Philipp Moritz; Jörg Reichardt; Nihat Ay

Discriminating between causal structures in Bayesian Networks given partial observations

Kybernetika, Vol. 50 (2014), No. 2, 284–295

Persistent URL: http://dml.cz/dmlcz/143794

KYBERNETIKA — VOLUME *50* (2014), NUMBER 2, PAGES 284–295

DISCRIMINATING BETWEEN CAUSAL STRUCTURES IN BAYESIAN NETWORKS GIVEN PARTIAL OBSERVATIONS

Philipp Moritz, Jörg Reichardt and Nihat Ay

Given a fixed dependency graph G that describes a Bayesian network of binary variables $X_1, \dots, X_n$, our main result is a tight bound on the mutual information $I_c(Y_1, \dots, Y_k) = \sum_{j=1}^{k} H(Y_j)/c - H(Y_1, \dots, Y_k)$ of an observed subset $Y_1, \dots, Y_k$ of the variables $X_1, \dots, X_n$. Our bound depends on certain quantities that can be computed from the connective structure of the nodes in G. Thus it allows to discriminate between different dependency graphs for a probability distribution, as we show from numerical experiments.

Keywords: Bayesian networks, causal Markov condition, information theory, information inequalities, common ancestors, causal inference

Classification: 60A08, 62B09

1. INTRODUCTION

Since Judea Pearl published his theory of causality [9], much progress has been made in applying and extending this framework. In its core, his theory is about inference and reasoning about causal structure specified by directed graphical models [7, 13]. The framework has for example been applied in

1. the study of genetic data from pedigrees, where causal relations are given by the inheritance structure [8] and
2. model-based approaches for inferring cellular networks from DNA microarray experiments [5].

One important concept he introduced is the *do-calculus*, which is a way to describe interventional experiments mathematically. Even if intervention is not possible, the causal graph of a distribution can sometimes be determined under additional model assumptions like additive noise [6]. These assumptions provide information beyond the independence structure of joint random variables and thus sometimes allow to determine causes and effects.

Following [12], we ask which assertions about the structure of possible causal graphs can be made if we have no additional information beyond the joint probability distribution of the observed variables. Given a system consisting of observable quantities $X_1, \dots, X_n$, a scientist may construct the causal model G of these observables by

DOI: 10.14736/kyb-2014-2-0284

systematic intervention [9]. If this is too expensive, experimentally not tractable or ethically questionable, observation alone must be employed to learn about G. Typically, only a subset $Y_1, \ldots, Y_k$ of the variables $X_1, \ldots, X_n$ can be observed. From these observations, it is possible to gain information about G, for example by *Reichenbach's common cause principle*: If two of the observed nodes are dependent, they must have a common ancestor in G. A quantitative version of this principle [2, 12] allows to infer certain aspects of the causal structure using information theoretic quantities. We extend this line of work by deriving a tight upper bound on the quantity $I_c(Y_1, \ldots, Y_k) = \sum_{j=1}^{k} H(Y_j)/c - H(Y_1, \ldots, Y_k)$, which is a way to quantify the mutual information of the observed variables $Y_1, \ldots, Y_k$. This bound, which is our main contribution, depends on the connective structure of the nodes in G to the root nodes of G and can thus be used to discriminate between causal models. The result is proved by inductively clustering the observed nodes in G by their root nodes and then applying the d-separation criterion to these sets. Compared with other constraints of probability distributions that arise from a given graphical model (like the implicitization approach for phylogenetic trees [1]) we only extract one scalar quantity I_c from the probability distribution. This limits the discriminating power of our method but allows for efficient computation.

The paper is organized as follows: In Section 2 we introduce the definitions used in the paper, in Section 3 we summarize existing work on the inference of common ancestors. Section 4 is the main part of the paper, the derivation of our bound on the mutual information I_c for fixed graph structure. In Section 5 we show how this bound can be used to discriminate between causal models.

2. DESCRIBING SYSTEMS WITH BAYESIAN NETWORKS

In this section, we define the terminology relevant for the rest of the paper: Discrete random variables, entropy and conditional independence. Conditional independence can conveniently be encoded in directed acyclic graphs using the d-separation criterion that we will describe.

