More on Markov Properties

Steffen Lauritzen, University of Oxford

CIMPA Summerschool, Hammamet 2011, Tunisia

September 7, 2011

The two conditions

$$A \perp\!\!\!\perp B, \quad A \perp\!\!\!\perp B \mid C$$

are *very different* and will typically not both hold unless we either have $A \perp\!\!\!\perp (B, C)$ or $(A, B) \perp\!\!\!\perp C$, i.e. if one of the variables are completely independent of both of the others.

This fact is a simple form of what is known as *Yule–Simpson paradox.*

It can be much worse than this: A *positive conditional association can turn into a negative marginal association* and vice-versa.

Florida murderers

Sentences in 4863 murder cases in Florida over the six years 1973-78

	Sentence	
Murderer	Death	Other
Black	59	2547
White	72	2185

The table shows a greater proportion of white murderers receiving death sentence than black (3.2% vs. 2.3%), although the difference is not big, the picture seems clear.

Controlling for colour of victim

		Sentence	
Victim	Murderer	Death	Other
Black	Black	11	2309
	White	0	111
White	Black	48	238
	White	72	2074

Now the table for given colour of victim shows a very different picture. In particular, note that 111 white murderers killed black victims and none were sentenced to death.

Formal definition

Random variables X and Y are *conditionally independent* given the random variable Z if

$$\mathcal{L}(X \mid Y, Z) = \mathcal{L}(X \mid Z).$$

We then write $X \perp\!\!\!\perp Y \mid Z$ (or $X \perp\!\!\!\perp_P Y \mid Z$)
Intuitively:
Knowing Z renders Y *irrelevant* for predicting X.
Factorisation of densities:

$$\begin{aligned} X \perp\!\!\!\perp Y \mid Z \quad &\Longleftrightarrow \quad f_{XYZ}(x,y,z) f_Z(z) = f_{XZ}(x,z) f_{YZ}(y,z) \\ &\Longleftrightarrow \quad \exists a, b : f(x,y,z) = a(x,z) b(y,z). \end{aligned}$$

For random variables X, Y, Z, and W it holds

(C1) If $X \perp\!\!\!\perp Y \mid Z$ then $Y \perp\!\!\!\perp X \mid Z$;

(C2) If $X \perp\!\!\!\perp Y \mid Z$ and $U = g(Y)$, then $X \perp\!\!\!\perp U \mid Z$;

(C3) If $X \perp\!\!\!\perp Y \mid Z$ and $U = g(Y)$, then $X \perp\!\!\!\perp Y \mid (Z, U)$;

(C4) If $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp W \mid (Y, Z)$, then $X \perp\!\!\!\perp (Y, W) \mid Z$;

If density w.r.t. product measure $f(x, y, z, w) > 0$ also

(C5) If $X \perp\!\!\!\perp Y \mid (Z, W)$ and $X \perp\!\!\!\perp Z \mid (Y, W)$ then $X \perp\!\!\!\perp (Y, Z) \mid W$.

An *independence model* $\perp_\sigma$ is a ternary relation over subsets of a finite set V. It is *graphoid* if for all subsets A, B, C, D:

(S1) if $A \perp_\sigma B \,|\, C$ then $B \perp_\sigma A \,|\, C$ (symmetry);

(S2) if $A \perp_\sigma (B \cup D) \,|\, C$ then $A \perp_\sigma B \,|\, C$ and $A \perp_\sigma D \,|\, C$ (decomposition);

(S3) if $A \perp_\sigma (B \cup D) \,|\, C$ then $A \perp_\sigma B \,|\, (C \cup D)$ (weak union);

(S4) if $A \perp_\sigma B \,|\, C$ and $A \perp_\sigma D \,|\, (B \cup C)$, then $A \perp_\sigma (B \cup D) \,|\, C$ (contraction);

(S5) if $A \perp_\sigma B \,|\, (C \cup D)$ and $A \perp_\sigma C \,|\, (B \cup D)$ then $A \perp_\sigma (B \cup C) \,|\, D$ (intersection).

Semigraphoid if only (S1)–(S4) holds. It is *compositional* if also

(S6) if $A \perp_\sigma B \,|\, C$ and $A \perp_\sigma D \,|\, C$ then $A \perp_\sigma (B \cup D) \,|\, C$ (composition).

Separation in undirected graphs

Let $\mathcal{G} = (V, E)$ be finite and simple undirected graph (no self-loops, no multiple edges).

For subsets A, B, S of V, let $A \perp_{\mathcal{G}} B \mid S$ denote that *S separates A from B in $\mathcal{G}$*, i.e. that all paths from A to B intersect S.

