Causal Discovery

Advanced Methods in Applied Statistics

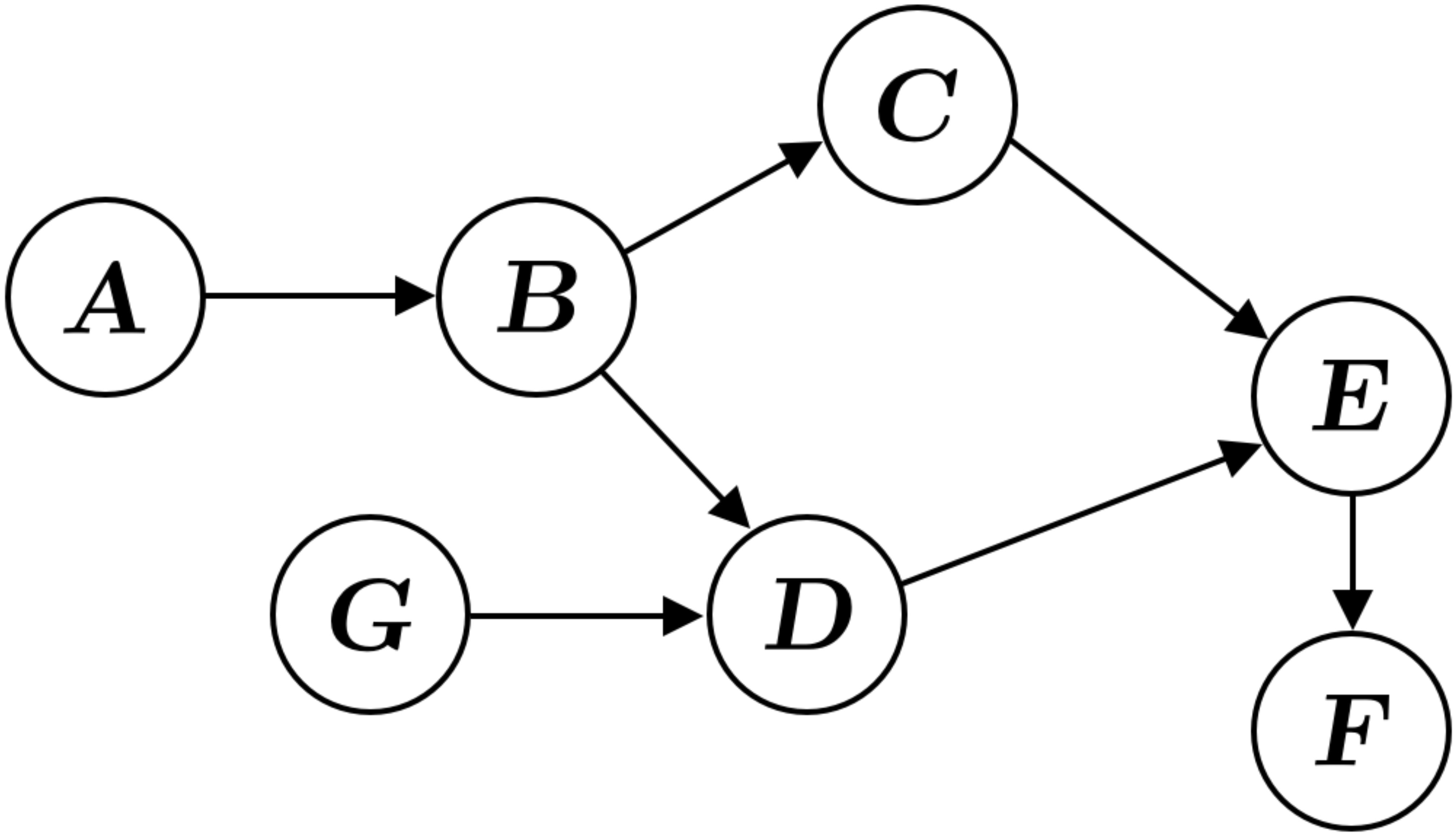
Jakob Harteg, March 9, 2023

- Correlation does not imply causation
- So what implies causation?
- If X causes Y:
  changing X $\longrightarrow$ change in Y
- What if we can't change X?
- Infer causality from data?

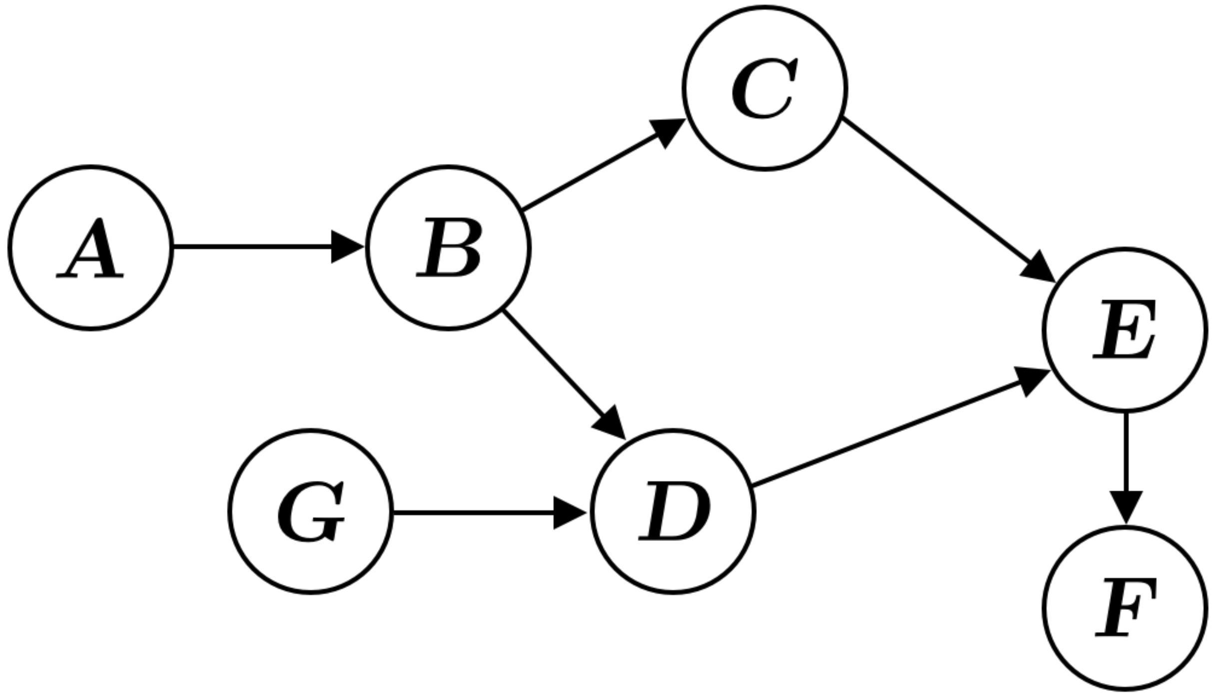
$\longrightarrow$ Graphs and Conditional independence

$\longrightarrow$ PC Algorithm and application

Graphs

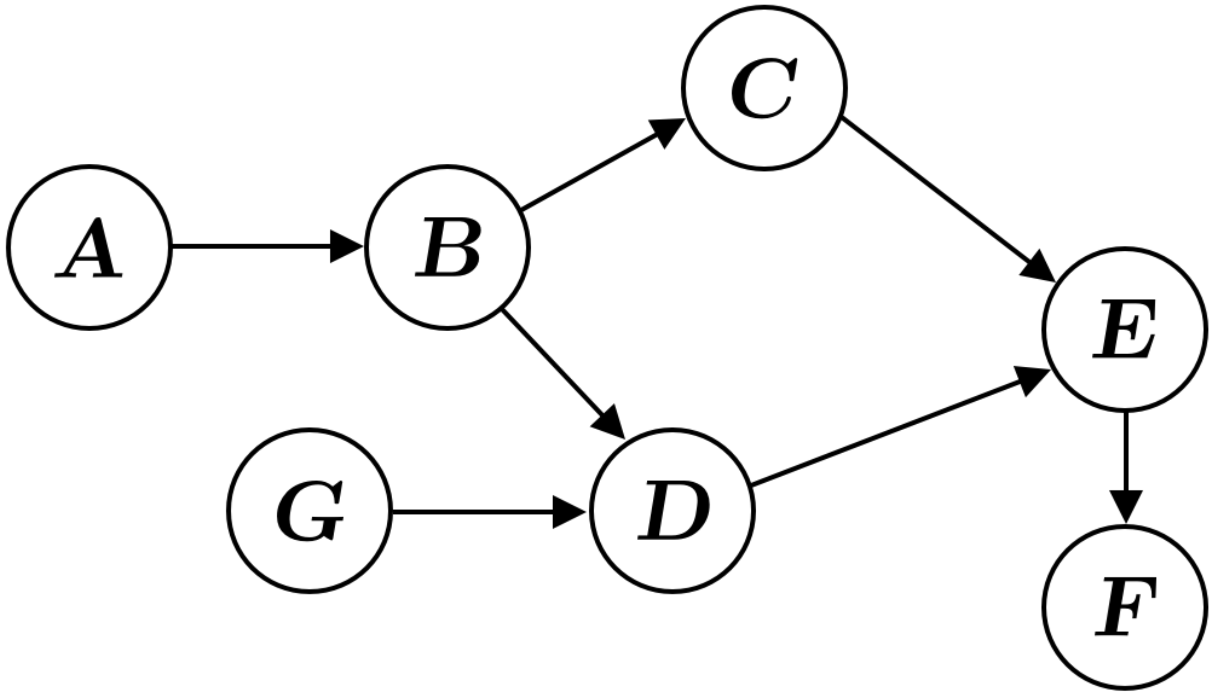


Graphs



A	B	C	D	E	F	G
0.285958	1.963270	3.330700	2.736891	7.842619	8.308887	0.520010
-1.086351	-1.637878	0.615447	-1.338098	-0.295491	0.562954	2.175564
0.672034	1.914775	2.847854	4.509284	6.700475	7.705827	0.656443
1.932187	3.252511	3.583642	5.817359	10.328517	12.750852	0.550836
0.462505	1.020929	2.003978	5.570001	6.815033	6.910609	2.332793
...	...	...	...	...	...	...

