

# **Causal inference and causal discovery-part(2)**

Shikui Tu

Department of Computer Science and  
Engineering, Shanghai Jiao Tong University

2019-06-04

# Outline

- **Discover causal structure by conditional independence**
  - PC algorithm
  - Markov equivalent class
- Pearl's do-calculus



ORIGINAL ARTICLE

## Association of Coffee Drinking with Total and Cause-Specific Mortality

Neal D. Freedman, Ph.D., Yikyung Park, Sc.D., Christian C. Abnet, Ph.D., Albert R. Hollenbeck, Ph.D., and Rashmi Sinha, Ph.D.

### BACKGROUND

Coffee is one of the most widely consumed beverages, but the association between coffee consumption and the risk of death remains unclear.

### METHODS

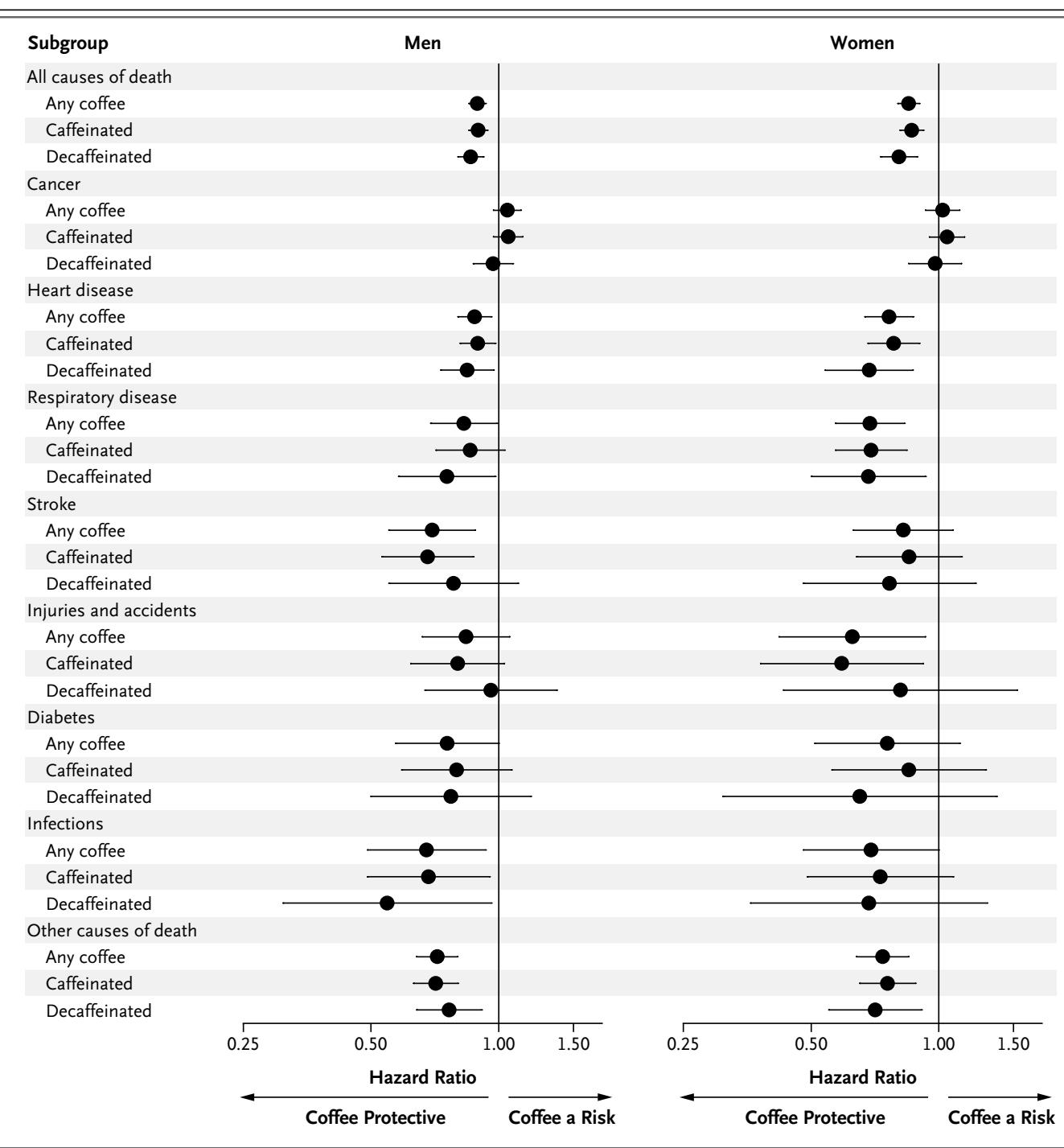
We examined the association of coffee drinking with subsequent total and cause-specific mortality among 229,119 men and 173,141 women in the National Institutes of Health–AARP Diet and Health Study who were 50 to 71 years of age at baseline. Participants with cancer, heart disease, and stroke were excluded. Coffee consumption was assessed once at baseline.

