Markov Equivalence in Bayesian Networks

Ildikó Flesch and Peter Lucas
Institute for Computing and Information Sciences
University of Nijmegen
Email: {ildiko,peterl}@cs.kun.nl

Abstract

Probabilistic graphical models, such as Bayesian networks, allow representing conditional independence information of random variables. These relations are graphically represented by the presence and absence of arcs and edges between vertices. Probabilistic graphical models are nonunique representations of the independence information of a joint probability distribution. However, the concept of Markov equivalence of probabilistic graphical models is able to offer unique representations, called essential graphs. In this survey paper the theory underlying these concepts is reviewed.

Contents

1 Introduction

During the past decade Bayesian-network structure learning has become an important area of research in the field of Uncertainty in Artificial Intelligence (UAI). In the early years of Bayesian-network research at the end of the 1980s and during the 1990s, there was considerable interest in the process of manually constructing Bayesian networks with the help of domain experts. In recent years, with the increasing availability of data and the associated rise of the field of data-mining, Bayesian networks are now seen by many researchers as promising tools for data analysis and statistical model building. As a consequence, a large number of papers discussing structure learning related topics have been published during the last couple of years, rendering it hard for the novice to the field to appreciate the relative importance of the various research contributions, and to develop a balanced view on the various aspects of the field. The present paper was written in an attempt to provide a survey of the issue lying at the very heart of Bayesian-network structure learning: (statistical) independence and its representation in graphical form.

The goal of structure learning is to find a good generative structure relative to the data and to derive the generative joint probability distribution over the random variables of this distribution. Nowadays, graphical models are widely used to represent generative joint probability distributions.

A *graphical model* is a knowledge-representation formalism providing a graph representation of structural properties of uncertain knowledge. Bayesian networks are special cases of graphical models; they offer a representation that is both powerful and easy to understand, which might explain their current high level of popularity among UAI, machine-learning and data-mining researchers.

Structure learning consists of two main, interrelated, problems:

(i) the evaluation problem, and

(ii) the identification problem.

The *evaluation problem* amounts to finding a suitable way of judging the quality of generated network structures. Using a scoring criterion we can investigate how well a structure with its associated constraints fits the data. Note that a scoring criterion allows comparing structures with each other in such way that a structure-ordering becomes possible. As we compare Bayesian networks comprising the same vertices, scoring criteria are based only on relationships modelled by means of arcs. One expects that the better the independence relations implied by the graph representation match the knowledge hidden in the data the higher the score obtained by a scoring criterion, and this should be taken as one of a set of requirements when designing a scoring criterion.

The *identification problem* concentrates on finding efficient methods to identify at least one, maybe more, network structures given a scoring criterion. The total number of possible graph representations for a problem grows superexponentially with the total number of random variables [12]. As a consequence, the application of brute-force algorithms is computationally speaking infeasible. Thus practical methods for learning Bayesian networks use heuristic search techniques to find a graph with a high score in the space of all possible network structures.

In this survey paper we do not go into details of network scoring criteria; rather we focus on another, however closely related, important element in all modern research regarding the learning of Bayesian networks: the identification of Bayesian networks that represent the same joint probability distribution, i.e. they are Markov equivalent. Exploiting the notion of Markov equivalence can yield computational savings by making the search space that must be explored more compact [4]. There are various proposals in the literature to represent Markov equivalent Bayesian networks. One of them, a proposal by Andersson et al, [1], uses a special type of graph, called an *essential graph*, to act as a class representative for Bayesian networks that encodes the same probabilistic independence information. Markov independence is therefore a key issue in learning Bayesian networks. This paper summarises the theory underlying equivalence of graphical models in terms of the underlying independence relationship.

The paper is organised as follows. In the next section, basic notions from graph theory and the logical notion of (statistical) independence are introduced. These act as the basis for the graph representation of independence information as described in Section 3. Equivalence of Bayesian networks is studied in depth in Section 4, where we are also concerned with the properties of essential graphs. The paper is rounded off with remarks with respect to the consequences of the theory summarised in this paper for the area of Bayesian-network structure learning.

2 Preliminaries

We start by introducing some elementary notions from graph theory in Section 2.1. Next, we review the foundation of the stochastic (or statistical) independence relation as defined in probability theory in Section 2.2. We assume that the reader has access to a basic textbook on probability theory (cf. [6]).

2.1 Basic concepts from graph theory

This subsection introduces some notions from graph theory based on Ref. [5]; it can be skipped by readers familiar with these notions.

Sets of objects will be denoted by bold, upright uppercase letters, e.g. $\mathbf{V}$. For singleton sets $\{X\}$, we will often only write the element X instead of the set $\{X\}$. A *graph* is defined as a pair $G = (\mathbf{V}(G), \mathbf{E}(G))$, with $\mathbf{V}(G)$ a finite set of *vertices*, where a vertex is denoted by an uppercase letter such as V, X and Y, and $\mathbf{E}(G) \subseteq \mathbf{V}(G) \times \mathbf{V}(G)$ is a finite set of *edges*. A graph $H = (\mathbf{V}(H), \mathbf{E}(H))$ is called an *induced subgraph* of graph $G = (\mathbf{V}(G), \mathbf{E}(G))$ if $\mathbf{V}(H) \subseteq \mathbf{V}(G)$ and $\mathbf{E}(H) = \mathbf{E}(G) \cap (\mathbf{V}(H) \times \mathbf{V}(H))$. A graph $G = (\mathbf{V}(G), \mathbf{E}(G))$ for which it holds that for each $(X, Y) \in \mathbf{E}(G)$: $(Y, X) \in \mathbf{E}(G)$, $X \neq Y$, is called an *undirected graph* (UG). An edge $(X, Y) \in \mathbf{E}(G)$ in an undirected graph is also denoted by $X - Y$ and called an *undirected edge*. However, we will usually refer to undirected edges simply as edges if this will not give rise to confusion. A graph $G = (\mathbf{V}(G), \mathbf{A}(G))$ is called a *directed graph* if it comprises a finite set of vertices $\mathbf{V}(G)$, but, in contrast to an undirected graph, contains a finite set of *arcs*, by some authors called *arrows* or *directed edges*, $\mathbf{A}(G) \subseteq \mathbf{V}(G) \times \mathbf{V}(G)$ for which it holds that for each $(X, Y) \in \mathbf{A}(G)$: $(Y, X) \notin \mathbf{A}(G)$. An arc $(X, Y) \in \mathbf{A}(G)$ is also denoted by $X \rightarrow Y$ in the following.

A *route* in a graph G is a sequence $V_1, V_2, \ldots, V_k$ of vertices in $\mathbf{V}(G)$, where either $V_i \rightarrow V_{i+1}$, or $V_i \leftarrow V_{i+1}$, and possibly $V_i - V_{i+1}$, for $i = 1, \ldots, k-1, k \geq 1$; k is the *length* of the route. Note that a vertex may appear more than once on a route. A *section* of a route $V_1, V_2, \ldots, V_k$ is a maximal undirected subroute $V_i - \cdots - V_j$, $1 \leq i \leq j \leq k, k \geq 1$, where V_i

resp. V_k is called a *tail terminal* if $V_{i-1} \leftarrow V_i$ resp. $V_k \rightarrow V_{k-1}$, and V_i resp. V_k is called a *head terminal* if $V_{i-1} \rightarrow V_i$ resp. $V_k \leftarrow V_{k-1}$. A *path* in a graph is a route, where vertices V_i and V_{i+1} are connected either by an arc $V_i \rightarrow V_{i+1}$ or by an edge $V_i - V_{i+1}$. A path is a *directed path*, if it contains at least one arc. A *trail* in a graph is a route where each arc appears at most once and, in addition, the vertices in each section of the trail appear at most once in the section. A *slide* is a special directed path with $V_1 \rightarrow V_2$ and $V_i - V_{i+1}$ for all $2 \leq i \leq k-1$. A graph has a *directed cycle* if it contains a directed path, which begins and ends at the same vertex, i.e. $V_1 = V_k$.

A graph $G = (\mathbf{V}(G), \mathbf{E}(G))$ is called a *chain graph* if it contains no directed cycles. An *acyclic directed graph* (ADG) is a chain graph that is a directed graph. Note that undirected graphs are special cases of chain graphs as well. Due to the acyclicity property of chain graphs, the vertex set of a chain graph can be partitioned into subsets $\mathbf{V}(1) \cup \mathbf{V}(2) \cup \cdots \cup \mathbf{V}(T), T \geq 1$, such that each partition only consists of edges and if $X \rightarrow Y$, then $X \in \mathbf{V}(i)$ and $Y \in \mathbf{V}(j)$, $i \neq j$. Based on this we can define a total order $\leq$ on vertices in a chain graph, such that if $X \in \mathbf{V}(i)$ and $Y \in \mathbf{V}(j)$, with $i < j$, then $X < Y$, and if $i = j$, then $X = Y$ (i.e. they are in the same $\mathbf{V}(i)$). This order can be generalised to sets such that $\mathbf{X} \leq \mathbf{Y}$ if for each $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$ we have that $X \leq Y$. Subsets $\mathbf{V}(1), \mathbf{V}(2), \ldots, \mathbf{V}(T)$ are called the *chain components* of the graph. A set of *concurrent variables* of $\mathbf{V}(t)$ is defined as $\mathbf{C}(t) = \mathbf{V}(1) \cup \mathbf{V}(2) \cup \cdots \cup \mathbf{V}(t), 1 \leq t \leq T$. Any vertex X in an ADG that is connected by a directed path to a vertex Y is called a *predecessor* of Y; the set of predecessors of Y is denoted by $pr(Y)$.

We say that the vertex $X \in \mathbf{V}(G)$ is a *parent* of $Y \in \mathbf{V}(G)$ if $X \rightarrow Y \in \mathbf{A}(G)$; the set of parents of Y is denoted by $\pi(Y)$. Furthermore, Y is then called X's *child*; the set of children of vertex X is denoted by $ch(X)$. Two vertices $X, Y \in \mathbf{V}(G)$ are *neighbours*, if there is an edge between these two vertices. The *boundary* of vertex $X \in \mathbf{V}(G)$, denoted by $bd(X)$, is the set of parents and neighbours of X, while the *closure* of X, denoted by $cl(X)$, is defined as $cl(X) = bd(X) \cup \{X\}$. Note that the boundary of a vertex X in an undirected graph is equal to its set of neighbours. The set of *ancestors* of a vertex X is the set of vertices $\alpha(X) \subseteq \mathbf{V}(G)$ where there exists a path from each $Y \in \alpha(X)$ to X, but there exists no path from X to Y, whereas the set of *descendants* of X, denoted by $\delta(X)$, is the set of vertices $\delta(X) \subseteq \mathbf{V}(G)$, where there exists a path from X to each $Y \in \delta(X)$, but no path from Y to X. The set of *non-descendants* of X, denoted by $\bar{\delta}(X)$, is equal to $\mathbf{V}(G) \setminus (\delta(X) \cup \{X\})$. If for some $\mathbf{W} \subseteq \mathbf{V}$ it holds that $bd(X) \subseteq \mathbf{W}$, for each $X \in \mathbf{W}$, then $\mathbf{W}$ is called an *ancestral* set. By $an(\mathbf{W})$ is denoted the *smallest* ancestral set containing $\mathbf{W}$.

From the chain graph G we can derive the *moral graph* G^m by the following procedure, called *moralisation*:

(i) add edges to all non-adjacent vertices, which have children in a common chain component, and

(ii) replace each arc with an edge in the resulting graph.

A moral graph is therefore an undirected graph.

A *chord* is an edge or arc between two non-adjacent vertices of a path. A graph is called *chordal* if every cycle of length $k \geq 4$ has a chord.

As mentioned above, two vertices can be connected by an arc or an edge. If two distinct vertices $X, Y \in \mathbf{V}(G)$ are connected but it is unknown whether by an edge or arc, we write $X \cdots Y$, where the symbol $\cdots$ denotes this connection.

2.2 Axiomatic basis of the independence relation

We continue by providing the basic definition of (conditional) independence that underlies almost all theory presented in this paper. The idea that conditional independence is a unifying notion of relationships among components of many mathematical structures was first expressed by Dawid [2].

Definition 1 ***(conditional independence)*** *Let* $\mathbf{V}$ *be a set of random variables with* $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$ *disjoint sets of random variables, and let* P *be a joint probability distribution defined on* $\mathbf{V}$*, then* $\mathbf{X}$ *is said to be* conditionally independent *of* $\mathbf{Y}$ given $\mathbf{Z}$*, denoted by* $\mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z}$*, if*

$$P(\mathbf{X} \mid \mathbf{Y}, \mathbf{Z}) = P(\mathbf{X} \mid \mathbf{Z}). \tag{1}$$

Conditional independence can be also interpreted as follows: learning about $\mathbf{Y}$ has no effect on our knowledge concerning $\mathbf{X}$ given our beliefs concerning $\mathbf{Z}$, and vice versa. If Definition 1 does not hold, then $\mathbf{X}$ and $\mathbf{Y}$ are said to be *conditionally dependent* given $\mathbf{Z}$, which is written as follows:

$$\mathbf{X} \not\!\perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z}. \tag{2}$$

We will often write $X \perp\!\!\!\perp_P Y \mid Z$ instead of $\{X\} \perp\!\!\!\perp_P \{Y\} \mid \{Z\}$.

Next, we will introduce the five most familiar axioms, called the *independence axioms* or *independence properties*, which the independence relation $\perp\!\!\!\perp_P$ satisfies. An example is provided for each of the axioms, freely following Ref. [3]. As the independence axioms are valid for many different mathematical structures, and we are concerned in this paper with independence properties represented by graphs—examples of such mathematical structures—we will use graphs to illustrate the various axioms. However, a discussion on how such graphs should be interpreted in the context of probability theory is postponed to the next section. In the example graphs of this paper, the first set $\mathbf{X}$ in the triple $\mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z}$ is coloured lightly grey, the second set $\mathbf{Y}$ is coloured medium grey and the set $\mathbf{Z}$ dark grey. If a vertex in the graph does not participate in an independence property illustrated by the example, it is left unshaded.

The independence relation $\perp\!\!\!\perp_P$ satisfies the following independence axioms or independence properties:[1]

- ***P1****: Symmetry.* Let $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$ be disjoint sets of random variables, then

  $$\mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z} \Longleftrightarrow \mathbf{Y} \perp\!\!\!\perp_P \mathbf{X} \mid \mathbf{Z}.$$

  If knowing $\mathbf{X}$ is irrelevant to our knowledge about $\mathbf{Y}$ given that we believe $\mathbf{Z}$, then the reverse also holds. An example is given in Figure 1(a).