Graphs

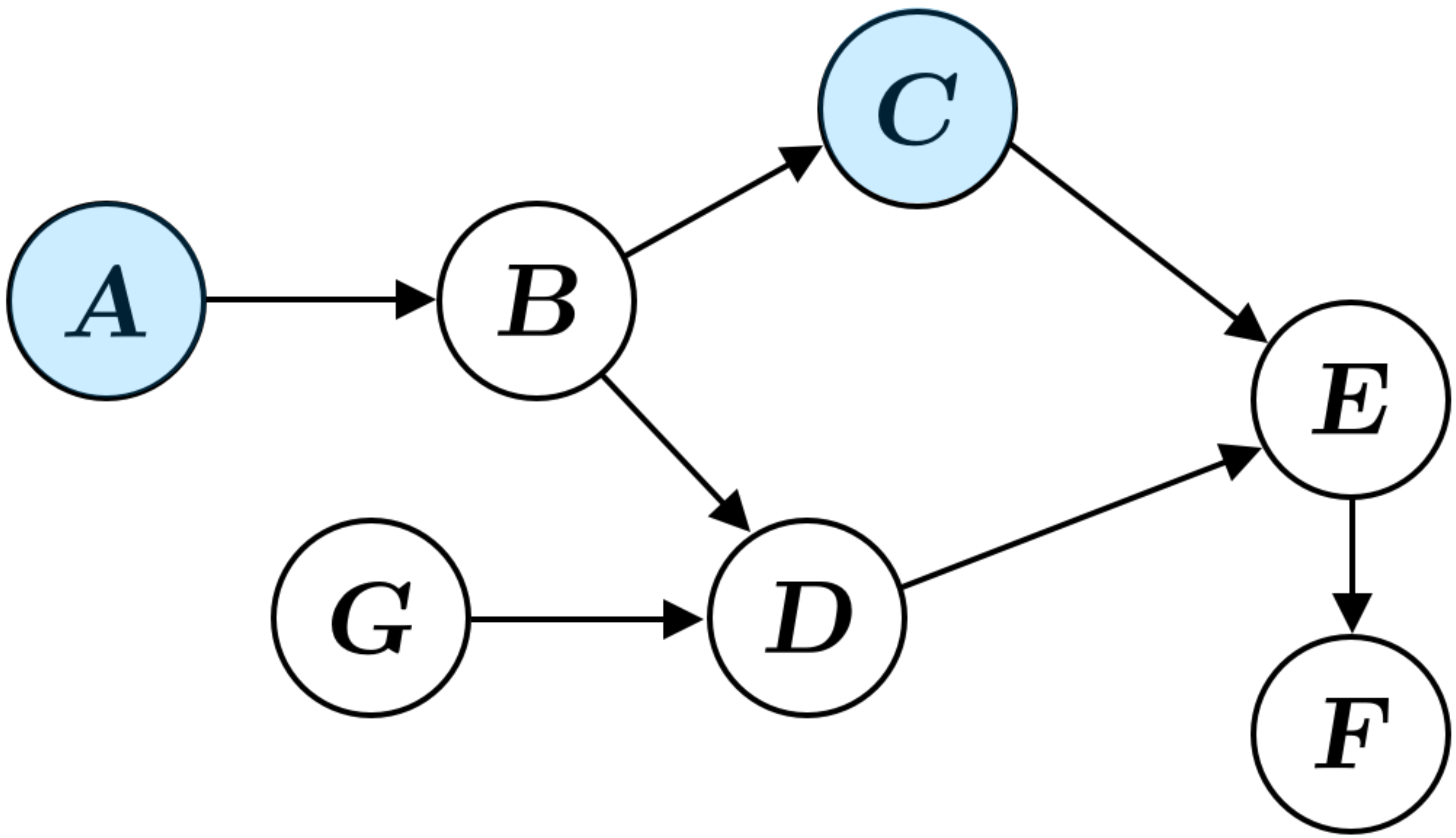



```
n = 10_000
A =              normal(scale=1,   size=n)
G =              normal(scale=1,   size=n)
B = A          + normal(scale=0.2, size=n)
C = B          + normal(scale=0.2, size=n)
D = G + B      + normal(scale=0.2, size=n)
E = C + D      + normal(scale=0.2, size=n)
F = E          + normal(scale=0.2, size=n)
```

A	B	C	D	E	F	G
0.285958	1.963270	3.330700	2.736891	7.842619	8.308887	0.520010
-1.086351	-1.637878	0.615447	-1.338098	-0.295491	0.562954	2.175564
0.672034	1.914775	2.847854	4.509284	6.700475	7.705827	0.656443
1.932187	3.252511	3.583642	5.817359	10.328517	12.750852	0.550836
0.462505	1.020929	2.003978	5.570001	6.815033	6.910609	2.332793
...	...	...	...	...	...	...

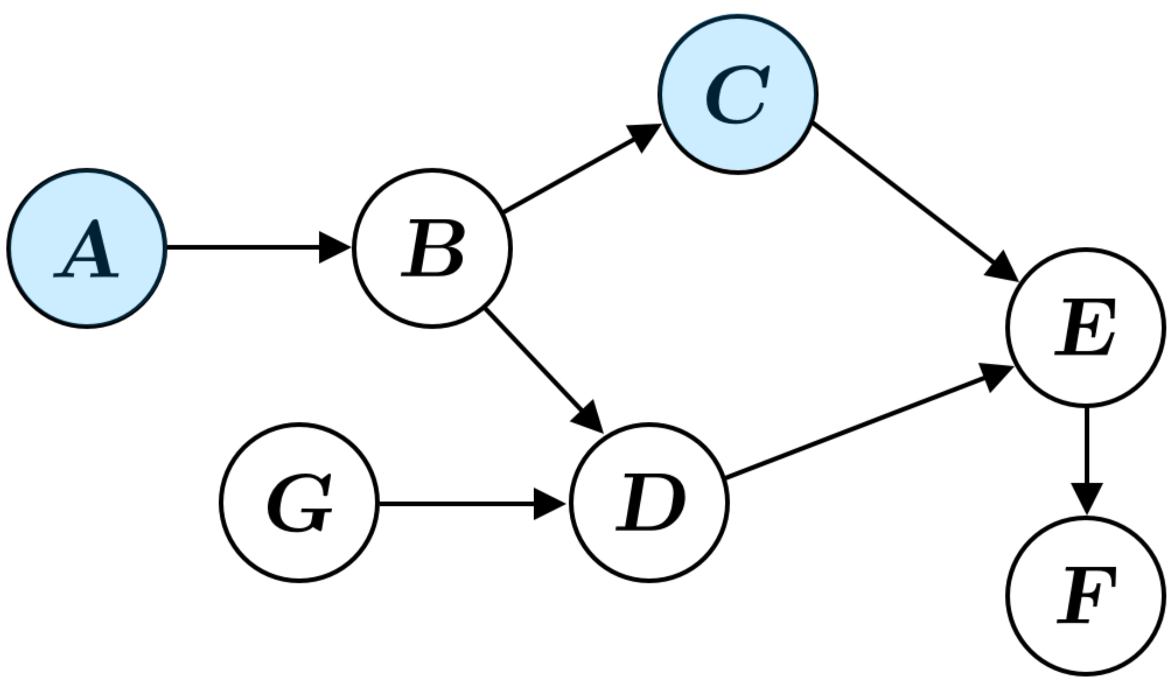
Graphs

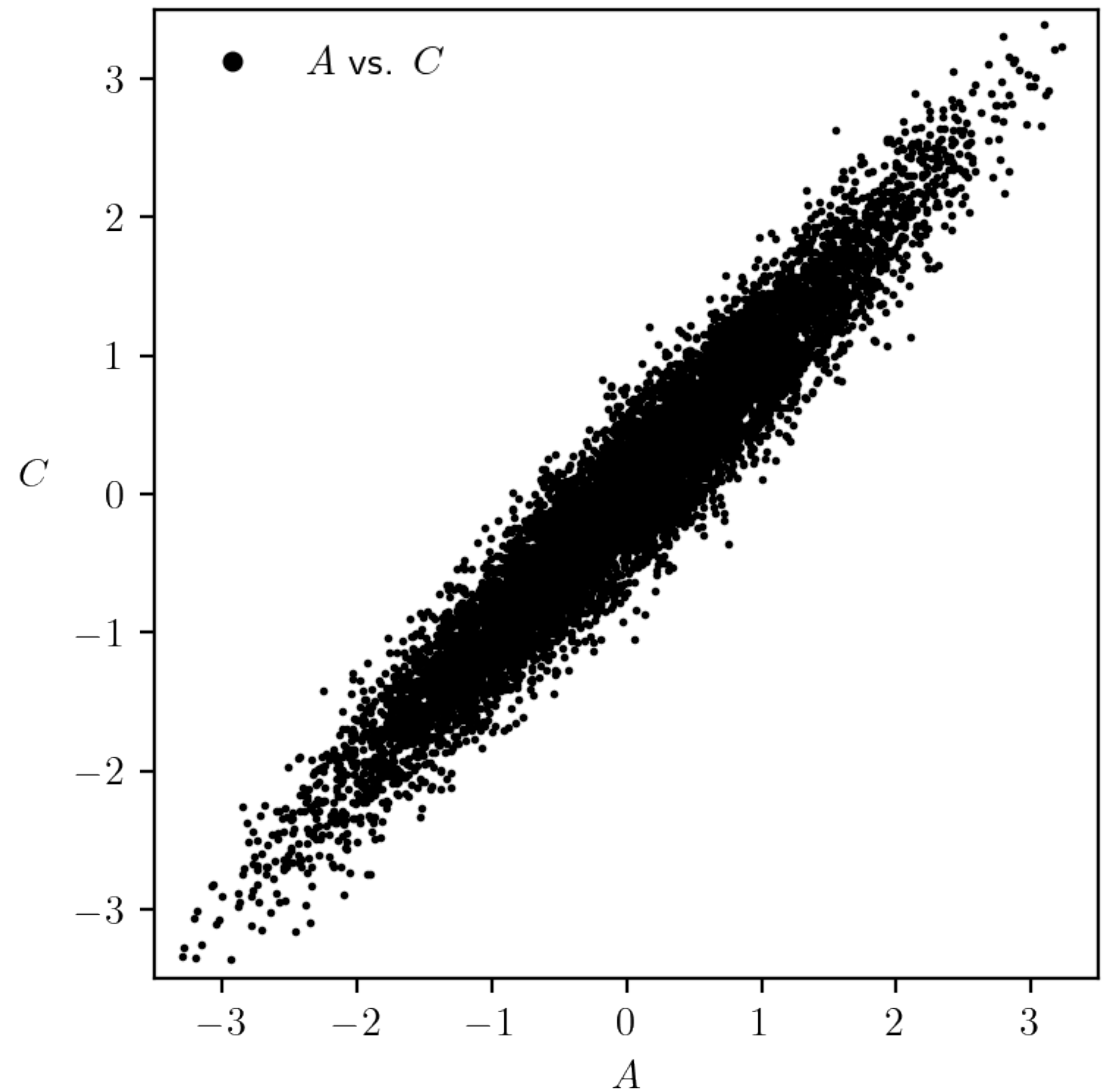
Q: Does A cause C?