Random variables. Throughout the paper we deal with a finite set of binary random variables denoted by upper-case roman characters such as X, Y, Z. Their values are denoted by lower-case roman characters, e. g. $X = x$ where $x \in \{0, 1\}$. The tuple of variables with indices from an index set $N = \{1, \ldots, n\}$ will be denoted by X_N. The probability distribution defined for variables from X_N will be denoted by $P(X_N)$, its marginal distribution for variables from X_A, $A \subseteq N$ will be denoted by $P(X_A)$. For A, B disjoint and nonempty, $P(X_A \mid X_B)$ will denote the conditional distribution of X_A given X_B. To simplify the reading, we write $p(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n)$.

Entropy. The *joint entropy* of $X_1, \ldots, X_n$ is defined as

$$H(X_1, \ldots, X_n) = -\sum_{x_1 \in \{0,1\}} \cdots \sum_{x_n \in \{0,1\}} p(x_1, \ldots, x_n) \log p(x_1, \ldots, x_n).$$

In the case $n = 1$ this reduces to the entropy of a single random variable. We have $H(X, Y) \leq H(X) + H(Y)$ with equality if X and Y are independent.

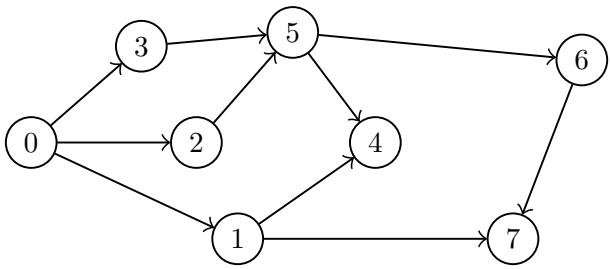


Fig. 1. For this example graph G, $\mathrm{roots}(G) = \{0\}$, the parents of node 5 are $\mathrm{pa}(5) = \{2, 3\}$ and its descendants are $\mathrm{de}(5) = \{4, 5, 6, 7\}$.

Conditional Independence. Let N be an index set and $A, B, C \subseteq N$ be nonempty. We then say X_A is conditionally independent of X_B given X_C, written as $X_A \perp\!\!\!\perp X_B \mid X_C$, if $P(X_A = x_A \mid X_B = x_B, X_C = x_C) = P(X_A = x_A \mid X_C = x_C)$ for all possible values x_A, x_B, x_C. This is succinctly written as $p(x_A \mid x_B, x_C) = p(x_A \mid x_C)$ if there is no danger of confusion.

Directed acyclic graphs. A *directed graph* is a tuple $G = (V, E)$ consisting of *nodes* V and *edges* $E \subseteq V \times V$. An edge $(u, v) \in E$ is interpreted as a directed connection between the nodes u and v, we write $u \to v$ if $(u, v) \in E$. A *directed path* between two nodes v_1 and v_n is a sequence $v_1, v_2, \ldots, v_n$ of distinct nodes v_j with $v_j \to v_{j+1}$ for $1 \leq j < n$. We write $v_1 \rightsquigarrow v_n$ if there exists a directed path from v_1 to v_n. We also admit paths of length 0, so $v \rightsquigarrow v$ for all $v \in V$. An *undirected path* between v_1 and v_n is a sequence $v_1, v_2, \ldots, v_n$ of distinct nodes v_j with $v_j \to v_{j+1}$ or $v_j \leftarrow v_{j+1}$ for $1 \leq j < n$. We call G *acyclic*, if no path $v \rightsquigarrow v$ is of length > 0. In addition we introduce the sets of

- *parents* $\mathrm{pa}(v) = \{u \in V : (u, v) \in E\}$ of $v \in V$,
- *root nodes* $\mathrm{roots}(G) = \{v \in V : \mathrm{pa}(v) = \varnothing\}$ of G (nodes without parents),
- *descendants* $\mathrm{de}(u) = \{v \in V : u \rightsquigarrow v\}$ of $u \in V$ and
- *non-descendants* $\mathrm{nd}(u) = V \setminus \mathrm{de}(u)$ of $u \in V$, as well as the
- *ancestral set* $\mathrm{an}(v) = \{u \in V : u \rightsquigarrow v\}$ of $v \in V$.

These concepts are illustrated in Figure 1.

The concept of d-separation. Let $G = (V, E)$ be a directed acyclic graph. An undirected path γ in G is *d-separated* by a set of nodes $C \subseteq V$ if and only if

- γ contains a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$ such that the middle node m is in C, or
- γ contains a collider $i \to m \leftarrow j$ such that the middle node m is not in C and such that no descendant of m is in C.