Fact: *The relation $\perp_{\mathcal{G}}$ on subsets of V is a compositional graphoid.*

This fact is the reason for choosing the name 'graphoid' for such independence model.

Systems of random variables

For a system V of *labeled random variables* $X_v, v \in V$, we use the shorthand

$$A \perp\!\!\!\perp B \,|\, C \iff X_A \perp\!\!\!\perp X_B \,|\, X_C,$$

where $X_A = (X_v, v \in A)$ denotes the variables with labels in A.

The properties (C1)–(C4) imply that $\perp\!\!\!\perp$ *satisfies the semi-graphoid axioms* for such a system, and the *graphoid axioms if the joint density of the variables is strictly positive.*

A regular *multivariate Gaussian distribution,* defines a *compositional graphoid independence model.*

$\mathcal{G} = (V, E)$ simple undirected graph; An independence model $\perp_\sigma$ satisfies

(P) *the pairwise Markov property w.r.t. $\mathcal{G}$* if

$$\alpha \not\sim \beta \Rightarrow \alpha \perp_\sigma \beta \,|\, V \setminus \{\alpha, \beta\};$$

(L) *the local Markov property w.r.t. $\mathcal{G}$* if

$$\forall \alpha \in V : \alpha \perp_\sigma V \setminus \mathrm{cl}(\alpha) \,|\, \mathrm{bd}(\alpha);$$

(G) *the global Markov property w.r.t. $\mathcal{G}$* if

$$A \perp_{\mathcal{G}} B \,|\, S \Rightarrow A \perp_\sigma B \,|\, S.$$

For any semigraphoid it holds that

$$(\mathrm{G}) \Rightarrow (\mathrm{L}) \Rightarrow (\mathrm{P})$$

If $\perp_\sigma$ satisfies graphoid axioms it further holds that

$$(\mathrm{P}) \Rightarrow (\mathrm{G})$$

so that *in the graphoid case*

$$(\mathrm{G}) \iff (\mathrm{L}) \iff (\mathrm{P}).$$

The latter holds in particular for $\perp\!\!\!\perp$, when $f(x) > 0$.

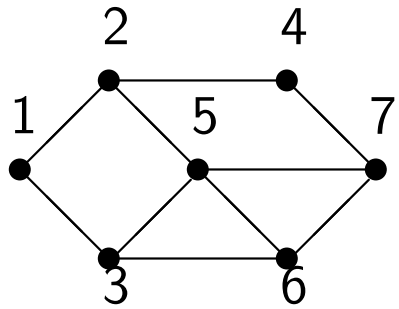
Assume density f w.r.t. product measure on $\mathcal{X}$.
For $a \subseteq V$, $\psi_a(x)$ denotes a function which depends on x_a only, i.e.

$$x_a = y_a \Rightarrow \psi_a(x) = \psi_a(y).$$

We can then write $\psi_a(x) = \psi_a(x_a)$ without ambiguity.
The distribution of X *factorizes w.r.t.* $\mathcal{G}$ or satisfies (F) if

$$f(x) = \prod_{a \in \mathcal{A}} \psi_a(x)$$

where $\mathcal{A}$ are *complete* subsets of $\mathcal{G}$.
Complete subsets of a graph are sets with all elements pairwise neighbours.



The *cliques* of this graph are the maximal complete subsets $\{1,2\}$, $\{1,3\}$, $\{2,4\}$, $\{2,5\}$, $\{3,5,6\}$, $\{4,7\}$, and $\{5,6,7\}$. A complete set is any subset of these sets.

The graph above corresponds to a factorization as

$$\begin{aligned} f(x) &= \psi_{12}(x_1,x_2)\psi_{13}(x_1,x_3)\psi_{24}(x_2,x_4)\psi_{25}(x_2,x_5) \\ &\times \psi_{356}(x_3,x_5,x_6)\psi_{47}(x_4,x_7)\psi_{567}(x_5,x_6,x_7). \end{aligned}$$

Let (F) denote the property that f factorizes w.r.t. $\mathcal{G}$ and let (G), (L) and (P) denote Markov properties w.r.t. $\perp\!\!\!\perp$. *It then holds that*

$$(\mathrm{F}) \Rightarrow (\mathrm{G})$$

and further: *If* $f(x) > 0$ *for all* x, $(\mathrm{P}) \Rightarrow (\mathrm{F})$.

The former of these is a simple direct consequence of the factorization whereas the second implication is more subtle and known as the *Hammersley–Clifford Theorem*.