Graphs

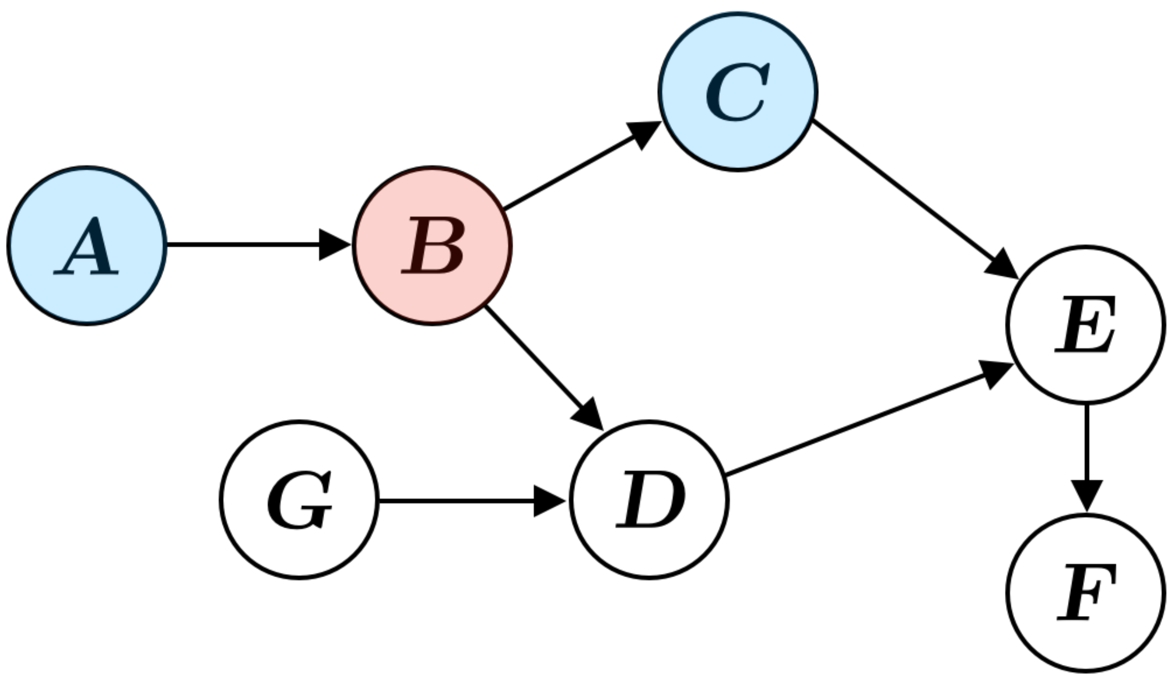
Q: Does A cause C?

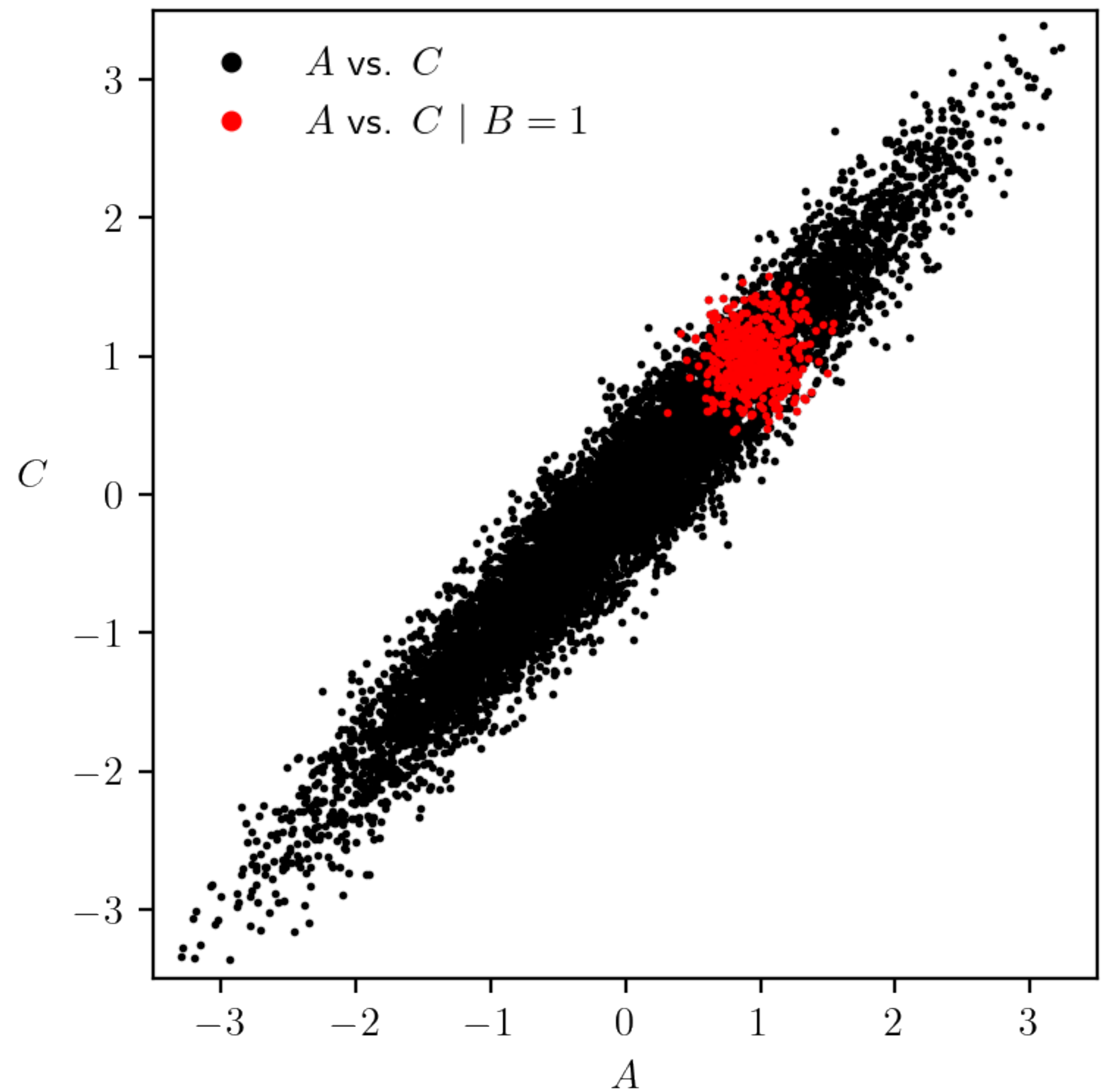




Graphs

Q: Does A cause C?



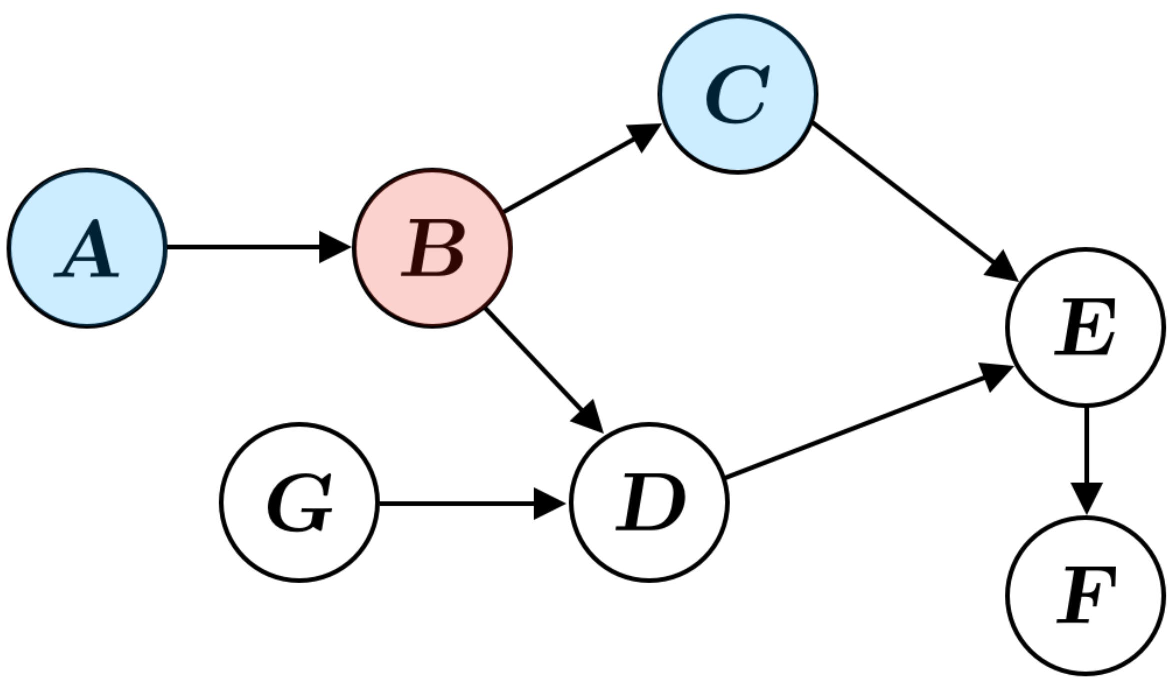


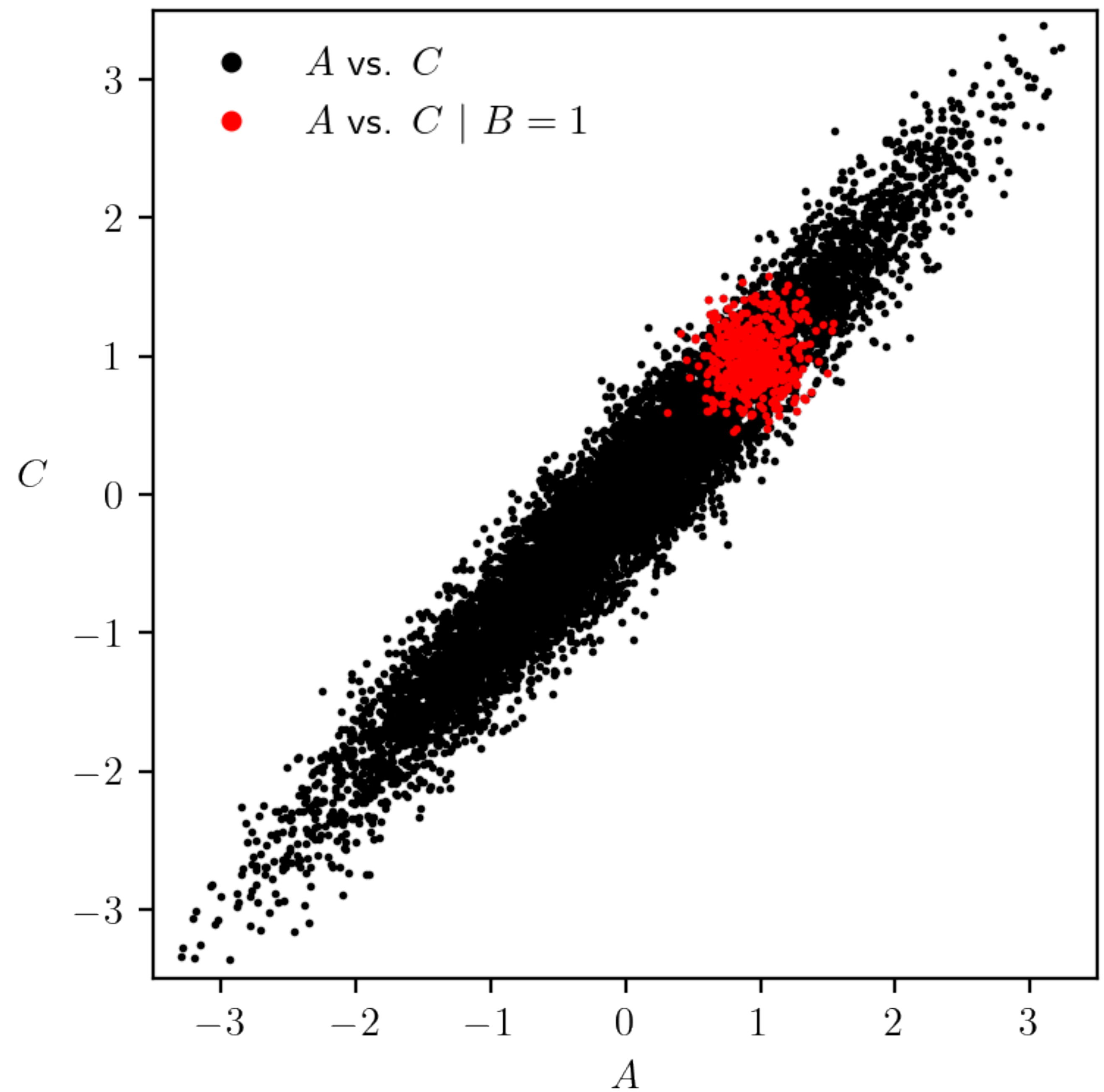
Graphs

Q: Does A cause C?