The set C d-separates [9] sets of nodes $A \subseteq V$ and $B \subseteq V$ if and only if every undirected path between a node in A and a node in B is d-separated by C.

Bayesian networks. Now we show how conditional independence of random variables can be described using graph-theoretic concepts. Consider a graph with vertices $V = \{1, \ldots, n\}$ and edges $E \subseteq V \times V$. We say that the joint distribution of the variables $X_V = (X_v : v \in V)$ for $V = \{1, \ldots, n\}$ *factorizes according to the directed acyclic graph* $G = (V, E)$ if

$$p(x_V) = \prod_{v=1}^{n} p(x_v \mid x_{\mathrm{pa}(v)}) \tag{1}$$

for all possible combinations of values. Equivalent to this is the so called *local Markov condition* which postulates that $X_v \perp\!\!\!\perp X_{\mathrm{nd}(v)} \mid X_{\mathrm{pa}(v)}$ for all $v \in V$. Another still equivalent condition is the *global Markov condition* which postulates that for disjoint $A, B, C \subseteq V$, X_A is independent of X_B given X_C whenever C d-separates A and B.

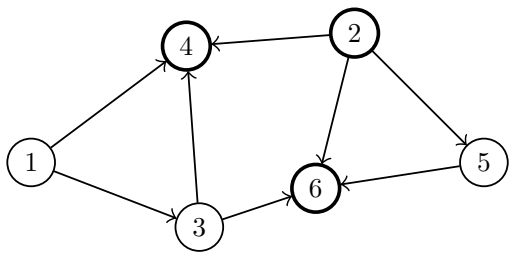


Fig. 2. Example of a system with $n = 6$ nodes, of which the $k = 3$ nodes $Y_1 = X_2$, $Y_2 = X_4$ and $Y_3 = X_6$ are observed.

Partially observed systems. Let $X_1, \ldots, X_n$ be random variables that factorize according to the directed acyclic graph G. A subset $Y_1, \ldots, Y_k$ of these variables is *observed*. This is defined as follows: Write $Y_1 = X_{\pi(1)}, \ldots, Y_k = X_{\pi(k)}$ for an injective function $\pi : \{1, \ldots, k\} \to \{1, \ldots, n\}$, then

$$P(Y_1 = y_1, \ldots, Y_k = y_k) = \sum_{(x_1, \ldots, x_n) \in A_y} P(X_1 = x_1, \ldots, X_n = x_n)$$

where $A_y = \{(x_1, \ldots, x_n) \in \{0,1\}^n : x_{\pi(1)} = y_1, \ldots, x_{\pi(k)} = y_k\}$. These definitions are illustrated in Figure 2, in this example we would have

$$p(x_2, x_4, x_6) = \sum_{x_1 \in \{0,1\}} \sum_{x_3 \in \{0,1\}} \sum_{x_5 \in \{0,1\}} p(x_1, x_2, x_3, x_4, x_5, x_6).$$

A *common cause* or *common ancestor* of the observed nodes $Y_1, \ldots, Y_k$ is a node X_j with

$$X_j \in \bigcap_{1 \le i \le k} \mathrm{an}(Y_i).$$

In the example of Figure 2, X_2 is a common ancestor of X_2, X_4 and X_6.

3. INFERENCE OF COMMON ANCESTORS

In this section we briefly summarize the existing common cause principles and explain how they can be used to discriminate between partially observed Bayesian networks.

Reichenbach's principle of common cause. Reichenbach formulates this most elementary common cause principle in [10]: *"If an improbable coincidence has occured, there must exist a common cause"*. For example if all electrical devices and lights in the room suddenly go out, this coincidence can be explained by a common cause, namely the breakdown of the power supply. A more formal version of the principle states that if we observe the dependence of two jointly distributed random variables X and Y, one of the following must be true: X causes Y or Y causes X or there is a common cause of X and Y. In our framework, this can be understood in the following way: If X and Y are part of a larger system, modeled by a dependency graph G and they are stochastically dependent, then their ancestral sets must be overlapping. Otherwise they would be d-separated by the empty set (which means $X \perp\!\!\!\perp Y \mid \varnothing$) and thus be independent [12].