Thus in the case of positive density (but typically only then), *all the properties coincide:*

$$(\mathrm{F}) \iff (\mathrm{G}) \iff (\mathrm{L}) \iff (\mathrm{P}).$$

Any joint probability distribution P of $X = (X_v, v \in V)$ has a *dependence graph* $G = G(P) = (V, E(P))$.

This is defined by letting $\alpha \not\sim \beta$ in $G(P)$ exactly when

$$\alpha \perp\!\!\!\perp_P \beta \mid V \setminus \{\alpha, \beta\}.$$

X will then satisfy the pairwise Markov w.r.t. $G(P)$ and $G(P)$ is smallest with this property, i.e. *P is pairwise Markov w.r.t. $\mathcal{G}$ iff*

$$G(P) \subseteq \mathcal{G}.$$

If $f(x) > 0$ for all x, *P is also globally Markov w.r.t. $G(P)$.*

Let $\mathcal{A}$ denote an arbitrary set of subsets of V. A density f (or function) *factorizes* w.r.t. $\mathcal{A}$ if there exist functions $\psi_a(x)$ which depend on x_a only and

$$f(x) = \prod_{a \in \mathcal{A}} \psi_a(x).$$

Similar to factorization w.r.t. graph, but *$\mathcal{A}$ are not necessarily complete subsets of a graph.*

The set of distributions $\mathcal{P}_{\mathcal{A}}$ which factorize w.r.t. $\mathcal{A}$ is the *hierarchical log–linear model* generated by $\mathcal{A}$.

To avoid redundancy, it is common to assume the sets in $\mathcal{A}$ to be incomparable in the sense that no subset in $\mathcal{A}$ is contained in any other member of $\mathcal{A}$. $\mathcal{A}$ is the *generating class* of the log–linear model.

The logarithm of the factors $\phi_a = \log \psi_a$ are known as *interaction terms of order* $|a| - 1$ or $|a|$*-factor interactions*.

Interaction terms of 0th order are called *main effects*.
We also refer to the factors themselves (rather than their logarithms) using the same terms.

The *dependence graph $G(\mathcal{P})$ for a family $\mathcal{P}$* is the smallest graph $\mathcal{G}$ so that all $P \in \mathcal{P}$ are pairwise Markov w.r.t. $\mathcal{G}$:

$$\alpha \perp\!\!\!\perp_P \beta \mid V \setminus \{\alpha, \beta\} \text{ for all } P \in \mathcal{P}.$$

For any generating class $\mathcal{A}$ we construct the dependence graph $G(\mathcal{A}) = G(\mathcal{P}_\mathcal{A})$ of the log–linear model $\mathcal{P}_\mathcal{A}$.

The dependence graph is determined by the relation

$$\alpha \sim \beta \iff \exists a \in \mathcal{A} : \alpha, \beta \in a.$$

For sets in $\mathcal{A}$ are clearly complete in $G(\mathcal{A})$ and therefore *distributions in $\mathcal{P}_\mathcal{A}$ do factorize according to $G(\mathcal{A})$*. On the other hand, any graph with fewer edges would not suffice.

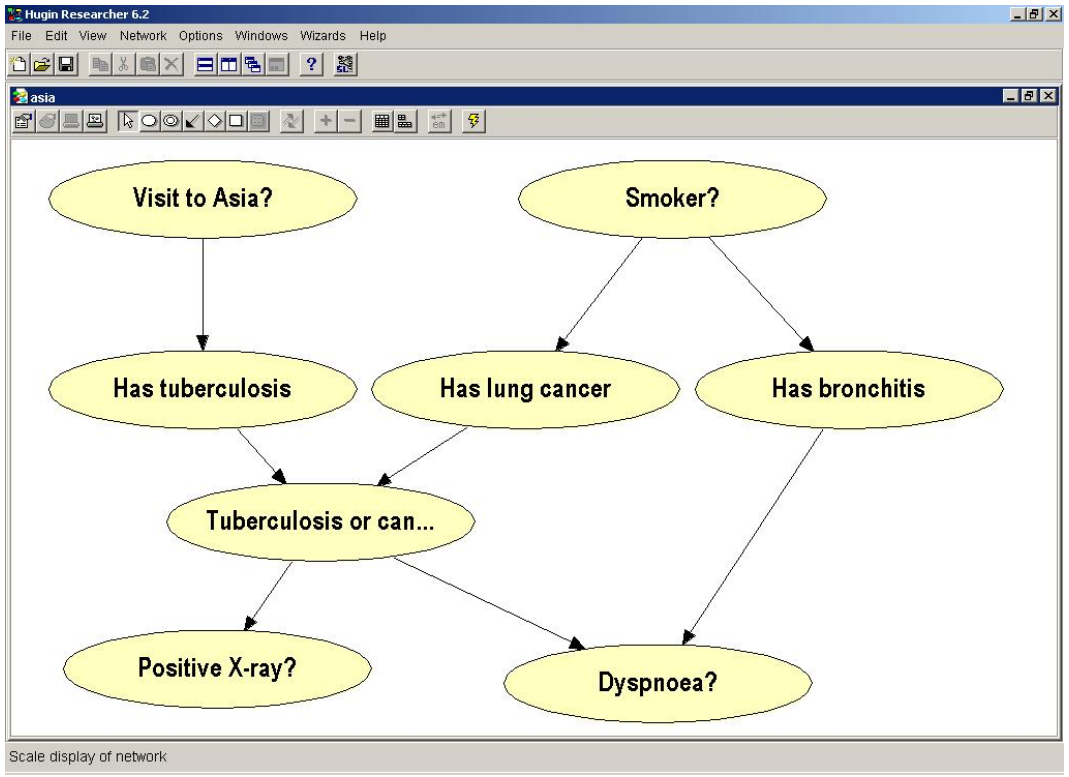
They are thus also global, local, and pairwise Markov w.r.t. $G(\mathcal{A})$.

