Towards causal inference

Antal Péter

ComBine Lab

Artificial Intelligence group

Department of Measurement and Information Systems

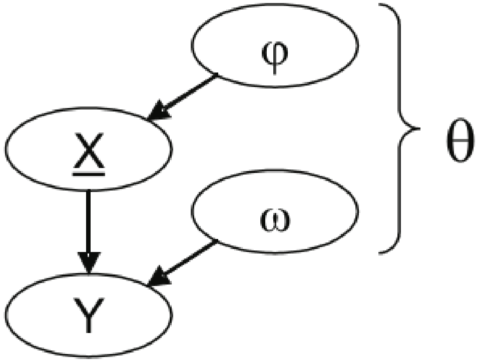
MŰEGYETEM 1782

Agenda

- Reminder:
  - Last lecture: general Bayesian networks
  - Now: axiomatic approaches to causal inference (~causality)
- Limits of conditional predictive machine learning
- From associations to direct dependencies
- The ultimate limit of observational learning
- Causal inference
- Learning causal relations and models
- Counterfactual inference

Limits of conditional machine learning

On the validity of conditional (Bayesian) approach



$$
\begin{aligned}
p(\theta|x,y) \triangleq p(\omega,\phi|x,y) &\propto p(x,y|\omega,\phi)p(\omega,\phi) \\
&= p(y|x,\omega)p(x|\phi)p(\omega|\phi)p(\phi) \\
&= p(y|x,\omega)p(\omega)p(x|\phi)p(\phi) \\
&\propto p(\omega|x,y)p(\phi|x).
\end{aligned}
$$

$$
\begin{aligned}
p(\omega|x,y) &\propto \int_\phi p(y|x,\omega,\phi)p(x|\phi)p(\omega|\phi)p(\phi)\,\mathrm{d}\phi \\
&= p(y|x,\omega)p(\omega)
\end{aligned}
$$

Limits of conditional modelling

- Incomplete data (incomplete input)
- Unsupervised learning (missing output)
  - Learning the joint distribution
  - Dimensionality reduction
  - Clustering
- Multitask learning (complete or partial output)
- Transfer learning
- Interpretation (feature subset selection, effect strength)
  - Prior incorporation
  - Bias (ethical machine learning?)
  - Confounding: effect strength (causal factors?)
- Interventionist data, mixed observational and interventionist data (health!)
- Structured data: temporal, dyadic, relational

Inference by enumeration

Every question about a domain can be answered by the joint distribution.

Typically, we are interested in the posterior joint distribution of the query variables **Y** given specific values **e** for the evidence variables **E**

Let the hidden variables be **H** = **X** – **Y** – **E**

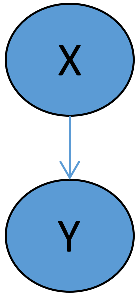
Then the required summation of joint entries is done by summing out the hidden variables:

$\mathbf{P}(\mathbf{Y} \mid \mathbf{E} = \mathbf{e}) = \alpha\mathbf{P}(\mathbf{Y},\mathbf{E} = \mathbf{e}) = \alpha\Sigma_h\mathbf{P}(\mathbf{Y},\mathbf{E}= \mathbf{e}, \mathbf{H} = \mathbf{h})$

- The terms in the summation are joint entries because **Y**, **E** and **H** together exhaust the set of random variables
- Obvious problems:
  1. Worst-case time complexity $O(d^n)$ where d is the largest arity
  2. Space complexity $O(d^n)$ to store the joint distribution
  3. How to find the numbers for $O(d^n)$ entries?

Motivation: from observational inference…

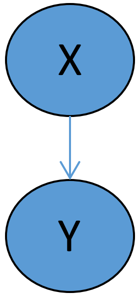
- In a joint distribution , any query can be answered corresponding to passive observations: p(Q=q|E=e).
  - What is the (conditional) probability of *Q=q* given that *E=e.*
  - Note that Q can preceed temporally E.



- Specification: p(X), p(Y|X)
- Joint distribution: p(X,Y)
- Inferences: p(X), p(Y), p(Y|X), p(X|Y)

Motivation: to interventional inference…

- Perfect intervention: do(X=x) as set X to x.
- What is the relation of p(Q=q|E=e) and p(Q=q|do(E=e))?



- Specification: p(X), p(Y|X)
- Joint distribution: p(X,Y)
- Inferences:
  - p(Y|X=x)=p(Y|do(X=x))
  - p(X|Y=y)≠p(X|do(Y=y))

- What is a formal knowledge representation of a causal model?
- What is the formal inference method?

Motivation: and to counterfactual inference

- Imagery observations and interventions:
  - We observed X=x, but imagine that x’ would have been observed: denoted as X’=x’.
  - We set X=x, but imagine that x’ would have been set: denoted as do(X’=x’).
- What is the relation of
  - Observational p(Q=q|E=e, X=x)
  - Interventional p(Q=q|E=e, do(X=x’))
  - Counterfactual p(Q’=q’|Q=q, E=e, do(X=x), do(X’=x’))

- O: What is the probability that the patient recovers if he takes the drug *x*.
- I: What is the probability that the patient recovers if we prescribe* the drug *x’*.
- C: Given that the patient had not recovered for the drug x, what would have been the probability that patient recovers if we had prescribed* the drug *x’*, instead of *x*.

- *: Assume that the patient is fully compliant.
- **” expected to neither he will.

Challenges in a complex domain

The domain is defined by the joint distribution

$$P(X_1,\ldots, X_n|\text{Structure,parameters})$$

1. Representation of parameteres

   „small number of parameters”

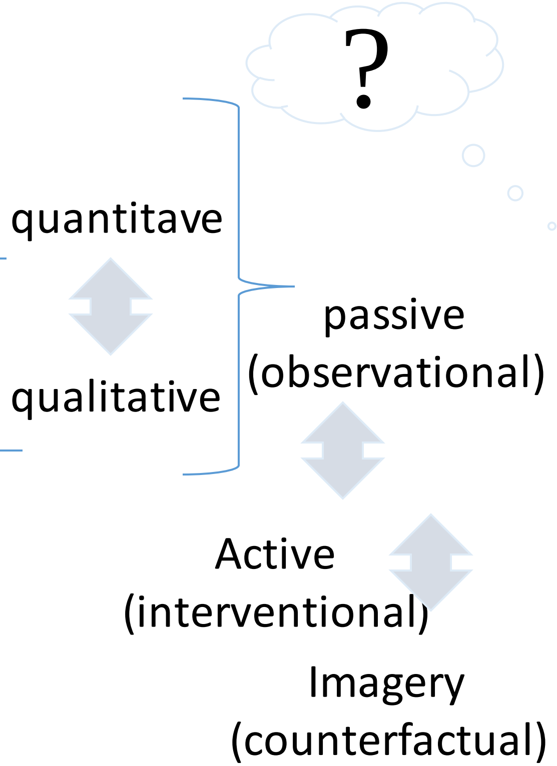
2. Representation of independencies

   „what is relevant for diagnosis”

3. Representation of causal relations

   „what is the effect of a treatment”

4. Representation of possible worlds



Probabilistic inference

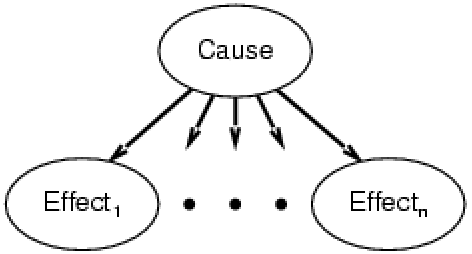
Reminder: Naive Bayesian network

- Definition: conditional independence of „effects” X_i given „cause” *Y.*
- Properties:
  - Number of parameters (~model complexity): linear
  - Complexity of inference: linear

Naive Bayesian network

Assumptions:

1, Two types of nodes: a cause and effects.

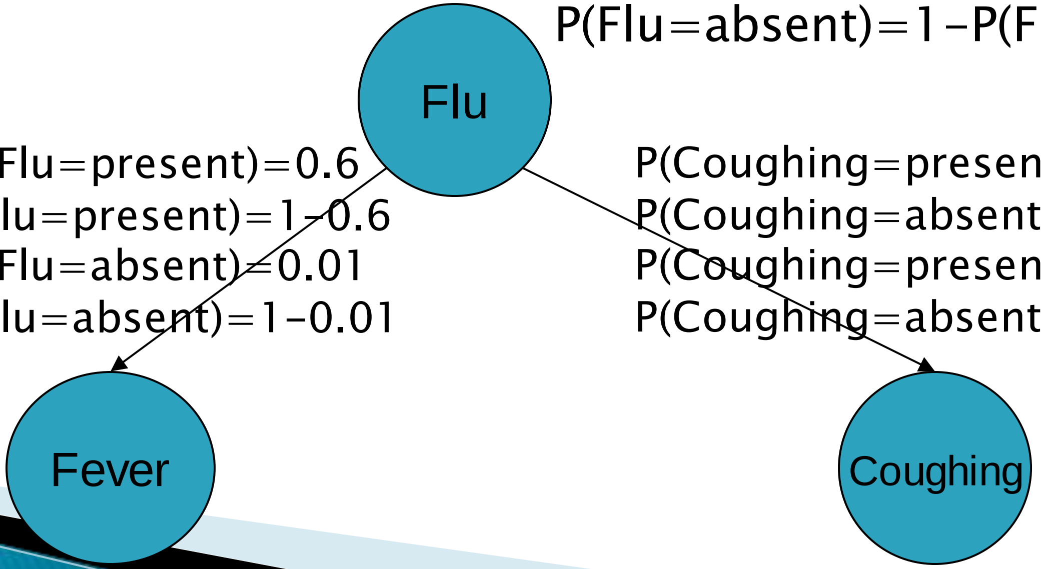


2, Effects are conditionally independent of each other given their cause.

Variables (nodes)

Flu: present/absent
FeverAbove38C: present/absent
Coughing: present/absent

Model

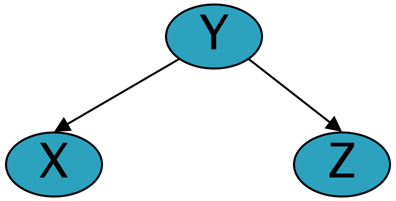


P(Flu=present)=0.001
P(Flu=absent)=1−P(Flu=present)

P(Fever=present|Flu=present)=0.6
P(Fever=absent|Flu=present)=1−0.6
P(Fever=present|Flu=absent)=0.01
P(Fever=absent|Flu=absent)=1−0.01

P(Coughing=present|Flu=present)=0.3
P(Coughing=absent|Flu=present)=1−0.7
P(Coughing=present|Flu=absent)=0.02
P(Coughing=absent|Flu=absent)=1−0.02

The independence map of a N-BN



If P(Y,X,Z) is a naive Bayesian network, then
M_P={D(X;Y), D(Y;Z), I(X;Z|Y)}
Normally/almost always: D(X;Z)
Exceptionally: I(X;Z)

Independence models

Conditional independence

$I_P(X;Y|Z)$ or $(X \perp\!\!\!\perp Y|Z)_P$ denotes that X is independent of Y given Z: $P(X;Y|z)=P(Y|z)\ P(X|z)$ for all z with $P(z)>0$.

(Almost) alternatively, $I_P(X;Y|Z)$ iff

$P(X|Z,Y)= P(X|Z)$ for all z,y with $P(z,y)>0$.

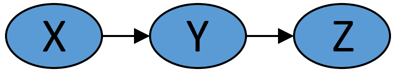
Other notations: $D_P(X;Y|Z)$ =def= $\neg\, I_P(X;Y|Z)$

Contextual independence: for not all z.