- ***P2****: Decomposition.* Let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W} \subseteq \mathbf{V}$ be disjoint sets of random variables, then

  $$\mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z} \;\Rightarrow\; \mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z} \,\wedge\, \mathbf{X} \perp\!\!\!\perp_P \mathbf{W} \mid \mathbf{Z}.$$

[1] Here and in the following we will always assume that the sets participating in the various independence relations $\perp\!\!\!\perp$ are disjoint, despite the fact that this is not strictly necessary. However, as disjoint and non-disjoint sets bear a completely different meaning, and it does not appear to be a good idea to lump these two meanings together, we have decided to restrict our treatment to disjoint sets, as this seems to offer the most natural interpretation of (in)dependence.

This property states that if both $\mathbf{Y}$ and $\mathbf{W}$ are irrelevant with regard to our knowledge of $\mathbf{X}$ assuming that we believe $\mathbf{Z}$, then they are also irrelevant separately. See the example in Figure 1(b).

- ***P3:*** *Weak union.* Let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W} \subseteq \mathbf{V}$ be disjoint sets of random variables, then

  $$\mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z} \Rightarrow \mathbf{X} \perp\!\!\!\perp_P \mathbf{W} \mid \mathbf{Y} \cup \mathbf{Z} \ \ \wedge \ \ \mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z} \cup \mathbf{W}.$$

  It expresses that when learning about $\mathbf{Y}$ and $\mathbf{W}$ is irrelevant with respect to our knowledge about $\mathbf{X}$ given our beliefs about $\mathbf{Z}$, then $\mathbf{W}$ will remain irrelevant when our beliefs do not only include $\mathbf{Z}$ but also $\mathbf{Y}$ (the same holds for $\mathbf{W}$). The weak union relation is illustrated by Figure 1(c).

- ***P4:*** *Contraction.* Let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W} \subseteq \mathbf{V}$ be disjoint sets of random variables, then

  $$\mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z} \ \ \wedge \ \ \mathbf{X} \perp\!\!\!\perp_P \mathbf{W} \mid \mathbf{Y} \cup \mathbf{Z} \Rightarrow \mathbf{X} \perp\!\!\!\perp_P \mathbf{W} \cup \mathbf{Y} \mid \mathbf{Z}.$$

  Contraction expresses the idea that if learning about $\mathbf{Y}$ is irrelevant to our knowledge about $\mathbf{X}$ given that we believe $\mathbf{Z}$ and in addition learning about $\mathbf{W}$ does not change our knowledge with respect to $\mathbf{X}$ either, then the irrelevance of $\mathbf{W}$ with respect to $\mathbf{X}$ is not dependent on our knowledge of $\mathbf{Y}$, but only on $\mathbf{Z}$. The notion of contraction is illustrated by Figure 1(d).

- ***P5:*** *Intersection.* Let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W} \subseteq \mathbf{V}$ be disjoint sets of random variables, then

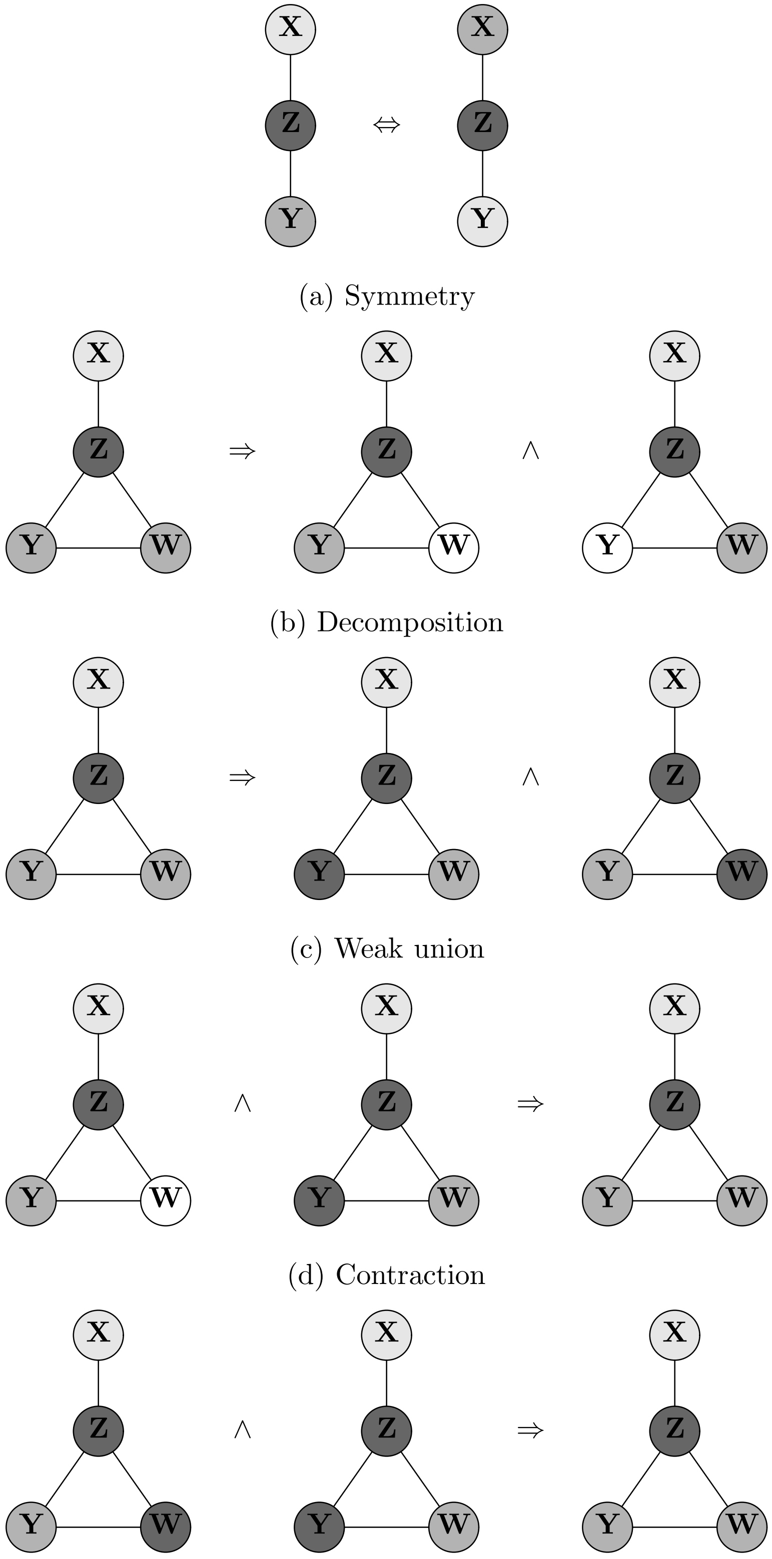
  $$\mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z} \cup \mathbf{W} \ \wedge \ \mathbf{X} \perp\!\!\!\perp_P \mathbf{W} \mid \mathbf{Z} \cup \mathbf{Y} \Rightarrow \mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \cup \mathbf{W} \mid \mathbf{Z}.$$

  The intersection property states that if learning about $\mathbf{Y}$ has no effect on our knowledge about $\mathbf{X}$ assuming that we believe $\mathbf{Z}$ and $\mathbf{W}$, knowing, in addition, that our knowledge of $\mathbf{W}$ does not affect our knowledge concerning $\mathbf{X}$ if we also know $\mathbf{Y}$, then learning about $\mathbf{Y}$ and $\mathbf{W}$ together has also no effect on $\mathbf{X}$. This property only holds for *strictly positive* joint probability distributions. An example of the intersection property is shown in Figure 1(e).

Any model satisfying the independence axioms **P1** to **P4** is called a *semi-graphoid*, whereas any model of the axioms **P1** to **P5** is called a *graphoid*. Any joint probability distribution P satisfies axioms **P1** to **P4**, while a joint probability distribution only satisfies **P5** if its co-domain is restricted to the open interval $(0, 1)$, i.e. it is a joint probability distribution that does not represent *logical relationships*. A counterexample of the intersection property is shown in Table 1. Here, each random variable can take values a or b. There are only four possibilities, each having a probability equal to $\frac{1}{4}$, and the other possibilities have probability equal to 0. It holds that $X \perp\!\!\!\perp_P W \mid Z \cup Y$ and $X \perp\!\!\!\perp_P Y \mid Z \cup W$, however $X \not\perp\!\!\!\perp_P Y \cup W \mid Z$.

The independence axioms **P1** to **P4** were first introduced by Pearl (cf. [10]); he claimed that they offered a finite characterisation of the independence relation (Pearl's famous "completeness conjecture"). This statement, however, was shown to be incorrect by Studený after he discovered an axiom which indeed appeared to be a property of the independence relation, yet could not be deduced from axioms **P1** to **P4** [13]. Subsequently, Studený proved that no finite axiomatisation of the independence relation exists [14].

The five axioms mentioned above are well known by researchers in probabilistic graphical models; however, there are a number of other axioms which are also worth mentioning. We mention four of these axioms:


X Z Y ⇔ X Z Y
X Z Y W ⇒ X Z Y W ∧ X Z Y W
X Z Y W ⇒ X Z Y W ∧ X Z Y W
X Z Y W ∧ X Z Y W ⇒ X Z Y W
X Z Y W ∧ X Z Y W ⇒ X Z Y W

(a) Symmetry

(b) Decomposition

(c) Weak union

(d) Contraction

(e) Intersection

Figure 1: Example graphs illustrating the following independence axioms: (a) Symmetry, (b) Decomposition, (c) Weak union, (d) Contraction and (e) Intersection.

X	Y	W	Z
a	a	a	a
a	a	a	a
b	a	a	a
b	b	b	a

Table 1: Counterexample to the intersection property.

- ***P6***: *Strong union.* Let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W} \subseteq \mathbf{V}$ be disjoint sets of random variables, then

  $$\mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z} \Rightarrow \mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z} \cup \mathbf{W}.$$

  This property says that if learning about $\mathbf{Y}$ does no convey any knowledge with regard to $\mathbf{X}$ given our beliefs concerning $\mathbf{Z}$, then this knowledge concerning $\mathbf{Y}$ remains irrelevant if our beliefs also include $\mathbf{W}$. An example is shown in Figure 2(a).

  It holds that strong union implies weak union [3].

- ***P7***: *Strong transitivity.* Let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W} \subseteq \mathbf{V}$ be disjoint sets of random variables, then

  $$\mathbf{X} \not\perp\!\!\!\perp_P \mathbf{W} \mid \mathbf{Z} \ \wedge \ \mathbf{Y} \not\perp\!\!\!\perp_P \mathbf{W} \mid \mathbf{Z} \Rightarrow \mathbf{X} \not\perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z}.$$

  This property says that if based on our beliefs concerning $\mathbf{Z}$, observing $\mathbf{W}$ will learn us something about both $\mathbf{X}$ and $\mathbf{Y}$, then our beliefs concerning $\mathbf{Z}$ already made $\mathbf{X}$ and $\mathbf{Y}$ relevant to each other. Applying the equivalence $a \Rightarrow b \equiv \neg b \Rightarrow \neg a$, strong transitivity can be rewritten to

  $$\mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z} \Rightarrow \mathbf{X} \perp\!\!\!\perp_P \mathbf{W} \mid \mathbf{Z} \ \vee \ \mathbf{Y} \perp\!\!\!\perp_P \mathbf{W} \mid \mathbf{Z}.$$

  For an example see Figure 2(b).

- ***P8***: *Weak transitivity.* Let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W} \subseteq \mathbf{V}$ be disjoint sets of random variables, then

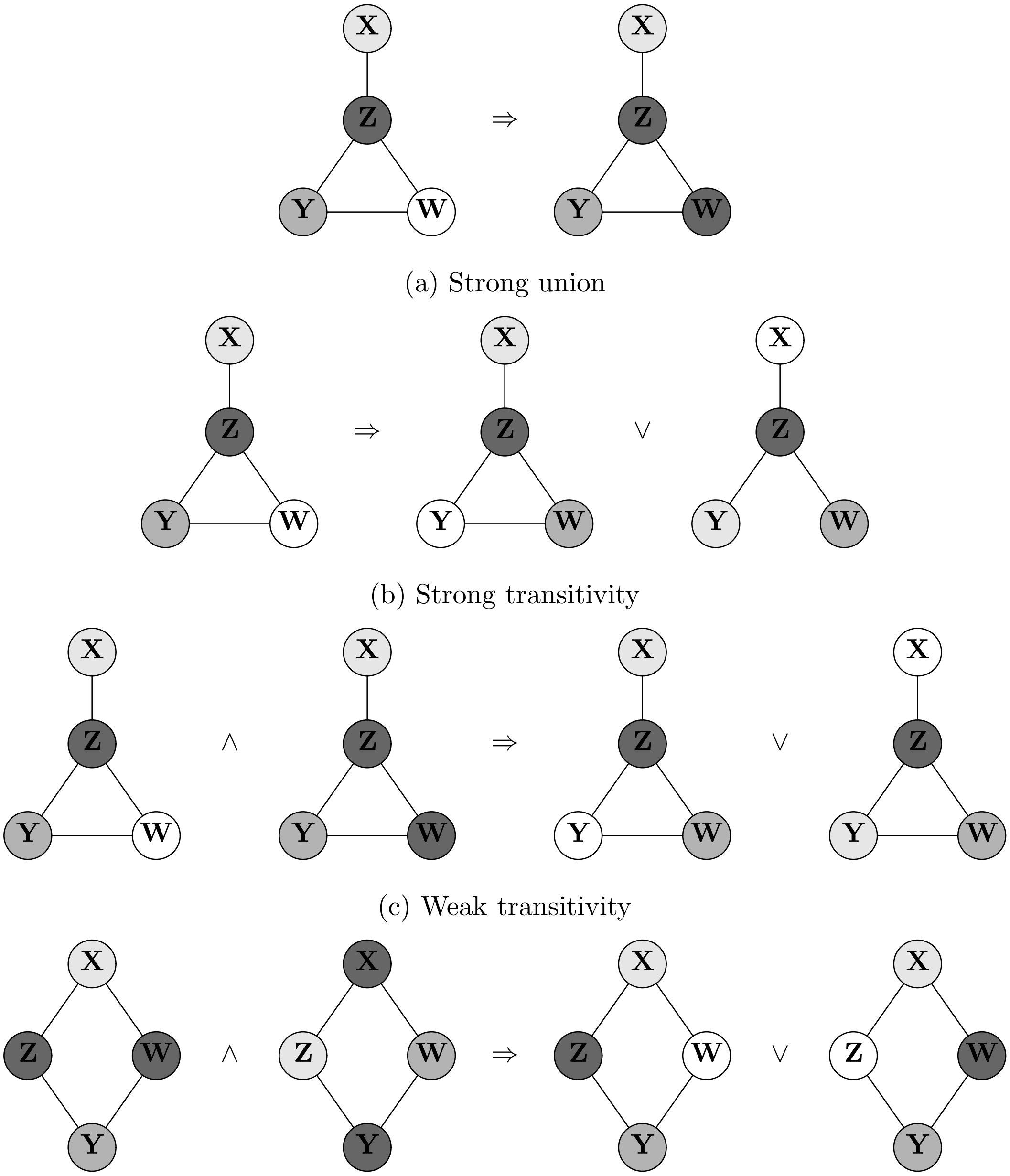
  $$\mathbf{X} \not\perp\!\!\!\perp_P \mathbf{W} \mid \mathbf{Z} \ \wedge \ \mathbf{Y} \not\perp\!\!\!\perp_P \mathbf{W} \mid \mathbf{Z} \Rightarrow \mathbf{X} \not\perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z} \ \vee \ \mathbf{X} \not\perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z} \cup \mathbf{W}.$$

  Weak transitivity is an extension of strong transitivity and states that if $\mathbf{X}$ and $\mathbf{Y}$ are separately dependent of $\mathbf{W}$ given our beliefs about $\mathbf{Z}$, then it holds that knowledge exchange between $\mathbf{X}$ and $\mathbf{Y}$ is accomplished via $\mathbf{Z}$ *or* $\mathbf{Z}$ and $\mathbf{W}$. Applying the equivalence $a \Rightarrow b \equiv \neg b \Rightarrow \neg a$ the above mentioned dependence relation can also be written as

  $$\mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z} \ \ \wedge \ \mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z} \cup \mathbf{W} \Rightarrow \mathbf{X} \perp\!\!\!\perp_P \mathbf{W} \mid \mathbf{Z} \ \vee \ \mathbf{Y} \perp\!\!\!\perp_P \mathbf{W} \mid \mathbf{Z}.$$

  This property is illustrated in Figure 2(c).

