)$ are

$$x^* = -\frac{\beta}{3} \quad \text{if } \Delta = p = q = 0 \qquad\qquad x^* = \frac{\beta^3 - 4\beta\gamma + 9\delta}{3\gamma - \beta^2} \quad \text{if } \Delta = 0 \text{ and } p^2 + q^2 > 0$$

If $\underline{\Delta < 0}$, then x^* is one of the following three values

$$-\sqrt{-\frac{4p}{3}} \cdot \cos\Big(\frac{\pi}{3} + \frac{1}{3}\arccos\Big(-\frac{q}{2} \cdot \sqrt{-\frac{27}{p^3}}\Big)\Big) - \frac{\beta}{3}; \qquad \sqrt{-\frac{4p}{3}} \cdot \cos\Big(\frac{1}{3}\arccos\Big(-\frac{q}{2} \cdot \sqrt{-\frac{27}{p^3}}\Big)\Big) - \frac{\beta}{3};$$
$$-\sqrt{-\frac{4p}{3}} \cdot \cos\Big(-\frac{\pi}{3} + \frac{1}{3}\arccos\Big(-\frac{q}{2} \cdot \sqrt{-\frac{27}{p^3}}\Big)\Big) - \frac{\beta}{3}.$$

If it were always the case that $\Delta \geq 0$, then there would be some hope to intuitively interpret the maximum entropy function $P^\dagger$. However, we have found examples in which $\Delta < 0, \Delta = 0$ and $\Delta > 0$, respectively, and we do not have much intuition about the roots involving arccos.

$P^\dagger$ can now quickly be computed. Simply check which of the six possible values for x^* is a root of $\mathcal{P}(x)$ and gives rise to a probability function which solves (2). In particular, we find $P^\dagger$ without solving an optimisation problem.

Computational Complexity. β, γ, δ can be computed by performing simple arithmetic on rational numbers, the $\mathcal{B}_i$ do not use non-rational conditional probabilities. In general, computing the six possible values of x^* requires the use of non-rational real numbers. In practise, only a small number of digits is required to decide which possible value is the root x^*. The overall computational complexity is then low in terms of the desired precision.

COLLIDERS

We now study a more general case with N datasets. First, construct a graph $\mathcal{G}$ representing the independence structure of $P^\dagger$, as we did for the general algorithm. Our aim is to apply the algorithm for maximising entropy over a triangle. We hence restrict our attention to triangulated and connected $\mathcal{G}$. We make one further restriction: whenever there is an edge between two variables A, B in $\mathcal{G}$, then there exists a dataset DS_i which measures A and B. This condition guarantees that for all such A, B $P_i^*(A|B)$ is defined. By consistency, if there are two or more such datasets then these marginalised distributions have to agree. For the remainder of this section fix such a connected, undirected and triangulated graph $\mathcal{G}$.

We now characterise a rich class of such graphs for which an OBN can be found by independently maximising entropy over multiple triangles. As we saw above, every such problem can be solved by checking which of six values is a root of a polynomial of degree three and gives rise to a probability function solving (2).

An orientation of the edges of $\mathcal{G}$ is consistent with the standard enumeration if we can apply the edge-orientation algorithm from [3, §5.7] to yield this orientation. $\mathcal{G}$ is called *cc* if and only if there exists an acyclic orientation which is consistent with standard enumeration such that every collider and the set of its parents form a triangle and no collider is a parent. This DAG is called $\mathcal{G}^{\rightarrow}$.

Theorem 4. *$\mathcal{G}$ is cc if and only if every triangle in $\mathcal{G}$ contains a vertex of degree two.*

Proof: $\mathcal{G}$ is cc. Assume for contradiction that there exists a triangle in which no vertex has degree two. Since all three vertices are a member of this triangle their degree has to be at least two. So, let us assume that there exists some triangle in which all vertices have a degree of 3 or greater.

Since the orientation is acyclic, there has to exist a vertex A in this triangle to which two edges of this triangle point. Since the degree of A is at least three there has to be another edge incident on A. This third edge cannot point to A, because A has exactly two parents. Since A is a childless collider, this third edge cannot originate from A. Contradiction.

Every triangle contains a vertex of degree two. In every triangle pick a vertex which has degree two. Remove all these vertices from $\mathcal{G}$ and obtain $\mathcal{G}'$.