A *directed acyclic graph* $\mathcal{D}$ over a finite set V is a simple graph with all edges directed and *no directed cycles.* We use DAG for brevity.

Absence of directed cycles means that, *following arrows in the graph, it is impossible to return to any point.*

Graphical models based on DAGs have proved fundamental and useful in a wealth of interesting applications, including expert systems, genetics, complex biomedical statistics, causal analysis, and machine learning.

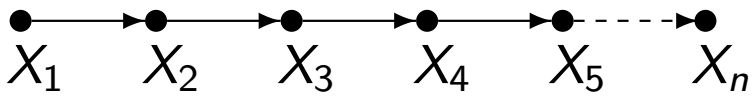
Example of a directed graphical model



An independence model $\perp_\sigma$ satisfies *the local Markov property* (L) w.r.t. a directed acyclic graph $\mathcal{D}$ if

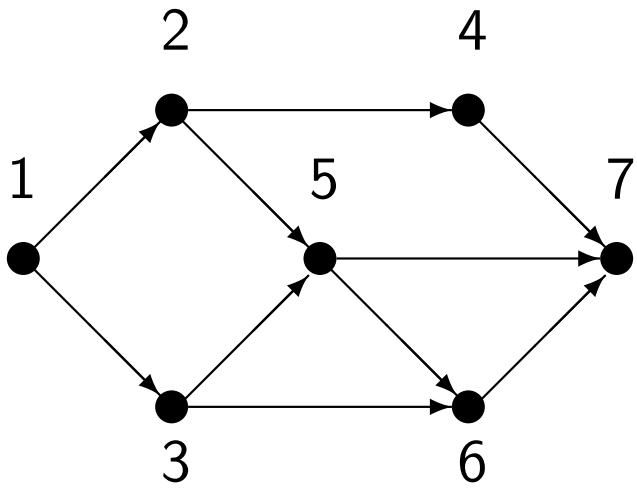
$$\forall \alpha \in V : \alpha \perp_\sigma \{\mathsf{nd}(\alpha) \setminus \mathsf{pa}(\alpha)\} \mid \mathsf{pa}(\alpha).$$

Here $\mathsf{nd}(\alpha)$ are the *non-descendants* of α.
A well-known example is a Markov chain:



with $X_{i+1} \perp\!\!\!\perp (X_1, \ldots, X_{i-1}) \mid X_i$ for $i = 3, \ldots, n$.

Local directed Markov property



For example, the local Markov property says

$4 \perp_\sigma \{1, 3, 5, 6\} \,|\, 2,$

$5 \perp_\sigma \{1, 4\} \,|\, \{2, 3\}$

$3 \perp_\sigma \{2, 4\} \,|\, 1.$

Suppose the vertices V of a DAG $\mathcal{D}$ are *well-ordered* in the sense that they are linearly ordered in a way which is compatible with $\mathcal{D}$, i.e. so that