No, A and C are conditionally independent on B:

$$A \perp C \mid B$$



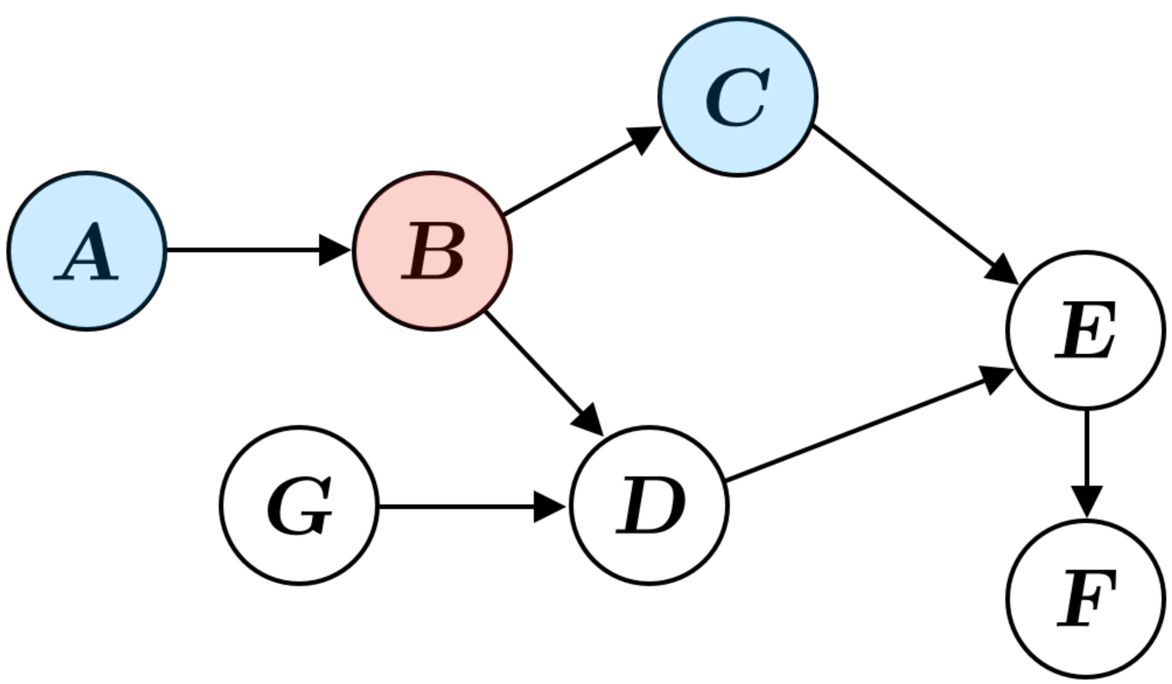


Graphs

Q: Does A cause C?

No, A and C are conditionally independent on B:

$$A \perp C \mid B$$



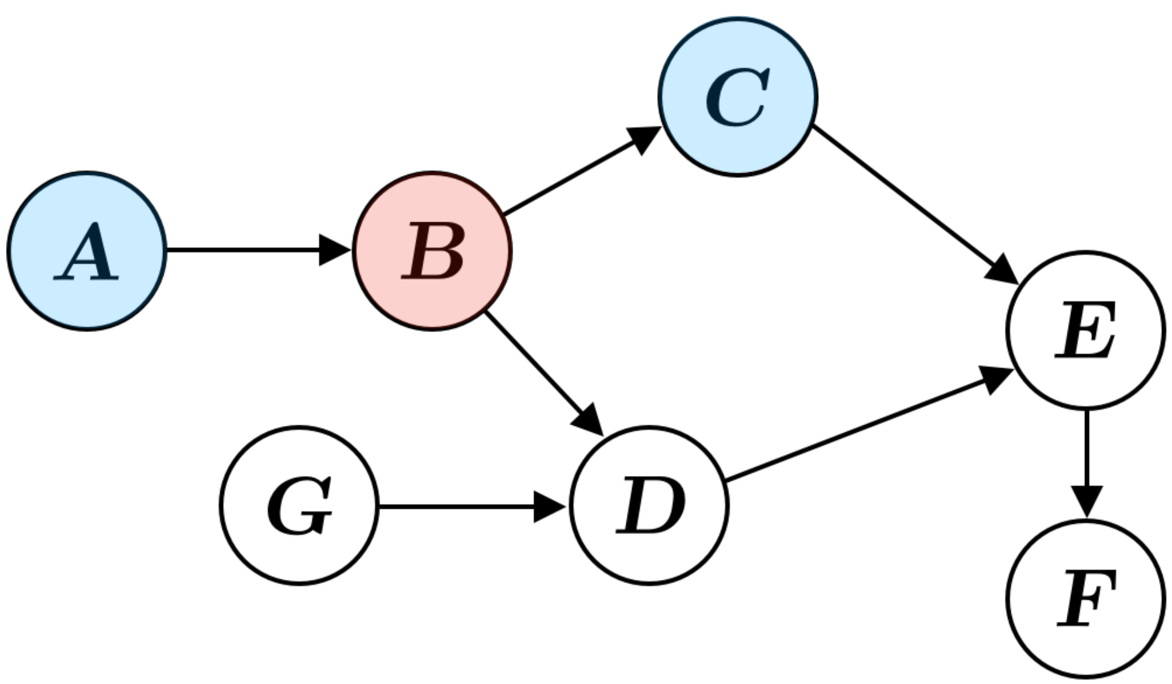
We can infer the **<u>absence</u>** of a causal relationship with conditional independence tests

Graphs

Q: Does A cause C?

No, A and C are conditionally independent on B:

$$A \perp C \mid B$$



We can infer the **absence** of a causal relationship with conditional independence tests

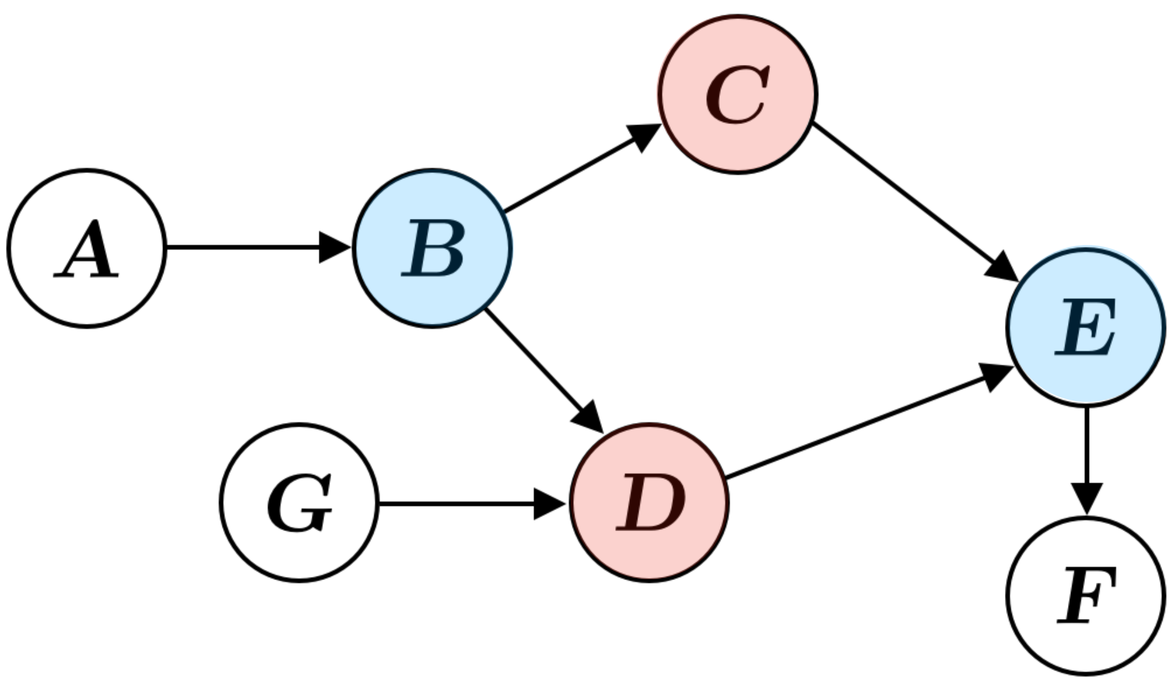
$$\Downarrow$$

If we are unable to find a set of nodes $\mathbf{Z}$ such that $X \perp Y \mid \mathbf{Z}$, then X and Y are adjacent.

Graphs

Q: Does B cause E?

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



We can infer the **<u>absence</u>** of a causal relationship with conditional independence tests

$\Downarrow$

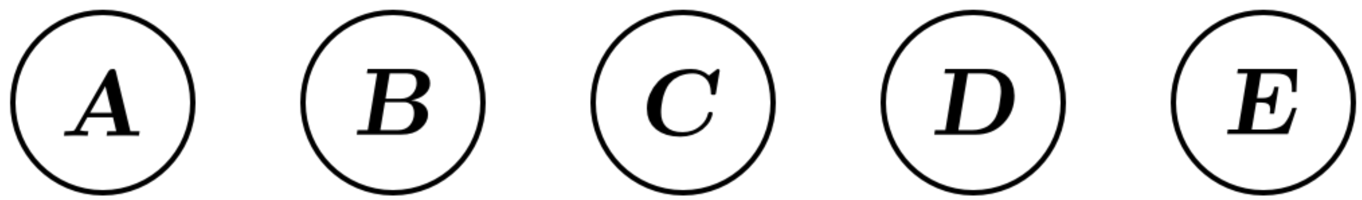
If we are unable to find a set of nodes $\mathbf{Z}$ such that $X \perp Y \mid \mathbf{Z}$, then X and Y are adjacent.