The independence model of a distribution

The independence map (model) M of a distribution P is the set of the valid independence triplets:

$M_P=\{I_{P,1}(X_1;Y_1|Z_1),\ldots, I_{P,K}(X_K;Y_K|Z_K)\}$



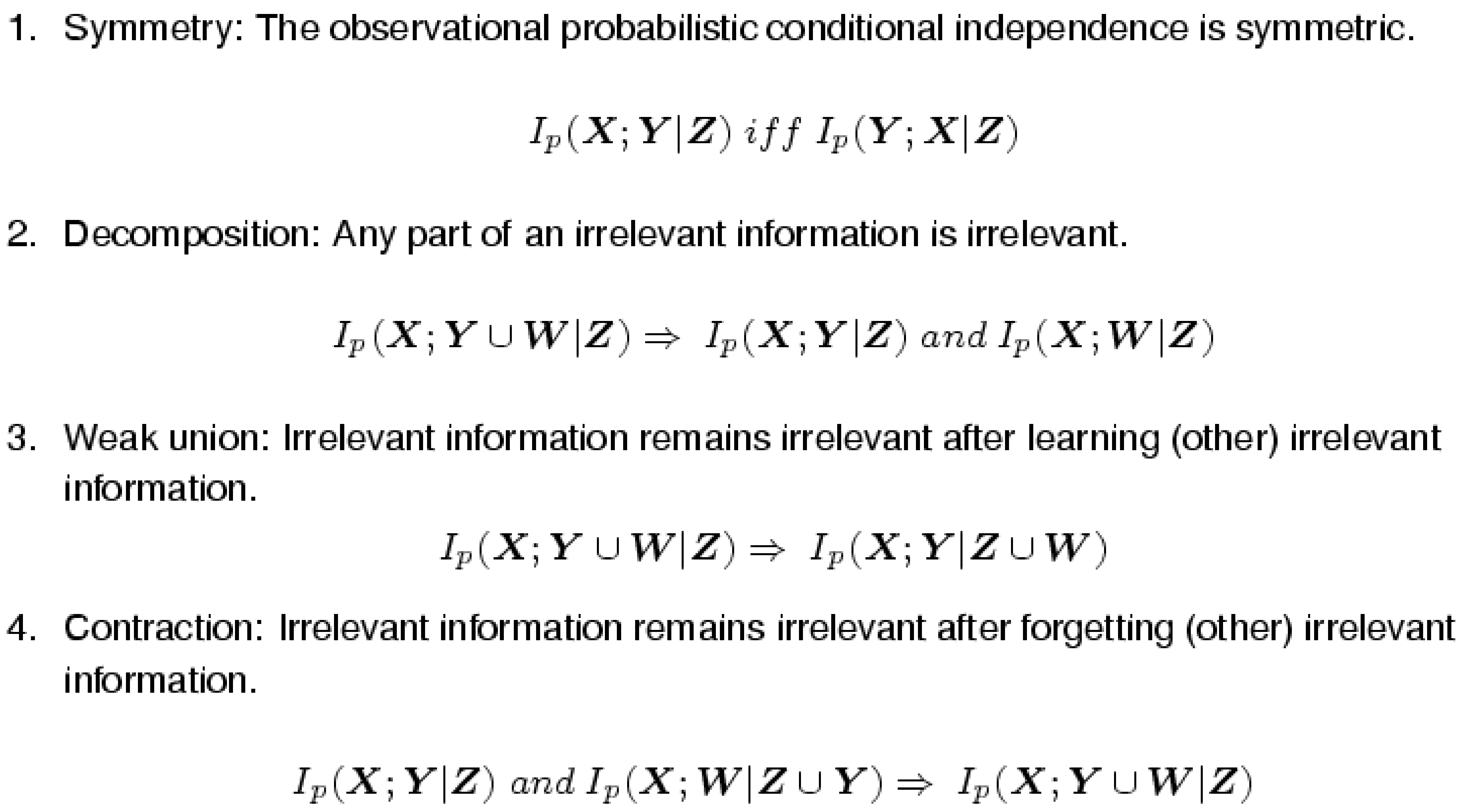
If P(X,Y,Z) is a Markov chain, then

$M_P=\{D(X;Y), D(Y;Z), I(X;Z|Y)\}$

Normally/almost always: D(X;Z)

Exceptionally: I(X;Z)

The semi-graphoid axioms



1. Symmetry: The observational probabilistic conditional independence is symmetric.

$$I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})\ iff\ I_p(\boldsymbol{Y};\boldsymbol{X}|\boldsymbol{Z})$$

2. Decomposition: Any part of an irrelevant information is irrelevant.

$$I_p(\boldsymbol{X};\boldsymbol{Y}\cup\boldsymbol{W}|\boldsymbol{Z})\Rightarrow\ I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})\ and\ I_p(\boldsymbol{X};\boldsymbol{W}|\boldsymbol{Z})$$

3. Weak union: Irrelevant information remains irrelevant after learning (other) irrelevant information.

$$I_p(\boldsymbol{X};\boldsymbol{Y}\cup\boldsymbol{W}|\boldsymbol{Z})\Rightarrow\ I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}\cup\boldsymbol{W})$$

4. Contraction: Irrelevant information remains irrelevant after forgetting (other) irrelevant information.

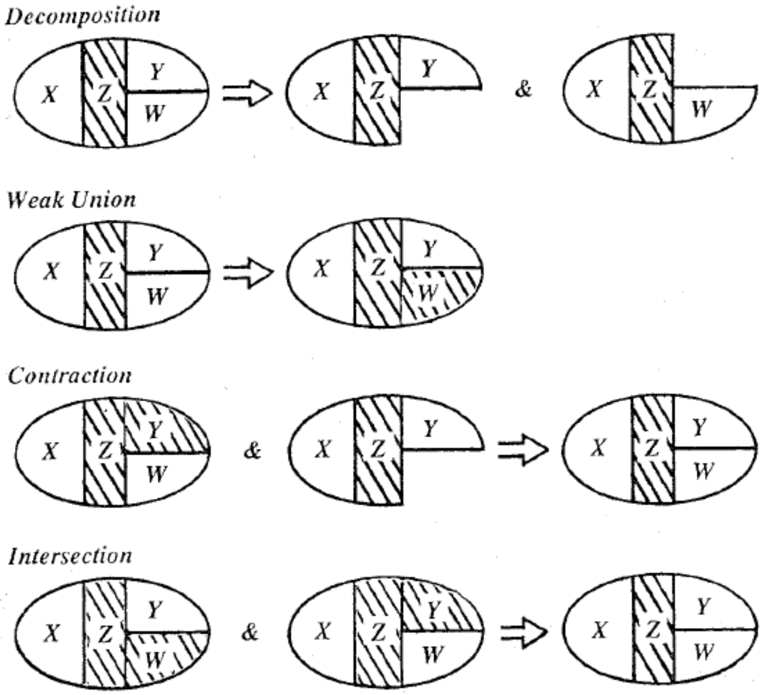
$$I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})\ and\ I_p(\boldsymbol{X};\boldsymbol{W}|\boldsymbol{Z}\cup\boldsymbol{Y})\Rightarrow\ I_p(\boldsymbol{X};\boldsymbol{Y}\cup\boldsymbol{W}|\boldsymbol{Z})$$

Graphoids

Graphoids: Semi-graphoids+Intersection (holds only in strictly positive distribution)

Intersection: Symmetric irrelevance implies joint irrelevance if there are no dependencies.

$$I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}\cup\boldsymbol{W})\ and\ I_p(\boldsymbol{X};\boldsymbol{W}|\boldsymbol{Z}\cup\boldsymbol{Y})\Rightarrow\ I_p(\boldsymbol{X};\boldsymbol{Y}\cup\boldsymbol{W}|\boldsymbol{Z})$$

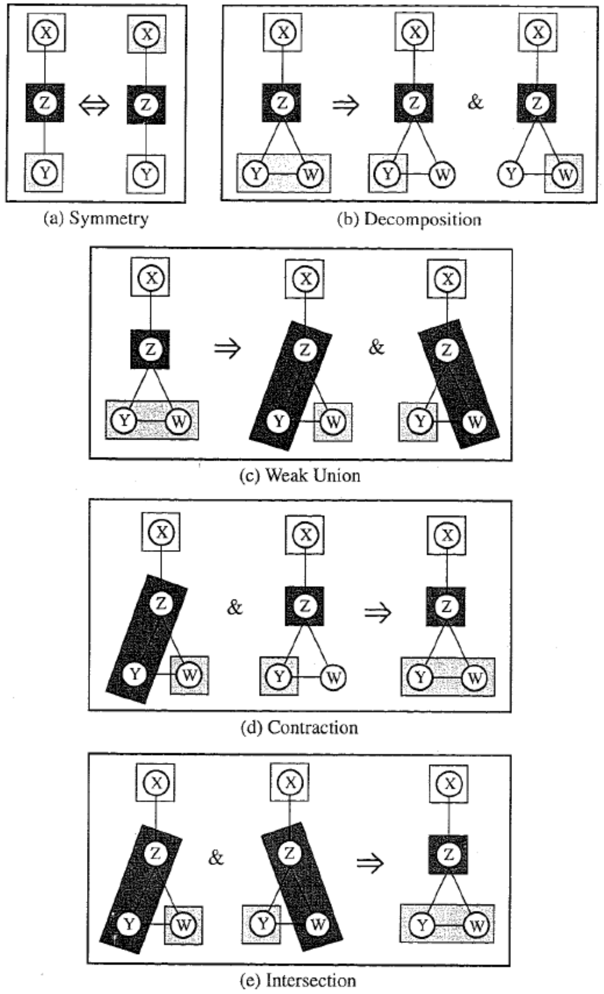


Decomposition

Weak Union

Contraction

Intersection



(a) Symmetry

(b) Decomposition

(c) Weak Union

(d) Contraction

(e) Intersection

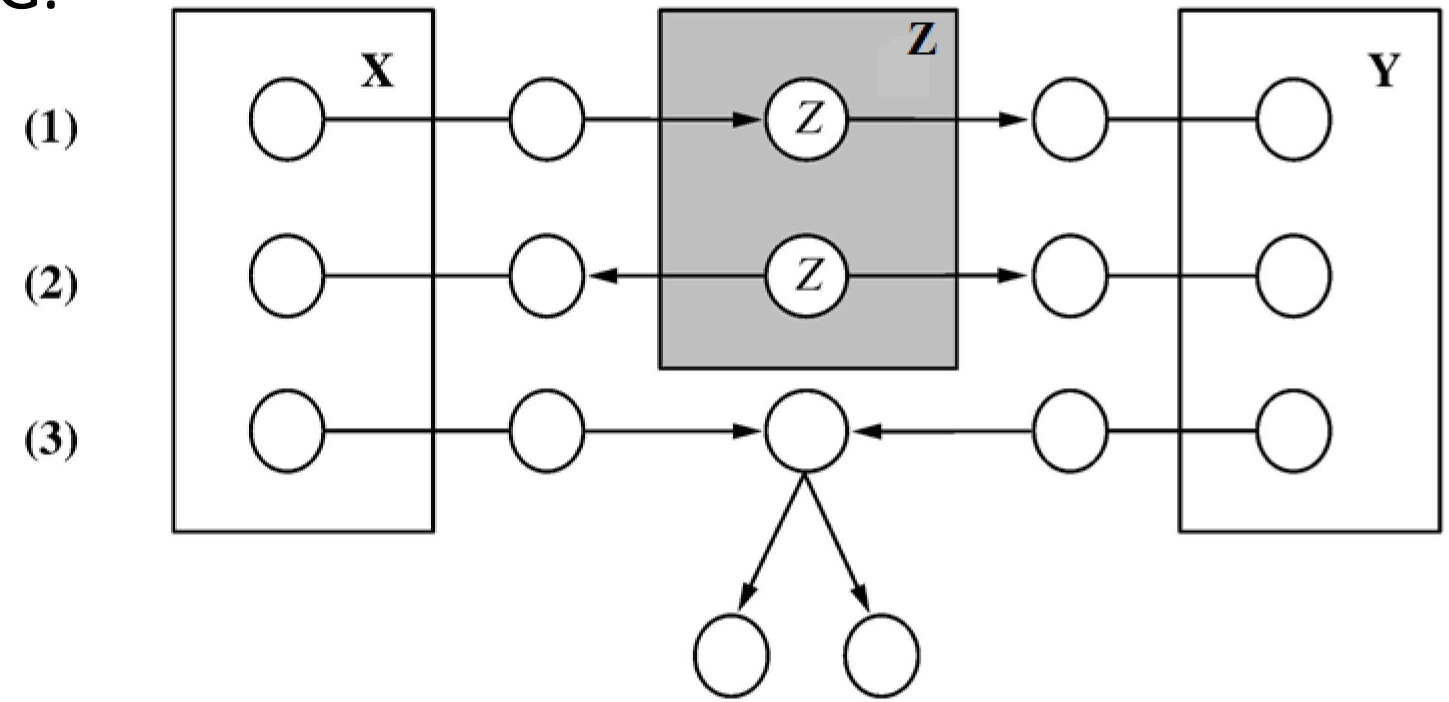
J.Pearl: Probabilistic Reasoning in intelligent systems, 1998

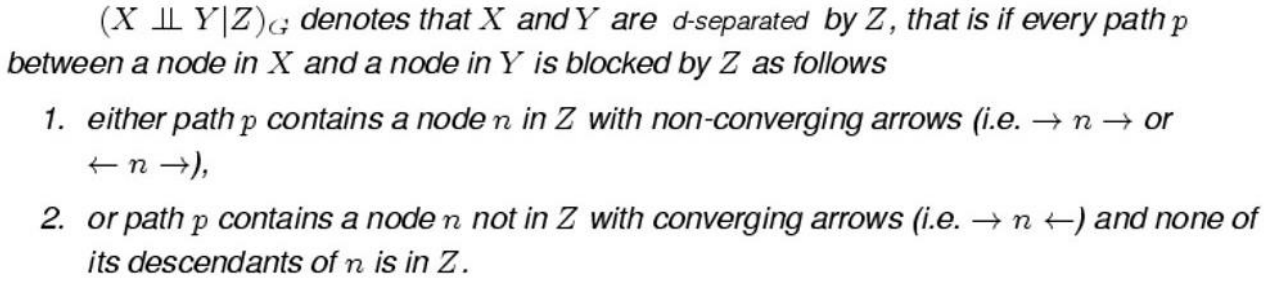
Separation in undirected graphs

$I_G(X;Y|Z)$ denotes that X is separated from Y by Z in undirected graph G, i.e. every path between X and Y is blocked by Z (it contains a node from Z).