  It holds that strong transitivity implies weak transitivity [3].



(a) Strong union

(b) Strong transitivity

(c) Weak transitivity

(d) Chordality

Figure 2: Example graphs illustrating the following independence axioms: (a) Strong union, (b) Strong transitivity, (c) Weak transitivity and (d) Chordality.

- ***P9:*** *Chordality.* Let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W} \subseteq \mathbf{V}$ be disjoint sets of random variables, then

  $$\mathbf{X} \not\perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z} \,\wedge\, \mathbf{X} \not\perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{W} \Rightarrow \mathbf{X} \not\perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z} \cup \mathbf{W} \,\vee\, \mathbf{Z} \not\perp\!\!\!\perp_P \mathbf{W} \mid \mathbf{X} \cup \mathbf{Y}.$$

  It implies that if learning about $\mathbf{Y}$ yields knowledge about $\mathbf{X}$, having beliefs concerning $\mathbf{Z}$, and the same holds when we have beliefs about $\mathbf{W}$, then our knowledge about $\mathbf{Y}$ is still relevant to our knowledge about $\mathbf{X}$ if we know both $\mathbf{Z}$ and $\mathbf{W}$, or our knowledge about both $\mathbf{Z}$ and $\mathbf{W}$ makes $\mathbf{Z}$ and $\mathbf{W}$ exchange knowledge. It is equivalent to

  $$\mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z} \cup \mathbf{W} \ \wedge\ \mathbf{Z} \perp\!\!\!\perp_P \mathbf{W} \mid \mathbf{X} \cup \mathbf{Y} \Rightarrow \mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z} \,\vee\, \mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{W}.$$

  An example of chordality is depicted in Figure 2(d).

3 Graphical representation of independence

In this section, we discuss the representation of the independence relation by means of graphs, the rest of this paper will be devoted to this topic. In the previous section, the conditional independence relationship was defined in terms of a joint probability distribution P. In Section 3.1 closely related notions of graph separation are defined and informally linked to conditional independence. In Section 3.2, various special Markov properties are introduced and discussed, building upon the separation criteria from Section 3.1. Finally, in Section 3.3, possible relationships between conditional (in)dependences in joint probability distributions and the graph separation properties introduced earlier are established formally. This provides a semantic foundation for the various types of graphs in terms of the theory of statistical independence. Let $G = (\mathbf{V}(G), \mathbf{E}(G))$ be an undirected graph, and let $\mathbf{V}$ be a set of random variables, such that there is a one-to-one correspondence $\mathbf{V} \leftrightarrow \mathbf{V}(G)$ between $\mathbf{V}$ and $\mathbf{V}(G)$. Due to this one-to-one correspondence we will normally not sharply distinguish between random variables and vertices; the context will make clear whether random variables or vertices are meant.

3.1 Graph separation and conditional independence

The independence relation defined earlier can be represented as a graphical model, where arcs and edges represent the dependences, and absence of arcs and edges represents the (conditional) independences. Arcs and edges represent roughly the same (in)dependence information; however, there are some differences between the meaning of arcs and edges. The actual interpretation is subtle, and is the topic of this and subsequent sections. In this section, we provide the foundation for representing conditional independence statements by graphs, and we cover the similarities between these principles for undirected, acyclic directed as well as for chain graphs.

In an undirected graph $G = (\mathbf{V}(G), \mathbf{E}(G))$ two vertices $X, Y \in \mathbf{V}(G)$ are dependent if $X - Y \in \mathbf{E}(G)$; if X and Y are connected by a single path containing an intermediate vertex $Z \in \mathbf{V}(G)$, $Z \neq X, Y$, then X and Y are conditionally independent given Z. This is the underlying idea of the following separation criterion (cf. [3]):

Definition 2 *(**u-separation**) Let $G = (\mathbf{V}(G), \mathbf{E}(G))$ be an undirected graph, and $\mathbf{X}, \mathbf{Y}, \mathbf{S} \subseteq \mathbf{V}(G)$ be disjoint sets of vertices. Then if each path between a vertex in $\mathbf{X}$ and a vertex*

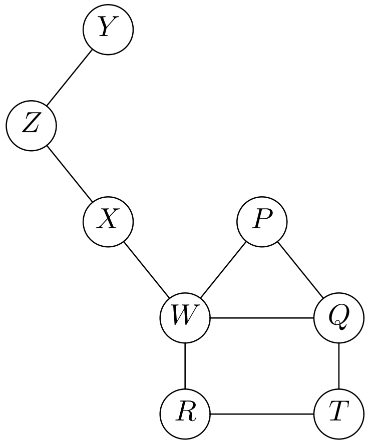


Figure 3: Graphical illustration of u-separation. Vertex P and vertices $\{R, T\}$ are u-separated by vertices $\{Q, W\}$, while vertex P and vertices $\{R, T\}$ are u-connected by Q.

in $\mathbf{Y}$ *contains a vertex in* $\mathbf{S}$*, then it is said that* $\mathbf{X}$ *and* $\mathbf{Y}$ *are* u-separated *by* $\mathbf{S}$*, denoted by* $\mathbf{X} \perp\!\!\!\perp_G \mathbf{Y} \mid \mathbf{S}$*. Otherwise, it is said that* $\mathbf{X}$ *and* $\mathbf{Y}$ *are* u-connected *by* $\mathbf{S}$*, denoted by* $\mathbf{X} \not\!\perp\!\!\!\perp_G \mathbf{Y} \mid \mathbf{S}$*.*

The basic idea of u-separation can be illustrated by Figure 3; for example, P is u-separated from $\{R, T\}$ by $\{Q, W\}$, i.e. $P \perp\!\!\!\perp_G \{R, T\} \mid \{Q, W\}$, whereas P and $\{R, T\}$ are u-connected by vertex Q, i.e. $P \not\!\perp\!\!\!\perp_G \{R, T\} \mid \{Q\}$.

The independence relation represented by means of an ADG can be uncovered by means of one of the following two procedures:

- *d-separation*, as introduced by Pearl (cf. [10]);
- *moralisation*, as introduced by Lauritzen (cf. [7]).

First we discuss d-separation based on Ref. [9]. Let the distinct vertices $X, Y, Z \in \mathbf{V}(G)$ constitute an induced subgraph of the ADG $G = (\mathbf{V}(G), \mathbf{A}(G))$, with $(X \cdots Z), (Y \cdots Z) \in \mathbf{A}(G)$ and X and Y are non-adjacent. Because the direction of the arcs between X, Z and Y, Z is unspecified, there are four possible induced subgraphs, which we call *connections*, illustrated in Figure 4.[2] These four possible connections offer the basis for the representation of conditional dependence and independence in ADGs. The two serial connections shown in Figure 4(a) and Figure 4(b) represent exactly the same independence information; this is also the case for the divergent connection represented in Figure 4(c). Figure 4(d) illustrates the situation where random variables X and Y are initially independent, but become dependent once random variable Z is instantiated.

Let $\mathbf{S} \subseteq \mathbf{V}(G)$, and $X, Y \in (\mathbf{V}(G) \setminus \mathbf{S})$ be distinct vertices, which are connected to each other by the trail τ. Then τ is said to be *blocked* by $\mathbf{S}$ if one of the following conditions is satisfied:

- $Z \in \mathbf{S}$ appears on the trail τ, and the arcs of τ meeting at Z constitute a serial or divergent connection;

[2] The terminology used in Figure 4 varies in different papers. Here the meaning of serial connection corresponds to *head-to-tail meeting*, divergent connection to *tail-to-tail* meeting and convergent connection to *head-to-head* meeting.

- $Z \notin \mathbf{S}$, $\delta(Z) \cap \mathbf{S} = \varnothing$ and the arcs meeting at Z on τ constitute a convergent connection, i.e. if Z appears on the trail τ then neither Z nor any of its descendants occur in $\mathbf{S}$.

The notion of d-separation exploits this notion of blocking, taking into account that vertices can be connected by more than one trail:

Definition 3 ***(d-separation)*** *Let $G = (\mathbf{V}(G), \mathbf{A}(G))$ be an ADG, and let $\mathbf{X}, \mathbf{Y}, \mathbf{S} \subseteq \mathbf{V}(G)$ be disjoint sets of vertices. Then $\mathbf{X}$ and $\mathbf{Y}$ are said to be* d-separated *by $\mathbf{S}$, denoted by $\mathbf{X} \perp\!\!\!\perp_G^d \mathbf{Y} \mid \mathbf{S}$, if each trail τ in G between each $X \in \mathbf{X}$ and each $Y \in \mathbf{Y}$ is blocked by $\mathbf{S}$; otherwise, $\mathbf{X}$ and $\mathbf{Y}$ are said to be* d-connected *by $\mathbf{S}$, denoted by $\mathbf{X} \not\perp\!\!\!\perp_G^d \mathbf{Y} \mid \mathbf{S}$.*

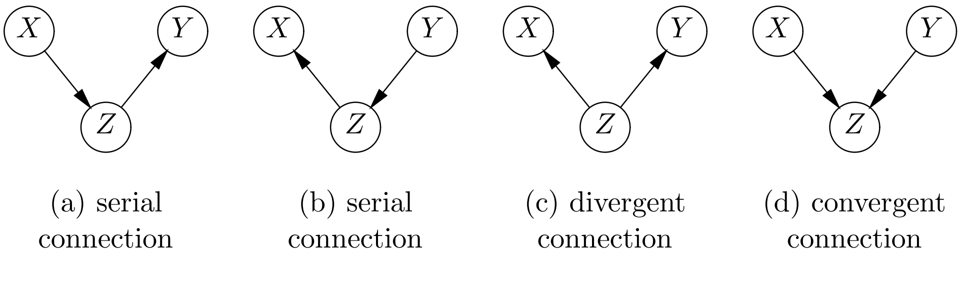
As an example, consider the graph in Figure 5(a), where the vertices Z and P are connected by the following three trails: $\tau_1 = Z \to X \to W \leftarrow P$; $\tau_2 = Z \to X \to W \to Q \leftarrow P$ and $\tau_3 = Z \to X \to W \to R \to T \leftarrow Q \leftarrow P$. Then trail τ_1 is blocked by $\mathbf{S} = \{X, Y\}$; since Y does not appear on this trail and the arcs on τ_1 meeting at X form a serial connection. As X blocks τ_2 and τ_3 following Definition 3, we conclude that $\mathbf{S}$ d-separates Z and P. On the other hand, neither $\mathbf{S}' = \{Y, W\}$ nor $\mathbf{S}'' = \{Y, T\}$ block τ_1, because $X \to W \leftarrow P$ is a convergent connection, $W \in \mathbf{S}'$; and T is a descendant of vertex W which occurs in $\mathbf{S}''$; it also participates in a convergent connection with respect to τ_3. Thus not every trail between Z and P in G is blocked by $\mathbf{S}'$ or $\mathbf{S}''$, and Z and P are d-connected by $\mathbf{S}'$ or $\mathbf{S}''$.

Next, we discuss the procedure of moralisation. Recall that the procedure of moralisation of a graph G consists of two steps:

(i) non-adjacent parents of a common chain component become connected to each other by an edge, and

(ii) each arc becomes an edge by removing its direction, resulting in an undirected graph.

An example of moralisation is presented in Figure 5. Since acyclic directed graphs are chain graphs, moralisation can also be applied to ADGs, where each chain component contains exactly one vertex. Observe that during the first step of the moralisation procedure, there may be extra edges inserted into the graph. Since edges between vertices create a dependence between random variables, vertices which became connected in the first step have a dependence relation in the resulting undirected graph. For example, as X and P have a common child W, and R and Q have a common child T, the graph in Figure 5(a) is extended by two extra edges $X - P$ and $R - Q$. The resulting graph after the first step of moralisation is depicted in Figure 5(b). The moral graph, obtained by replacing arcs by edges, is shown in Figure 5(c).



(a) serial connection

(b) serial connection

(c) divergent connection

(d) convergent connection

Figure 4: The four possible connections for acyclic directed graph $G = (\mathbf{V}(G), \mathbf{A}(G))$ given vertices $X, Y, Z \in \mathbf{V}(G)$ with arcs $(X \cdots Z), (Y \cdots Z) \in \mathbf{A}(G)$.

Observe that moralisation transforms the independences and dependences represented by d-separation (d-connection) into u-separation (u-connection). In the resulting moral graph, the vertices X and P and the vertices R and Q have become dependent of one another, and thus, some independence information is now lost. This independence information, however, can still be represented in the underlying joint probability distribution such that it still holds that $Z \perp\!\!\!\perp_P P \mid Y$. *However, it is also possible to parametrise the moralisation procedure on the vertices which potentially gives rise to extra dependences.* This possibility is a consequence of the meaning of a convergent connection $X \rightarrow Z \leftarrow Y$, because X and Y are independent if Z is not instantiated, and only become dependent if we know Z. If this is the case, and if we also assume that we know X (or Y), the dynamically created dependence between X and Y gives rise to a type of reasoning known as *explaining away* [11]: Y (X) becomes less or more likely if we know for certain that X (Y) is the cause of Z.