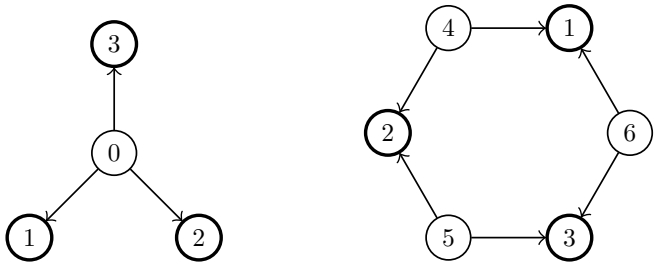
The extended common cause principle. We will now turn to a quantitative extension of the common cause principle, initially studied in [2] and later extended in [12]. Assume that we have a Bayesian network with variables $X_1, \ldots, X_n$ of which a subset $Y_1, \ldots, Y_k$ is observed. On these, we define the mutual information I_c as

$$I_c(Y_1, \ldots, Y_k) = \frac{1}{c} \sum_{j=1}^{k} H(Y_j) - H(Y_1, \ldots, Y_k), \quad \text{where } c > 0. \tag{2}$$

In the case $c = 1$, this is the regular definition of the mutual information from [4]. The quantity I_c is a measure of correlation of the $Y_1, \ldots, Y_k$ and allows the following quantitative extension of Reichenbach's principle of common cause, proven in [12].

Theorem 3.1. (Extended Common Cause Principle) Let $X_1, \ldots, X_n$ be a system with observed variables $Y_1, \ldots, Y_k$. If $I_c(Y_1, \ldots, Y_k) > 0$ then in *any* system containing the $Y_1, \ldots, Y_k$, there exists a common ancestor of strictly more than c variables out of the $Y_1, \ldots, Y_k$.

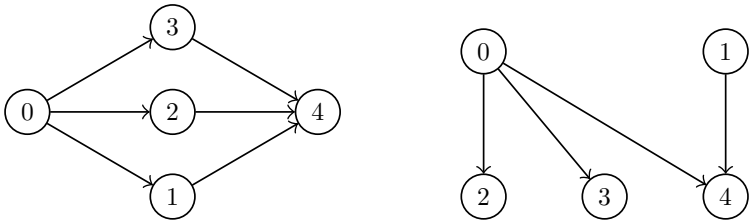
This extended common cause principle allows the discrimination between different causal models for a system by observation alone, even when Reichenbach's common cause principle would fail. In Figure 3 we show two systems from [12] where this is the case. The Reichenbach principle cannot distinguish between (a) and (b), because in both models the observed variables Y_1, Y_2 and Y_3 are not necessarily independent. If we however have $I_2(Y_1, Y_2, Y_3) > 0$, then model (b) can be refused on grounds of the extended common cause principle, because it does not contain a common ancestor of 3 nodes.



(a) one common ancestor of three variables

(b) three common ancestors of two varibles each

Fig. 3. Two possible Bayesian networks for observed variables Y_1, Y_2 and Y_3 (observed nodes are thick, unobserved ones thin). The Reichenbach principle of common cause cannot discriminate these.



(a) One root node

(b) Two root nodes

Fig. 4. Example graphs for the maximization of I_c.

4. A BOUND ON THE MUTUAL INFORMATION I_C

In this section, we derive an upper bound on $I_c(Y_1, \ldots, Y_k)$ over all probability distributions of $X_1, \ldots, X_n$ factorizing according to the dependency graph G.

Before we state the result, we want to illuminate the problem with two example networks. Both examples in Figure 4 refer to the fully observed case. For Figure 4 (a), the maximum $I_2 = (3/2) \cdot \log 2$ is achieved with

$$P(X_0 = 1) = 1/2, \quad X_1 = X_2 = X_3 = X_4 = X_0$$

as Theorem 4.3 will show. This can be achieved by setting $p(x_j = 1 \mid x_0 = 1) = 1$, $p(x_j = 1 \mid x_0 = 0) = 0$ for $1 \leq j \leq 3$ and $p(x_4 = 1 \mid x_1 = x_2 = x_3 = 1) = 1$, $p(x_4 = 1 \mid x_1 = x_2 = x_3 = 0) = 0$. For the example in Figure 4 (b), Theorem 4.3 yields the maximum $I_2 = \log 2$ with

$$P(X_0 = 1) = 1/2, \quad P(X_1 = 1) = 0, \quad X_2 = X_3 = X_4 = X_0.$$

We will now study the general case. The important concepts that are needed in the following theorem are summarized in Figure 5.