Directed separation in directed graphs

$I_G(X;Y|Z)$ denotes that X is d-separated from Y by Z in directed acyclic graph G.

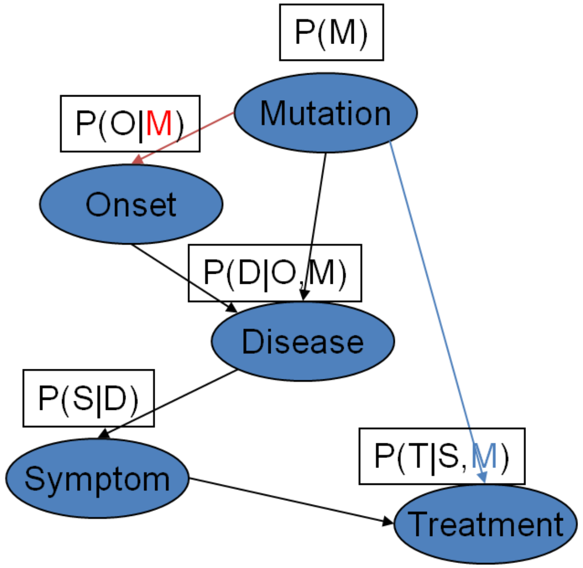




$(X \perp\!\!\!\perp Y|Z)_G$ denotes that X and Y are d-separated by Z, that is if every path p between a node in X and a node in Y is blocked by Z as follows

1. either path p contains a node n in Z with non-converging arrows (i.e. $\rightarrow n \rightarrow$ or $\leftarrow n \rightarrow$),
2. or path p contains a node n not in Z with converging arrows (i.e. $\rightarrow n \leftarrow$) and none of its descendants of n is in Z.

Bayesian networks: three facets



1. Causal model

$M_P=\{I_{P,1}(X_1;Y_1|Z_1),\ldots\}$

2. Graphical representation of (in)dependencies

3. Concise representation of joint distributions

$$P(M,O,D,S,T)=$$
$$P(M)P(O\mid M)P(D\mid O,M)P(S\mid D)P(T\mid S,M)$$

Bayesian networks

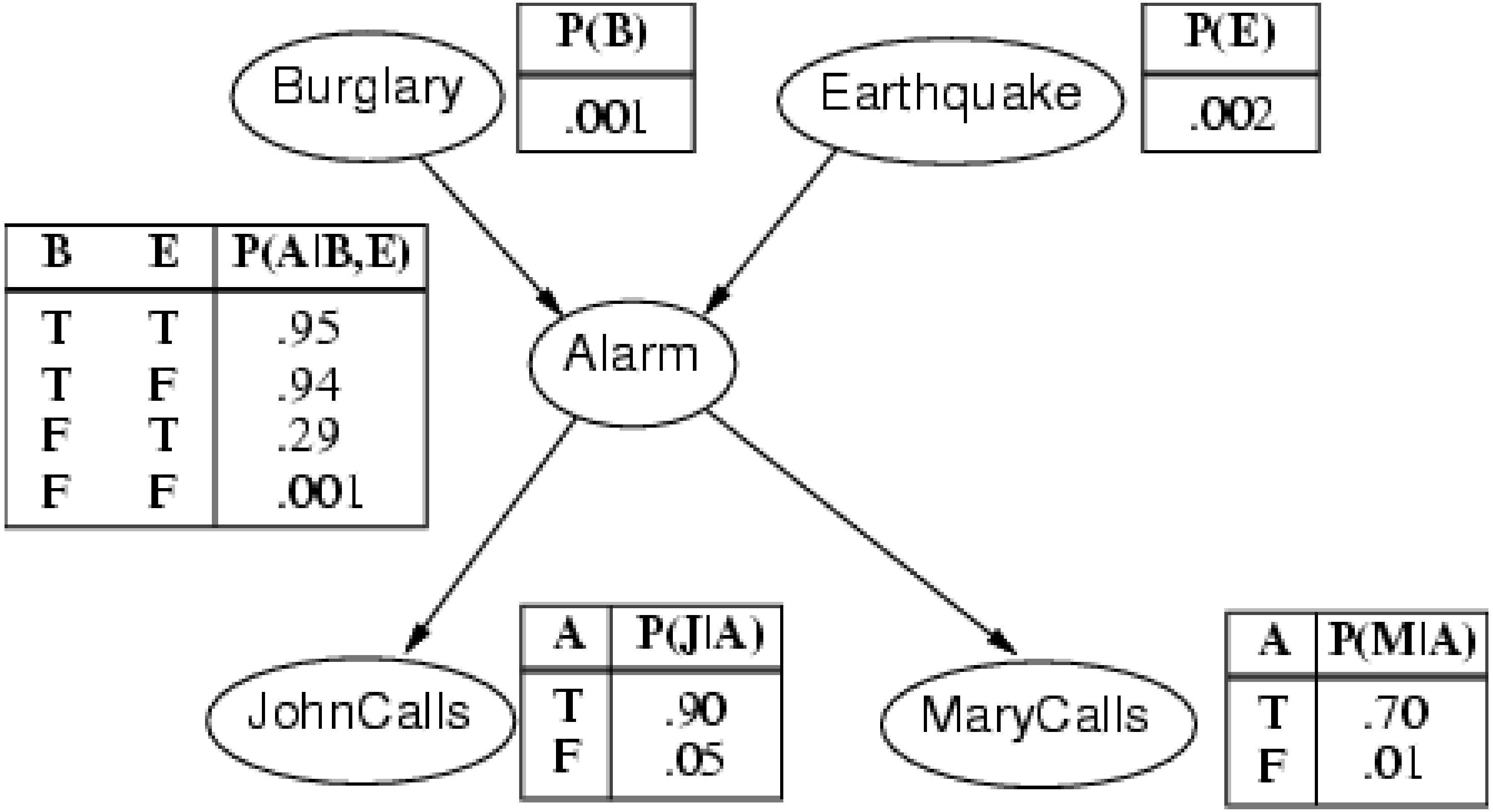
- A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions

- Syntax:
  - a set of nodes, one per variable
  -
  - a directed, acyclic graph (link ≈ "directly influences")
  - a conditional distribution for each node given its parents:

    $\mathbf{P}(X_i \mid \text{Parents}(X_i))$

- In the simplest case, conditional distribution represented as a conditional probability table (CPT) giving the distribution over X_i for each combination of parent values

Example contd.



P(B)
.001

P(E)
.002

B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01

Constructing Bayesian networks

- 1. Choose an ordering of variables $X_1, \ldots, X_n$
- 2. For $i = 1$ to n
  - add X_i to the network
  - select parents from $X_1, \ldots, X_{i-1}$ such that

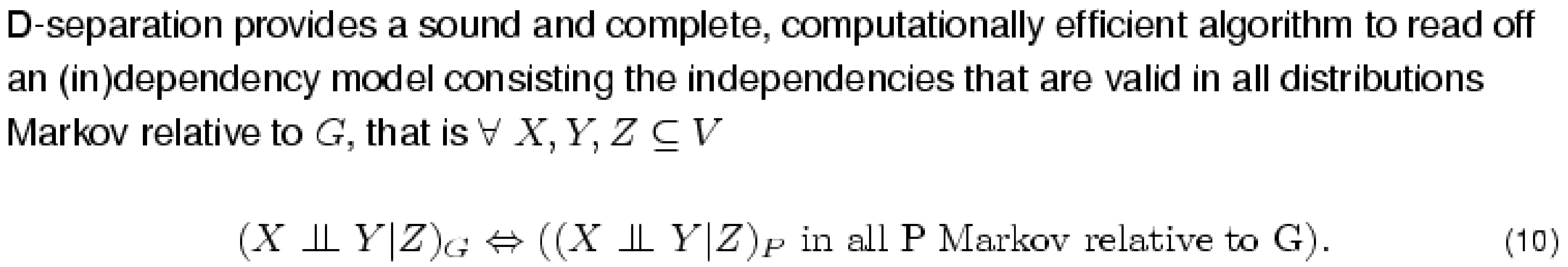
$$\mathbf{P}(X_i \mid Parents(X_i)) = \mathbf{P}(X_i \mid X_1, \ldots X_{i-1})$$

This choice of parents guarantees:

$$\mathbf{P}(X_1, \ldots, X_n) = \pi_{i=1}^{n} \mathbf{P}(X_i \mid X_1, \ldots, X_{i-1}) \quad //\text{(chain rule)}$$

$$= \pi_{i=1}^{n} \mathbf{P}(X_i \mid Parents(X_i)) \quad //\text{(by construction)}$$

Representation of independencies



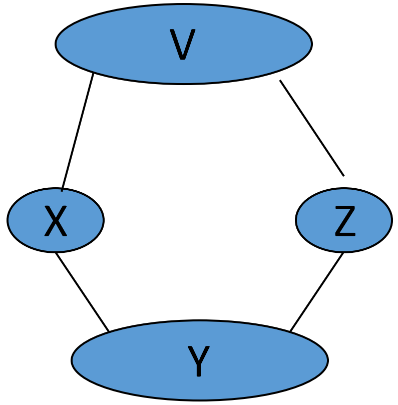
D-separation provides a sound and complete, computationally efficient algorithm to read off an (in)dependency model consisting the independencies that are valid in all distributions Markov relative to G, that is $\forall\ X, Y, Z \subseteq V$

$$(X \perp\!\!\!\perp Y|Z)_G \Leftrightarrow ((X \perp\!\!\!\perp Y|Z)_P \text{ in all P Markov relative to G}). \tag{10}$$

For certain distributions exact representation is not possible by Bayesian networks, e.g.:

1. Intransitive Markov chain: X➔Y➔Z
2. Pure multivariate cause: {X,Z}➔Y
3. Diamond structure:

P(X,Y,Z,V) with M_P={D(X;Z), D(X;Y), D(V;X), D(V;Z), I(V;Y|{X,Z}), I(X;Z|{V,Y}).. }.