The moralisation procedure takes the presence of created dependences into account by means of the ancestral set, introduced in Section 2.1. Hence, this form of moralisation preserves all relevant (independence) information represented in the original ADG. The correspondence between d-separation and u-separation after moralisation is established in the following proposition:

Proposition 1 *Let $G = (\mathbf{V}(G), \mathbf{A}(G))$ be an acyclic directed graph with disjoint sets of vertices $\mathbf{X}, \mathbf{Y}, \mathbf{S} \subseteq \mathbf{V}(G)$. Then $\mathbf{X}$ and $\mathbf{Y}$ are d-separated by $\mathbf{S}$ iff $\mathbf{X}$ and $\mathbf{Y}$ are u-separated in the moral graph $G^m_{an(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{S})}$, where $an(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{S})$ is the smallest ancestral set of $\mathbf{X} \cup \mathbf{Y} \cup \mathbf{S}$.*

Proof: See Ref. [5], page 72. □

Figure 6 illustrates Proposition 1 by means of the conditional independence $Z \perp\!\!\!\perp^d_G P \mid \{X, Y\}$ and the conditional dependence $Z \not\!\perp\!\!\!\perp^d_G P \mid \{Y, W\}$ represented in the graph shown in Figure 5(a). We start by investigating the conditional independence $Z \perp\!\!\!\perp^d_P P \mid \{X, Y\}$. The smallest ancestral set of $\{Z\} \cup \{P\} \cup \{X, Y\}$ is $an(\{Z, P, X, Y\}) = \{Z, P, X, Y\}$; the graph depicted in Figure 6(a) contains all vertices of $an(\{Z, P, X, Y\})$ for the graph shown in Figure 5(a). We see that vertex P is disconnected from the subgraph $Y \rightarrow Z \rightarrow X$. The moral graph of this smallest ancestral set is shown in Figure 6(b). Observe that in graph (b) the vertices Z and P are (unconditionally) independent, as there is no path between them. Therefore, $Z \perp\!\!\!\perp^d_G P \mid \{X, Y\}$ still holds, although now as $Z \perp\!\!\!\perp_{G^m} P \mid \{X, Y\}$, as in the original graph in Figure 5(a). The situation where we wish to keep the conditional dependence $Z \not\!\perp\!\!\!\perp^d_P P \mid \{Y, W\}$ is illustrated by Figures 6(c) and 6(d). In Figure 6(c) the subgraph associated with the smallest ancestral set $an(Z \cup P \cup \{Y, W\}) = \{Z, P, X, Y, W\}$ is shown, and Figure 6(d) gives the resulting moral graph of Figure 6(c). In the graph (d) we can see that vertices Z and P are connected by a path, therefore, the created dependence between X and P is now represented in the moral graph of G.

Moralisation can also be applied to chain graphs; however, there is also another read-off procedure, called *c-separation*, introduced by Studený and Bouckhaert [8]. The concept of c-separation generalises both u-separation and d-separation.

The concept of c-separation takes into account the chain graph property that vertices may be connected by either edges or arcs. Let $G = (\mathbf{V}(G), \mathbf{E}(G))$ be a chain graph and let σ denote a section of the trail τ in G. Then σ is *blocked* by $\mathbf{S} \subseteq \mathbf{V}(G)$, if one of the following conditions holds:

- $Z \in \mathbf{S}$ appears on the section σ, where σ has one head and one tail terminal and every

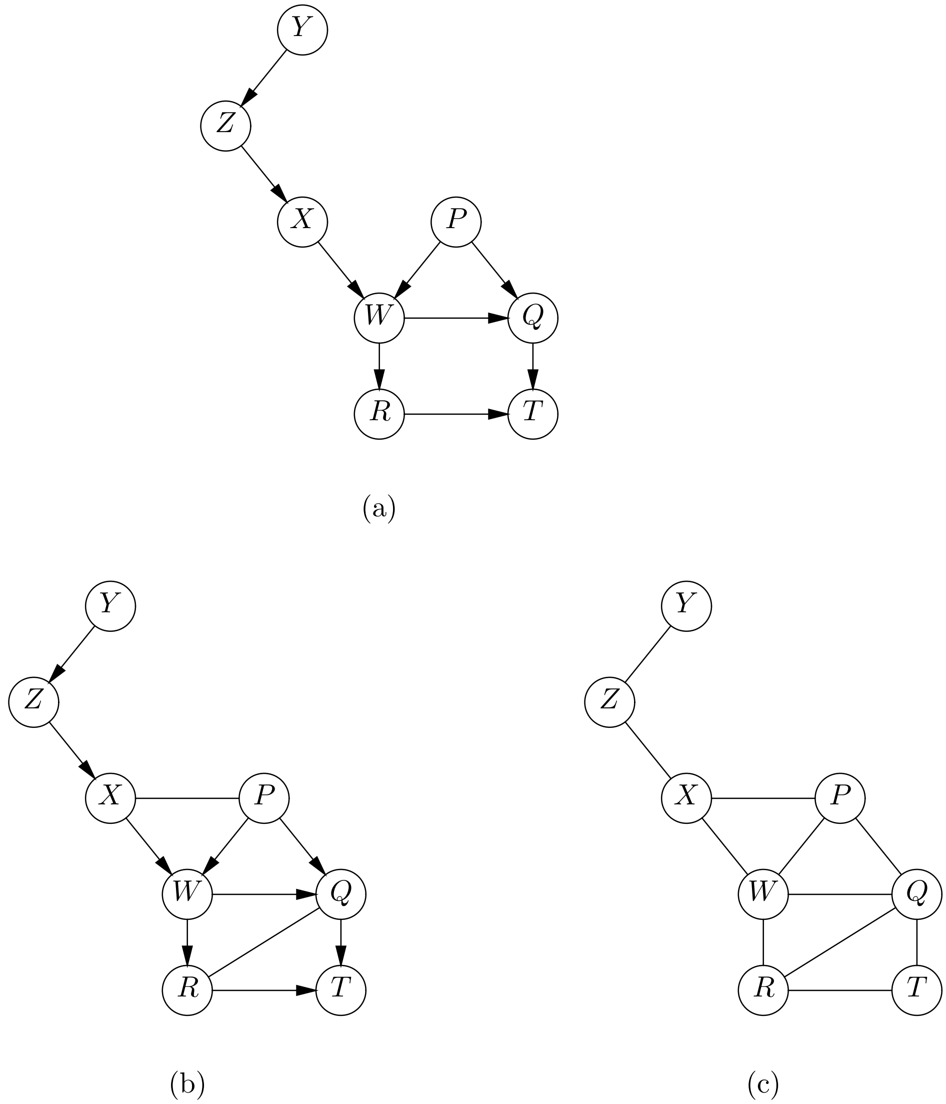
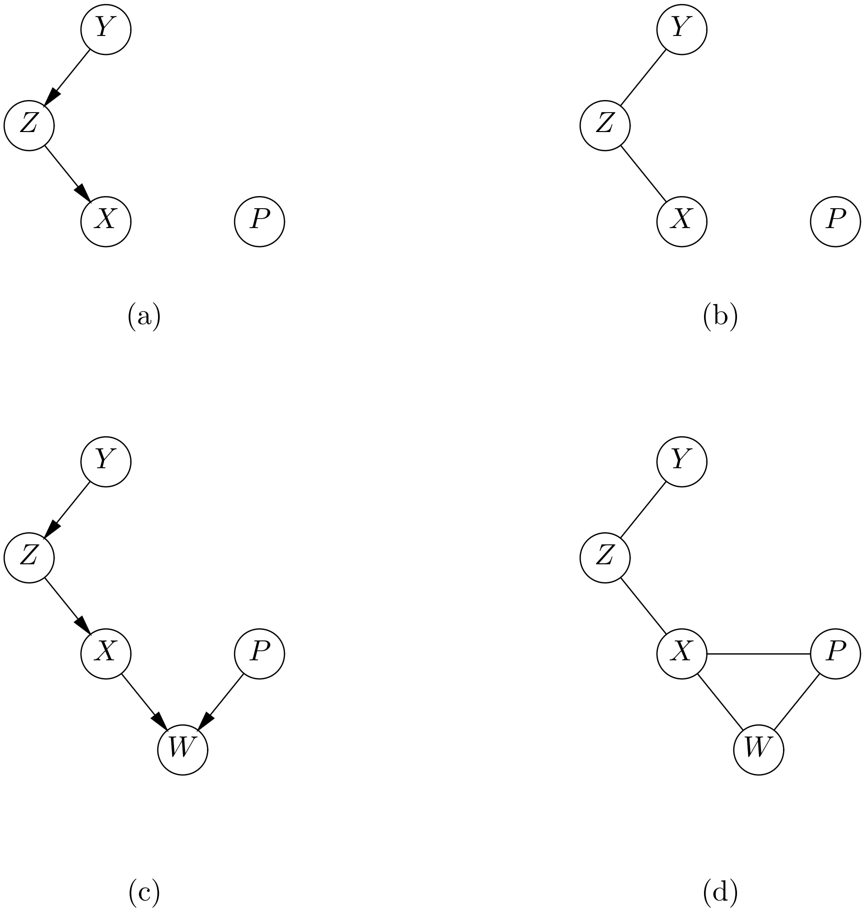


Figure 5: An example of the moralisation procedure as applied to the graph shown in Figure (a). Graph (b) depicts the resulting graph after application of the first step of moralisation. Note that the vertices X and P and the vertices R and Q are non-adjacent parents of the same child, therefore they became connected by an edge. Graph (c) results after changing the arcs in graph (b) into edges. Applying the definition of d-separation, it holds that $Z \perp\!\!\!\perp^d_G P \mid Y$ in graph (a); however for the moral graph in (c) we have that $Z \not\perp\!\!\!\perp_{G^m} P \mid Y$.



(a) (b) (c) (d)

Figure 6: Illustration of Proposition 1 with regard to the conditional independence $Z \perp\!\!\!\perp^d_G P \mid \{X, Y\}$ and the conditional dependence $Z \not\perp\!\!\!\perp^d_G P \mid \{Y, W\}$ which holds for the graph G shown in Figure 5(a). The induced subgraph H of this graph shown in Figure (a) above corresponds to $an(\{Z, P, X, Y\})$; in (b) its associated moral graph is represented. We see that the conditional independence $Z \perp\!\!\!\perp_{H^m} P \mid \{X, Y\}$ still holds. Figure (c) above shows the induced subgraph L of graph 5(a) corresponding to the smallest ancestral set of $\{Z\} \cup \{P\} \cup \{Y, W\}$. The graph L^m in (d) is the moral version of this graph. We see that the conditional dependence $Z \not\perp\!\!\!\perp_{L^m} P \mid \{Y, W\}$ holds for this moral graph. Hence, in both cases, all relevant (in)dependence information is preserved.

slide of the tail terminal is mediated by Z, or σ has two tail terminals and every slide of at least one of the two tail terminals is mediated by Z;

- $Z \in \mathbf{S}$ does *not* appear on the section σ, where σ has two head terminals and $Z \notin \delta(\sigma)$.

Based on these conditions we define the c-separation as follows.

Definition 4 ***(c-separation)*** *Let $G = (\mathbf{V}(G), \mathbf{E}(G))$ be a chain graph. Then two distinct vertices $X, Y \in \mathbf{V}(G)$ are* c-separated *by $\mathbf{S} \subseteq (\mathbf{V}(G)\backslash\{X, Y\})$, if at least one of the sections of each trail τ between the vertices X and Y is blocked by $\mathbf{S}$, written as $X \perp\!\!\!\perp^{\kappa}_{G} Y \mid \mathbf{S}$. Otherwise, X and Y are* c-connected *by $\mathbf{S}$ and we write $X \not\!\perp\!\!\!\perp^{\kappa}_{G} Y \mid \mathbf{S}$.*

As an example we use the chain graph presented in Figure 7(a). We examine whether $Z \perp\!\!\!\perp^{\kappa}_{G} T \mid \{X, Q, R\}$ (i.e. we have $\mathbf{S} = \{X, Q, R\}$). The following three trails between Z and T will be investigated: $\tau_1 = Z \rightarrow X - W \leftarrow P \rightarrow Q \rightarrow T$ with sections $\sigma_{11} = X - W$ and $\sigma_{12} = Q$; $\tau_2 = Z \rightarrow X - W - Q \rightarrow T$ with section $\sigma_{21} = X - W - Q$ and $\tau_3 = Z \rightarrow X - W \rightarrow R \rightarrow T$ with sections $\sigma_{31} = X - W$ and $\sigma_{32} = R$. In trail τ_1 section $X - W$ has two head-terminals and because $X \in \mathbf{S}$ section σ_{11} does not block trail τ_1. In contrast to σ_{11}, σ_{12} has one head and one tail terminal (the terminals are both equal to vertex Q) and slide $P \rightarrow Q$ is mediated and therefore blocked by Q. Since a trail is blocked if at least one of its sections is blocked by $\mathbf{S}$, we conclude that trail τ_1 is blocked by $\mathbf{S} = \{X, Q, R\}$. Section $X - W - Q$ in trail τ_2 has one head and one tail terminal, and satisfies the first blocking condition, because the slides $P \rightarrow Q$ and $Z \rightarrow X - W - Q$ are both mediated by $Q \in \mathbf{S}$. Therefore, trail τ_2 is also blocked by $\mathbf{S}$. This is also the case for trail τ_3 with section $X - W$, which has one head and one tail terminal and slides $Z \rightarrow X - W \rightarrow R$ and $P \rightarrow Q - W \rightarrow R$ are both mediated by $R \in \mathbf{S}$, thus τ_3 is also blocked by $\mathbf{S}$. There are also other trails between vertices Z and T (e.g. $Z \rightarrow X - W - Q \leftarrow P \rightarrow Q - W \rightarrow R \rightarrow T$), which are not mentioned here, because their sections are the same as in trails τ_1, τ_2 and τ_3. Therefore, these trails are also blocked by $\mathbf{S}$. Thus, following Definition 4, the conditional independence relation contains $Z \perp\!\!\!\perp^{\kappa}_{G} T \mid \{X, Q, R\}$.

3.2 Markov properties of graphical models

The dependence and independence relations determined by a joint probability distribution defined on the random variables corresponding to the vertices of a graph are graphically represented by the so-called *Markov properties.* We start by examining Markov properties for chain graphs, and next consider Markov properties for undirected and acyclic directed graphs, as these are special cases of chain graphs.

Each *chain Markov property* introduced below is illustrated by an example based on the chain graph shown in Figure 7(a). Vertices in the figures are presented using various shades depending on the role they play in visualising conditional independence properties. As before, we have that, for example, $X \in \mathbf{V}(G)$ is a vertex, while $\mathbf{X} \subseteq \mathbf{V}(G)$ represents a set of vertices.