For all $A, B \in \mathcal{G}'$ there exists a path from A to B in $\mathcal{G}$, since $\mathcal{G}$ is connected. If this path goes via some $C \in \mathcal{G} \setminus \mathcal{G}'$, then it also goes through two other vertices $D, E \in \mathcal{G}$ which, together with C, form a triangle. Neither D nor E can be part of another triangle and have degree two. Hence, $D, E \in \mathcal{G}'$. C, D, E form a triangle in $\mathcal{G}$, so there is an edge between D and E. So, there exists a path in $\mathcal{G}$ between A and B which does not contain $C \in \mathcal{G} \setminus \mathcal{G}'$. Eventually, we find a path between A and B in $\mathcal{G}'$. $\mathcal{G}'$ is connected.

The next step is to show that $\mathcal{G}'$ is a tree. Suppose for contradiction $\mathcal{G}'$ is not a tree. Then $\mathcal{G}'$ contains a simple cycle c of at least three vertices. We consider three cases.

If c consists of three vertices, then c is a triangle. c is also a triangle in $\mathcal{G}$. However, we removed a vertex from every triangle in $\mathcal{G}$ to obtain $\mathcal{G}'$. Contradiction.

If c consists of four vertices, then c is a cycle in $\mathcal{G}$ of length four. Furthermore, since $\mathcal{G}$ is triangulated, we have removed at least one edge linking members of c. Had we removed two edges, then the complete graph on four vertices, K_4, would be in $\mathcal{G}$. This contradicts that every triangle in $\mathcal{G}$ contains a vertex of degree two. Had we removed one edge, then the diamond graph $\mathcal{D}$ would be in $\mathcal{G}$. The diamond graph $\mathcal{D}$ is a graph on four vertices with five edges. But then we would have removed the two vertices in $\mathcal{D}$ with degree two in $\mathcal{G}$. The resulting subgraph in $\mathcal{G}'$ is then two vertices connected by an edge and not a circle of length four. Contradiction.

If c consists of five or more vertices, then c is a cycle in $\mathcal{G}$ of length five or greater. Since $\mathcal{G}$ is triangulated, three vertices of c from a triangle in $\mathcal{G}$ in which every vertex has degree three or greater. Contradiction.

Now choose an orientation on $\mathcal{G}$. First, pick a root vertex in $\mathcal{G}'$. Next, orientate all edges of $\mathcal{G}'$ such that there exists a unique directed path from the root to all other vertices of $\mathcal{G}'$. Finally, orientate the remaining edges of $\mathcal{G}$ such that they collide at the vertices in $\mathcal{G} \setminus \mathcal{G}'$. This orientation is acyclic and consistent with a standard enumeration. ■

An example of a cc graph $\mathcal{G}^{\rightarrow}$ (left in Figure 4) and examples non-cc graphs are on the right in Figure 4. The following

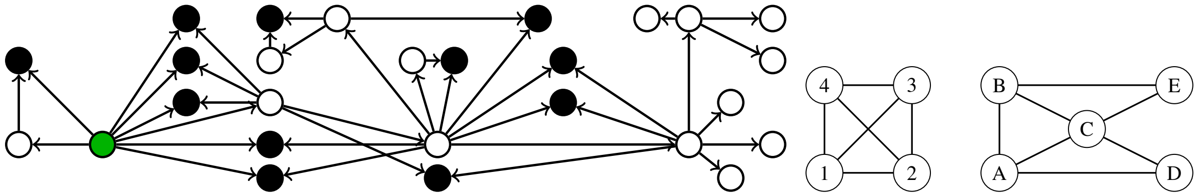

FIGURE 4: Left: Example of a complex graph $\mathcal{G}^{\rightarrow}$ which is cc, root vertex colored green, colliders are black, and all other vertices white. Middle: K_4 which is not cc. Right: Simplest graph which is not cc and not complete.

corollary follows an immediate consequence. The inclusion relation is taken with respect to the edge relation while keeping the set of vertices fixed.