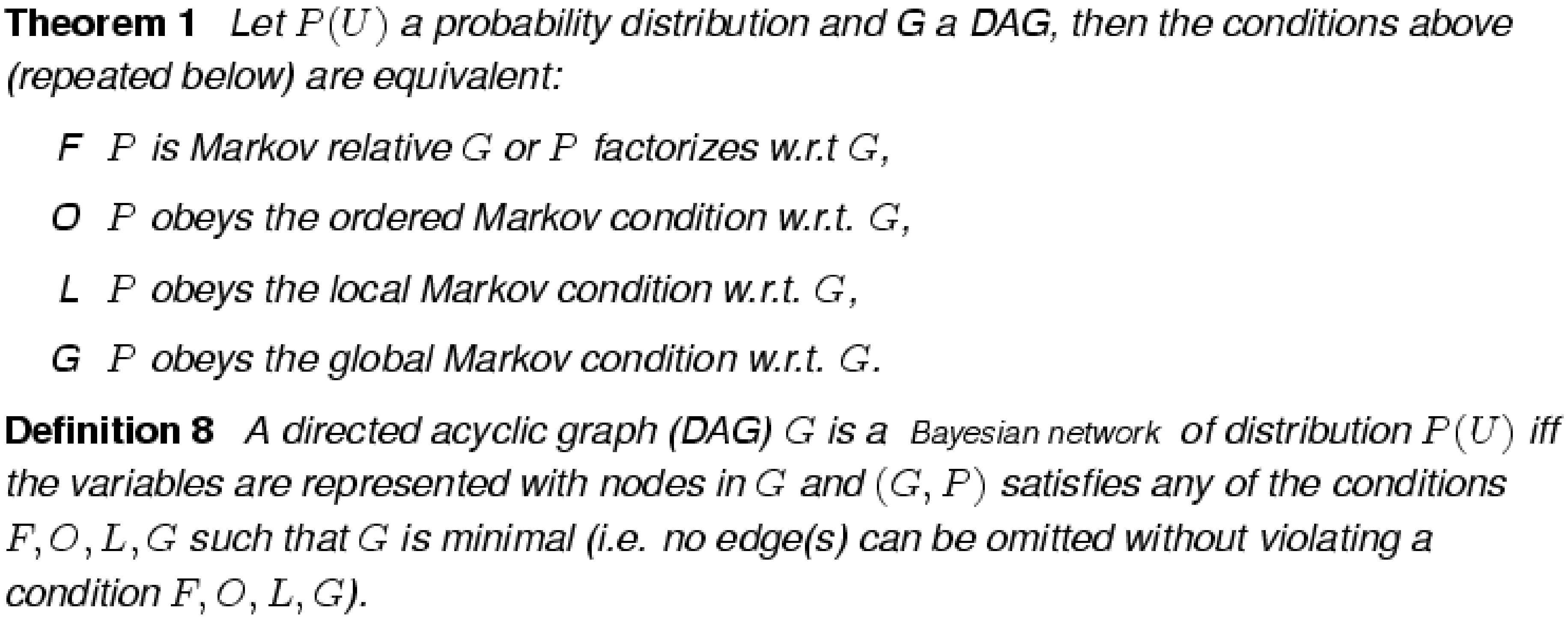


An almost always complete calculus for independencies

Counterexamples (parametrically encoded independencies)

- Binary XOR
- Intransitive Markov chain

Bayesian network definitions

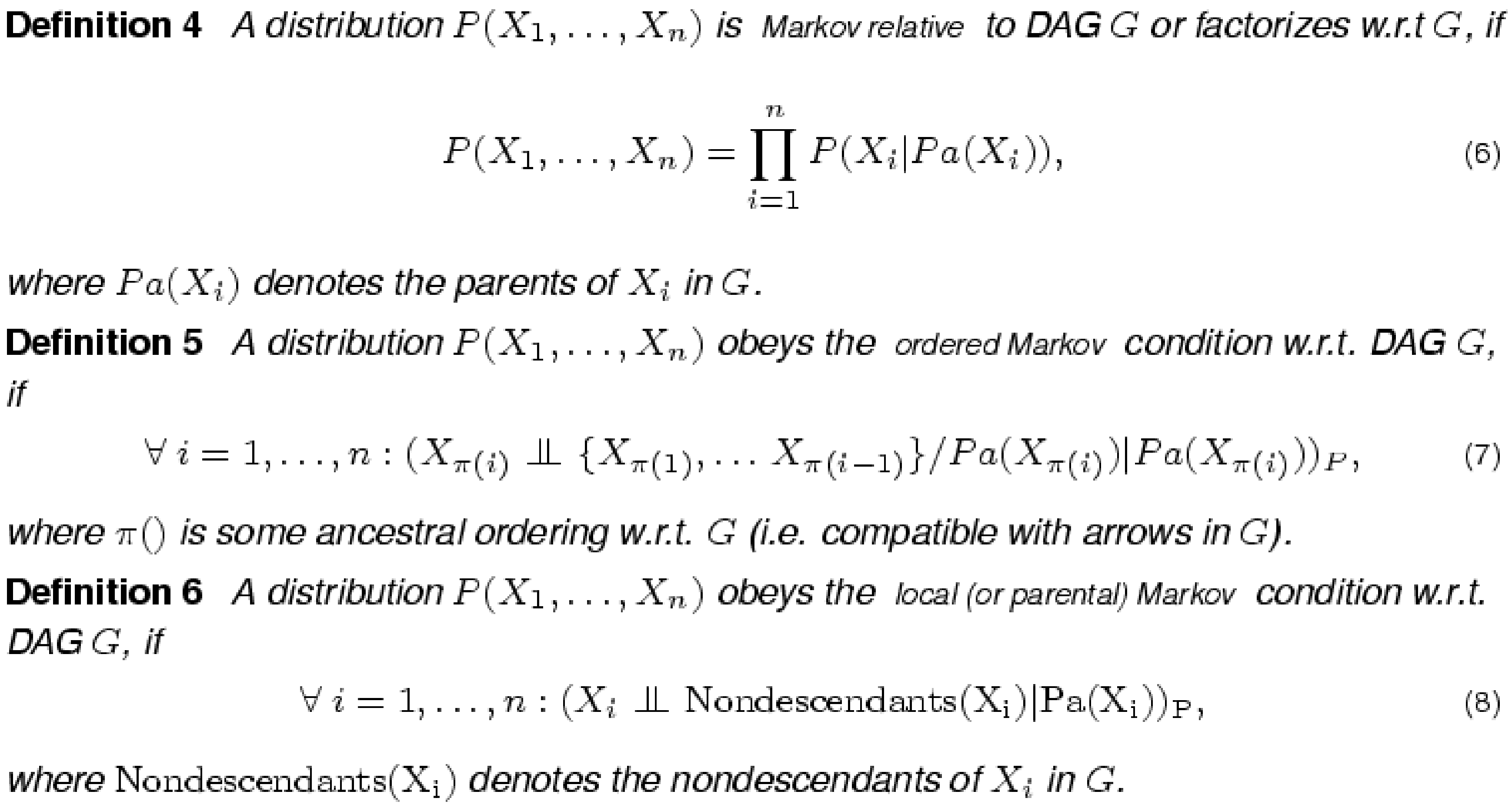


Theorem 1 *Let $P(U)$ a probability distribution and **G** a **DAG**, then the conditions above (repeated below) are equivalent:*

- ***F*** *P is Markov relative G or P factorizes w.r.t G,*
- ***O*** *P obeys the ordered Markov condition w.r.t. G,*
- ***L*** *P obeys the local Markov condition w.r.t. G,*
- ***G*** *P obeys the global Markov condition w.r.t. G.*

Definition 8 *A directed acyclic graph **(DAG)** G is a* Bayesian network *of distribution $P(U)$ iff the variables are represented with nodes in G and (G, P) satisfies any of the conditions F, O, L, G such that G is minimal (i.e. no edge(s) can be omitted without violating a condition F, O, L, G).*

Markov conditions



Definition 4 *A distribution $P(X_1, \dots, X_n)$ is* Markov relative *to DAG G or factorizes w.r.t G, if*

$$P(X_1, \dots, X_n) = \prod_{i=1}^{n} P(X_i | Pa(X_i)), \tag{6}$$

where $Pa(X_i)$ denotes the parents of X_i in G.

Definition 5 *A distribution $P(X_1, \dots, X_n)$ obeys the* ordered Markov *condition w.r.t. DAG G, if*

$$\forall\, i = 1, \dots, n : (X_{\pi(i)} \perp\!\!\!\perp \{X_{\pi(1)}, \dots X_{\pi(i-1)}\} / Pa(X_{\pi(i)}) | Pa(X_{\pi(i)}))_P, \tag{7}$$

where $\pi()$ is some ancestral ordering w.r.t. G (i.e. compatible with arrows in G).

Definition 6 *A distribution $P(X_1, \dots, X_n)$ obeys the* local (or parental) Markov *condition w.r.t. DAG G, if*

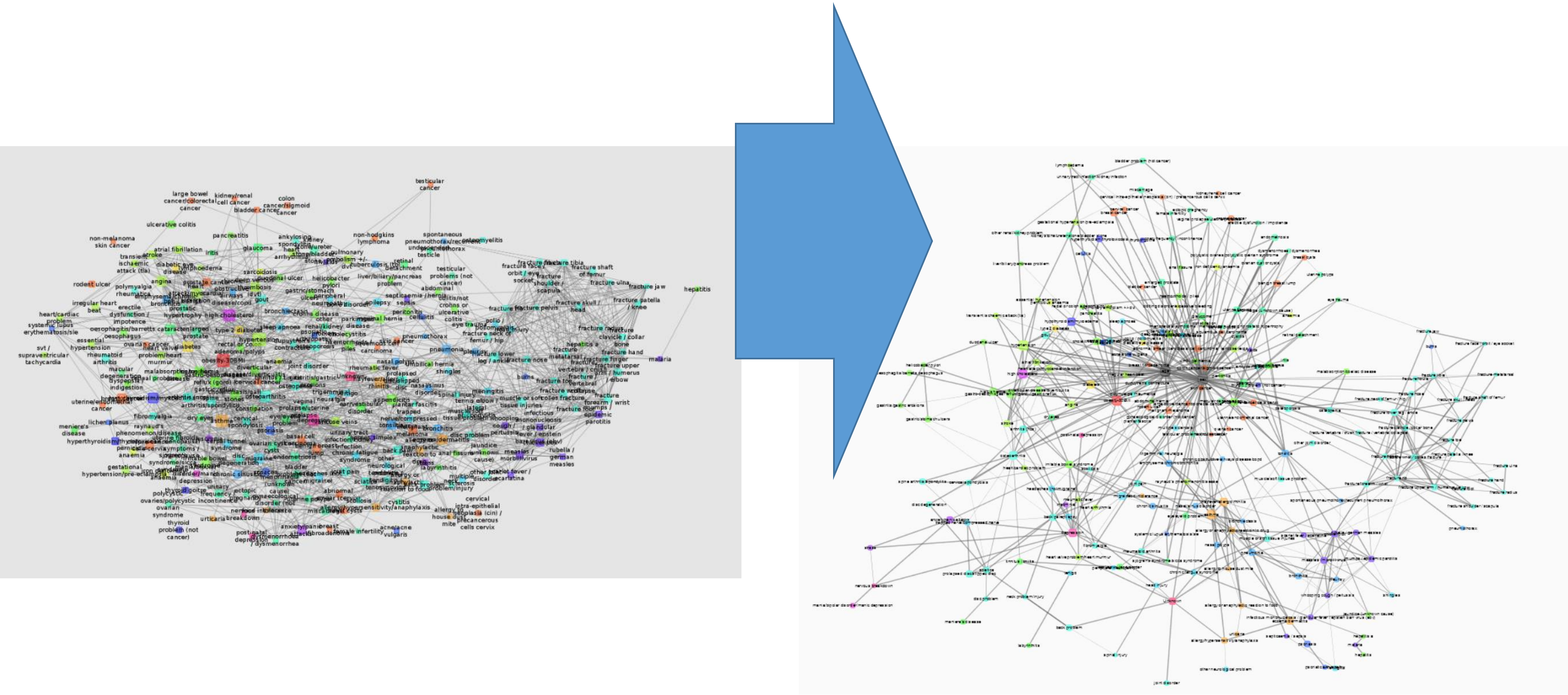
$$\forall\, i = 1, \dots, n : (X_i \perp\!\!\!\perp \mathrm{Nondescendants(X_i)} | \mathrm{Pa(X_i)})_{\mathrm{P}}, \tag{8}$$

where $\mathrm{Nondescendants(X_i)}$ denotes the nondescendants of X_i in G.