### RESULTS

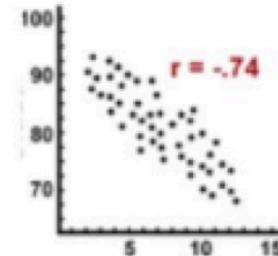
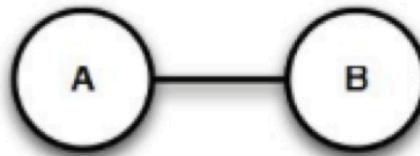
During 5,148,760 person-years of follow-up between 1995 and 2008, a total of 33,731 men and 18,784 women died. In age-adjusted models, the risk of death was increased among coffee drinkers. However, coffee drinkers were also more likely to smoke, and, after adjustment for tobacco-smoking status and other potential confounders, there was a significant inverse association between coffee consumption and mortality.

### CONCLUSIONS

In this large prospective study, coffee consumption was inversely associated with total and cause-specific mortality. Whether this was a causal or associational finding cannot be determined from our data.



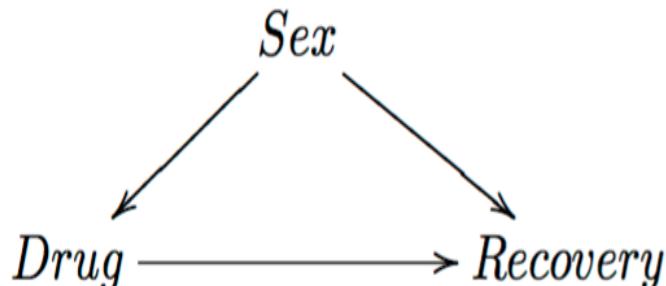
· Association



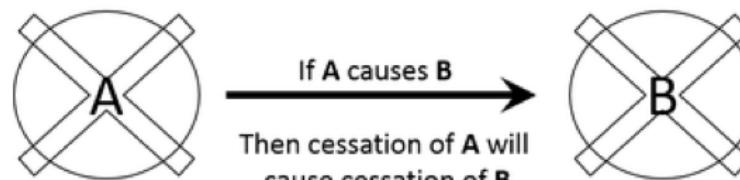
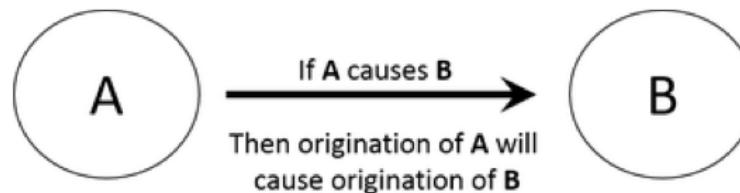
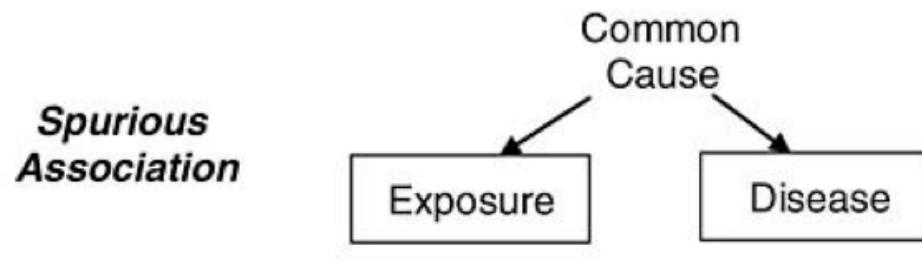
此悖论表明：

$X$  和  $Y$  边缘上正相关，  
但是,给定另外一个变量  $Z$  后，  
在  $Z$  的每一个水平上， $X$  和  $Y$   
可能负相关。

Yule-Simpson Paradox (Pearl, 2000)



合并表	康复	未康复	康复率
吃药	20	20	50%
安慰剂	16	24	40%
男性	康复	未康复	康复率
吃药	18	12	60%
安慰剂	7	3	70%
女性	康复	未康复	康复率
吃药	2	8	20%
安慰剂	9	21	30%



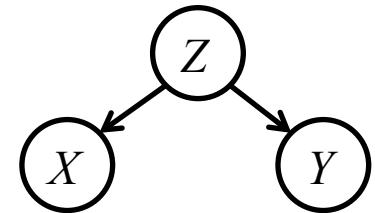
# Statistical Implications of Causality

Reichenbach's  
*Common Cause Principle*  
links **causality** and **probability**:

- (i) if  $X$  and  $Y$  are statistically dependent, then there is a  $Z$  causally influencing both;
- (ii)  $Z$  screens  $X$  and  $Y$  from each other (given  $Z$ , the observables  $X$  and  $Y$  become independent)



(Reichenbach 1956)



special cases:



## Independence of random variables

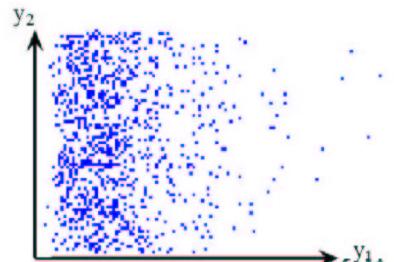
Two real-valued random variables  $X$  and  $Y$  are called *independent*,

$$X \perp\!\!\!\perp Y,$$

if for every  $a, b \in \mathbb{R}$ , the events  $\{X \leq a\}$  and  $\{Y \leq b\}$  are independent.