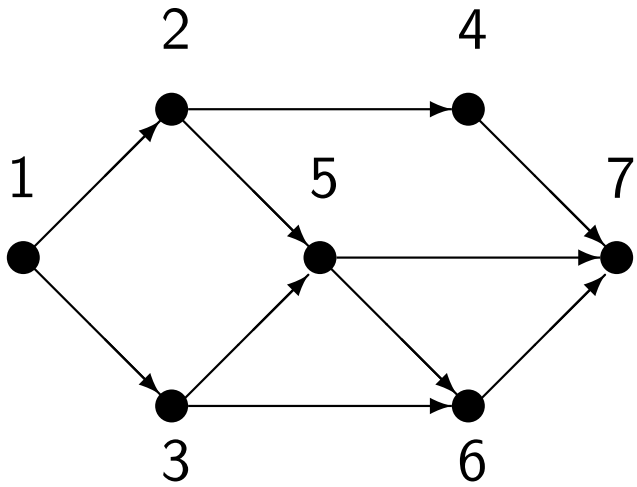
$$\alpha \in \mathsf{pa}(\beta) \Rightarrow \alpha < \beta.$$

We then say semigraphoid relation $\perp_\sigma$ satisfies the *ordered Markov property* (O) w.r.t. a well-ordered DAG $\mathcal{D}$ if

$$\forall \alpha \in V : \alpha \perp_\sigma \{\mathsf{pr}(\alpha) \setminus \mathsf{pa}(\alpha)\} \mid \mathsf{pa}(\alpha).$$

Here $\mathsf{pr}(\alpha)$ are the *predecessors* of α, i.e. those which are before α in the well-ordering..

Ordered Markov property



The numbering corresponds to a well-ordering. The ordered Markov property says for example

$4 \perp_\sigma \{1,3\} \,|\, 2,$

$5 \perp_\sigma \{1,4\} \,|\, \{2,3\}$

$3 \perp_\sigma \{2\} \,|\, 1.$

Separation in DAGs

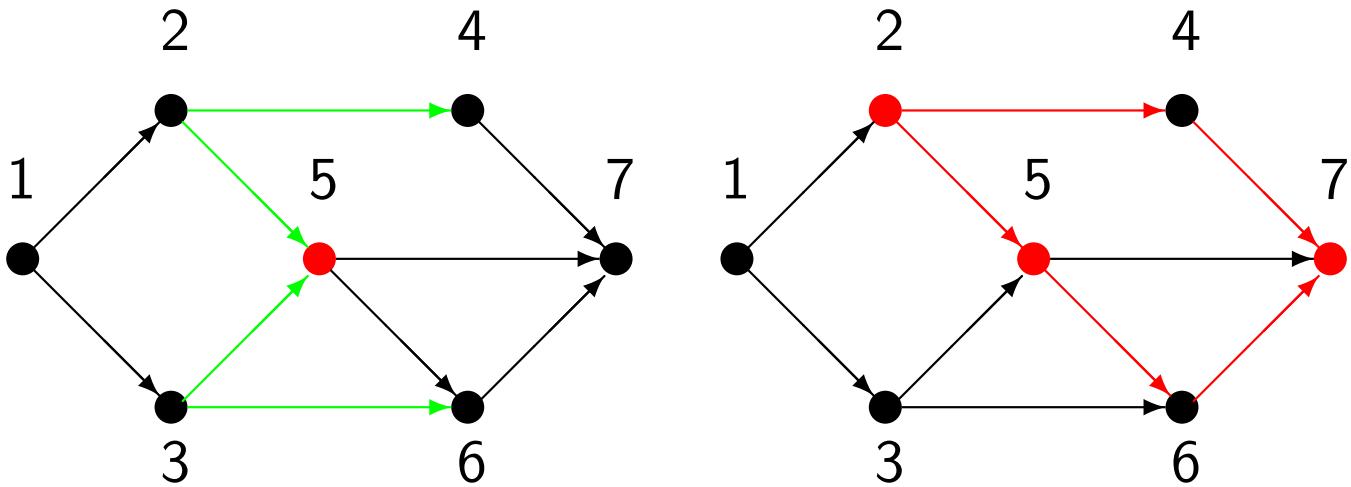
A node γ in a *trail* τ is a *collider* if edges meet head-to head at γ:

A trail τ from α to β in $\mathcal{D}$ is *active relative to* S if

- all its colliders are in $S \cup \mathrm{an}(S)$
- all its non-colliders are outside S