Relativity of the interpretations

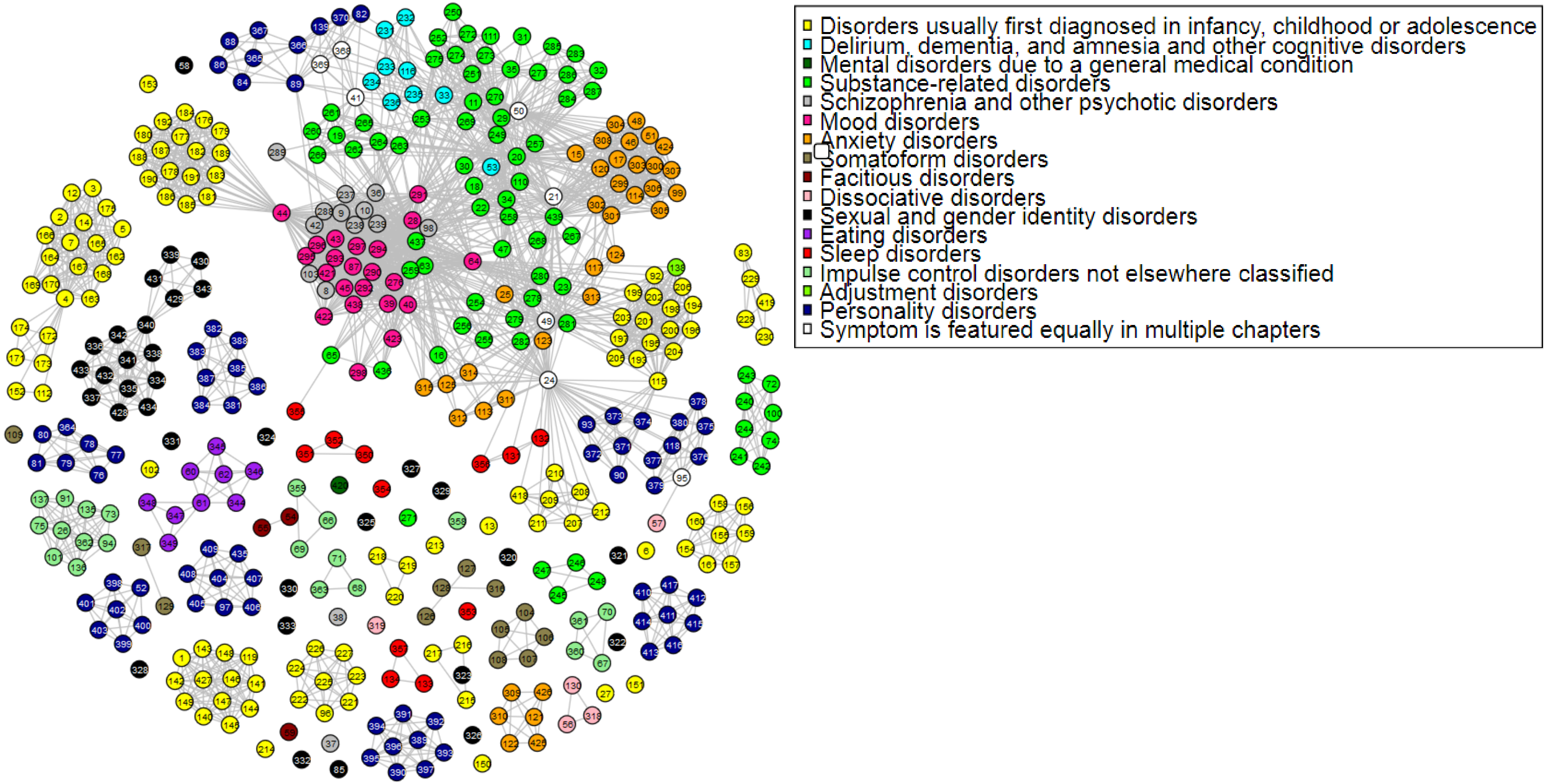
1. The presence of unobserved (hidden) variables as potential confounders.
2. Selection bias can occur if the observation depends on the joint combination of otherwise independent events, inducing non-causal dependencies between them.
3. The mixture of causal models, if conditionally both *X* causes *Y* and vice versa. A similar problem is the presence of feedback (and indirectly temporality).
4. Global physical and semantic constraints between the variables.
5. Stability can be also questioned, because of deterministic dependencies, resulting in the lack of guarantee for the uniqueness and exactness of the representation.
6. The (in)dependencies are relative to the set of variables and specifically, also to the values of the variables

Multimorbidity network



Marx, P., Antal, P., Bolgar, B., Bagdy, G., Deakin, B. and Juhasz, G., 2017. Comorbidities in the diseasome are more apparent than real: What Bayesian filtering reveals about the comorbidities of depression. *PLoS computational biology*, *13*(6), p.e1005487.

Probabilistic graphical models: Markov networks, pairwise Markov Random Fields

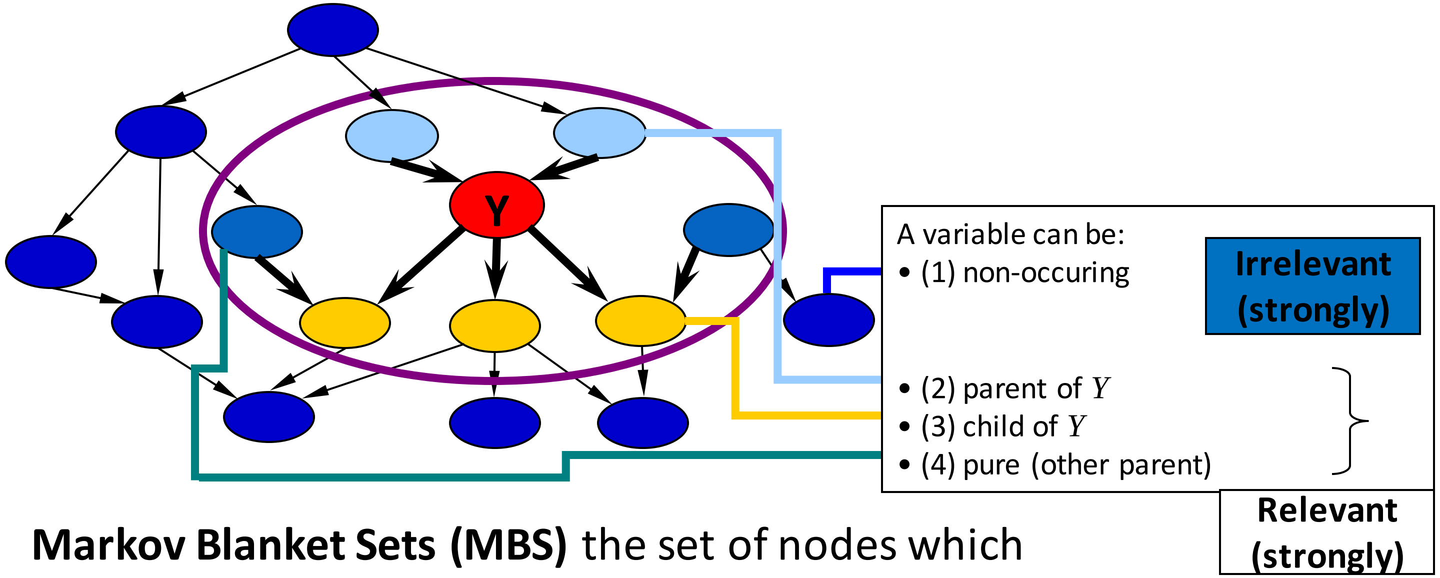


- ■ Disorders usually first diagnosed in infancy, childhood or adolescence
- ■ Delirium, dementia, and amnesia and other cognitive disorders
- ■ Mental disorders due to a general medical condition
- ■ Substance-related disorders
- ■ Schizophrenia and other psychotic disorders
- ■ Mood disorders
- ■ Anxiety disorders
- ■ Somatoform disorders
- ■ Facitious disorders
- ■ Dissociative disorders
- ■ Sexual and gender identity disorders
- ■ Eating disorders
- ■ Sleep disorders
- ■ Impulse control disorders not elsewhere classified
- ■ Adjustment disorders
- ■ Personality disorders
- □ Symptom is featured equally in multiple chapters

Borsboom, D. and Cramer, A.O., 2013. Network analysis: an integrative approach to the structure of psychopathology. *Annual review of clinical psychology*, *9*, pp.91-121.

The Markov Blanket

A minimal sufficient set for prediction/diagnosis.



A variable can be:
- (1) non-occuring
- (2) parent of Y
- (3) child of Y
- (4) pure (other parent)

Irrelevant (strongly)

Relevant (strongly)

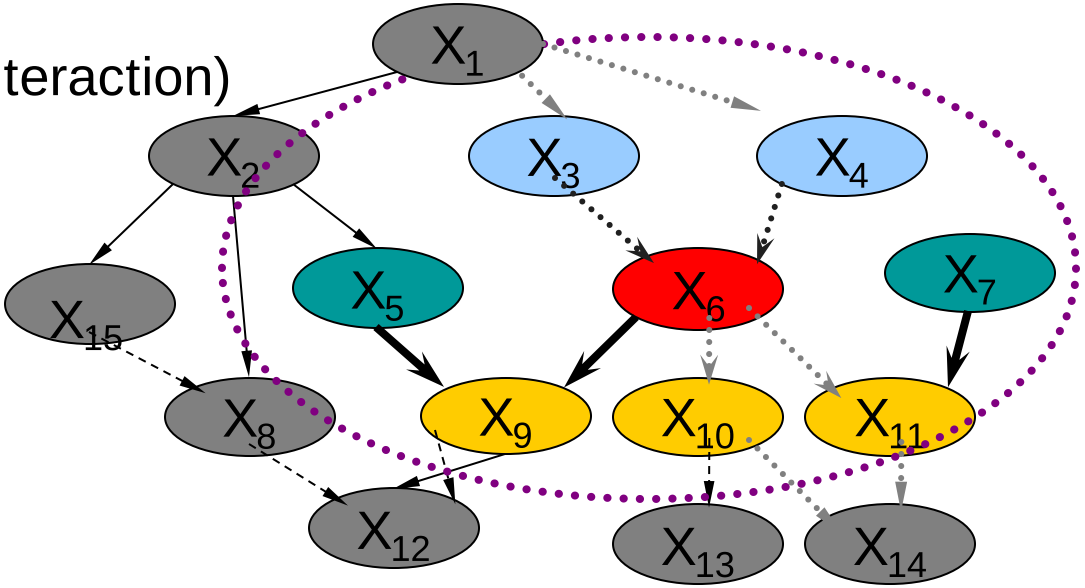
Markov Blanket Sets (MBS) the set of nodes which probabilistically isolate the target from the rest of the model

Markov Blanket Membership (MBM)