A chain graph $G = (\mathbf{V}(G), \mathbf{E}(G))$ is said to obey:

- the *pairwise chain Markov property*, relative to G, if for any non-adjacent disjoint pair $X, Y \in \mathbf{V}(G)$ with $Y \in \bar{\delta}(X)$:

$$X \perp\!\!\!\perp^{\kappa}_{G} Y \mid \bar{\delta}(X) \setminus \{Y\}. \tag{3}$$

Recall that $\bar{\delta}(X)$ is the set of non-descendants of X. For the corresponding example shown in Figure 7(b) it holds that $\bar{\delta}(X) = \{Z, Y, P, W\}$ and therefore this property expresses that $X \perp\!\!\!\perp_G^{\kappa} Y \mid \{Z, P, W\}$.

- the *local chain Markov property*, relative to G, if for any vertex $X \in \mathbf{V}(G)$:

  $$X \perp\!\!\!\perp_G^{\kappa} \bar{\delta}(X) \setminus bd(X) \mid bd(X). \tag{4}$$

  Figure 7(c) illustrates this property by $X \perp\!\!\!\perp_G^{\kappa} \{Y, P\} \mid \{Z, W\}$.

- the *global chain Markov property*, relative to G, if for any triple of disjoint sets $\mathbf{X}, \mathbf{Y}$, $\mathbf{Z} \subseteq \mathbf{V}(G)$:

  $$\mathbf{X} \perp\!\!\!\perp_G^{\kappa} \mathbf{Y} \mid \mathbf{Z}. \tag{5}$$

  Figure 7(d) includes the following example of the global chain Markov property: $\{X, W, P, Q, R, T\} \perp\!\!\!\perp_G^{\kappa} Y \mid Z$.

- the *block-recursive chain Markov property*, relative to G, if for any non-adjacent disjoint pair $X, Y \in \mathbf{V}(G)$:

  $$X \perp\!\!\!\perp_G^{\kappa} Y \mid \mathbf{C}(t^*) \setminus \{X, Y\}, \tag{6}$$

  where $\mathbf{C}(t)$ is the set of concurrent variables of $\mathbf{V}(t)$ and t^* is the smallest t with $X, Y \in \mathbf{C}(t)$. This is shown in Figure 7(e). The well-ordered partitioning of a chain graph is not unique. In our example we take the following order of the partitioning: $\{Y\} < \{Z\} < \{P\} < \{X, W, Q\} < \{R\} < \{T\}$, where corresponding to Section 2.1 $\mathbf{V}(1) = \{Y\}$, $\mathbf{V}(2) = \{Z\}, \ldots, \mathbf{V}(6) = \{T\}$ (hence, by this ordering $t = 6$). Then based on this partitioning, the block-recursive chain Markov property states for example that it holds that $X \perp\!\!\!\perp_G^{\kappa} R \mid \{Y, Z, W, P, Q\}$ with $t^* = 5$.

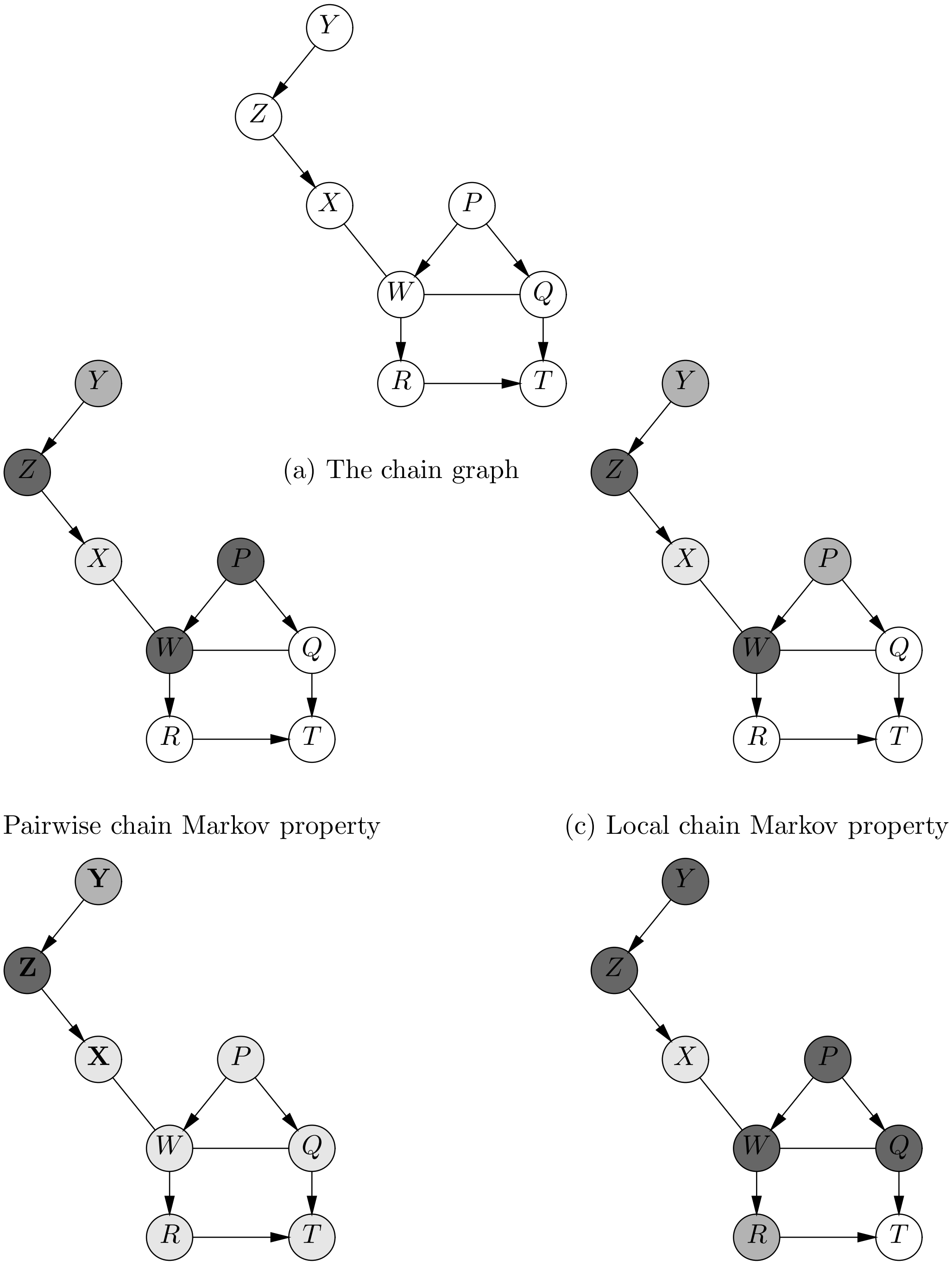
Based on the Markov properties for chain graphs, we will derive the related properties for undirected graphs, followed by acyclic directed graphs. As before, the undirected Markov properties are illustrated by means of figures. Here the graph in Figure 8(a) is taken as the example undirected graph.

Let $G = (\mathbf{V}(G), \mathbf{E}(G))$ be an undirected graph. Due to the fact that undirected graphs do not include arcs we cannot distinguish between ancestors and descendants of vertices; non-descendants $\bar{\delta}(X)$ in the chain Markov properties have to be replaced by the entire vertex set $\mathbf{V}(G)$. In addition, the block-recursive chain Markov property makes no sense for the undirected graphs, because they do not have directionality. The undirected graph G is said to obey:

- the *pairwise undirected Markov property*, relative to G, if for any non-adjacent vertices $X, Y \in \mathbf{V}(G)$:

  $$X \perp\!\!\!\perp_G Y \mid \mathbf{V}(G) \setminus \{X, Y\}. \tag{7}$$

  In this case the set of non-descendants $\bar{\delta}(X)$ from the chain property case is replaced by $\mathbf{V}(G)$. The example in Figure 8(b) shows that $X \perp\!\!\!\perp_G Y \mid \{Z, W, P, Q, R, T\}$.



(a) The chain graph

(b) Pairwise chain Markov property

(c) Local chain Markov property

(d) Global chain Markov property

(e) Block-recursive Markov property

Figure 7: Graphical illustration of the chain Markov properties, taking the chain graph in (a) as an example. Shown are (b): the pairwise chain Markov property $X \perp\!\!\!\perp_G^{\kappa} Y \mid \{Z, P, W\}$; (c): the local chain Markov property $X \perp\!\!\!\perp_G^{\kappa} \{Y, P\} \mid \{Z, W\}$; (d): the global chain Markov property $\{X, W, P, Q, R, T\} \perp\!\!\!\perp_G^{\kappa} Y \mid Z$; (e): the block-recursive Markov property $X \perp\!\!\!\perp_G^{\kappa} R \mid \{Y, Z, W, P, Q\}$.

- the *local undirected Markov property* relative to G, if for any $X \in \mathbf{V}(G)$:

  $$X \perp\!\!\!\perp_G \mathbf{V}(G) \setminus \mathit{cl}(X) \mid \mathit{bd}(X), \tag{8}$$

  where $\mathit{bd}(X)$ is the boundary or *undirected Markov blanket* (the minimal boundary) of X and $\mathit{cl}(X)$ is the closure of X defined in Section 2.1. As was mentioned above $\bar{\delta}(X)$ of the chain property is replaced by $\mathbf{V}(G)$. Observe that in the local chain Markov property the expression $\bar{\delta}(X) \setminus \mathit{bd}(X)$ does not contain the random variable X, which would not be the case for $\mathbf{V}(G) \setminus \mathit{bd}(X)$. Therefore, the boundary set $\mathit{bd}(X)$ is replaced by $\mathit{cl}(X)$. Graph (c) in Figure 8 depicts $X \perp\!\!\!\perp_G \{Y, P, Q, R, T\} \mid \{Z, W\}$, i.e. $\mathit{bd}(X) = \{Z, W\}$.

- the *global undirected Markov property*, relative to G, if for any triple of disjoint sets $\mathbf{X}, \mathbf{Y}$, $\mathbf{Z} \subseteq \mathbf{V}(G)$:

  $$\mathbf{X} \perp\!\!\!\perp_G \mathbf{Y} \mid \mathbf{Z}. \tag{9}$$

  For this property no changes need to be made with regard to the corresponding chain Markov property. An example for this property is given in Figure 8(d); here: $\{X, W, P, Q, R, T\} \perp\!\!\!\perp_G Y \mid Z$.

Finally we consider Markov properties for ADGs, i.e. *directed Markov properties*. These are visualised using the ADG shown in Figure 9(a) as a basis. For the acyclic directed graph $G = (\mathbf{V}(G), \mathbf{A}(G))$ the local and global directed Markov properties are derived from the local, respectively global chain Markov properties, replacing the boundary by the parents of a vertex. Furthermore, the local chain Markov property generalises the blanket directed Markov property, and the ordered directed Markov property is derived from the block-recursive chain Markov property. The ADG G is said to obey:

- the *local directed Markov property*, relative to G, if for any $X \in \mathbf{V}(G)$:

  $$X \perp\!\!\!\perp_G^d (\bar{\delta}(X) \setminus \pi(X)) \mid \pi(X). \tag{10}$$

  Note that the set $\mathit{bd}(X)$ from the chain property is replaced by $\pi(X)$ and, in addition, the expression $\bar{\delta}(X) \setminus \mathit{bd}(X)$ in the local chain Markov property is simplified to $\bar{\delta}(X) \setminus \pi(X)$. This property is illustrated in Figure 9(b); it expressed the conditional independence $X \perp\!\!\!\perp_G^d \{Y, P\} \mid Z$.

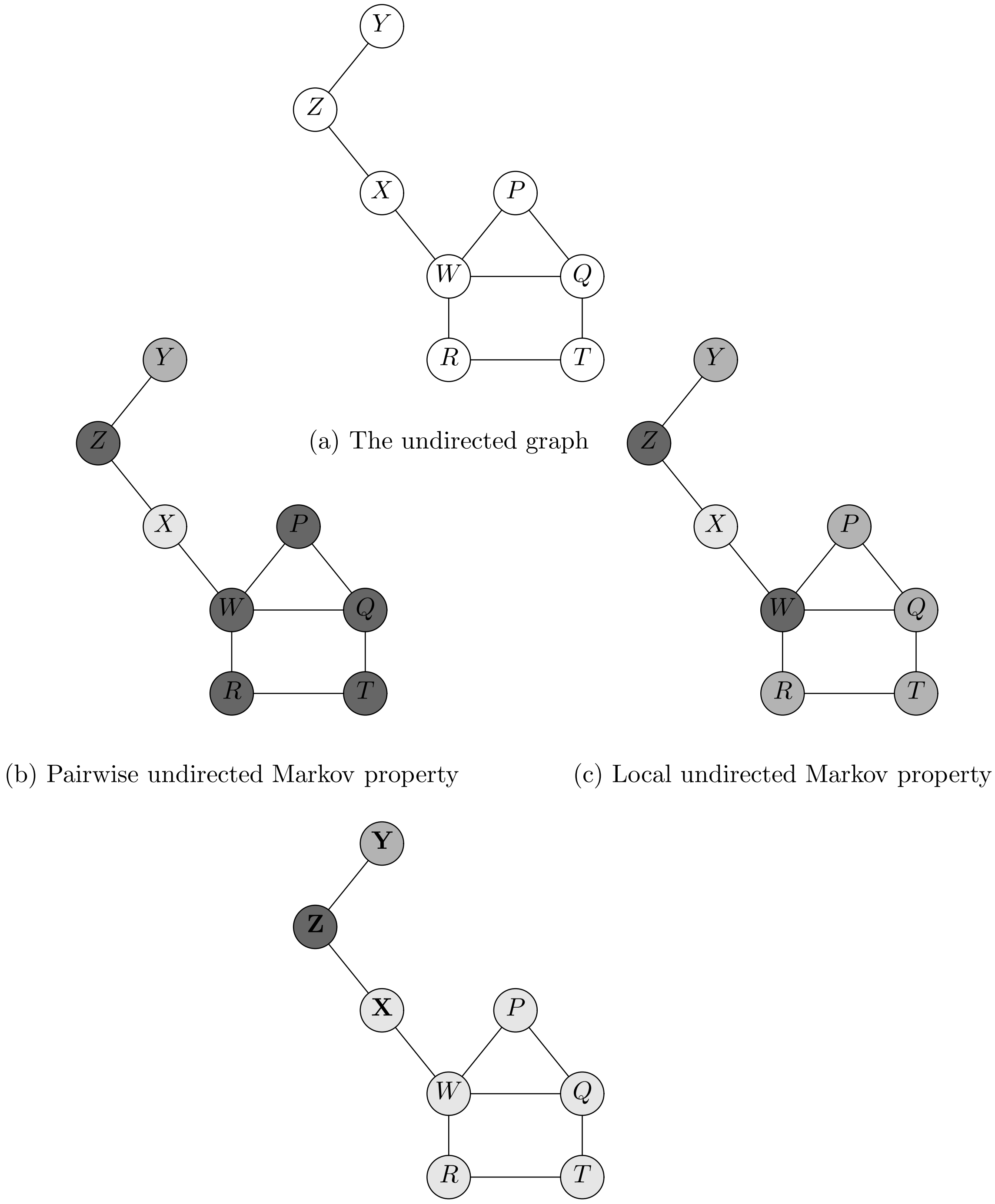
- the *blanket directed Markov property*, relative to G, which is derived from the local Markov property for chain graphs if we assume that for any $X \in \mathbf{V}(G)$:

  $$X \perp\!\!\!\perp_G^d \mathbf{V}(G) \setminus (\beta(X) \cup X) \mid \beta(X), \tag{11}$$

  where $\beta(X)$ is the *directed Markov blanket*, defined as follows:

  $$\beta(X) = \pi(X) \cup \mathit{ch}(X) \cup \{Y : \mathit{ch}(Y) \cap \mathit{ch}(X) \neq \varnothing; Y \in \mathbf{V}(G)\}. \tag{12}$$

  This property can be derived from the blanket undirected Markov property easily, as X's children, parents and children's parents constitute the directed Markov blanket. An example is given in Figure 9(c); here we have for example $X \perp\!\!\!\perp_G^d \{Y, Q, R, T\} \mid \{Z, W, P\}$.