Corollary 5. *If $\mathcal{G}$ is cc, then so is every subgraph $\mathcal{G}' \subset \mathcal{G}$. If $\mathcal{G}$ is non-cc, then so is every supergraph $\mathcal{G}' \supset \mathcal{G}$.*

Proposition 6. *If $\mathcal{G}$ is cc, then there exists a unique directed path in $\mathcal{G}^{\rightarrow}$ from the root to all other non-collider vertices.*

Proof: Since $\mathcal{G}^{\rightarrow}$ is connected, every vertex not contained in the clique D_1 has a parent. Furthermore, all arrows between a vertex $A \in D_1$ and a vertex $B \notin D_1$ point to B. Since $\mathcal{G}^{\rightarrow}$ is acyclic, there exists a unique vertex A_1 in $\mathcal{G}^{\rightarrow}$ which is in D_1 which does not have a parent.

Every non-collider vertex $A_i \neq A_1$ has at least one parent. Since A_i is not a collider it has to have a unique parent A_i'. Tracing back the ancestors of A_i we eventually arrive at A_1. It follows that A_1 is the unique root of $\mathcal{H}$.

If there were multiple directed paths from A_1 to some other non-collider vertex A', then there would have to be a collider on this path, possibly A'. A' is not a collider by assumption. If some other vertex on these paths is a collider, then it has to have a child. Contradiction. ■

Proposition 7. *If $\mathcal{G}$ is cc and if there exists a directed path from A_{i_1} to A_{i_l} in $\mathcal{G}^{\rightarrow}$, then for all $v@\{A_{i_1}, \dots, A_{i_l}\}$*

$$P^{\dagger}(a^v_{i_1} \dots a^v_{i_l}) = P^*(a^v_{i_1}) \prod_{j=2}^{l} P^*(a^v_{i_j} | a^v_{i_{j-1}}). \tag{4}$$

Proof: There is no collider on the direct path $A_{i_1}, \dots, A_{i_l}$, with the possible exception of the endpoint, since all other vertices on the path have children. So, all vertices on this directed path, with the possible exception of the endpoint and the root, have exactly one parent. All conditional probabilities of the form $P(a^v_{i_j} | a^v_{i_{j-1}})$ are given by some P^*_i. Since $P^{\dagger}$ has to agree with the P^*_i and all vertices on the path have a single parent we obtain for all $v@\{A_{i_1}, \dots, A_{i_n}\}$, using the chain rule and (1), $P^{\dagger}(a^v_{i_1} \dots a^v_{i_n}) = P^*(a^v_{i_1}) \prod_{j=2}^{n} P^*(a^v_{i_j} | a^v_{i_{j-1}})$. ■

Theorem 8. *If $\mathcal{G}$ is cc, then the problem of computing an OBN reduces to the problem of computing the entropy maximisers of the triangles in $\mathcal{G}$ independently.*

Proof: Let Co denote the set of colliders in $\mathcal{G}^{\rightarrow}$. Using (1), we can now write the entropy $H(P)$ of a probability function P which is consistent with $\mathcal{G}^{\rightarrow}$ as follows

$$-\sum_{v@A_1} y^v_1 \log(y^v_1) - \sum_{\substack{2\le i\le N\\ A_i\notin Co}} \sum_{v@Anc'_i} y^v_i \log(y^v_i) \prod_{\substack{1\le j\le i-1\\ A_j\in Anc_i}} y^v_j - \sum_{\substack{3\le i\le N\\ A_i\in Co}} \sum_{v@Anc'_i} u^v_i \log(u^v_i) \cdot \prod_{\substack{1\le j\le i-1\\ A_j\in Anc_i}} y^v_j. \tag{5}$$

The terms $\prod_{\substack{1\le j\le i-1\\ A_j\in Anc_i}} y^v_j$ are given by (4). The only undetermined values in (5) are those u^v_i, for which there does not exist a dataset which contains the collider A_i and both parents of A_i. Attention is now restricted to such colliders.

Note that since no collider has a child, it follows that no two collider vertices are linked by an edge in $\mathcal{G}$. Every pair of colliders is thus separated by the parents of either one of the colliders in $\mathcal{G}^{\rightarrow}$, and hence (c.f., our explanations to the general algorithm) the colliders are conditionally independent given the parents of one of the colliders. Maximising

(5) amounts to independently maximising $\sum_{v@Anc'_i} u_i^v \log(u_i^v) \cdot \prod_{\substack{1\leq j\leq i-1 \\ A_j \in Anc_i}} y_j^v$ for all $i \geq 3$ with u_i^v subject to the constraints applying to A_i.