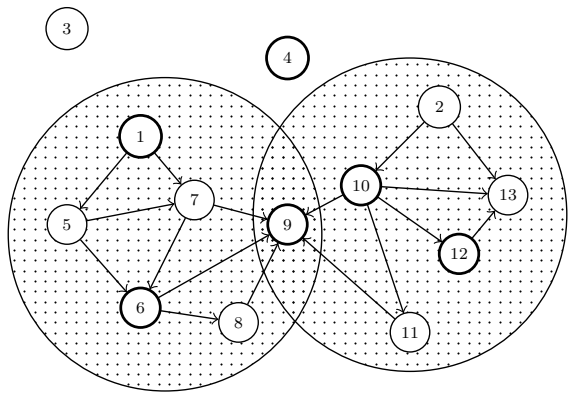


Fig. 5. The nodes X_1, X_2, X_3 and X_4 are the roots, the descendants of X_1 are contained in the dotted circle on the left, the descendants of X_2 in the one on the right. The node X_9 is descendant of both X_1 and X_2. Observed nodes $Y_1, \ldots, Y_k$ are thick, unobserved ones thin.

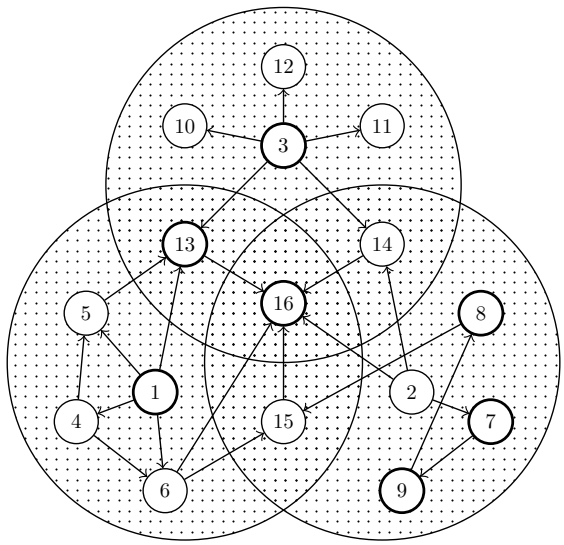


Fig. 6. The set $A_1 = \{X_7, X_8, X_9, X_{16}\}$ contains the largest number of observed variables. Then $A_2 \setminus A_1 = \{X_3, X_{13}\}$ contains only 2 variables and $A_3 \setminus (A_1 \cup A_2) = \{X_1\}$ only one. The ordering is not unique, we could also interchange the names of A_2 and A_3.

Definition 4.1. (Redundancy r and number a of essential nodes) Let G be the dependency graph of a Bayesian network with nodes $X_1, \ldots, X_n$ such that $\{Y_1, \ldots, Y_k\} \subseteq \{X_1, \ldots, X_n\}$ are observed, $X_1, \ldots, X_s$ are the roots and $c > 0$ a fixed integer. The *redundancy* r and the *number of essential nodes* a are obtained by the following procedure.

- Let the set $A_j = \mathrm{de}(X_j) \cap \{Y_1, \ldots, Y_k\} \subseteq \{X_1, \ldots, X_n\}$ for $1 \leq j \leq s$ contain the observed nodes from $\mathrm{de}(X_j)$ (the sets A_j can be overlapping)
- Relabel the first s indices such that $|A_1| \geq |A_2 \setminus A_1| \geq |A_3 \setminus (A_1 \cup A_2)| \geq \ldots$
- The *redundancy* r is the number of sets with $|A_j \setminus (A_1 \cup \cdots \cup A_{j-1})| \geq c$.
- The number of essential nodes is $a = |A_1 \cup \cdots \cup A_r|$

That is, A_{r+1} is the first set in the above order with $|A_{r+1} \setminus (A_1 \cup \cdots \cup A_r)| < c$ (if $r < s$). Note that the numbers a and r depend on c and may not be unique. The names 'redundancy' and 'number of essential nodes' are inspired by the networks that achieve the upper bound on I_c: The observed descendants of root nodes X_j with $H(X_j) \neq 0$ in our construction are the 'essential nodes'. The more root nodes with nonzero marginal entropy, the more failure tolerant the network would be against setting the marginal entropy of root nodes to zero, thus the name 'redundancy'.