(a) The undirected graph

(b) Pairwise undirected Markov property

(c) Local undirected Markov property

(d) Global undirected Markov property

Figure 8: Graphical illustration of the undirected Markov properties, taking the UG from (a) as an example. Shown are: (b): the pairwise undirected Markov property $X \perp\!\!\!\perp_G Y \mid \{Z, W, P, Q, R, T\}$; (c): the local undirected Markov property as $X \perp\!\!\!\perp_G \{Y, P, Q, R, T\} \mid \{Z, W\}$; (d): the global chain Markov property $\{X, W, P, Q, R, T\} \perp\!\!\!\perp_G Y \mid Z$.

- the *global directed Markov property*, relative to G, if for any triple of disjoint sets $\mathbf{X}, \mathbf{Y}$, $\mathbf{Z} \subseteq \mathbf{V}(G)$:

$$\mathbf{X} \perp\!\!\!\perp_G^d \mathbf{Y} \mid \mathbf{Z}. \tag{13}$$

  This property need not be changed. Graph (d) in Figure 9 illustrates this property; for example, we have: $\{X, W, P, Q, R, T\} \perp\!\!\!\perp_G^d Y \mid Z$.

- the *ordered directed Markov property*, relative to G, if for any $X \in \mathbf{V}(G)$:

$$X \perp\!\!\!\perp_G^d (pr(X) \setminus \pi(X)) \mid \pi(X), \tag{14}$$

  where $pr(X)$ denotes the predecessor set of X. This property can be derived from the block-recursive chain property by the following idea: the acyclicity of graph G provides a well-ordering of its vertex set, in which each vertex can be seen as a chain component containing exactly one element. Figure 9(e) gives an example; based on the well-ordering $Y < Z < P < X < W < Q < R < T$ it holds that $X \perp\!\!\!\perp_G^d \{Y, P\} \mid Z$.

3.3 D-map, I-map and P-map

In a graphical model it is not always the case that all independence information is represented, and it may also not be the case that all dependence information is represented. In this section the relationship between the representation of conditional dependence and independence by joint probability distributions and graphs is explored.

Let $\perp\!\!\!\perp_P$ be an independence relation defined on $\mathbf{V}$ for joint probability distribution P, then for each $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$, where $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$ are disjoint:

- G is called an undirected *dependence map*, *D-map* for short, if

$$\mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z} \;\Rightarrow\; \mathbf{X} \perp\!\!\!\perp_G \mathbf{Y} \mid \mathbf{Z},$$

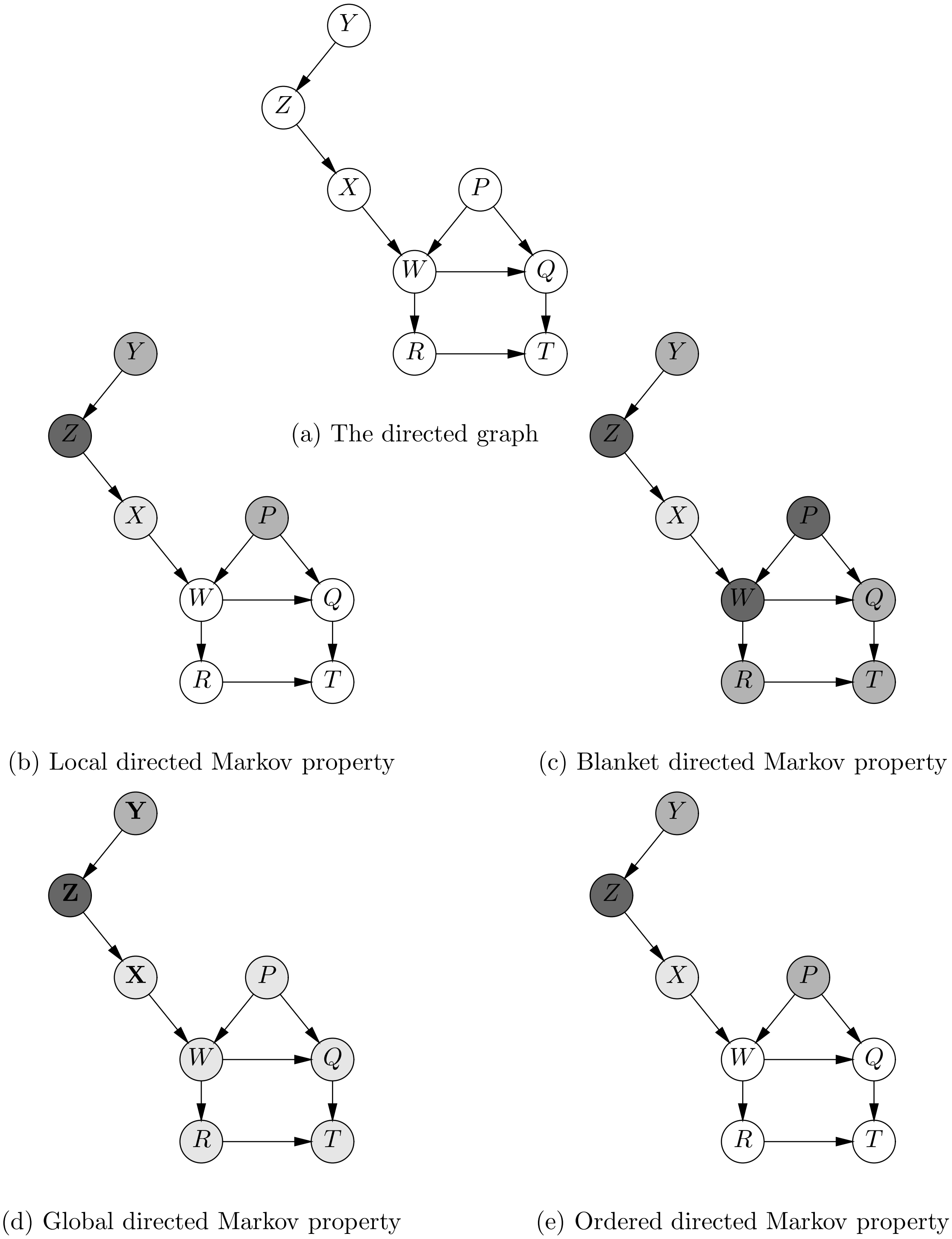
- G is called an undirected *independence map*, *I-map* for short, if

$$\mathbf{X} \perp\!\!\!\perp_G \mathbf{Y} \mid \mathbf{Z} \;\Rightarrow\; \mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z}.$$

- G is called an undirected *perfect map*, or *P-map* for short, if G is both a D-map and an I-map, or, equivalently

$$\mathbf{X} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z} \iff \mathbf{X} \perp\!\!\!\perp_G \mathbf{Y} \mid \mathbf{Z}.$$

Observe that in a D-map each independence encoded in the joint probability distribution P has to be represented in the graph G. Using the equivalence $a \Rightarrow b \equiv \neg b \Rightarrow \neg a$, it holds for D-maps that each dependence encoded by the graph G has to be represented in the joint probability distribution P. This does not mean that each dependence represented in the joint probability distribution P is also discerned in the D-map. In contrast to D-maps, in I-maps each independence relationship modelled in the graph G has to be consistent with the joint probability distribution P and each dependence relationship represented in the joint



(a) The directed graph

(b) Local directed Markov property

(c) Blanket directed Markov property

(d) Global directed Markov property

(e) Ordered directed Markov property

Figure 9: Graphical illustration of the acyclic directed Markov properties, taking the ADG shown in (a) as an example. Shown are (b): the local directed Markov property $X \perp\!\!\!\perp^d_G \{Y, P\} \mid Z$; (c): the blanket directed Markov property $X \perp\!\!\!\perp^d_G \{Y, Q, R, T\} \mid \{Z, W, P\}$; (d): the global directed Markov property $\{X, W, P, Q, R, T\} \perp\!\!\!\perp^d_G Y \mid Z$; (e): the ordered directed Markov property $X \perp\!\!\!\perp^d_G \{Y, P\} \mid Z$.

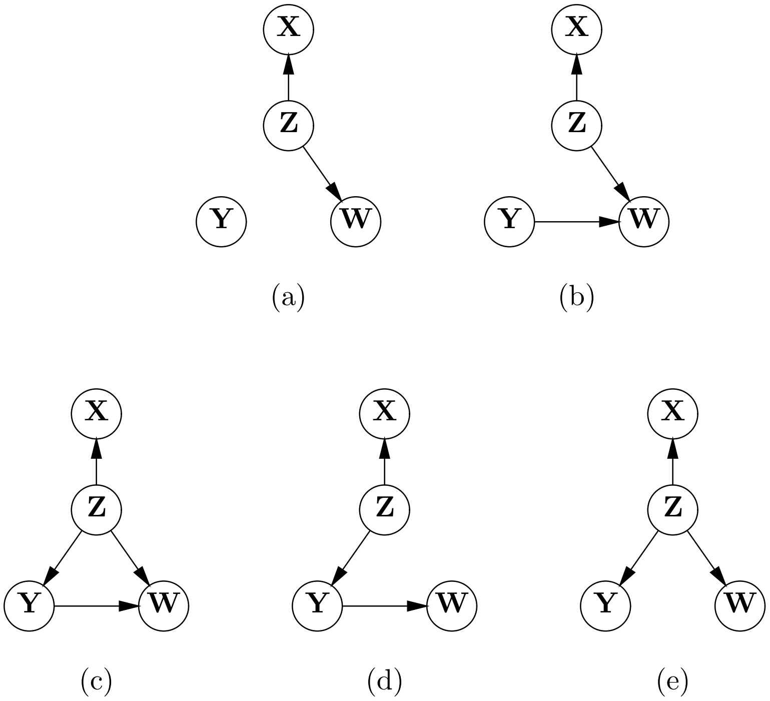


Figure 10: Given the joint probability distribution $P(X, Y, Z, W) = P(X \mid Z)\ P(Y \mid Z)P(W \mid Z)P(Z)$ with conditional independence set: $X \perp\!\!\!\perp_P \{Y, W\} \mid Z$, $Y \perp\!\!\!\perp_P \{X, W\} \mid Z$ and $W \perp\!\!\!\perp_P \{X, Y\} \mid Z$, graph (a) is a D-map, graph (b) is neither a D-map nor an I-map, graph (c) is an I-map, graph (d) is neither a D-map nor an I-map and graph (e) is a perfect map.

probability distribution P has to be present in the graph representation G. Clearly, a perfect map is just a combination of a D-map and an I-map.

The notions of D-map, I-map and P-map can easily be adapted to similar notions for ADGs and chain graphs, and thus we will not include the definitions here. Consider the following example, illustrated by Figure 10. Let $\mathbf{V} = \{X, Y, Z, W\}$ be the set of random variables with joint probability distribution: $P(X, Y, Z, W) = P(X \mid Z)P(Y \mid Z)P(W \mid Z)P(Z)$. The associated conditional independence set consists of three members: $X \perp\!\!\!\perp_P \{Y, W\} \mid Z$, $Y \perp\!\!\!\perp_P \{X, W\} \mid Z$ and $W \perp\!\!\!\perp_P \{X, Y\} \mid Z$. Then, the graph in Figure 10(a) is a D-map of P whereas graph (b) is not a D-map of P, since it describes a dependence $Y \rightarrow W$, which is not in P. Graph (b) is also not an I-map, since it does not include the arc $Z \rightarrow Y$. Graph (c) is an I-map of P but not a perfect map, because it includes the dependence $Y \rightarrow W$, which is not part of P. Graph (d) is not an I-map by the fact that it does not represent the dependence between vertices Z and W (i.e. it does not contain arc $Z \rightarrow W$) and it is also not a D-map. Graph (e) is a perfect map of the joint probability distribution P. In the remainder of this section we investigate the correspondence between the above-mentioned properties of conditional independence and undirected, respectively, directed perfect maps. The following theorem establishes conditions for the existence of an undirected perfect map for any joint probability distribution.

Theorem 1 *The conditional independence relations associated with a joint probability distribution P need to satisfy the necessary and sufficient conditions of (i) symmetry, (ii) decomposition, (iii) intersection, (iv) strong union, and (v) strong transitivity, to allow their* *representation as an undirected perfect map.*

As mentioned above, any joint probability distribution obeys the semi-graphoid properties (symmetry, decomposition, weak union and contraction). According to Theorem 1 in addition to the properties of symmetry and decomposition, the properties of intersection, strong union and strong transitivity should hold, which however, are not semi-graphoid properties. Thus, not every joint probability distribution will have a corresponding undirected graphical representation as a perfect map. Furthermore, for directed perfect maps we have a number of necessary conditions, but these are not always sufficient.

Theorem 2 *Necessary conditions for the conditional independence relations associated with a joint probability distribution P to allow representation as a directed perfect map are: (i) symmetry, (ii) contraction, (iii) decomposition, (iv) weak union, (v) intersection, (vi) weak transitivity, and (vii) chordality.*

Theorem 2 indicates that similar to the undirected case, the independence relations corresponding to a joint probability distribution need not always allow representation as a directed perfect map. In many practical situations, it will not be possible to find a perfect map of a joint probability distribution. Therefore we wish to focus on graphical representations that are as sparse as possible, and thus do not encode spurious dependences, which is something offered by *minimal* I-maps.