The PC-algorithm

- Step 1) Find causal connections with CI tests
- Step 2) Identify causal direction

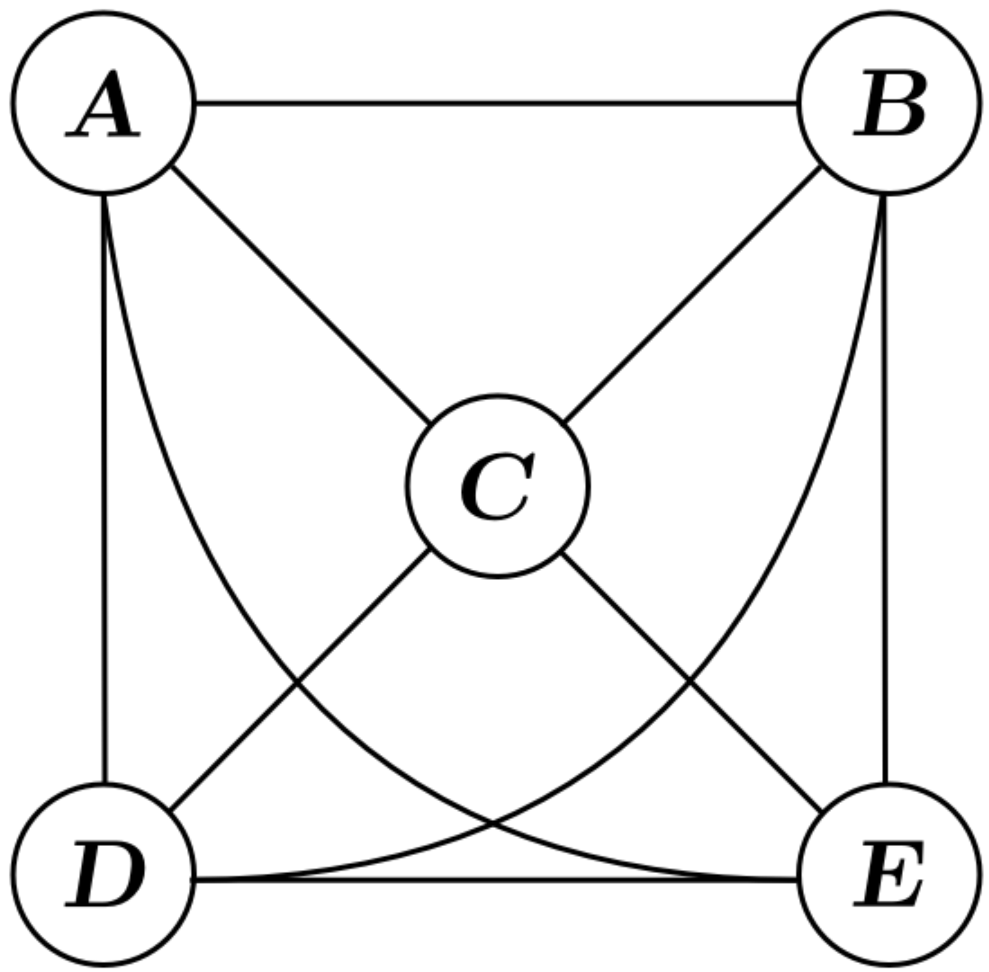
The PC-algorithm

Step 1 : Learn skeleton



The PC-algorithm

Step 1 : Learn skeleton



Start with *complete* graph

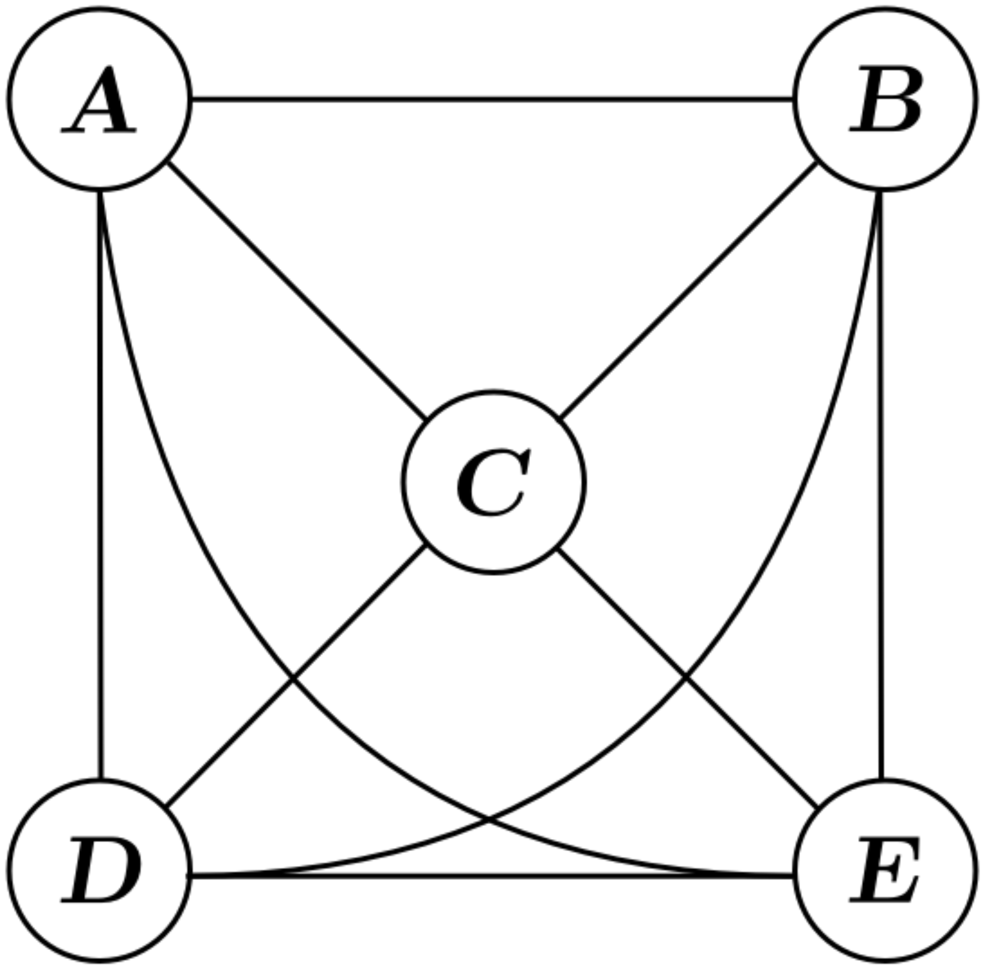
Remove edges with CI test

Starting with $\mathbf{Z} = \{\}$:

Plot X-Y regression
if no correlation: remove edge

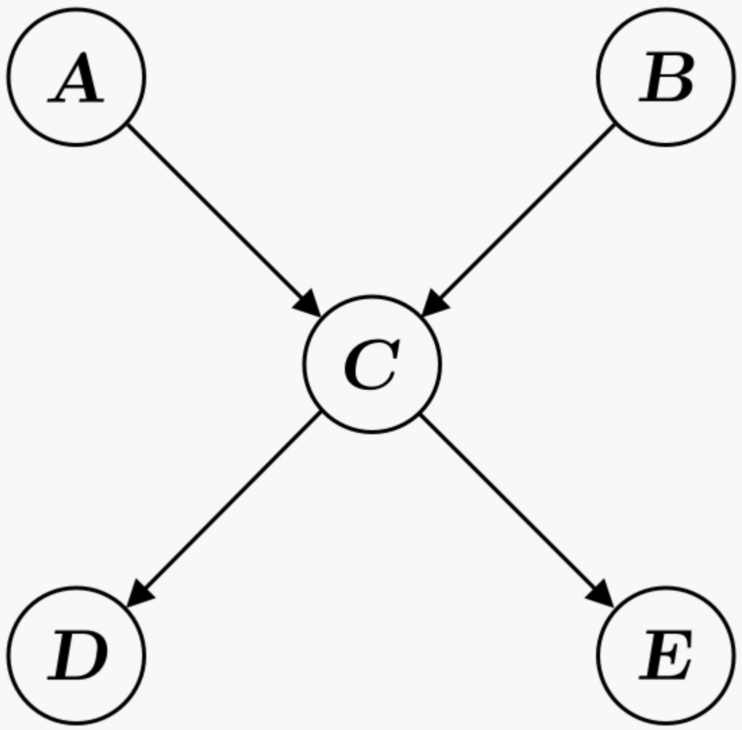
The PC-algorithm

Step 1 : Learn skeleton



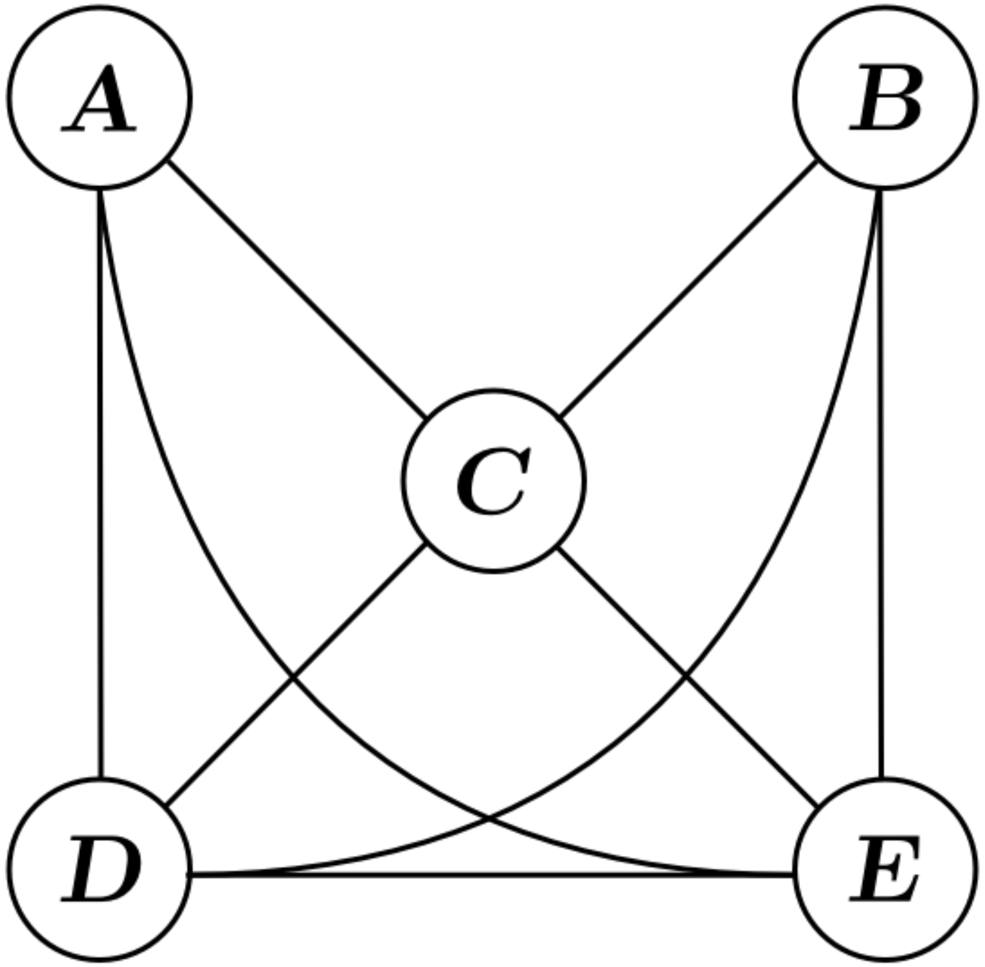
$$X \perp Y \mid \mathbf{Z} = \{\}$$

True graph of the data



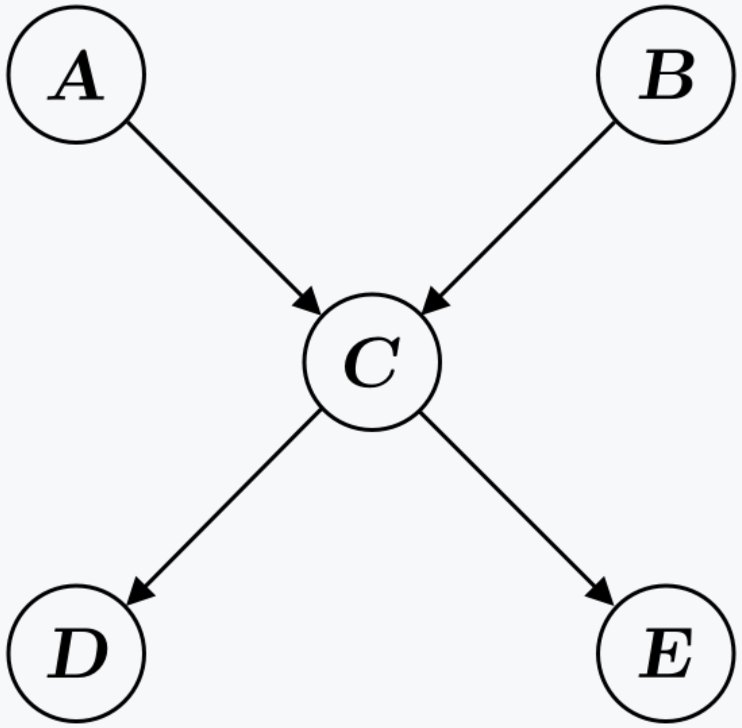
The PC-algorithm

Step 1 : Learn skeleton



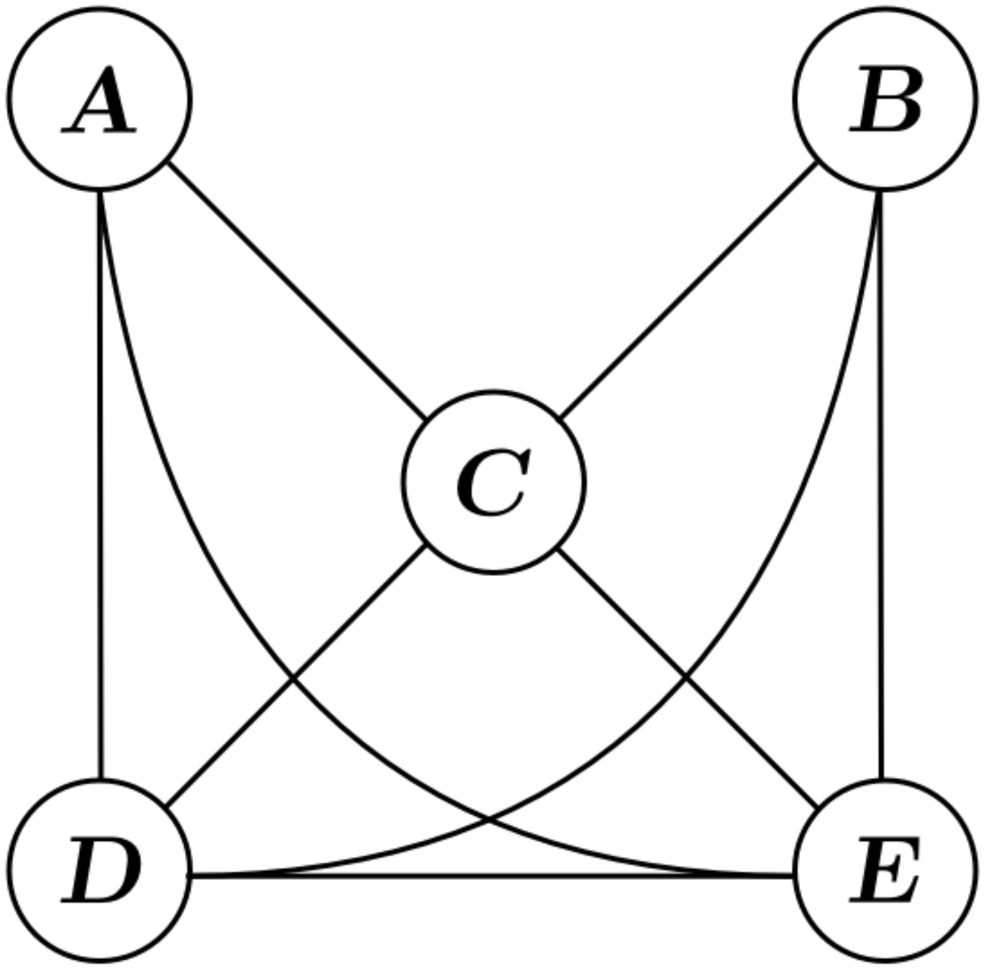
$$X \perp Y \mid \mathbf{Z} = \{\}$$

True graph of the data



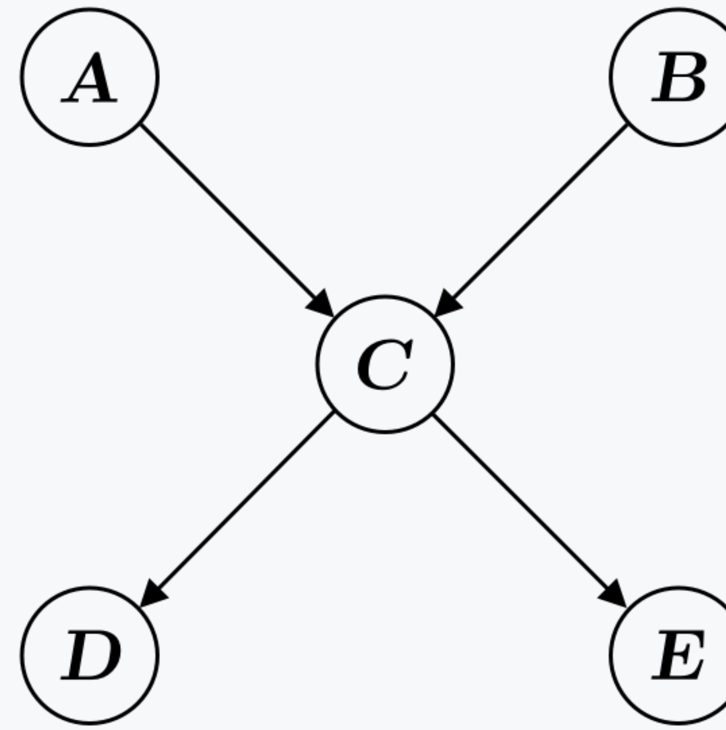
The PC-algorithm

Step 1 : Learn skeleton



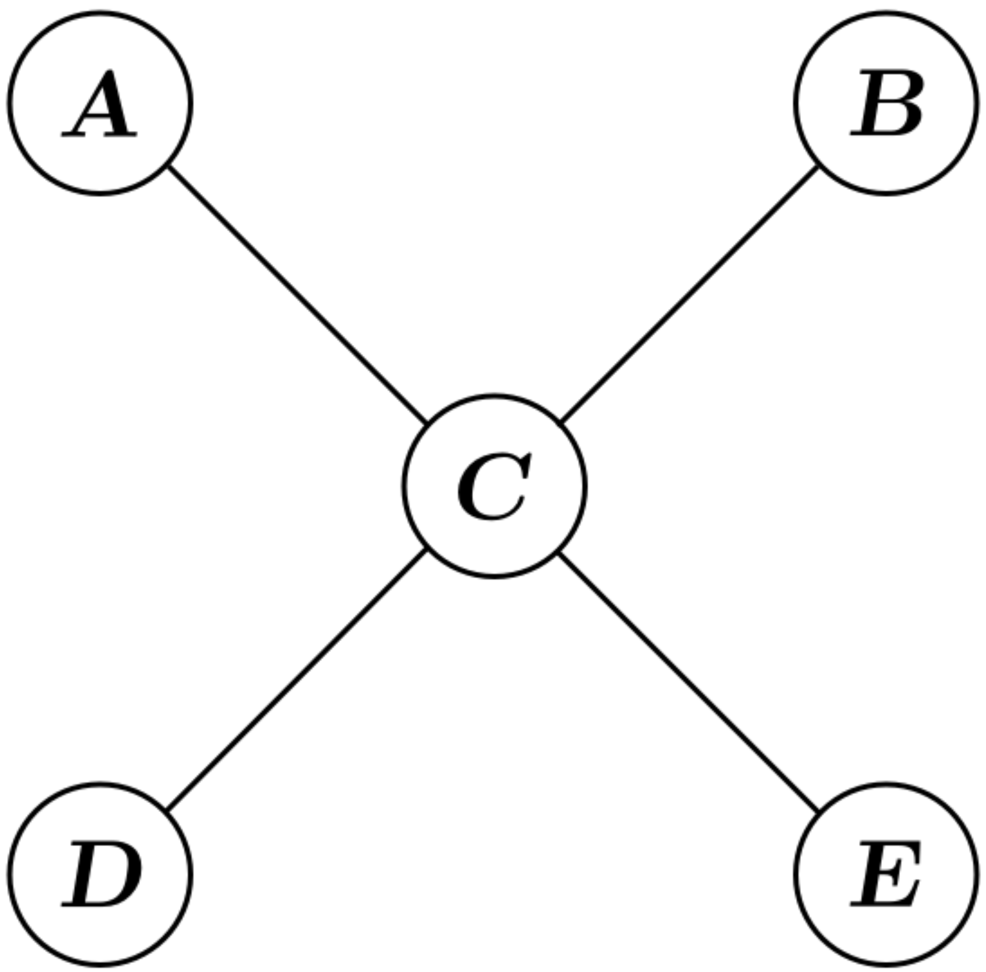
$$X \perp Y \mid Z$$