Equivalently, in terms of densities: for all  $x, y$ ,

$$p(x, y) = p(x)p(y)$$



Note:

If  $X \perp\!\!\!\perp Y$ , then  $E[XY] = E[X]E[Y]$ , and  $\text{cov}[X, Y] = E[XY] - E[X]E[Y] = 0$ .

The converse is not true:  $\text{cov}[X, Y] = 0 \not\Rightarrow X \perp\!\!\!\perp Y$ .

However, we have, for large  $\mathcal{F}$ :  $(\forall f, g \in \mathcal{F} : \text{cov}[f(X), g(Y)] = 0) \Rightarrow X \perp\!\!\!\perp Y$

## Conditional Independence of random variables

Two real-valued random variables  $X$  and  $Y$  are called *conditionally independent* given  $Z$ ,

$$(X \perp\!\!\!\perp Y) | Z \text{ or } X \perp\!\!\!\perp Y | Z \text{ or } (X \perp\!\!\!\perp Y | Z)_p$$

if

$$p(x, y | z) = p(x | z)p(y | z)$$

for all  $x, y$ , and for all  $z$  s.t.  $p(z) > 0$ .

Note: conditional independence neither implies nor is implied by independence.

I.e., there are  $X, Y, Z$  such that we have only independence or only conditional independence.

# Conditional independence tests

- discrete case: contingency tables

		不吸烟	吸烟	合计
年龄 < 40	呼吸正常	567	874	1441
	呼吸不正常	14	28	42
年龄 40-59	呼吸正常	328	780	1108
	呼吸不正常	2	68	70

$$\left( \frac{p(x, y)}{p(x)p(y)} - 1 \right)$$

C	A	B		
		B <sub>1</sub>	...	B <sub>c</sub>
C <sub>1</sub>	A <sub>1</sub>	p <sub>111</sub>	...	p <sub>1cl</sub>
	:	:	...	:
C <sub>t</sub>	A <sub>r</sub>	p <sub>r11</sub>	...	p <sub>rc1</sub>
	:	:	...	:
C <sub>t</sub>	A <sub>1</sub>	p <sub>11t</sub>	...	p <sub>1ct</sub>
	:	:	...	:
C <sub>t</sub>	A <sub>r</sub>	p <sub>r1t</sub>	...	p <sub>rc1t</sub>

- multivariate gaussian case: covariance matrix

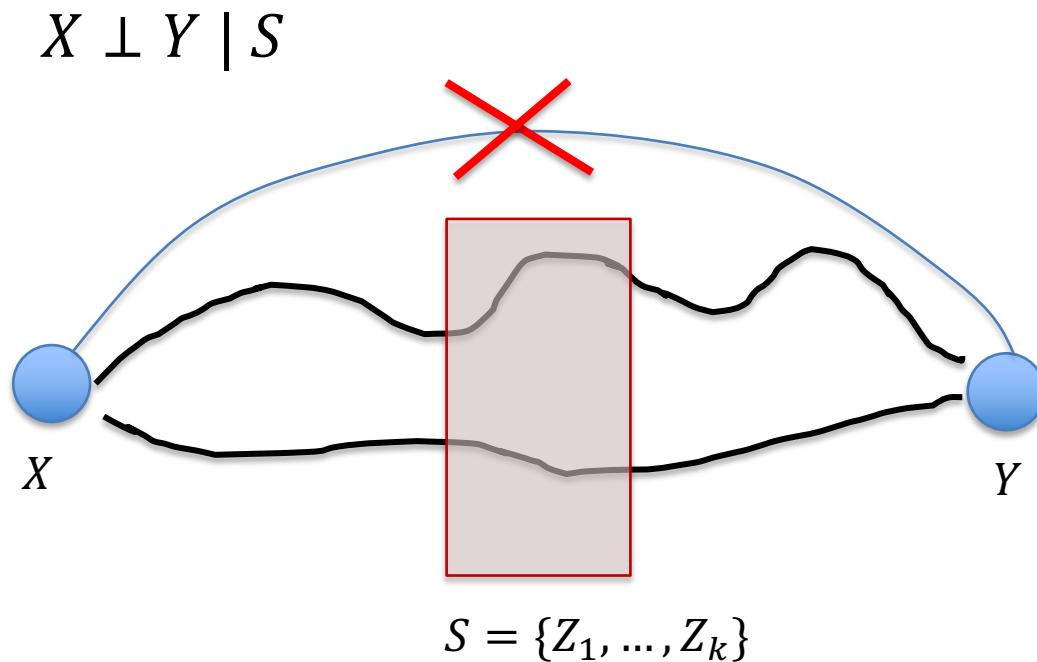
$$\Sigma_s = \begin{array}{c|cc} \sigma_{ww} & \sigma_{1w} & \sigma_{2w} \\ \hline \sigma_{w1} & & \\ \sigma_{w2} & & \end{array} \quad \Rightarrow \quad \sigma_{ij} - \sigma_{wi}\sigma_{jw} / \sigma_{ww} = 0, \quad i \neq j \text{ and } i, j = 1, 2,$$