Definition 5 ***(minimal I-map)*** *A graph is called a* minimal I-map *of the set of independence relations of the joint probability distribution P, if it is an I-map and removing any arc of the graph will yield a graph which is no longer an I-map.*

Minimising the number of arcs in a graphical model is not only important for representation reasons, i.e. in order to keep the amount of probabilistic information that has to be specified to the minimum, but also for computational reasons. It has been shown that every joint probability distribution P for which the conditional independence relations satisfy the conditions of symmetry, decomposition, and intersection has a minimal undirected I-map, whereas any joint probability distribution P with associated conditional independence relations satisfying the conditions of symmetry, decomposition, weak union and contraction has a minimal directed I-map representation [3]. This implies that each graphoid has a corresponding minimal undirected I-map, as well as a minimal directed I-map, and each semi-graphoid has a minimal directed I-map as graphical representation. As for every joint probability distribution the semi-graphoid properties hold, we can conclude that each joint probability distribution has a directed minimal I-map.

4 Equivalence of Bayesian networks

In this section we return to the question which acted as the main motivation for writing this paper: how can equivalence of Bayesian networks be characterised best? It appears that in particular the concept of essential graphs plays a pivotal role in this. Before discussing essential graphs, we start by reviewing the definition of a Bayesian network in Section 4.1. Subsequently, in Sections 4.2 and 4.3, the equivalence relation on Bayesian networks which forms the basis for the concept of essential graphs will be studied.

4.1 Bayesian networks

Formally, a *Bayesian network* is a pair $\mathcal{B} = (G, P)$, where $G = (\mathbf{V}(G), \mathbf{A}(G))$ is an acyclic directed graph and P is a joint probability distribution defined on a set of random variables $\mathbf{X}$. As before, we assume that there is a one-to-one correspondence between the vertices $\mathbf{V}(G)$ and the random variables $\mathbf{X}$ associated with P, i.e. $\mathbf{V}(G) \leftrightarrow \mathbf{X}$. Therefore, we will indistinguishably use $\mathbf{V}(G)$ both to refer to vertices of the graph G and random variables of the associated joint probability distribution P of $\mathcal{B}$. Which of these two interpretations is intended will become clear from the context.

As mentioned above, the set of arcs $\mathbf{A}(G)$ describes the dependence and independence relationships between groups of vertices in $\mathbf{V}(G)$ corresponding to random variables. If a joint probability distribution P admits a recursive factorisation then P can be defined on the set of random variables $\mathbf{V}(G)$ as follows:

$$P(\mathbf{V}(G)) = \prod_{V \in \mathbf{V}(G)} P(V \mid \pi(V)). \tag{15}$$

Equation (15) implies that a joint probability distribution over a set of random variables can be defined in terms of local (conditional) joint probability distributions $P(V \mid \pi(V))$. Considerable research efforts have been made to exploit the structure of such a joint probability distribution for achieving computational savings. A Bayesian network is by definition a directed I-map.

What is interesting about Bayesian networks, and which is a main difference between directed and undirected graphical models, is that by instantiating vertices in the directed structure independences may change to dependences, i.e. stochastic independence has specific *dynamic* properties. In Section 3.1 we have called the type of reasoning associated with this 'explaining away'. This dynamic property is illustrated by Figure 4(d), where random variables X and Y are independent of one another if random variable Z is unknown, but as soon as Z becomes instantiated, a dependence between X and Y is created. However, similar to undirected graphs, part of the independence information represented in the graphical part of a Bayesian network is *static*. The structure of a Bayesian network allows reading off independence statements, essentially by using the notions of d-separation and moralisation treated in the previous section.

Our motivation to study the Markov properties associated with graphs arises from our wish to understand the various aspects regarding the representation of independence in Bayesian networks. The following proposition establishes a very significant relationship between Markov properties on the one hand, and joint probability distributions on the other hand; it is due to Lauritzen [5]:

Proposition 2 *If the joint probability distribution admits a recursive factorisation according to the acyclic directed graph $G = (\boldsymbol{V}(G), \boldsymbol{A}(G))$, it factorises according to the moral graph G^m and therefore obeys the global Markov property.*

Proof: See Ref. [5], page 70. □

Proposition 2 implies an important correspondence between a recursive factorisation according to graph G and the global Markov property. This proposition can be extended resulting in the following theorem, also by Lauritzen [5]:

Theorem 3 *Let $G = (\mathbf{V}(G), \mathbf{A}(G))$ be an acyclic directed graph. For the joint probability distribution P the following conditions are equivalent:*

- *P admits a recursive factorisation according to G;*
- *P obeys the global directed Markov property, relative to G;*
- *P obeys the local directed Markov property, relative to G;*
- *P obeys the ordered directed Markov property, relative to G.*

Proof: See Ref. [5], page 74. □

Theorem 3 establishes the relation between a recursive factorisation of joint probability distribution P and the directed Markov properties introduced in Section 3.2, and therefore explains why the Markov properties and their relations are relevant in the context of Bayesian networks and thus to structure learning.

4.2 The equivalence relation on acyclic directed graphs

In this section we introduce some notions required to study equivalence among Bayesian networks. We start by the definition of Markov constraints [15].

Definition 6 ***(Markov constraints)*** *Let $G = (\boldsymbol{V}(G), \boldsymbol{A}(G))$ be an ADG. Then the* Markov independence constraints, Markov constraints *for short, are the set of independence relations defined by the global directed Markov property.*

The Markov independence constraints allow us to define an equivalence relation on ADGs, as follows:

Definition 7 ***(Markov equivalent)*** *Two ADGs are* Markov equivalent *if they have the same set of Markov constraints.*

However, this definition is far removed from a procedural recipe: it is difficult to imagine how we can actually determine whether two ADGs are equivalent without enumerating all triples in the independence relations defined using these graphs. However, the following two definitions allow us to look at the problem from a different, and practically more useful, angle.

Definition 8 ***(skeleton)*** *Let G be an ADG. The undirected version of G is called the* skeleton *of G.*

For example, the graph in Figure 11(b) is the skeleton of graph (a).

Definition 9 ***(immorality)*** *An induced subgraph in an ADG G with $X, Y, Z \in \mathbf{V}(G)$ is called an* immorality, *if the graph contains the arcs $X \rightarrow Z$ and $Y \rightarrow Z$, and vertices X and Y are non-adjacent.*

Definition 9 implies that the concept of immorality is equivalent to that of convergence connection (cf. Figure 4(d)); both describe conditional dependence between random variables. (Immorality is synonymous with *v-structure* introduced by Verma and Pearl (cf. [15]).) However, immoralities are also the smallest induced subgraphs for the representation of conditional

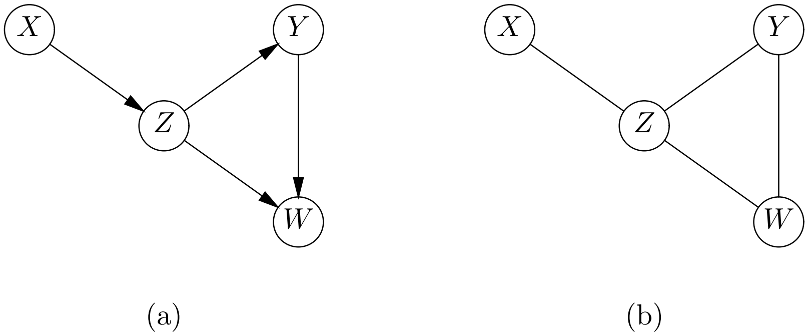


Figure 11: An acyclic directed graph (a) and its skeleton (b).

dependence. Observe that if the direction of one or both arcs of an immorality is reversed, the conditional dependence would turn into conditional independence, and thus would destroy the original meaning of the graph. Therefore to keep the independence relation defined on the random variables unchanged, it is not allowed to reverse the direction of these arcs.

Definition 10 *(**essential arcs**) Arcs that cannot be reversed without changing the conditional dependence and independence relations are called* essential arcs.

By Definition 10 both arcs of an immorality are essential arcs.

Applying Definition 8 and Definition 9, Markov equivalence is redefined in terms of the concepts of skeleton and immoralities by the following theorem, originally introduced by Verma and Pearl, which establishes the connection between these notions (cf. [15]):

Theorem 4 *Two ADGs are* Markov equivalent *with each other if and only if they have the same skeleton and they consist of the same set of immoralities.*

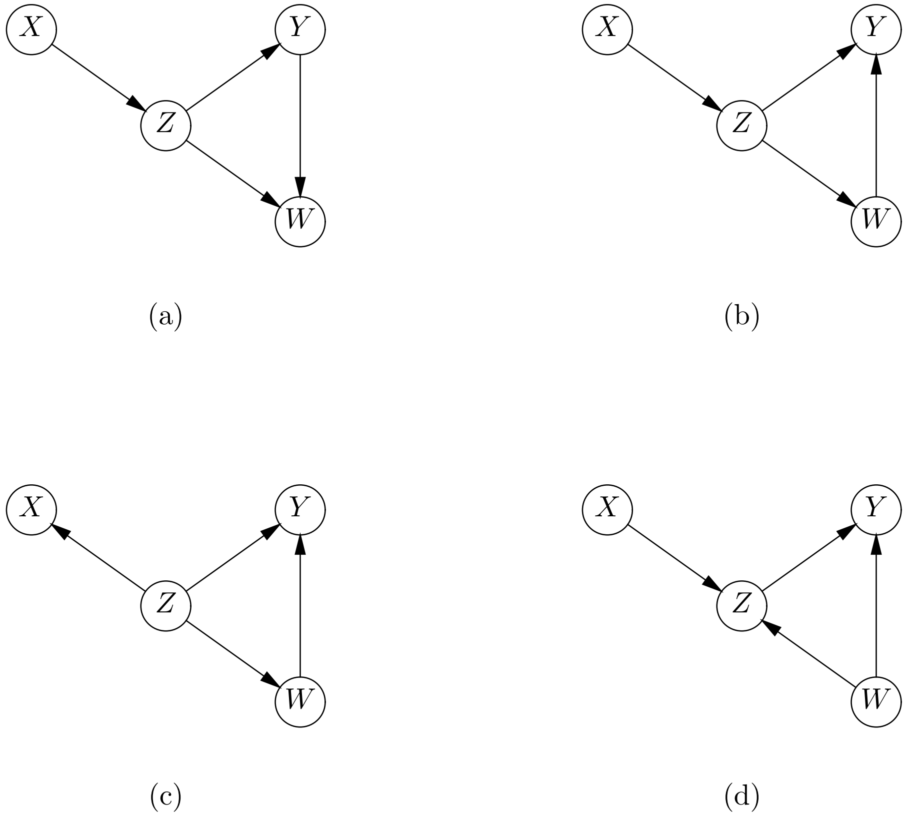
An example of Markov equivalence is given in Figure 12. Graph (a), (b) and (c) are equivalent by Theorem 4, but graph (d) is not equivalent to graphs (a), (b) and (c) since it contains, in contrast to the other graphs, the immorality $X \rightarrow Z \leftarrow W$.

Let us try to explain why Theorem 4 plays a significant role in the field of structure learning. Recall that an immorality describes an independence relationship between random variables and it is also the smallest induced subgraph reflecting conditional dependence. The purpose of structure learning is to find the relations between the random variables of the problem domain based on the data. Thus, if we have the entire set of independence relationships (the Markov constraints) or the entire set of dependence relationships over the random variables our aim has been achieved. In the graphical representation of dependence there are two kinds of dependences that can be distinguished:

(i) static, and

(ii) dynamic dependence.

By static dependences we mean the existence of a direct connection (i.e. an arc or edge) between vertices. Since the joint probability distribution on two static dependent random variables $X \rightarrow Y$ is the same as $Y \rightarrow X$, according to Bayes' theorem, this dependence can be represented by an arc. In contrast to static dependences, dynamic dependences are conditionally dependent on the instantiation of random variables associated with vertices with convergent connections. Therefore, these arcs have to preserve their direction. This is exactly what is said by Theorem 4.



(a) (b) (c) (d)

Figure 12: An example of Markov equivalence. Graph (a), (b) and (c) are equivalent since they have the same skeleton and the same set of immoralities. Graph (d) has also the same skeleton as graph (a), (b) and (c), but graph (d) also contains an immorality $X \rightarrow Z \leftarrow W$ which does not occur in the other graphs. Therefore graph (d) is not equivalent to graph (a), (b) and (c).

4.3 Essential graphs

Taking Theorem 4 as a foundation, in this section we will study the important problem of equivalence of ADGs. Recall that equivalent ADGs have the same immoralities, and these immoralities consist of essential arcs, which in each equivalent ADG have the same direction. In contrast, if one wishes to build an ADG from a skeleton and a collection of immoralities, there are normally different choices possible for edges which do not participate in an immorality, to the extent that choices that give rise to a directed cycle or to a new immorality are not allowed. We can therefore conclude that the difference between equivalent ADGs is entirely based on the difference in the direction of their non-essential arcs.

It now appears that classes of Markov equivalent ADGs can be uniquely described by means of chain graphs, called essential graphs, which thus act as class representatives [1]; they are defined as follows:

Definition 11 ***(essential graph)*** *Let* $[G]$ *denote the equivalence class of ADGs that are Markov equivalent. The* essential graph G^* *is then the smallest graph larger than any of the ADGs* G *in the equivalence class* $[G]$*; formally:*

$$G^* := \bigcup \{G \mid G \in [G]\}. \tag{16}$$

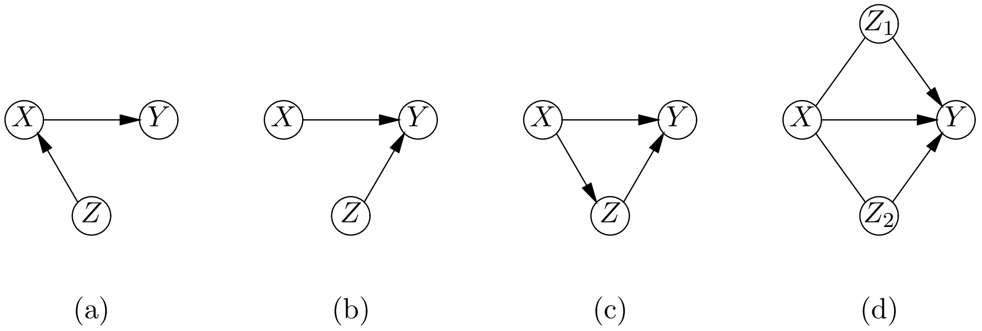


Figure 13: The four possible induced subgraphs, where arc $X \to Y$ is strongly protected.