For all colliders A_c there is an edge between the parents A_a, A_b, since $\mathcal{G}$ is cc. W.l.o.g, this edge points from A_a to A_b. Note that A_b cannot have any further parents, since it would otherwise be a collider which has a child, A_c.

Since A_a is not a collider, there exists a unique path from A_1 to A_a (Proposition 6). The graph structure is depicted in the top of Figure 5. Since we are considering this optimisation problem in isolation, we can choose any standard enumeration which suits us best. We hence re-enumerate as depicted in Figure 5. For this re-enumeration the optimisation problem is to maximise

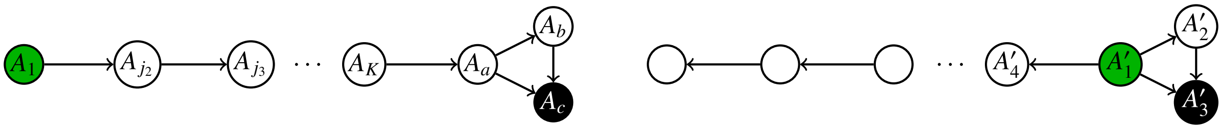

FIGURE 5: Graph with original enumeration and re-enumeration.

$$-\sum_{v@\{A'_1\}} z_1^v \log(z_1^v) - \sum_{v@\{A'_1,A'_2\}} z_1^v z_2^v \log z_2^v - \sum_{v@\{A'_1,A'_2,A'_3\}} z_1^v z_2^v z_3^v \log z_3^v \quad - \sum_{1\leq l\leq K} \sum_{v@\{A'_1,A'_4,\ldots,A'_{3+l}\}} z_1^v z_{3+l}^v \log(z_{3+l}^v) \prod_{m=1}^{l-1} z_{3+m}^v. \quad (6)$$

The only unknowns in this equation are the z_3^v. We are thus left with the problem of maximising entropy of a triangle. ■

Computational Complexity. Computing an orientation consistent with the standard enumeration is achievable in polynomial time [9]. There are at most $|V| - 2$-many triangles in a cc graph $\mathcal{G}$. An OBN is determined by computing the maximum entropy over all triangles separately. Computing the maximum entropy of a single triangle is fast in the desired precision.

CONCLUSIONS AND FUTURE RESEARCH

We have explored the question of how to determine an OBN from consistent datasets. For centred collections of datasets (in particular for any pair of datasets) we showed how to obtain an OBN given Bayesian nets representing the P_i^* without solving an optimisation problem. In these cases, all one needs to do to obtain an OBN is to perform triangulations and calculate some conditional probabilities from the $\mathcal{B}_i$.

For binary variables, we first showed how to find an OBN for a triangle. Again, our approach does not require solving an optimisation problem. All one needs is to do is some algebra to calculate six values and check which of them is a root of polynomial of degree three and corresponds to a probability function solving (2). This approach turned out to be all that is required to find an OBN in case $\mathcal{G}$ is cc.

Three main avenues for further research are: i) to identify further cases in which an OBN can be computed efficiently, ii) to extend the methodology to include inconsistent datasets and iii) to implement algorithms on a computer to demonstrate their feasibility and correctness in practice.

REFERENCES

[1] J. Williamson, *In defence of objective Bayesianism* (Oxford University Press, Oxford, 2010).
[2] J. B. Paris, *The uncertain reasoner's companion* (Cambridge University Press, Cambridge, 1994).
[3] J. Williamson, *Bayesian Nets and Causality* (Oxford University Press, 2005).
[4] R. E. Neapolitan, *Learning Bayesian Networks* (Pearson, 2003).
[5] I. Tsamardinos, L. Brown, and C. F. Aliferis, Machine Learning **65**, 31–78 (2006).
[6] A. Berry, J. R. S. Blair, P. Heggernes, and B. W. Peyton, Algorithmica **39**, 287–298 (2004).
[7] D. R. Fulkerson and O. Gross, Pacific Journal of Mathematics **15**, 835–855 (1965).
[8] A. Berry and R. Pogorelcnik, Information Processing Letters **111**, 508–511 (2011).
[9] D. Dor and M. Tarsi, "A simple algorithm to construct a consistent extension of a partially oriented graph," Tech. Rep. (UCLA, 1992).