$$\int \log \left( \frac{p(x, y)}{p(x)p(y)} \right) dp(x, y)$$

- non-Gaussian continuous case: via reproducing kernel Hilbert spaces

# Discover causal structure by conditional independence

- Basic idea:

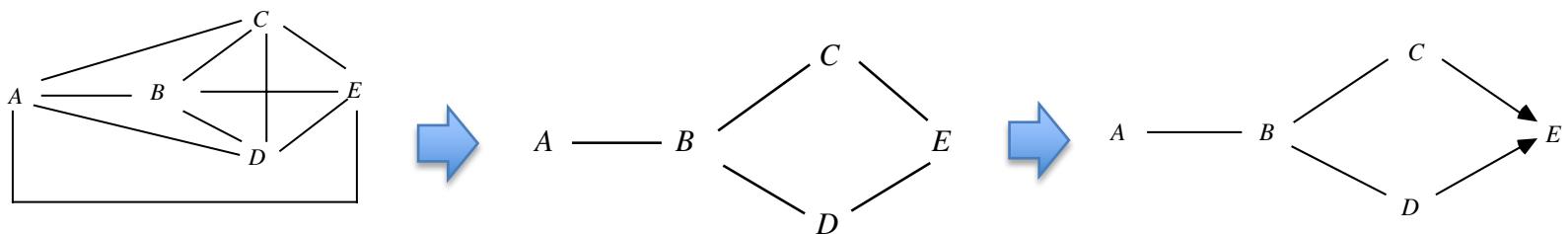


$X$  and  $Y$  are  $d$ -separated by  $S$ .

# PC Algorithm

[Spirtes, Glymour, and Scheines 1991]

- Four steps.
  - A.) Form the complete undirected graph;
  - B.) Remove edges according to n-order conditional independence relations;
  - C.) Orient edges by v-structures
  - D.) Orient edges



# Step B: Remove edges

B.)

$n = 0.$

repeat

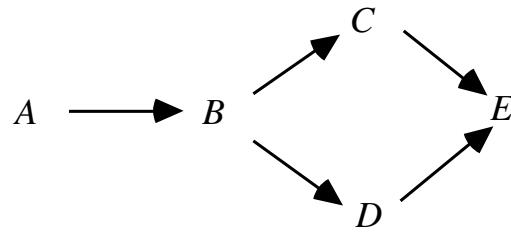
repeat

select an ordered pair of variables  $X$  and  $Y$  that are adjacent in  $C$  such that **Adjacencies**( $C,X\backslash\{Y\}$ ) has cardinality greater than or equal to  $n$ , and a subset  $S$  of **Adjacencies**( $C,X\backslash\{Y\}$ ) of cardinality  $n$ , and if  $X$  and  $Y$  are d-separated given  $S$  delete edge  $X - Y$  from  $C$  and record  $S$  in **Sepset**( $X,Y$ ) and **Sepset**( $Y,X$ );

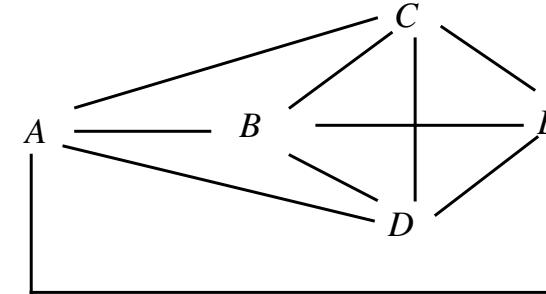
until all ordered pairs of adjacent variables  $X$  and  $Y$  such that **Adjacencies**( $C,X\backslash\{Y\}$ ) has cardinality greater than or equal to  $n$  and all subsets  $S$  of **Adjacencies**( $C,X\backslash\{Y\}$ ) of cardinality  $n$  have been tested for d-separation;

$n = n + 1;$

until for each ordered pair of adjacent vertices  $X, Y$ , **Adjacencies**( $C,X\backslash\{Y\}$ ) is of cardinality less than  $n$ .