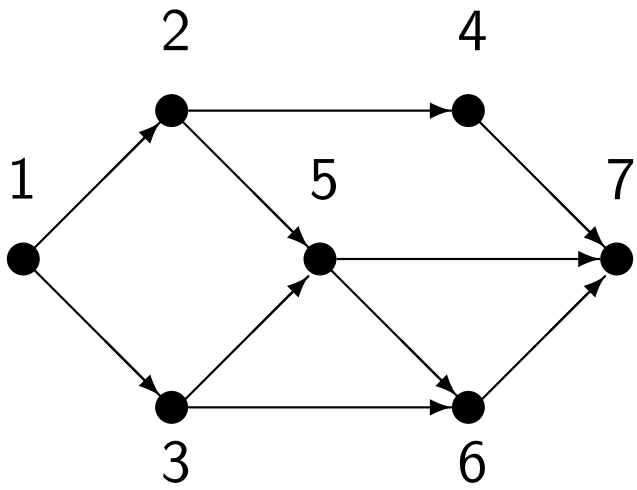
A trail that is not active is *blocked.* Two subsets A and B of vertices are *d-separated* by S if all trails from A to B are blocked by S. We write $A \perp_{\mathcal{D}} B \mid S$ and $\perp_{\mathcal{D}}$ *is a compositional graphoid for all* $\mathcal{D}$.

Separation by example



For $S = \{5\}$, the trail $(4, 2, 5, 3, 6)$ is *active*, whereas the trails $(4, 2, 5, 6)$ and $(4, 7, 6)$ are *blocked.*
For $S = \{3, 5\}$, they are all blocked.

Returning to example



Hence $4 \perp_{\mathcal{D}} 6 \,|\, 3, 5$, but it is *not* true that $4 \perp_{\mathcal{D}} 6 \,|\, 5$ nor that $4 \perp_{\mathcal{D}} 6$.

Equivalence of Markov properties

A semigraphoid relation $\perp_\sigma$ satisfies the *global Markov property* (G) w.r.t. $\mathcal{D}$ if

$$A \perp_{\mathcal{D}} B \,|\, S \Rightarrow A \perp_\sigma B \,|\, S.$$

It holds for any DAG $\mathcal{D}$ and any semigraphoid relation $\perp_\sigma$ that all directed Markov properties are equivalent:

$$(\mathrm{G}) \iff (\mathrm{L}) \iff (\mathrm{O}).$$

There is also a pairwise property (P), but it is less natural than in the undirected case and only equivalent to the others for graphoids.

A probability distribution P over $\mathcal{X} = \mathcal{X}_V$ *factorizes* over a DAG $\mathcal{D}$ if its density or probability mass function f has the form

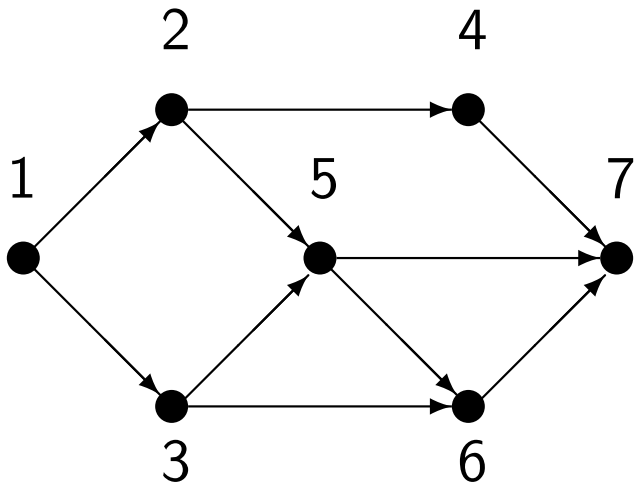
$$(\mathrm{F}): \quad f(x) = \prod_{v \in V} k_v(x_v \mid x_{\mathrm{pa}(v)})$$

where $k_v \geq 0$ and $\int_{\mathcal{X}_v} k_v(x_v \mid x_{\mathrm{pa}(v)})\, \mu_v(dx_v) = 1$.
(F) *is equivalent to* (F^*), where

$$(\mathrm{F}^*): \quad f(x) = \prod_{v \in V} f(x_v \mid x_{\mathrm{pa}(v)}),$$

i.e. it follows from (F) that k_v *in fact are conditional densities/pmf's.* Proof by induction!