(symmetric) pairwise relationship induced by MBS

A more detailed language for associations: typed relevance

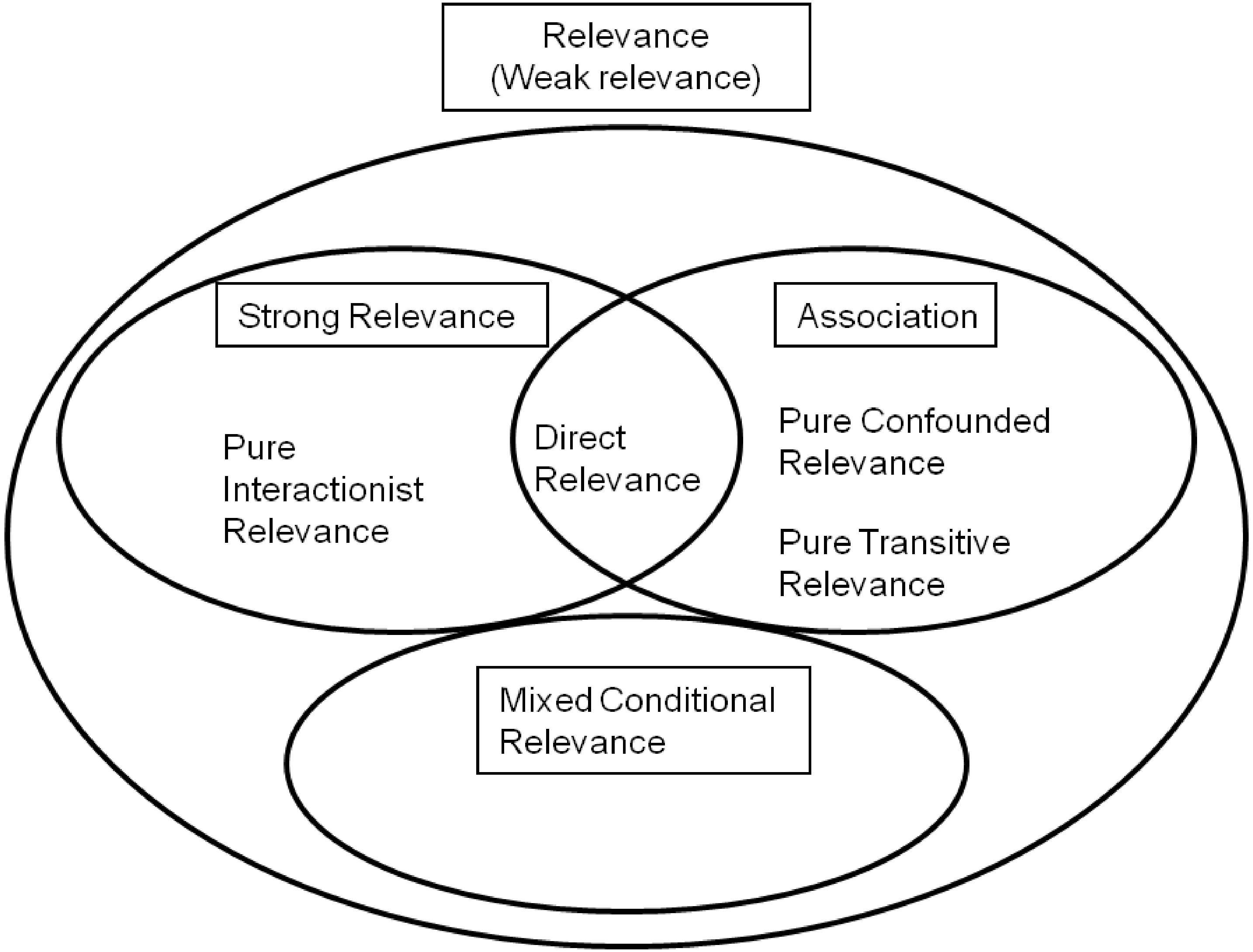
- Weak relevance
- Strong relevance
- Conditional relevance (pure interaction)
- Direct relevancia
  - With hidden variable
  - No hidden variable
- Causal relevance
- Effect modifier
  - Probabilistic, direct, causal



- Typed relevance
  - Parent, Child
  - Direct=Parent or Child
  - Ascendant=Parent+, Descendant=Child+
  - Markovian=Parent, or Child or Pure interaction
  - Confounded
  - Associated= Ascendant or Descendant or Confounded

Antal, Péter, et al. "A Bayesian view of challenges in feature selection: feature aggregation, multiple targets, redundancy and interaction." *New Challenges for Feature Selection in Data Mining and Knowledge Discovery*. 2008.

A more detailed language for associations: typed relevance



Relevance
(Weak relevance)

Strong Relevance

Pure
Interactionist
Relevance

Direct
Relevance

Association

Pure Confounded
Relevance

Pure Transitive
Relevance

Mixed Conditional
Relevance

Towards causal inference

Principles of causality

- strong association,
- X precedes temporally Y,
- plausible explanation without alternative explanations based on confounding,
- necessity (generally: if cause is removed, effect is decreased or actually: y would not have been occurred with that much probability if x had not been present),
- sufficiency (generally: if exposure to cause is increased, effect is increased or actually: y would have been occurred with larger probability if x had been present).

- Autonomous, transportable mechanism.

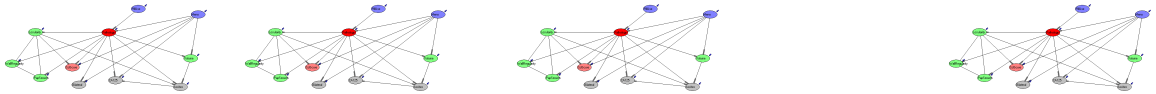
- The probabilistic definition of causation formalizes many, but for example not the counterfactual aspects.

Questions

- Can we represent exactly (in)dependencies by a BN?
  - From a causal model? Suff.&nec.?
- Can we interpret
  - edges as causal relations
    - with no hidden variables?
    - in the presence of hidden variables?
  - local models as autonomous mechanisms?
- Can we infer the effect of interventions?

Observational equivalence of causal models

Causal models:



J.Pearl:
~„3D objects"

From passive observations:

$P(X_1,\ldots, X_n)$

$M_P=\{I_{P,1}(X_1;Y_1|Z_1),\ldots, I_{P,K}(X_K;Y_K|Z_K)\}$

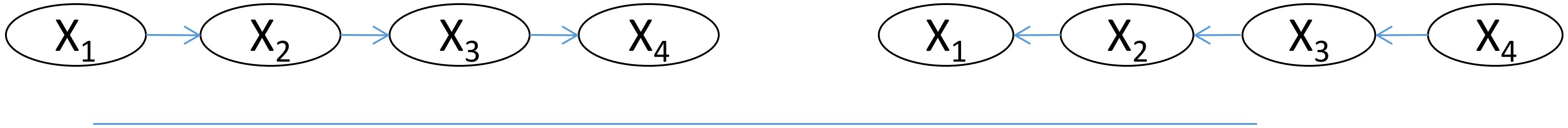
„2D projection"

Different causal models can have the same independence map!

Typically causal models cannot be identified from passive observations, they are ***observationally equivalent***.

Association vs. Causation: Markov chain

Causal models:



Markov chain

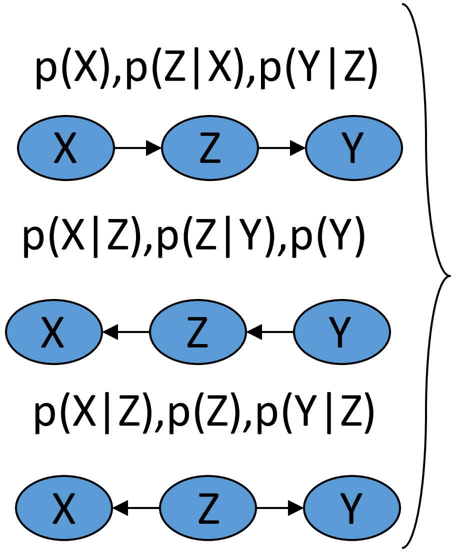
$P(X_1,\ldots)$

$M_P=\{I(X_{i+1};X_{i-1}|X_i)\}$

„first order Markov property”

Flow of time?

The building block of causality: v-structure (arrow of time)

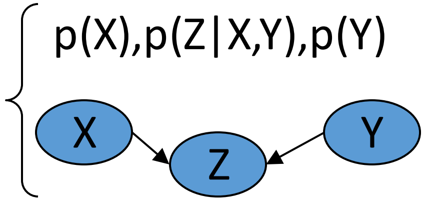


p(X),p(Z|X),p(Y|Z)

p(X|Z),p(Z|Y),p(Y)

p(X|Z),p(Z),p(Y|Z)

“transitive” M ≠ „intransitive” M



p(X),p(Z|X,Y),p(Y)

„v-structure”

M_P={D(X;Z), D(Z;Y), D(X,Y), I(X;Y|Z)}

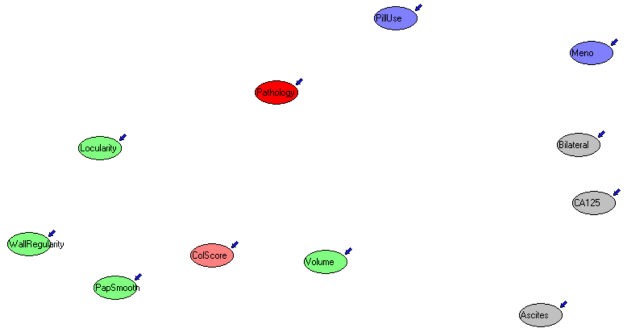
M_P={D(X;Z), D(Y;Z), I(X;Y), D(X;Y|Z) }

Often: present knowledge renders future states conditionally independent.
(confounding)

Ever(?): present knowledge renders past states conditionally independent.
(backward/atemporal confounding)

Observational equivalence: total independence

„Causal” model:



One-to-one relation

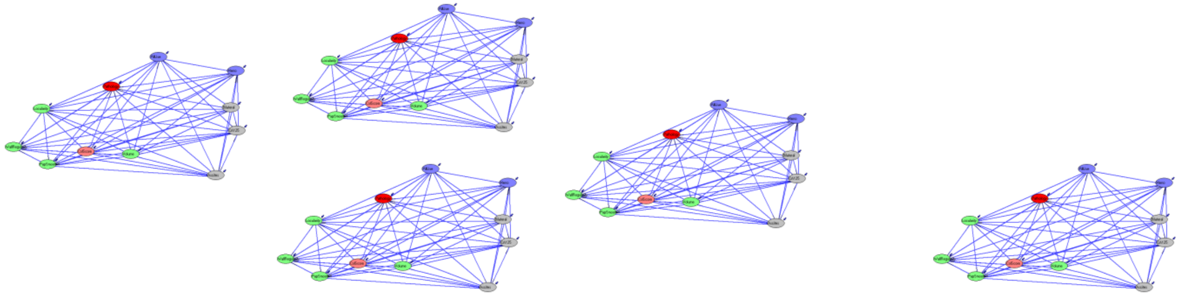
Dependency map:

$P(X_1,\ldots, X_n)$

$M_P=\{I_{P,1}(X_1;X_2),\ldots\}$

Observational equivalence: full dependence

„Causal” models (there is a DAG for each ordering, i.e. n! DAGs):



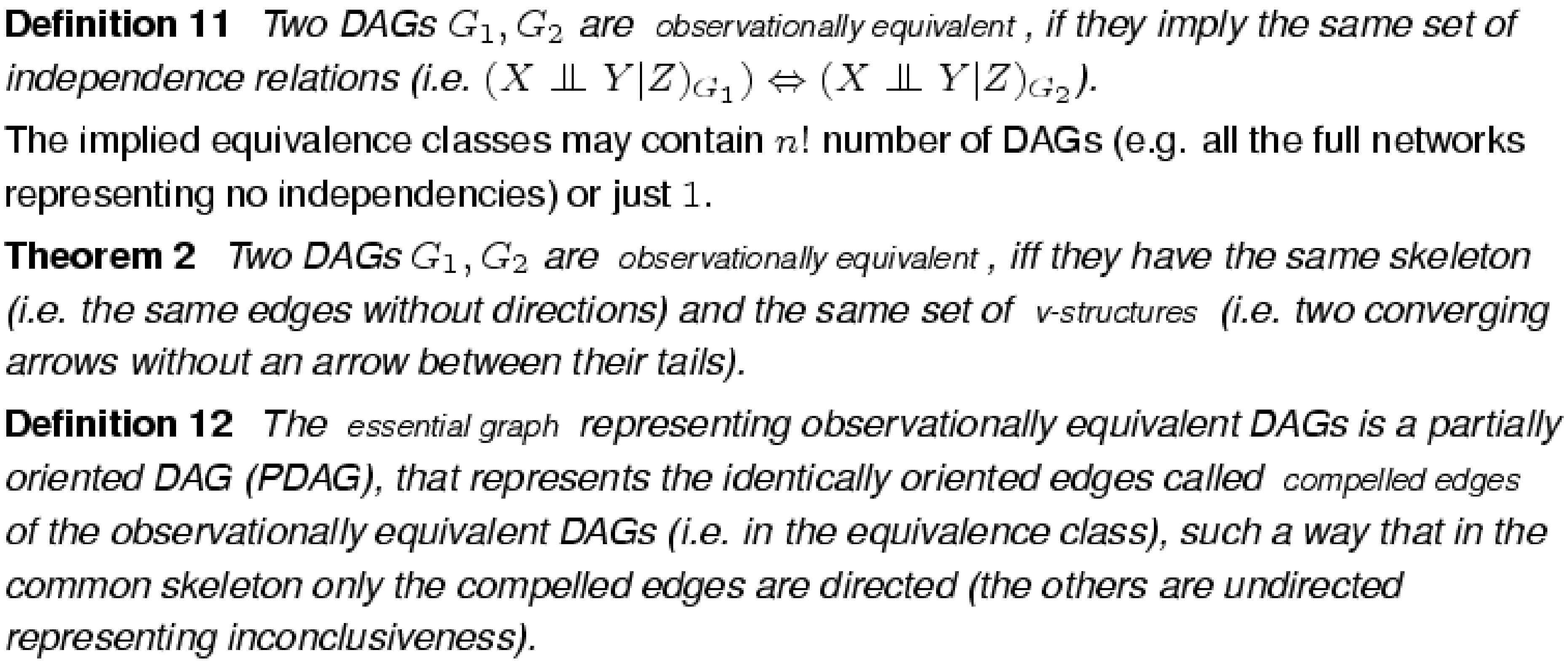
One-to-many relation

Dependency map:

$P(X_1,\ldots, X_n)$

$M_P=\{D_{P,1}(X_1;X_2),\ldots\}$

Observational equivalence of causal models



Definition 11 *Two DAGs G_1, G_2 are observationally equivalent , if they imply the same set of independence relations (i.e. $(X \perp\!\!\!\perp Y|Z)_{G_1}) \Leftrightarrow (X \perp\!\!\!\perp Y|Z)_{G_2}$).*

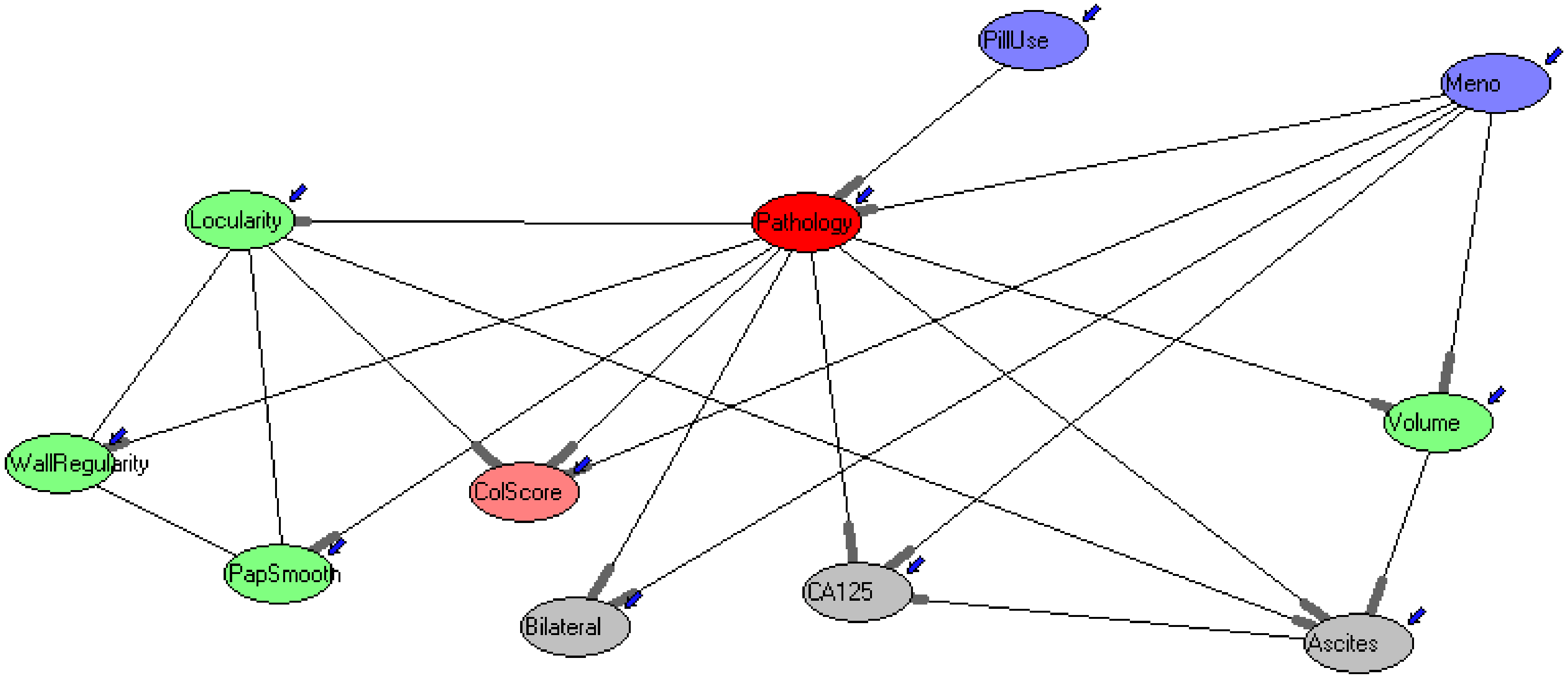
The implied equivalence classes may contain $n!$ number of DAGs (e.g. all the full networks representing no independencies) or just 1.

Theorem 2 *Two DAGs G_1, G_2 are observationally equivalent , iff they have the same skeleton (i.e. the same edges without directions) and the same set of v-structures (i.e. two converging arrows without an arrow between their tails).*

Definition 12 *The essential graph representing observationally equivalent DAGs is a partially oriented DAG (PDAG), that represents the identically oriented edges called compelled edges of the observationally equivalent DAGs (i.e. in the equivalence class), such a way that in the common skeleton only the compelled edges are directed (the others are undirected representing inconclusiveness).*

Compelled edges and PDAG

(“can we interpret edges as causal relations?”➔compelled edges)



The Causal Markov Condition

- A DAG is called a *causal structure* over a set of variables, if each node represents a variable and edges direct influences. A *causal model* is a causal structure extended with local probabilistic models.
- A causal structure *G* and distribution *P* satisfies the Causal Markov Condition, if P obeys the local Markov condition w.r.t. G.
- The distribution P is said to stable (or faithful), if there exists a DAG called *perfect map* exactly representing its (in)dependencies (i.e. $I_G(X;Y|Z) \Leftrightarrow I_P(X;Y|Z)\ \forall\ X,Y,Z \subseteq V$).

- CMC: **sufficiency** of G (there are no extra, acausal dependencies)
- Faithfulness/stability: **necessity** of G (there are no extra, parametric independencies)

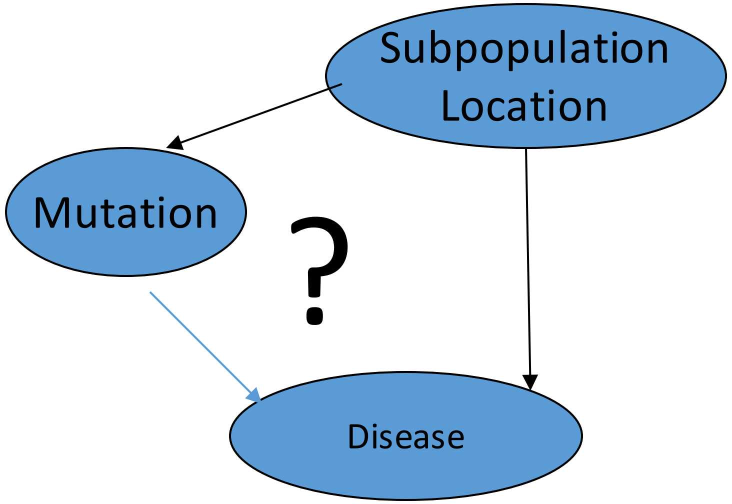
Interventional inference in causal Bayesian networks

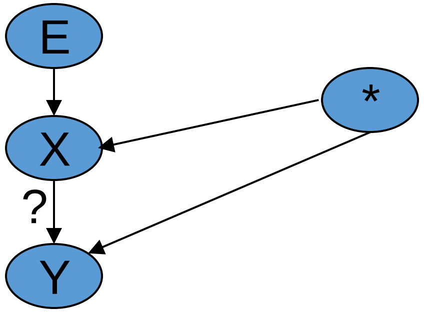
- (Passive, observational) inference
  - P(Query|Observations)
- **Interventionist inference**
  - **P(Query|Observations, Interventions)**
- Counterfactual inference
  - P(Query| Observations, Counterfactual conditionals)

Interventions and graph surgery

If G is a causal model, then compute p(Y|do(X=x)) by

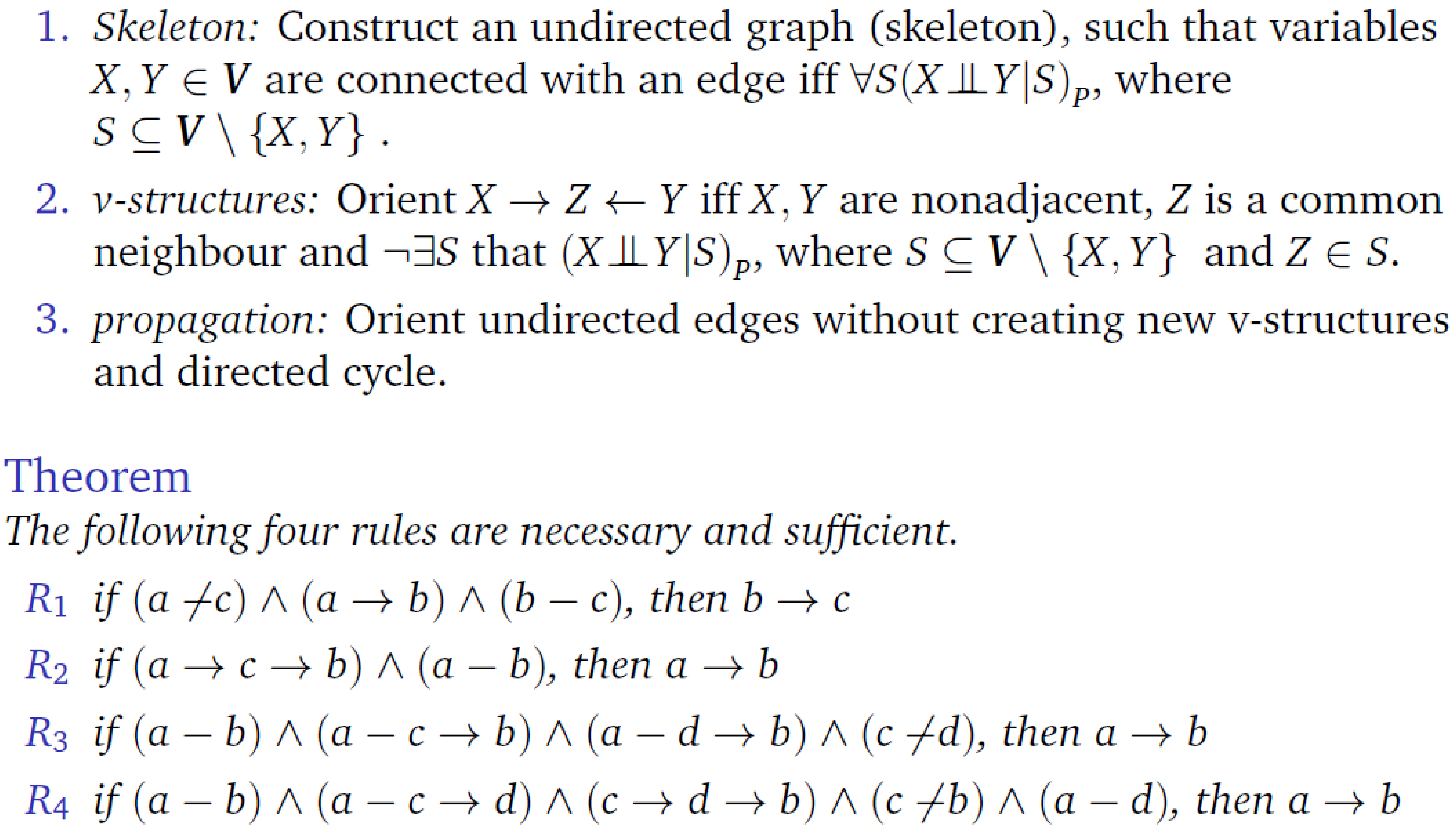
1. deleting the incoming edges to X
2. setting X=x
3. performing standard Bayesian network inference.