**True Graph**



**Complete Undirected Graph**

---

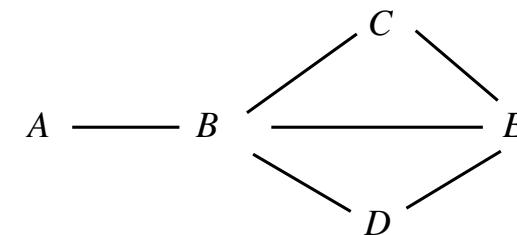
$n = 0$       No zero order independencies.

---

$n = 1$       First order independencies.

$$\begin{array}{ll} A \perp C \mid B & A \perp D \mid B \\ A \perp E \mid B & C \perp D \mid B \end{array}$$

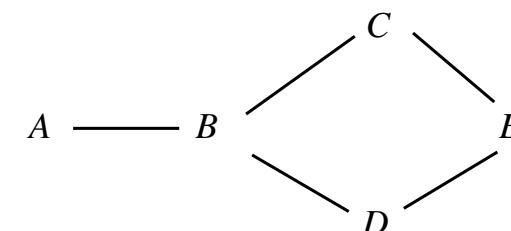
**Resulting Adjacencies**



$n = 2$       Second order independencies.

$$B \perp E \mid \{C, D\}$$

**Resulting Adjacencies**



# Step C-D: Orient edges

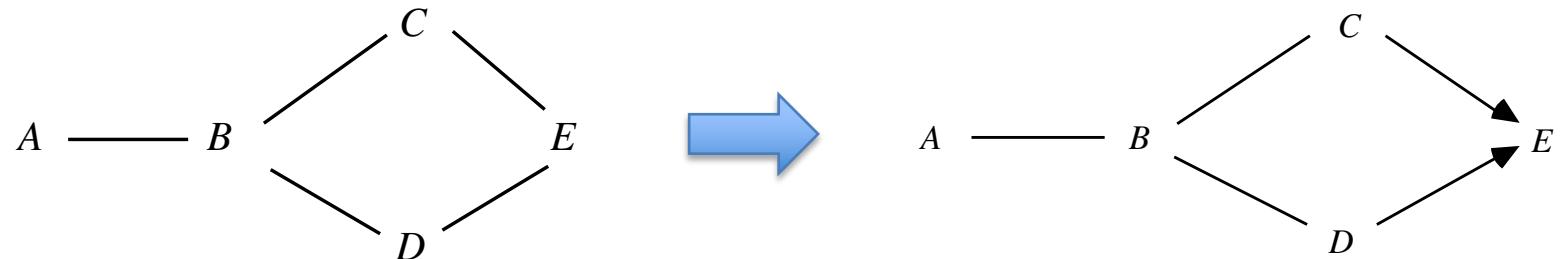
C.) For each triple of vertices  $X, Y, Z$  such that the pair  $X, Y$  and the pair  $Y, Z$  are each adjacent in  $C$  but the pair  $X, Z$  are not adjacent in  $C$ , orient  $X - Y - Z$  as  $X \rightarrow Y \leftarrow Z$  if and only if  $Y$  is not in  $\text{Sepset}(X, Z)$ .

D.) repeat

If  $A \rightarrow B$ ,  $B$  and  $C$  are adjacent,  $A$  and  $C$  are not adjacent, and there is no arrowhead at  $B$ , then orient  $B - C$  as  $B \rightarrow C$ .

If there is a *directed* path from  $A$  to  $B$ , and an edge between  $A$  and  $B$ , then orient  $A - B$  as  $A \rightarrow B$ .

until no more edges can be oriented.



The triples of variables with only two adjacencies among them are:

$A - B - C;$

$C - B - D;$

$B - D - E;$

$A - B - D;$

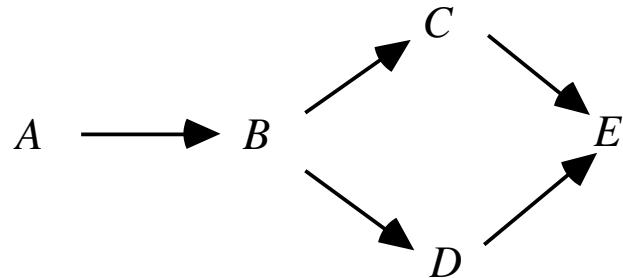
$B - C - E;$

$C - E - D$

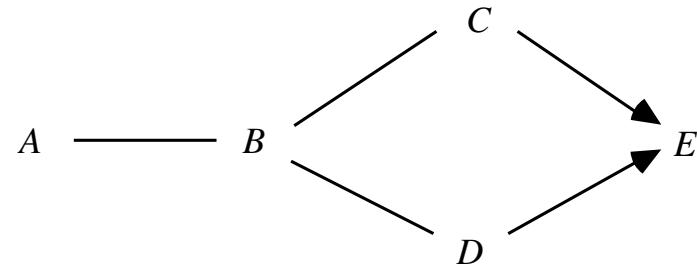
$E$  is not in  $\text{Sepset}(C, D)$

so  $C - E$  and  $E - D$  collide at  $E$ .