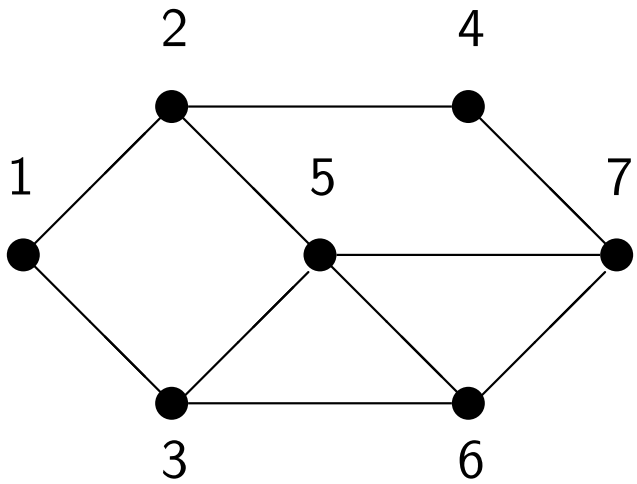
Example of DAG factorization



The above graph corresponds to the factorization

$$\begin{aligned} f(x) &= f(x_1)f(x_2 \mid x_1)f(x_3 \mid x_1)f(x_4 \mid x_2) \\ &\times f(x_5 \mid x_2, x_3)f(x_6 \mid x_3, x_5)f(x_7 \mid x_4, x_5, x_6). \end{aligned}$$

Contrast with undirected factorization



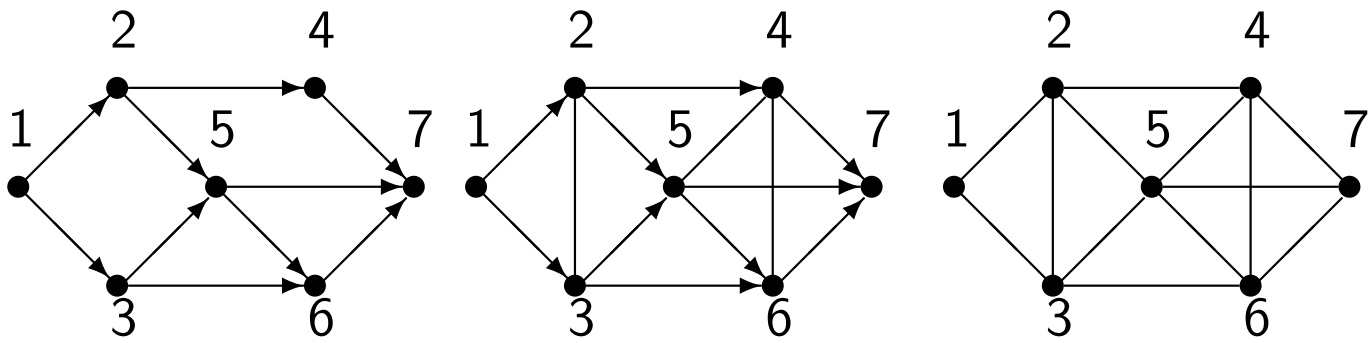
Factors ψ are typically not normalized as conditional probabilities:

$$\begin{aligned} f(x) &= \psi_{12}(x_1, x_2)\psi_{13}(x_1, x_3)\psi_{24}(x_2, x_4)\psi_{25}(x_2, x_5) \\ &\times \psi_{356}(x_3, x_5, x_6)\psi_{47}(x_4, x_7)\psi_{567}(x_5, x_6, x_7). \end{aligned}$$

In the directed case it is essentially *always true that (F) holds if and only if* $\perp\!\!\!\perp_P$ *satisfies (G)*,
so all directed Markov properties are equivalent to the factorization property!

$$(\mathrm{F}) \iff (\mathrm{G}) \iff (\mathrm{L}) \iff (\mathrm{O}).$$

The *moral graph* $\mathcal{D}^m$ of a DAG $\mathcal{D}$ is obtained by adding undirected edges between unmarried parents and subsequently dropping directions, as in the example below:



Undirected factorizations

If P factorizes w.r.t. $\mathcal{D}$, it factorizes w.r.t. the moralised graph $\mathcal{D}^m$.

This is seen directly from the factorization:

$$f(x) = \prod_{v \in V} f(x_v \mid x_{\mathrm{pa}(v)}) = \prod_{v \in V} \psi_{\{v\} \cup \mathrm{pa}(v)}(x),$$

since $\{v\} \cup \mathrm{pa}(v)$ are all complete in $\mathcal{D}^m$.