In the example of Figure 5, one possible choice for the A_j would be $A_1 = \{X_1, X_6, X_9\}$, $A_2 = \{X_9, X_{10}, X_{12}\}$, $A_3 = \{X_4\}$ and $A_4 = \varnothing$, thus for $c = 3$ we have $r = 1$. In Figure 6 we could choose $A_1 = \{X_7, X_8, X_9, X_{16}\}$, $A_2 \setminus A_1 = \{X_3, X_{13}\}$ and $A_3 \setminus (A_1 \cup A_2) = \{X_1\}$, thus for $c = 2$ we have $r = 2$.

The following Lemma gives a preliminary bound on I_c for binary random variables without constraints.

Lemma 4.2. For any binary random variables $X_1, \ldots, X_n$ we have the bound

$$I_c(X_1, \ldots, X_n) \leq \left(\frac{n}{c} - 1\right) \cdot \log 2 \quad \text{if } n \geq c > 0. \tag{3}$$

Proof. Without restriction assume $H(X_1) \geq H(X_2), H(X_3), \ldots$. The chain rule yields $H(X_1, \ldots, X_n) = \sum_{k=1}^{n} H(X_k \mid X_1, \ldots, X_{k-1}) \geq H(X_1)$ and thus

$$I_c(X_1, \ldots, X_n) \leq \left(\frac{1}{c} - 1\right) \cdot H(X_1) + \frac{H(X_2) + \cdots + H(X_n)}{c}.$$

The bound $H(X_k) \leq \log 2$ for $1 \leq k \leq n$ then proves the result. □

We now have prepared all the necessary tools for our main theorem, which relates the structure of G with the maximum of I_c.

Theorem 4.3. Let $\mathcal{S}$ be the set of all probability distributions on binary random variables $X_1, \ldots, X_n$ that factorize according to the dependency graph G, so

$$\mathcal{S} = \left\{ p : \{0,1\}^n \to [0,1] \mid p(x_1, \ldots, x_n) = \textstyle\prod_{1 \leq j \leq n} p(x_j \mid x_{\mathrm{pa}(X_j)}) \right\}.$$

(i) For any subset $Y_1, \dots, Y_k$ of observed nodes we have

$$\sup_{p \in \mathcal{S}} I_c(Y_1, \dots, Y_k) = \left(\frac{a}{c} - r\right) \cdot \log 2 \tag{4}$$

where $c > 0$ and r, a are from Definition 4.1.

(ii) Certain *deterministic networks* factoring w. r. t. G, with $H(X_j \mid \mathrm{pa}(X_j)) = 0$ for all non-root nodes X_j and a specific probability distribution of the root nodes, attain this supremum.

P r o o f . The nodes are ordered as in Definition 4.1. First of all we construct a probability distribution to show (ii).

Set $P(X_j = 0) = 1/2$ for $1 \le j \le r$, where r is the redundancy, and for all non-root descendents of these X_j, choose the probability distribution such that they copy the value of X_j deterministically (if a node is descendent of two roots, choose one to copy from). For all the remaining nodes set $P(X_j = 0 \mid \mathrm{pa}(X_j)) = 1$. The joint probability distribution $P(X_1, \dots, X_n)$ consists of 2^r equiprobable events, these are the events for $(X_1, \dots, X_r) \in \{0,1\}^r$. Because in each A_j for $1 \le j \le r$ there is at least one observed node, the marginalized distribution $P(Y_1, \dots, Y_k)$ also consists of 2^r equiprobable events, so we have

$$H(Y_1, \dots, Y_k) = -\sum_{j=1}^{2^r} \frac{1}{2^r} \log \frac{1}{2^r} = r \log 2.$$

On the other hand, $H(X) = \log 2$ for $X \in \mathrm{de}(X_1) \cup \dots \cup \mathrm{de}(X_r)$, all other nodes have zero entropy by construction. So we conclude

$$\sum_{j=1}^{k} H(Y_j) = a \cdot \log 2,$$

and $I_c(Y_1, \dots, Y_k)$ from (4) is achieved.