# Results up to an indistinguishable class



True graph



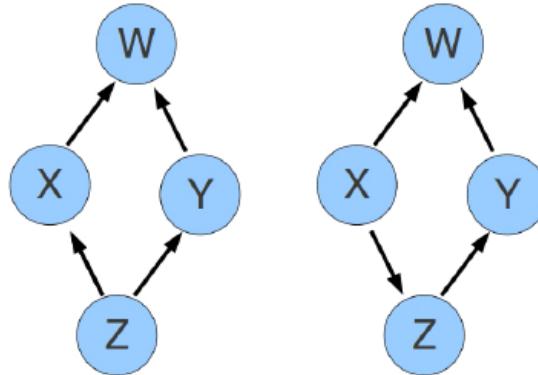
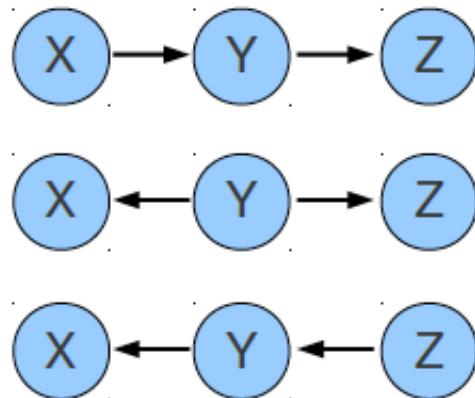
Result by PC algorithm

Every orientation of the undirected edges in the result by PC algorithm is permissible that does not include a collision at  $B$ .

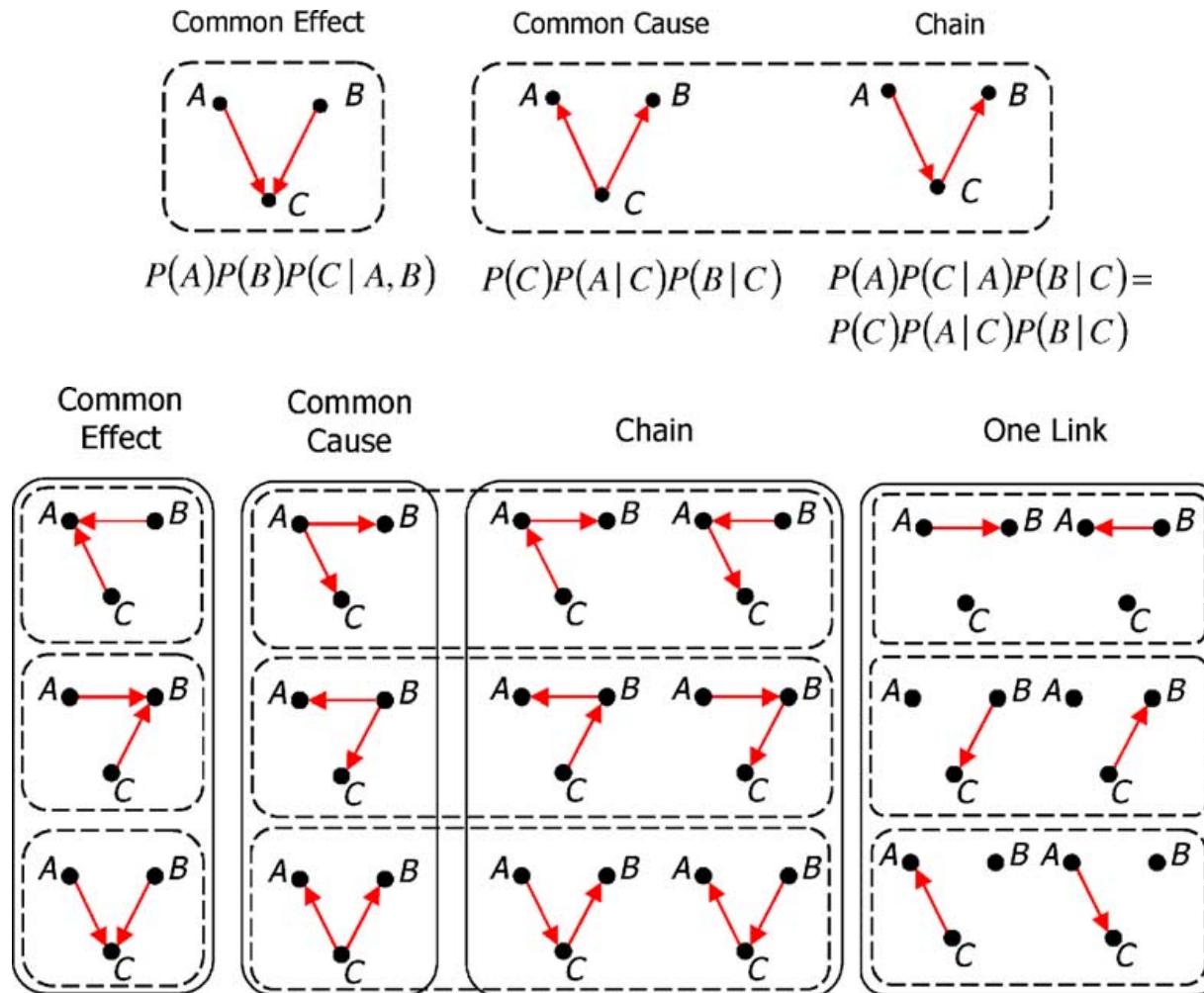
What is the complexity of the PC algorithm?