Learning causal relations and models

Inductive Causation (asymptotic, no hidden)



1. *Skeleton:* Construct an undirected graph (skeleton), such that variables $X, Y \in \mathbf{V}$ are connected with an edge iff $\forall S (X \perp\!\!\!\perp Y | S)_P$, where $S \subseteq \mathbf{V} \setminus \{X, Y\}$.
2. *v-structures:* Orient $X \rightarrow Z \leftarrow Y$ iff X, Y are nonadjacent, Z is a common neighbour and $\neg \exists S$ that $(X \perp\!\!\!\perp Y | S)_P$, where $S \subseteq \mathbf{V} \setminus \{X, Y\}$ and $Z \in S$.
3. *propagation:* Orient undirected edges without creating new v-structures and directed cycle.

Theorem

The following four rules are necessary and sufficient.

R_1 *if* $(a \not- c) \wedge (a \rightarrow b) \wedge (b - c)$, *then* $b \rightarrow c$

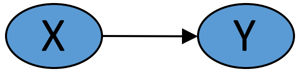
R_2 *if* $(a \rightarrow c \rightarrow b) \wedge (a - b)$, *then* $a \rightarrow b$

R_3 *if* $(a - b) \wedge (a - c \rightarrow b) \wedge (a - d \rightarrow b) \wedge (c \not- d)$, *then* $a \rightarrow b$

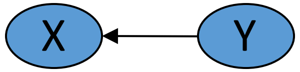
R_4 *if* $(a - b) \wedge (a - c \rightarrow d) \wedge (c \rightarrow d \rightarrow b) \wedge (c \not- b) \wedge (a - d)$, *then* $a \rightarrow b$

Association vs. Causation

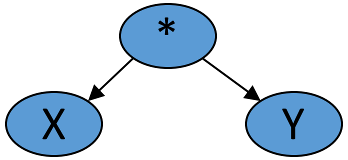
Causal models:



X causes Y

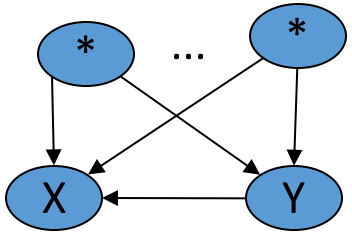


Y causes X



There is a common cause (pure confounding)

…

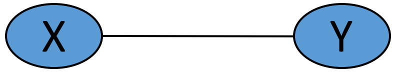


Causal effect of Y on X is confounded by many factors

From passive observations:

P(X,Y)

M_P={D(X;Y)}



„X and Y are associated”

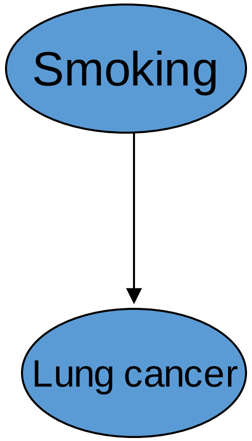
Reichenbach's Common Cause Principle:

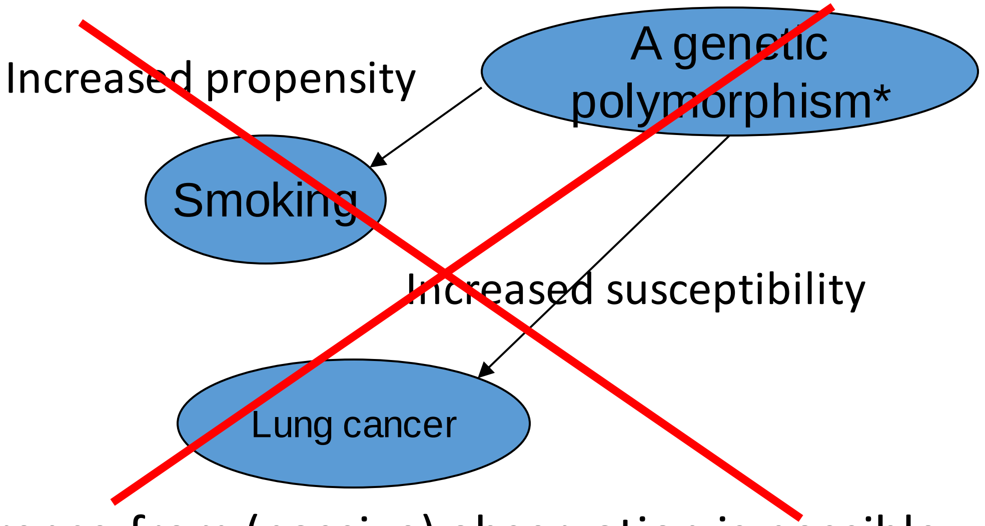
a correlation between events *X* and *Y* indicates either that *X* causes *Y*, or that *Y* causes *X*, or that *X* and *Y* have a common cause.

Local Causal Discovery

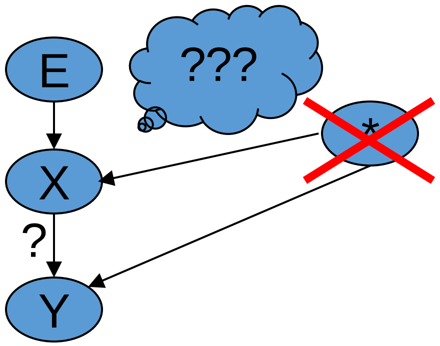
“can we interpret edges as causal relations in the presence of hidden variables?”

- Can we learn causal relations from observational data in presence of confounders???





- Automated, tabula rasa causal inference from (passive) observation is possible, i.e. hidden, confounding variables can be excluded



A deterministic concept of causation

- H.Simon
  - $X_i=f_i(X_1,..,X_{i-1})$ for i=1..n
  - In the linear case the sytem of equations indicates a natural causal ordering (flow of time?)

					X
				X	X
			X	X	X
		X	X	X	X
	....				

The probabilistic conceptualization is its generalization:

$$P(X_i,|X_1,..,X_{i-1}) \sim X_i=f_i(X_1,..,X_{i-1})$$

A posteriori probability of a „causal” ordering…

Towards counterfactual inference

Functional (causal) Bayesian network

The axiomatic foundation for the graph surgery semantics of the $P(.|do(.),.)$ notation.

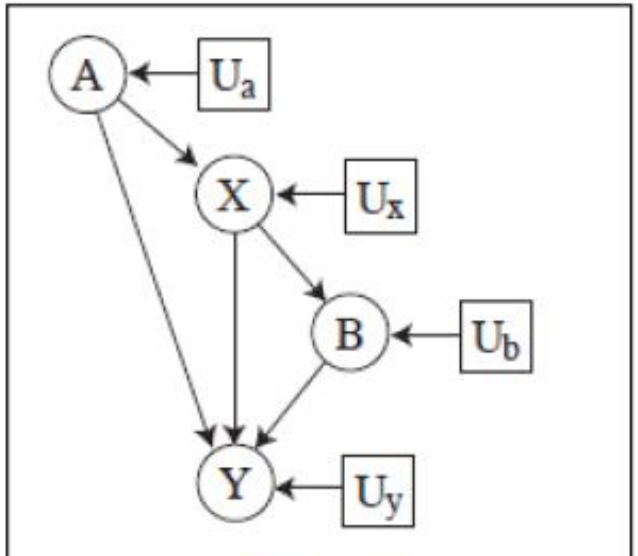
Definition

Let $p(\boldsymbol{V}|do(\boldsymbol{x})$ denote an interventional distribution corresponding to setting variable(s) $\boldsymbol{X} \subseteq \boldsymbol{V}$ to value $\boldsymbol{x}$ and P_* the set of all interventional distributions (including $p(\boldsymbol{V}|do(0))$ the observational target distribution without intervention). A DAG G is said to be a causal Bayesian network compatible with P_* iff for each $p(\boldsymbol{V}|do(\boldsymbol{x})) \in P_*$ the following three conditions hold

1. $p(\boldsymbol{V}|do(\boldsymbol{x})$ is Markov relative to G,
2. $\forall X_i \in \boldsymbol{X}\ p(x_i|do(\boldsymbol{x}) = 1$ if value x_i and $\boldsymbol{x}$ is compatible,
3. $\forall X_i \notin \boldsymbol{X}\ p(x_i|pa_i, do(\boldsymbol{x}) = p(x_i|pa_i)$ if value(s) pa_i and $\boldsymbol{x}$ is compatible.

Counterfactuals I.

- Observe $X = x$ and $Y = y$
- What is the probability, that Y would have attained the value y' if X had been x'? (here y and y' can be equal but $x' \neq x$)
- Variables A and B can be either observed or hidden, but the full model (graph, functions, and $P(\mathbf{U})$) is assumed to be known



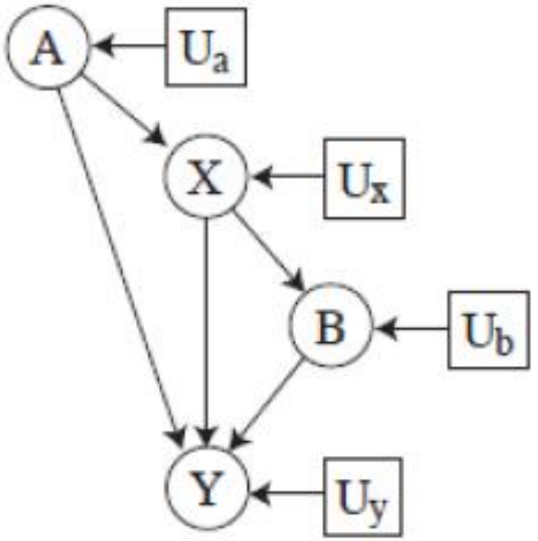
$$a := f_a(u_a)$$
$$x := f_x(a, u_x)$$
$$b := f_b(x, u_b)$$
$$y := f_y(a, x, b, u_y)$$

- Interpreting the question: We assume a minimal change of mechanism, i.e. we set X into state x' without changing anything else, i.e: $\mathrm{do}(X = x')$
- Interpreting the question: We assume that the disturbance variables $\mathbf{U} \backslash U_x = \{U_a, U_b, U_y\}$ are persistent, i.e. do not change

Counterfactuals II.

- With these specifications, we have a well defined probability:
  $P(Y_{x'} = y' \mid X = x, Y = y)$

  $Y_x(u) =$ 'the value of Y, when the disturbance variables attain the values u and X is set to equal x'.



$$a := f_a(u_a)$$
$$x := f_x(a, u_x)$$
$$b := f_b(x, u_b)$$
$$y := f_y(a, x, b, u_y)$$

Counterfactuals III.

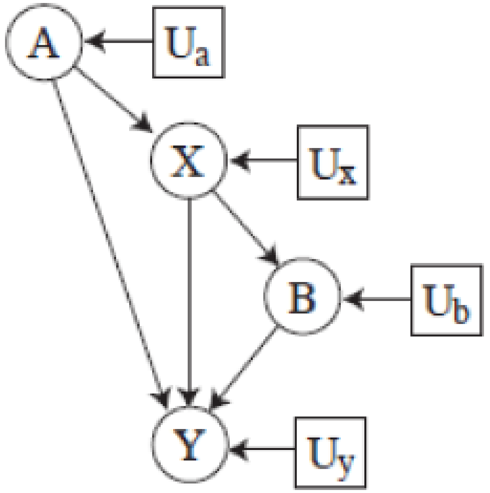
- But, how to calculate $P(Y_{x'} = y' \mid X = x, Y = y)$?

  (Pearl theorem 7.1.7)
  Three steps:

  i. ('abduction'): Calculate the probability distribution over all disturbances, given the evidence e, i.e. $P(\mathbf{U} \mid e)$

  ii. ('action'): Change the model by the intervention $\mathrm{do}(X = x')$, i.e. remove all arrows into X and set its value to x'

  iii. ('prediction') Using the updated model, and the probability distribution $P(\mathbf{U} \mid e)$, calculate $P(Y = y')$



$$a := f_a(u_a)$$
$$x := f_x(a, u_x)$$
$$b := f_b(x, u_b)$$
$$y := f_y(a, x, b, u_y)$$

Summary

- Independence models
- Probabilistic graphical models
    - Bayesian networks
    - Causal interpretation
- Causal inference
    - do-operator
    - local causal inference
- Counterfactual inference