For (i), we use induction on the number of roots. For a single root, the bound follows from Lemma 4.2. The induction step then proceeds as follows. The nodes in $A_1, A_2, \dots$ are partitioned as in Figure 7, so by the chain rule

$$H(Y_A, Y_B, Y_R) = H(Y_A, Y_R) + H(Y_B \mid Y_A, Y_R) \ge H(Y_A, Y_R),$$

and because X_A and X_R are independent and then also Y_A and Y_R, it follows that $H(Y_A, Y_B, Y_R) \ge H(Y_A) + H(Y_R)$. By our induction hypothesis and Lemma 4.2 we then have

$$\begin{aligned} I_c(Y_1, \dots, Y_k) &\le \sum_{j \in B} \frac{H(Y_j)}{c} + \sum_{j \in A} \frac{H(Y_j)}{c} - H(Y_A) + \sum_{j \in R} \frac{H(Y_j)}{c} - H(Y_R) \\ &\le \frac{|B|}{c} \cdot \log 2 + \underbrace{\left(\frac{|A|}{c} - 1\right)}_{\ge 0} \cdot \log 2 + I_c(Y_R) \le \left(\frac{a}{c} - r\right) \cdot \log 2, \end{aligned}$$

which is the claimed bound. □

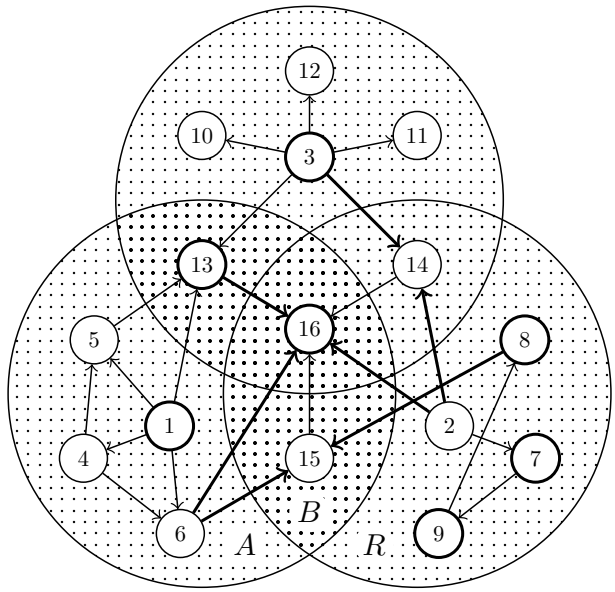
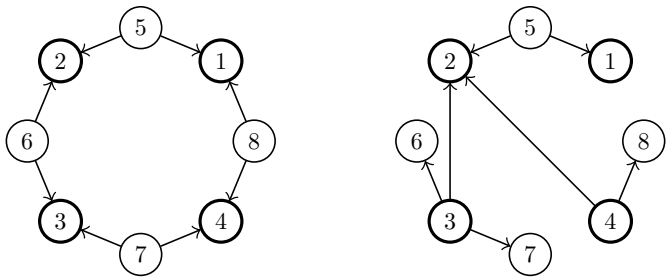


Fig. 7. For $R_j = \mathrm{de}(X_j)$, that means $A_j \subseteq R_j$ (A_j contains only the observed nodes from R_j), the set $R_1 \cup R_2 \cup \cdots$ is partitioned into $A = R_1 \setminus (R_2 \cup R_3 \cup \cdots)$, $B = R_1 \cap (R_2 \cup R_3 \cup \cdots)$ and the rest $R = (R_2 \cup R_3 \cup \cdots) \setminus R_1$. Note that X_A and X_R are independent because they are d-separated by the empty set, namely $X_A \perp\!\!\!\perp X_R \mid \varnothing$ (the thick arrows are all pointing in).

5. DISCRIMINATING BAYESIAN NETWORKS

Now we describe how this theorem can be used to discriminate between two causal hypotheses. Take the Bayesian networks from Figure 8 (a) and (b) as an example. In both cases, there are common ancestors of at most two observed variables. Thus with a straightforward application of the extended common cause principle we cannot distinguish them. However, from Definition 4.1 for $c = 1$ we get $r = 2$ and $a = 4$ for (a) and $r = 3$ and $a = 4$ for (b). Thus $I_1 \leq 2 \log 2$ for (a) and $I_1 \leq \log 2$ for (b) and we can reject hypothesis (b) on the grounds of Theorem 4.3 if I_1 is in the range from $\log 2$ to $2 \log 2$.

How effective is this procedure? We elucidate this with the following toy numerical experiment: Generate random pairs of directed Erdös–Rényi graphs $G_{n,p}$ [3] and remove cycles by considering only edges (u, v) with $u < v$. Then test if the two hypotheses could be distinguished by the extended common cause principle from Theorem 3.1 or the result from Theorem 4.3 with the above method. The results are shown in Table 1. For graphs $G_{n,p}$ with $p = 0.15$ and $n = 10$, the second method is significantly more powerful than the method that employs the extended common cause principle. In this case we have $np = 1.5$. It is conjectured that for random Boolean networks, $np = 2$ (the critical regime) is of greatest interest for real biological systems [11]. In this regime, the method from Theorem 4.3 yields the largest improvement over the result from Theorem 3.1.