Hence if *P satisfies any of the directed Markov properties w.r.t. $\mathcal{D}$, it satisfies all Markov properties for $\mathcal{D}^m$.*

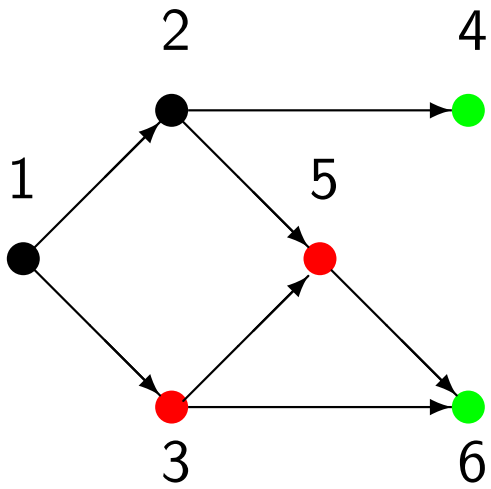
Alternative equivalent separation

To resolve query involving three sets A, B, S:

1. Reduce to subgraph induced by ancestral set $\mathcal{D}_{\mathrm{An}(A \cup B \cup S)}$ of $A \cup B \cup S$;
2. Moralize to form $(\mathcal{D}_{\mathrm{An}(A \cup B \cup S)})^m$;
3. Say that *S m-separates A from B* and write $A \perp_m B \mid S$ if and only if S separates A from B in this undirected graph.

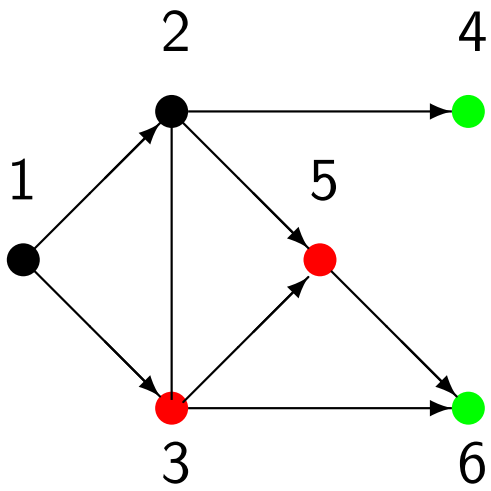
It then holds that $A \perp_m B \mid S$ *if and only if* $A \perp_{\mathcal{D}} B \mid S$.
Proof in Lauritzen (1996) needs to allow self-intersecting paths to be correct.

Forming ancestral set



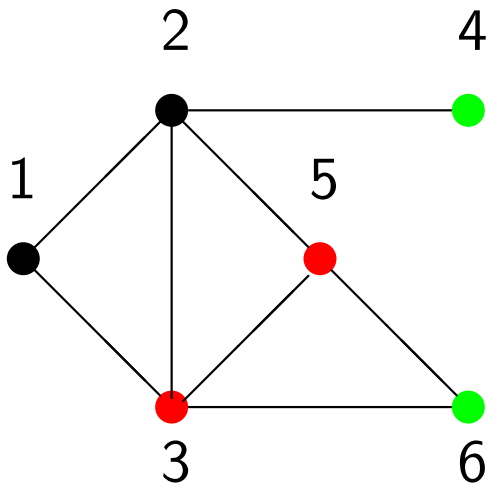
The subgraph induced by all ancestors of nodes involved in the query $4 \perp_m 6 \,|\, 3, 5$?

Adding links between unmarried parents



Adding an undirected edge between 2 and 3 with common child 5 in the subgraph induced by all ancestors of nodes involved in the query $4 \perp_m 6 \,|\, 3, 5$?

Dropping directions



Since $\{3, 5\}$ separates 4 from 6 in this graph, we can conclude that $4 \perp_m 6 \,|\, 3, 5$

Two DAGS $\mathcal{D}$ and $\mathcal{D}'$ are *Markov equivalent* if the independence models $\perp_{\mathcal{D}}$ and $\perp_{\mathcal{D}'}$ are identical.

$\mathcal{D}$ and $\mathcal{D}'$ are equivalent if and only if:

1. $\mathcal{D}$ and $\mathcal{D}'$ have same *skeleton*
2. $\mathcal{D}$ and $\mathcal{D}'$ have same unmarried parents

so

Contrast with undirected case, where *two undirected graphs are Markov equivalent if and only if they are identical.*

The *skeleton* $\sigma(\mathcal{D})$ of a DAG is the undirected graph obtained by ignoring directions.

Markov equivalence of directed and undirected graphs

A DAG $\mathcal{D}$ is *Markov equivalent* to an undirected $\mathcal{G}$ if the independence models $\perp_{\mathcal{D}}$ and $\perp_{\mathcal{G}}$ are identical.

This happens if and only if $\mathcal{D}$ is *perfect*, i.e. all parents are married, and $\mathcal{G} = \sigma(\mathcal{D})$. So, these are all equivalent

but not equivalent to

If $\mathcal{D}$ is perfect, $\sigma(\mathcal{D})$ is chordal.