This definition implies that any of the non-essential arcs in any of the ADGs $G \in [G]$ is replaced by an edge (which means that to the arc $(X, Y) \in \mathbf{A}(G)$ an arc (Y, X) is added), and this explains why an essential graph is as large or larger than any of the members of the equivalence class $[G]$ which it represents. Of course, as an essential graph is a chain graph, it may not be (and usually is not) an ADG, and therefore usually not a member of the equivalence class it represents.

It has been established that if an arc is part of a particular subgraph with a specific structure, then we know that the arc must be essential. There are four different (sub)graphs where $X \to Y$ will always be an essential arc; these are shown in Figure 13. As mentioned above, a serial or divergent connection mirrors conditional independence, while a convergent connection reflects a potential dependence relationship between random variables (see Figure 4). Clearly, it is not allowed to express a dependence represented in the ADGs of an equivalence class as an independence in the associated essential graph, and vice versa. This is illustrated by the subgraphs (a) and (b) in Figure 13. Case (a) means that we have a serial connection, which would be turned into convergent connection if the direction of $X \to Y$ is reversed. Therefore $X \to Y$ is an essential arc. In contrast, changing the direction of $X \to Y$ in case (b) would destroy an immorality, as a convergent connection would be changed into a serial connection. Even though any graph $G \in [G]$ is acyclic, reversing an arc might create a directed cycle. Clearly, reversing the direction of such arcs is not allowed, i.e. it is also an essential arc. This is shown in Figure 13(c). Finally, in case (d) $X \to Y$ is an essential arc and the two other essential arcs $Z_1 \to Y$ and $Z_2 \to Y$ are participating in the immorality $Z_1 \to Y \leftarrow Z_2$ (i.e. they are irreversible), the direction of the arc $X \to Y$ cannot be reversed to ensure that vertices Z_1 and Z_2 will not become dependent when conditioning on X.

Definition 12 ***(strongly protected)*** *An arc $X \to Y$ is called* strongly protected *if it is part of one of the four induced subgraphs shown in Figure 13.*

In the next part of this section we turn our attention to the characterisation of essential graphs G^*. First of all we consider two of the most significant properties of essential graphs.

Lemma 1 *The essential graph G^* representing the equivalence class $[G]$ is a chain graph, i.e. G^* comprises no directed cycles.*

Proof: In this proof we suppose that the essential graph G^* has a directed cycle and then show that this assumption results in a contradiction.

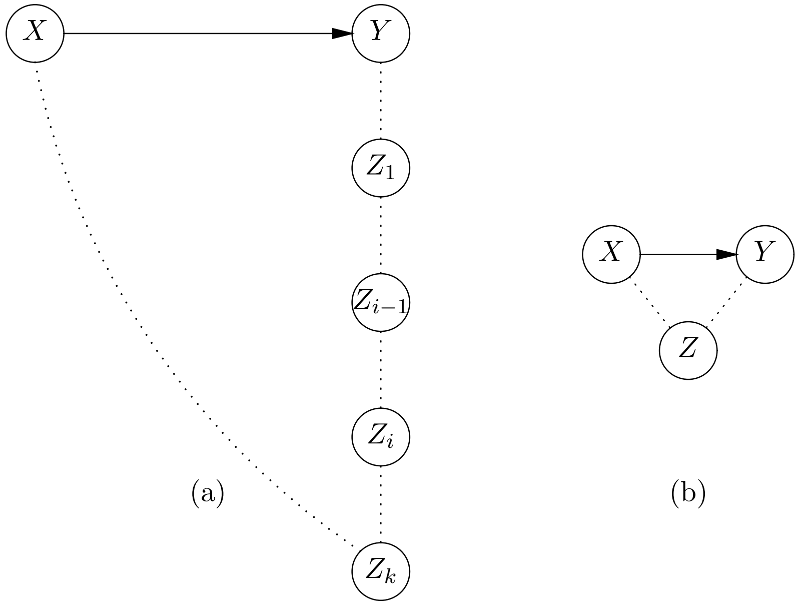


Figure 14: The directed cycle (a) and the reduced variant (b).

Suppose that G^* has a directed cycle $X, Y, Z_1, \ldots, Z_k \equiv X$ with $k \geq 2$. Observe that this directed cycle has at least: (i) one arc, which follows from the definition of directed cycle and (ii) one edge, otherwise some $G \in [G]$ would have a directed cycle which does not fit their property of acyclicity. Thus the cycle can be written as $X \to Y, Z_1, \ldots, Z_k \equiv X$, with $k \geq 2$, as shown in Figure 14(a). Because there is at least one edge inside this cycle, assume that this is the edge $Z_{i-1} - Z_i$ in G^* with $i \leq k$. Due to the fact that Z_{i-2} and Z_{i-1} can be connected by an edge (i.e. $Z_{i-2} - Z_{i-1}$) or by an arc directioned into Z_{i-1} (i.e. $Z_{i-2} \to Z_{i-1}$) there must exist at least one $G \in [G]$ with substructure $Z_{i-2} \to Z_{i-1} \leftarrow Z_i$ (deduced from $Z_{i-2} \to Z_{i-1} - Z_i$). As this cannot be an immorality there has to be a connection $Z_{i-2} \cdots Z_i$ (hence, the open possibility for either an edge or an arc directed to Z_i is denoted by $\cdots$). But this means that there is a smaller cycle such that $X \to Y, Z_1, \ldots, Z_{i-2}, Z_i, \ldots, Z_k \equiv X$, with $k \geq 2$. If we continue to reduce this directed cycle using the same idea, we observe that $X \to Y \cdots Z_1 \cdots Z_2 \equiv X$ which is equivalent to $X \to Y \cdots Z \cdots X$ with $Z_1 = Z$. Figure 14(b) depicts the reduced variant of the original directed cycle in Figure 14(a).

Next we show that $X \to Y \cdots Z \cdots X$ cannot be an induced subgraph of the essential graph G^*. Our assumption says that G^* contains a directed cycle. Then there exist four possible structures in G^* deduced from $X \to Y \cdots Z \cdots X$, shown in Figure 15. For each of these cases there exists at least one $G \in [G]$ equivalent to G^* containing a directed cycle, thus contradicting the acyclicity property: case (a) $\exists G \in [G]$ containing arc $Y \to Z$; case (b) $\exists G \in [G]$ containing arc $Z \to X$; case (c) $\exists G \in [G]$ containing arcs $Y \to Z$ and $Z \to X$; case (d) is already a directed cycle.

Due to the fact that each $G \in [G]$ should be an ADG, G^* cannot contain any of the substructures from Figure 15; therefore, the essential graph is an chain graph, which completes our proof. □

The essential graph has another very important property which is stated in the following lemma.

Lemma 2 *Let G^* be the essential graph which represents the equivalence class $[G]$ and let $\mathbf{V}(t)$ be a the chain component of G^*, then $\mathbf{V}(t)$ is chordal.*

Proof: Suppose that there is an undirected cycle with $k \geq 4$ in chain component $\mathbf{V}(t), t \leq T$. Then there should exists at least one ADG $G \in [G]$, which by its property of being acyclic should consist of an immorality. Since G^* has to represent each immorality of $[G]$, thus the assumption above results in a contradiction. Therefore the chain components of the graph G^* are chordal. □

Lemma 1 and Lemma 2 concern two fundamental properties of essential graphs. In addition, an essential graph is meant to preserve dependence information from the ADGs it represents. As mentioned above, immoralities are meant to represent conditional dependence information. To preserve these immoralities in an essential graph, the concept of strongly protected arcs have been introduced. In the following lemma, this notion is used to further characterise essential graphs.

Lemma 3 *Let G^* be the essential graph corresponding to the equivalence class $[G]$. Then each arc in G^* is a strongly protected arc.*

Proof: Suppose that $X \to Y$ is not a strongly protected arc in G^*. This means that its direction is reversible. Thus, there exists a graph $G \in [G]$ with arc $X \leftarrow Y$. But then G^* should comprise $X - Y$, leading to a contradiction. □

As was discussed above in relationship to Figure 13(a) and 13(b) it is not permitted that an immorality is changed into a divergent or serial connection, and vice versa.

Lemma 4 *Let G^* be the essential graph corresponding to the equivalence class $[G]$. Then G^* cannot contain the structure $X \to Y - Z$ as an induced subgraph.*

Proof: Suppose $X \to Y - Z$ is an induced subgraph in G^*. Then, there exists a $G \in [G]$ such that $X \to Y \leftarrow Z$. But this is an immorality which should be included in G^*, leading to a contradiction; this completes the proof. □

Combining the lemmas (1), (2), (3) and (4) leads to a full characterisation of essential graphs in the next theorem.

Theorem 5 *Let G^* be the essential graph corresponding to the equivalence class $[G]$. Then G^* satisfies the following four conditions:*

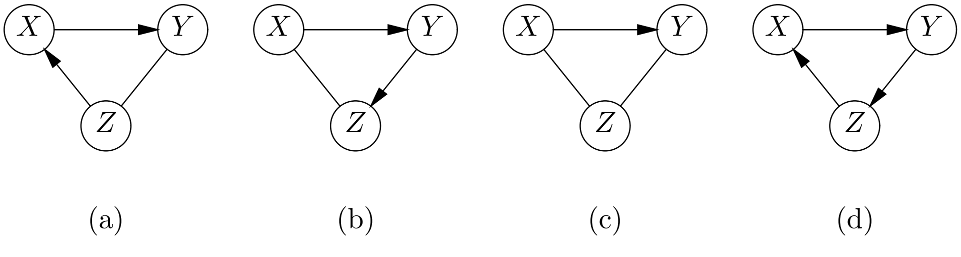


Figure 15: The possible graphs obtained by replacing the symbol $\cdots$ in $X \to Y \cdots Z \cdots X$ by an edge or an arc, taking the requirement into account that the resulting graph should contain a directed cycle.

- *G^* is a chain graph;*
- *each chain component of G^* is chordal;*
- *each arc in G^* is strongly protected;*
- *there exists no induced subgraph $X \rightarrow Y - Z$ in G^*.*

Proof: The lemmas 1, 2, 3 and 4 are used subsequently to prove the statements mentioned above, exactly in this order.

5 Conclusions

This paper is meant as a guide to the field of probabilistic graphical models, where in particular we have tried to offer a balanced view of the various issues involved in the study of stochastic dependence and independence, and its role in Bayesian-network structure learning. As there are many different ways in which (in)dependence information can be represented, e.g. as a joint probability distribution, as logical statements, or in the form of different types of graphs, we have focused on the relationships between these different representations.

There were a number of key results given attention to in the paper that are worth recalling. The independence relation may be looked upon as a logical relation, where special properties of the relation can be defined axiomatically. Unfortunately, the Independence relation does not permit finite axiomatisation. Nevertheless, there are a number of axioms that are worth knowing, as they support our understanding of the nature of independence; the most familiar axioms were covered in the paper.

The subtle differences between representing stochastic independence using undirected, acyclic directed and chain graphs was another related topic also studied in this paper. The process of moralisation transforms acyclic directed graphs and chain graphs into undirected graphs, which allows us to determine the semantic relationships between these different graphical ways to represent stochastic independence. Linked to this topic, a number of reading-off methods specific for particular types of graph were discussed, which supported reasoning about the independence information represented in a graph solely in terms of the graph structure.

Ways to identify and represent Markov equivalence in Bayesian networks were the last topics studied. In particular, the concept of the essential graph yields a significant insight into this matter, as an essential graph summarises a class of Markov equivalent networks, and thus renders it possible to determine which arcs in a Bayesian network are really significant. Bayesian networks contain static and dynamic dependences. For the case of static dependences changing directionality of arcs has no effect on the dependences in the entire network, as long as it does not give rise to the creation of immoralities. On the other hand, dynamic dependences are captured by the structure of the immoralities and as these cannot be changed without changing the meaning of a probabilistic graphical model, we have to maintain the direction of arcs in this case. Therefore, the equivalent relation on Bayesian networks is defined in terms of the structure of the skeleton and the associated set of immoralities contained in the graphs. The concept the of essential graph has given rise to much research activity, in particular in areas devoted to the development of algorithms for searching the equivalence space of Bayesian networks (instead of the entire space of Bayesian networks) to determine the Bayesian network that best fits the data from a given domain. What is clear is that probabilistic graphical models offer a rich and complicated landscape of probabilistic representations, which will remain a topic of research in the future.

References

[1] Andersson, S. A., Madigan, D., and Perlman, M. D. A Characterization of Markov Equivalence Classes for Acyclic Digraphs. *Annals of Statistic 25* (1997), 505–541.

[2] A.P.Dawid. Conditional Independence in statistical theory. *Journal of the Royal Statistical Society 41* (1979), 1–31.

[3] Castillo, E., Gutiérrez, J. M., and Hadi, A. S. *Expert Systems and Probabilistic Network Models*. Springer-Verlag New York, 1997.

[4] Chickering, D. M. Learning Equivalence Classes of Bayesian Network Structures. *Journal of Machine Learning Research 2* (2002), 445–498.

[5] Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. *Probabilistic Networks and Expert Systems*. Springer-Verlag New York, 1999.

[6] Kreyszig, E. *Introductory Mathematical Statistics*. Wiley, New York, 1970.

[7] Lauritzen, S. *Graphical models*. Clarendon Press, Oxford, 1996.

[8] M.Studený, and R.R.Bouckaert. On chain graph models for description of conditional independence structures. *Annals of Statistics 26*, 4 (1998), 1434–1495.

[9] Neapolitan, R. *Learning Bayesian Networks*. Northeastren Illinois University, Chicago, Illinois, 2003.

[10] Pearl, J. A constraint-propagation approach to probabilistic reasoning. In *Uncertainty in Artificial Intelligence* (Amsterdam, NorthHolland, 1986), L. N. Kanal and J. F. Lemmer (eds), pp. 3718–382.

[11] Pearl, J. *Probabilistic Reasoning in Intelligent Systems:Networks of Plausible Inference*. Morgan Kauffman, San Francisco, CA, 1988.

[12] Robinson, R. Counting unlabeled acyclic graphs. In *LNM 622* (1977), Springer, NY, pp. 220–227.

[13] Studený, M. Multiinformation and the Problem of Characterization of Independence Relations. *Problems of Control and Information Theory 3* (1989), 3–16.

[14] Studený, M. Conditional Independence Relations Have No Finite Complete Characterization. In *Information Theory, Statistical Decision Functions and Random Processes:Transactions of 11th Prague Conference* (1992), S. Kubíck and J. Vísek, Eds., Kluwer, Dordrecht, pp. 377–396.

[15] Verma, T., and Pearl, J. Equivalence and synthesis of causal models. In *Uncertainty in Artifical Intelligence Proceedings of the Sixth Conference* (San Francisco, CA, 1990), M. Henrion, R. Shachter, L. Kanal, and J. Lemmer, Eds., Morgan Kaufmann, pp. 220–227.