(a) First hypothesis (b) Second hypothesis

Fig. 8. Two Bayesian networks for the observations X_2, X_3, X_7, X_8, X_9.

Graph	Theorem 4.3	Theorem 3.1
$G_{10,0.05}$	4143 ± 47	4110 ± 52
$G_{10,0.10}$	5787 ± 51	5645 ± 54
$G_{10,0.15}$	6445 ± 51	6207 ± 50
$G_{10,0.20}$	6671 ± 41	6421 ± 33
$G_{10,0.25}$	6713 ± 42	6567 ± 61
$G_{10,0.30}$	6673 ± 35	6713 ± 61

Tab. 1. For each $G_{n,p}$ we sampled 10000 pairs of graphs and counted the number of pairs that could in principle be distinguished by the method described in section 5. The standard deviation was determined from 10 independent runs for each entry.

6. CONCLUSION

We derived a tight upper bound on the mutual information I_c in a partially observed Bayesian network factoring according to a dependence graph G. Our inequality and proof give insight in how the ancestral structure of a Bayesian network is related to the possible degree of correlation between the nodes of the network. We furthermore showed how this inequality can be used for discrimination between different causal hypotheses underlying a system and to what degree our method surpasses the extended common cause principle in this respect.

7. ACKNOWLEDGMENTS.

Philipp Moritz thanks his colleagues from the MPI for Mathematics in the Sciences in Leipzig (especially the Information Theory of Cognitive Systems Group, where most of this research was conducted) for many stimulating discussions and the participants of WUPES 2012 for new suggestions. Furthermore, he thanks the programme "FOKUS Physik" from the University of Würzburg for their support. The authors want to thank two anonymous reviewers for their valuable suggestions. Jörg Reichardt was supported by a Fellowship Computational Sciences

of the Volkswagen Foundation.

(Received February 27, 2013)

REFERENCES

[1] E. S. Allman and J. A. Rhodes: Reconstructing Evolution: New Mathematical and Computational Advances, chapter Phylogenetic invariants. Oxford University Press, 2007.

[2] N. Ay: A refinement of the common cause principle. Discrete Appl. Math. *157* (2009), 10, 2439–2457.

[3] B. Bollobás: Random Graphs. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2001.

[4] T. M. Cover and J. A. Thomas: Elements of Information Theory. Second edition. Wiley, 2006.

[5] N. Friedman: Inferring cellular networks using probabilistic graphical models. Science *303* (2004), 5659, 799–805.

[6] J. Peters, J. Mooij, D. Janzing, and B. Schölkopf: Causal discovery with continuous additive noise models. arXiv *1309.6779* (2013).

[7] S. L. Lauritzen: Graphical Models. Oxford Science Publications, Clarendon Press, 1996.

[8] S. L. Lauritzen and N. A. Sheehan: Graphical models for genetic analyses. Statist. Sci. *18* (2003), 489–514.

[9] J. Pearl: Causality: Models, Reasoning and Inference. Cambridge University Press, 2000.

[10] H. Reichenbach and M. Reichenbach: The Direction of Time. California Library Reprint Series, University of California Press, 1956.

[11] J. E. S. Socolar and S. A. Kauffman: Scaling in ordered and critical random boolean networks. Phys. Rev. Lett. *90* (2003), 068702.

[12] B. Steudel and N. Ay: Information-theoretic inference of common ancestors. CoRR, abs/1010.5720, 2010.

[13] M. Studený: Probabilistic Conditional Independence Structures. Information Science and Statistics. Springer, 2005.

Philipp Moritz, Institute for Theoretical Physics, University of Würzburg. Germany.
e-mail: philipp.moritz@physik.uni-wuerzburg.de

Jörg Reichardt, Institute for Theoretical Physics, University of Würzburg. Germany.
e-mail: reichardt@physik.uni-wuerzburg.de

Nihat Ay, MPI for Mathematics in the Sciences, Inselstraße 22, Leipzig, Germany and Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501. U. S. A.
e-mail: nay@mis.mpg.de