# Markov equivalent class

Theorem (Verma and Pearl, 1990): two DAGs are Markov equivalent iff they have the same skeleton and the same v-structures.  
skeleton: corresponding undirected graph  
v-structure: substructure  $X \rightarrow Y <- Z$  with no edge between X and Z.



# All three-node networks (1-2 arrows)

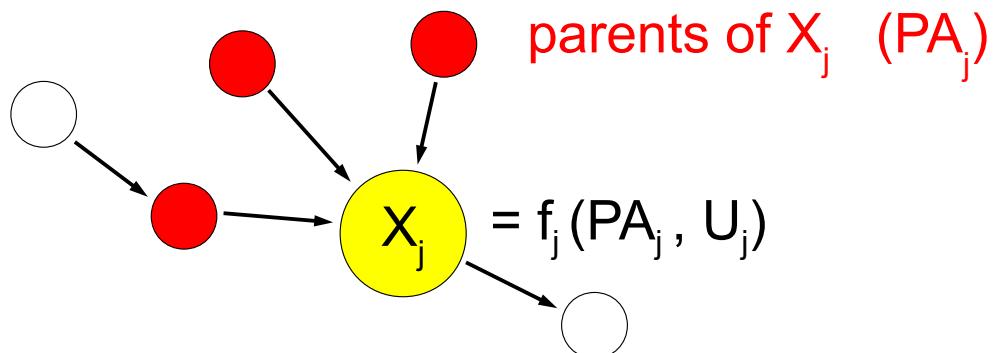


Solid lines group together networks of the same topological type.  
Dashed lines delineate Markov equivalence classes.



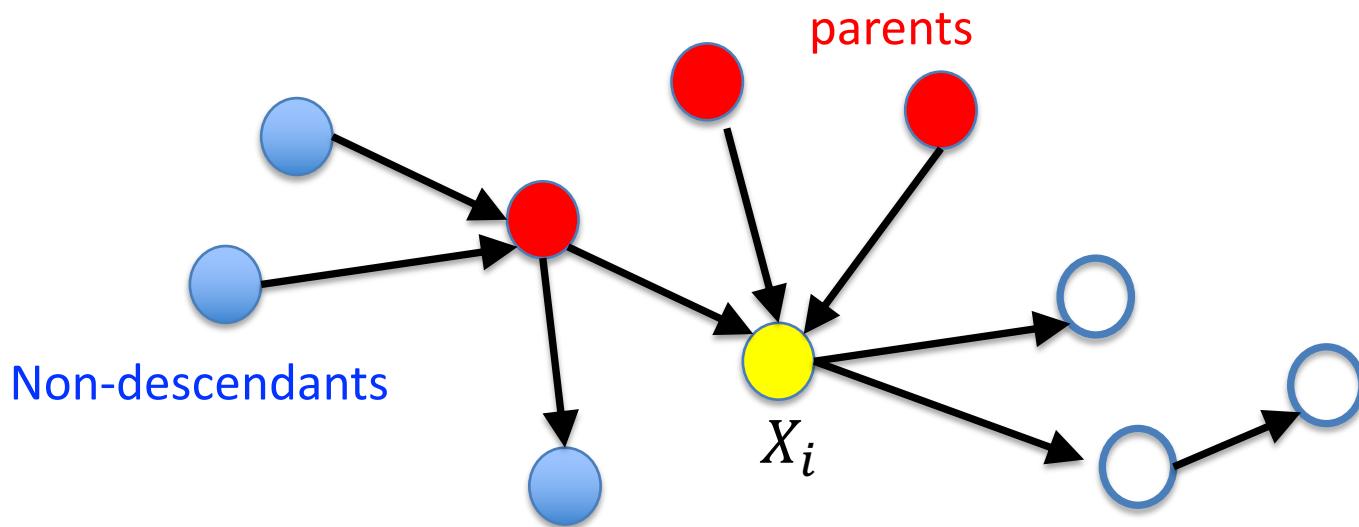
# Functional Causal Model (*Pearl et al.*)

- Set of observables  $X_1, \dots, X_n$
- directed acyclic graph  $G$  with vertices  $X_1, \dots, X_n$
- Semantics: parents = direct causes
- $X_i = f_i(\text{ParentsOf}_i, \text{Noise}_i)$ , with independent  $\text{Noise}_1, \dots, \text{Noise}_n$ .
- “Noise” means “unexplained” (or “exogenous”), we use  $U_i$
- Can add requirement that  $f_1, \dots, f_n, \text{Noise}_1, \dots, \text{Noise}_n$  “independent”  
(cf. Lemeire & Dirkx 2006, Janzing & Schölkopf 2010 — more below)



**Theorem:** the following are equivalent:

- Existence of a functional causal model
- Local Causal Markov condition:  $X_j$  statistically independent of **non-descendants**, given **parents** (i.e.: every information exchange with its non-descendants involves its parents)
- Global Causal Markov condition: d-separation (characterizes the set of independences implied by local Markov condition)
- Factorization  $P(X_1, \dots, X_n) = \prod_j P(X_j | \text{Parents}_j)$  (conditionals as causal mechanisms generating statistical dependence)



# Outline

- Discover causal structure by conditional independence
  - PC algorithm
  - Markov equivalent class
- **Pearl's do-calculus**

# Pearl's do-calculus

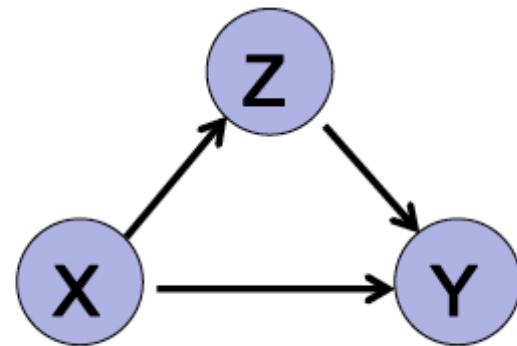
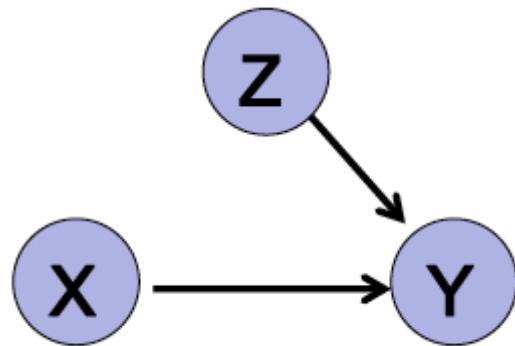
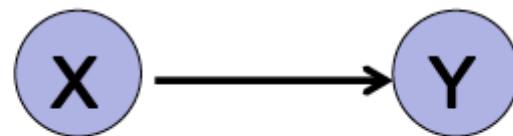
- Motivation: goal of causality is to infer the effect of interventions
- distribution of  $Y$  given that  $X$  is set to  $x$ :  
 $p(Y | \text{do } X = x)$  or  $p(Y | \text{do } x)$



根据 do 操作，便可以定义因果作用，比如二值的变量  $Z$  对于  $Y$  的平均因果作用定义为

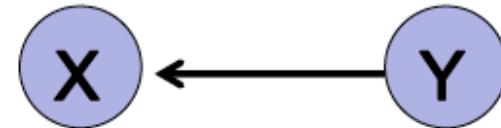
$$ACE(Z \rightarrow Y) = E(Y | \text{do}(Z) = 1) - E(Y | \text{do}(Z) = 0),$$

Examples for  $p(.|do x) = p(.|x)$

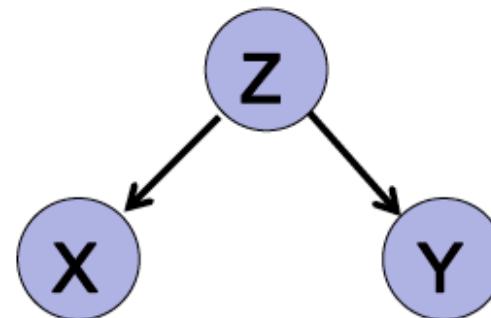


## Examples for $p(.|do x) \neq p(.|x)$

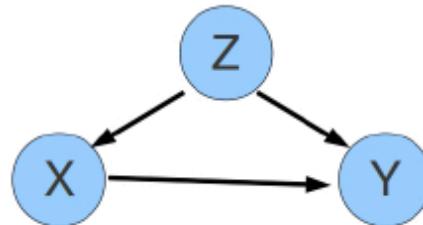
- $p(Y|do x) = P(Y) \neq P(Y|x)$



- $p(Y|do x) = P(Y) \neq P(Y|x)$



## Example: controlling for confounding



$X \not\perp\!\!\!\perp Y$  partly due to the  $Z$  and partly due to  $X \rightarrow Y$

- causal factorization

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

- replace  $P(X|Z)$  with  $\delta_{Xx}$

$$p(Y, Z|do x) = p(Z) \delta_{Xx} p(Y|X, Z)$$

- marginalize

$$p(Y|do x) = \sum_z p(z)p(Y|x, z) \neq \sum_z p(z|x)p(Y|x, z) = p(Y|x)$$

Thank you!