True graph of the data



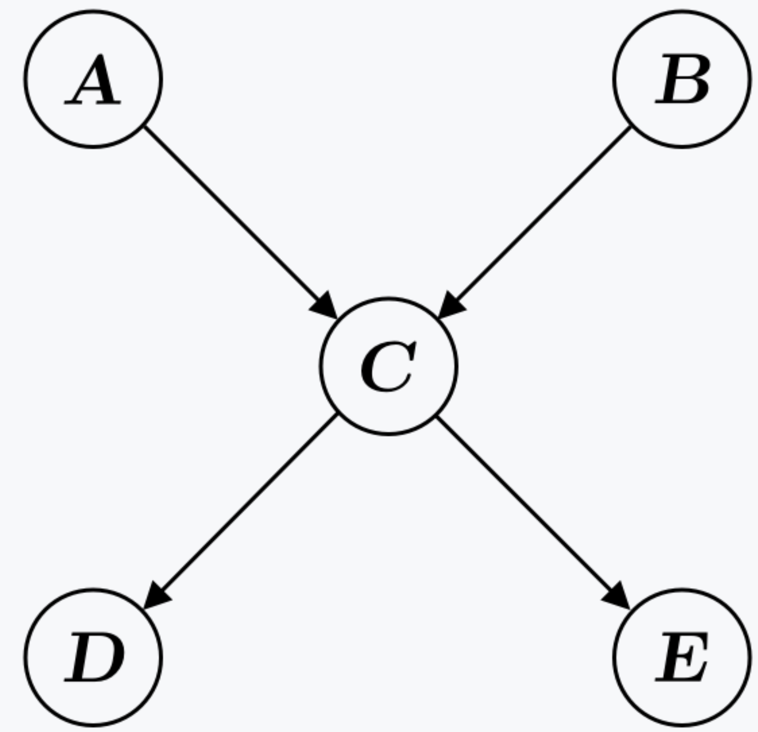
The PC-algorithm

Step 1 : Learn skeleton



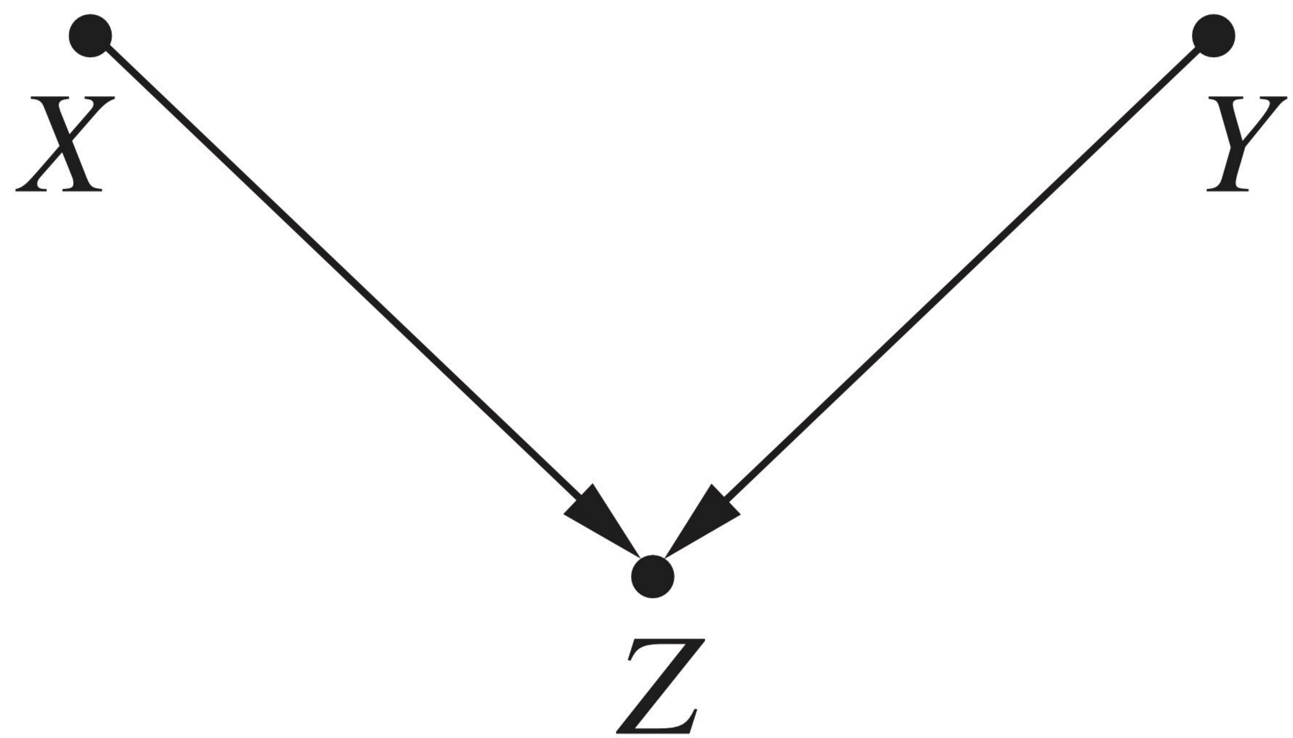
$$X \perp Y \mid Z$$

True graph of the data



The PC-algorithm

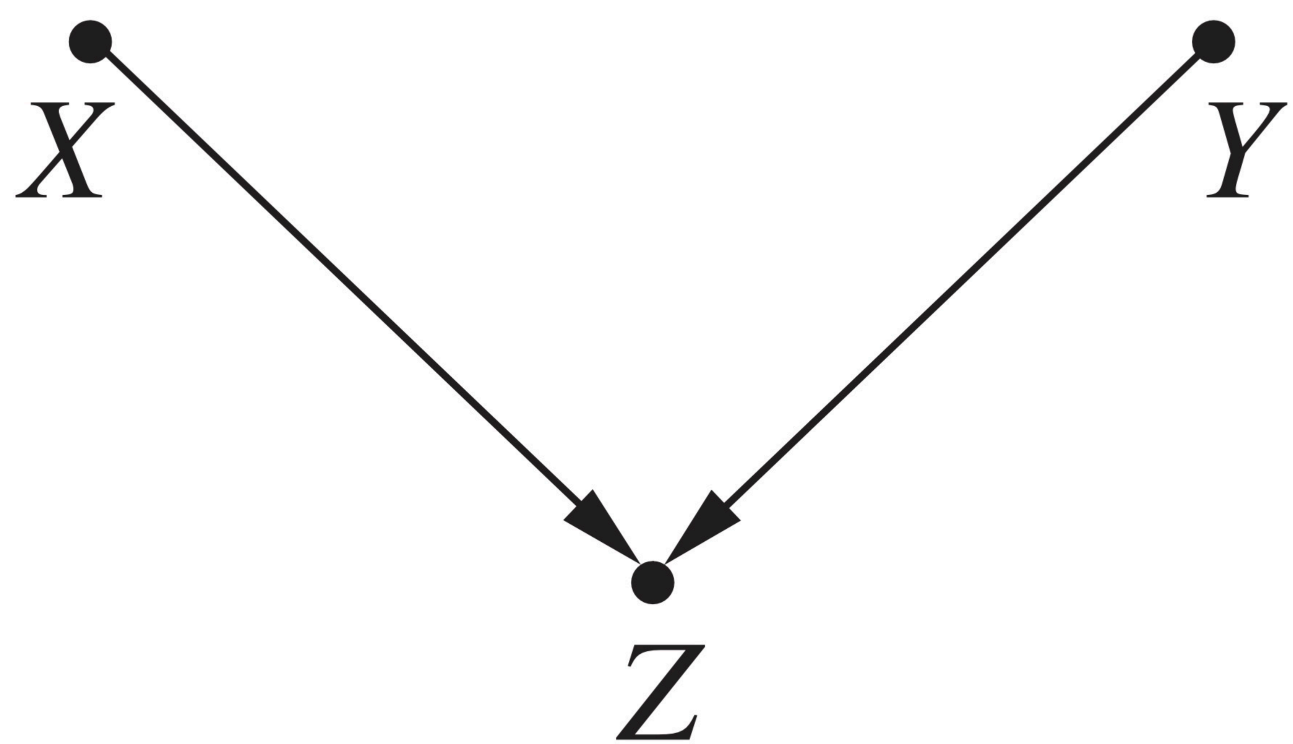
Step 2a : Identify colliders



X and Y are independent, unless we condition on Z

The PC-algorithm

Step 2a : Identify colliders



X and Y are independent, unless we condition on Z

If Z fixed, changing X changes Y

The PC-algorithm

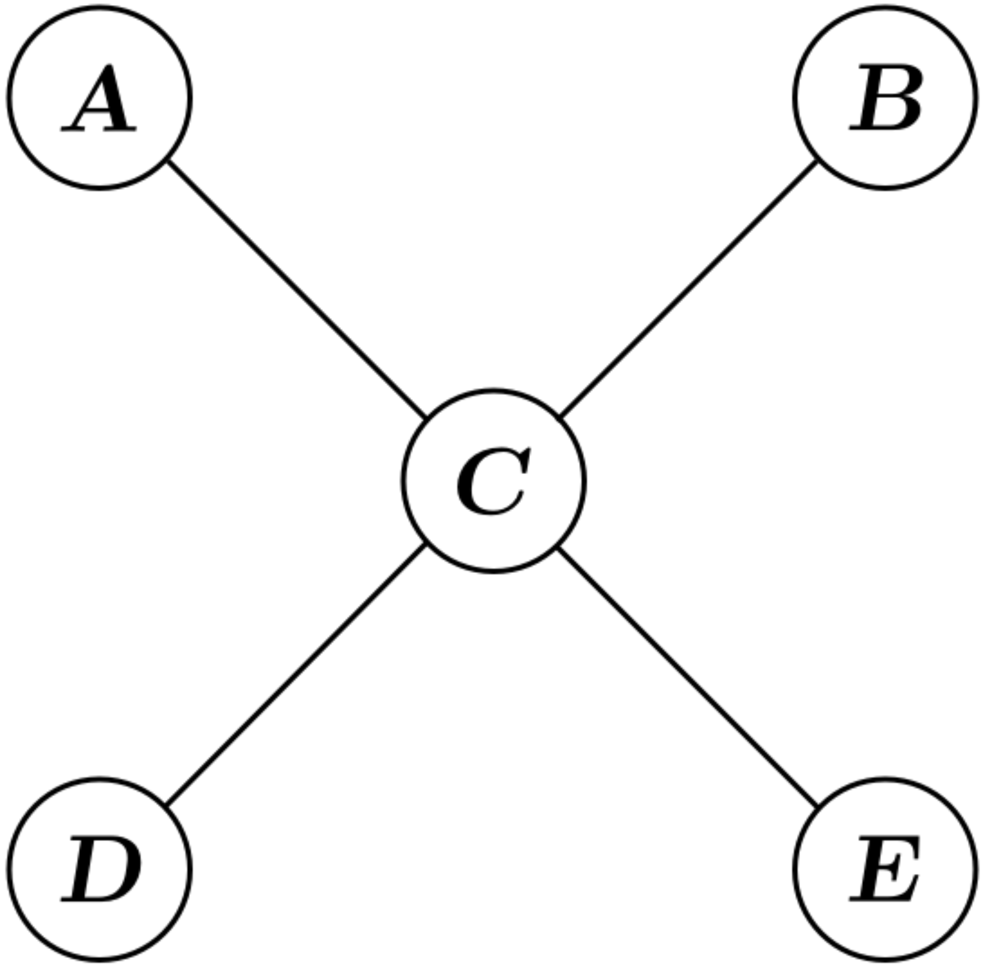
Step 2a : Identify colliders

A path X — Z — Y is a collider if:

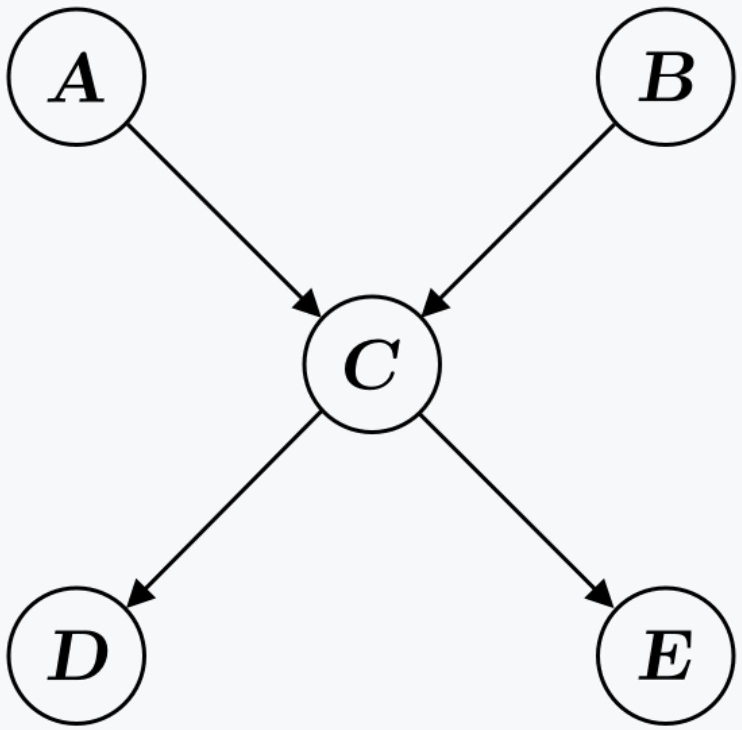
- X and Y are independent, and
- Z was not in the conditioning set, that made X and Y independent

The PC-algorithm

Step 2a : Identify colliders

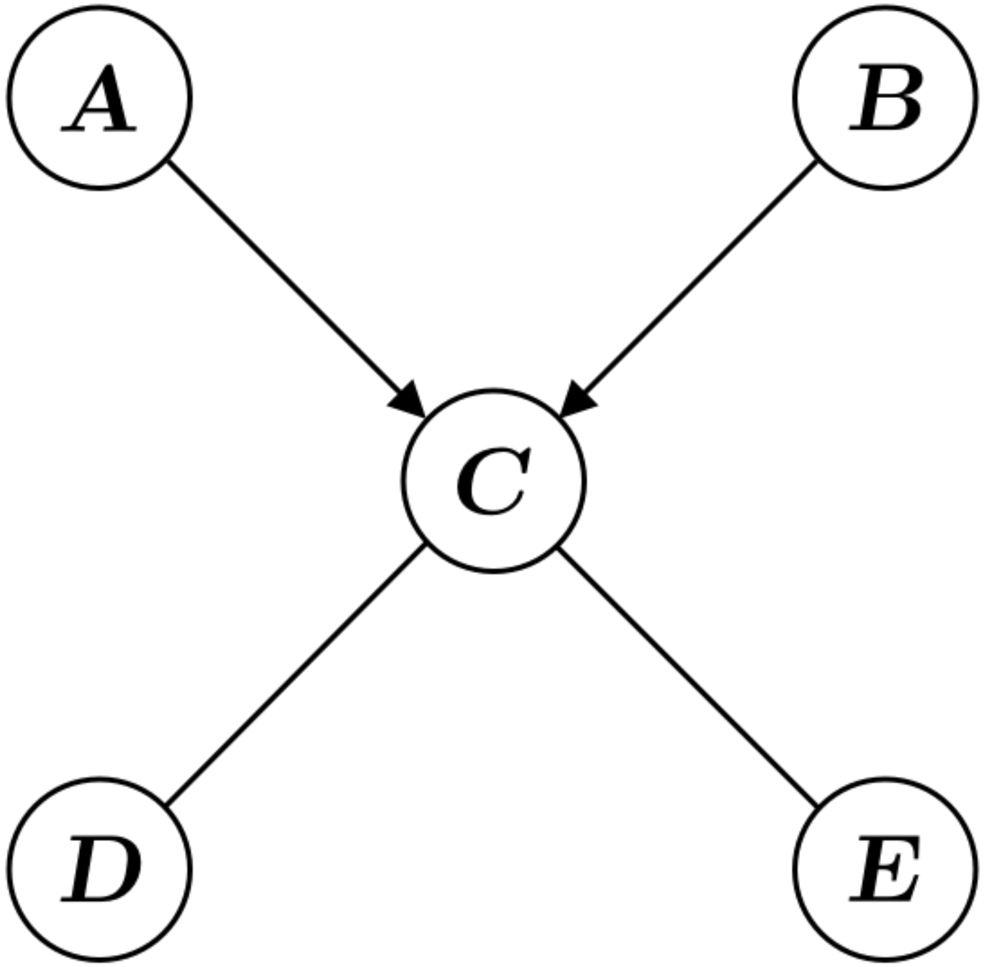


True graph
of the data

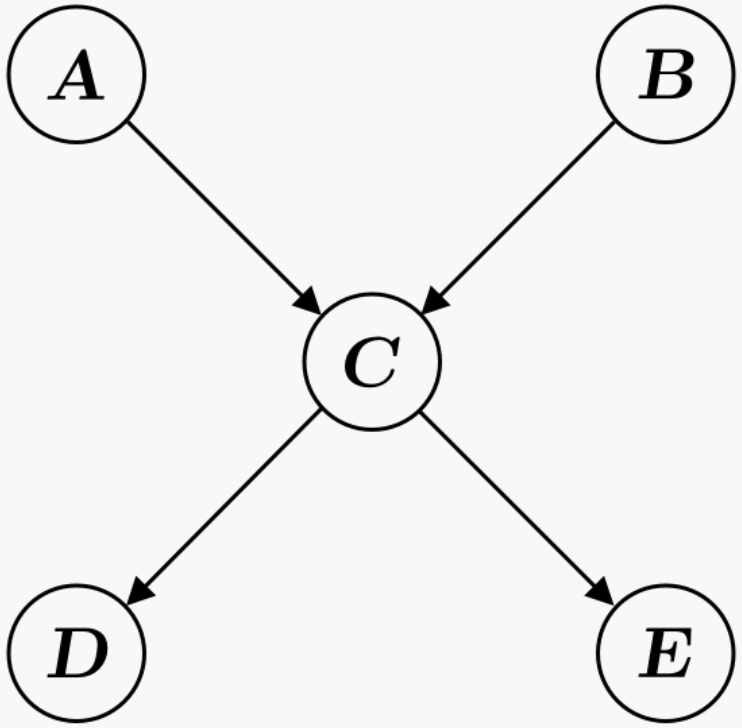


The PC-algorithm

Step 2a : Identify colliders

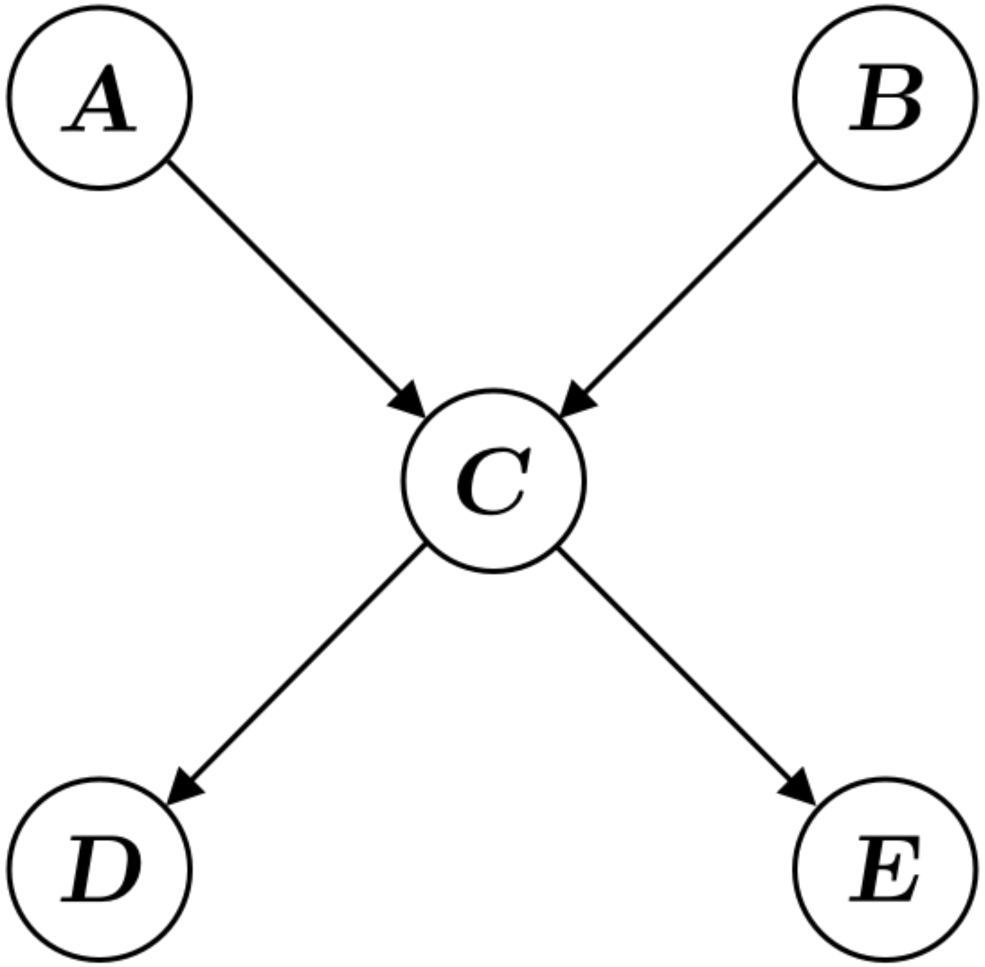


True graph
of the data

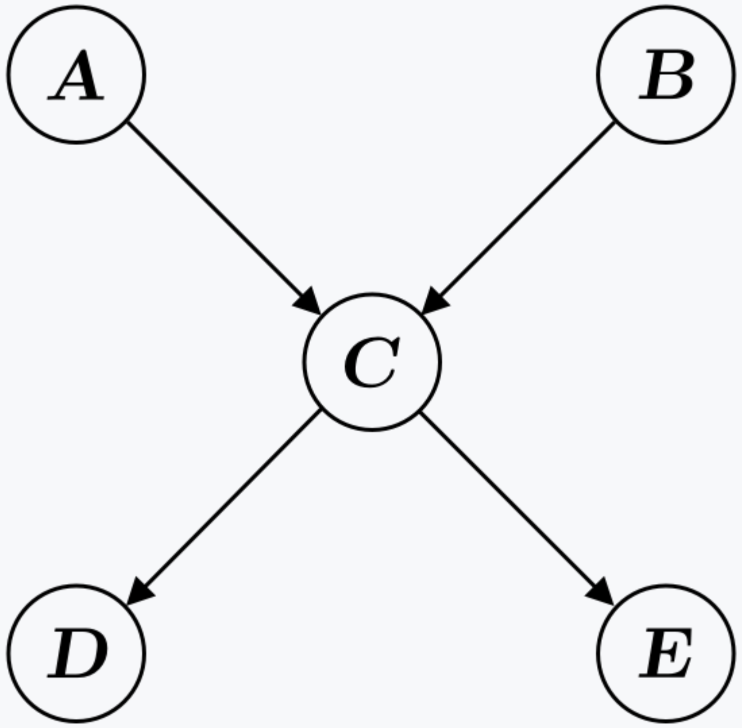


The PC-algorithm

Step 2b: Orient remaining edges



True graph of the data



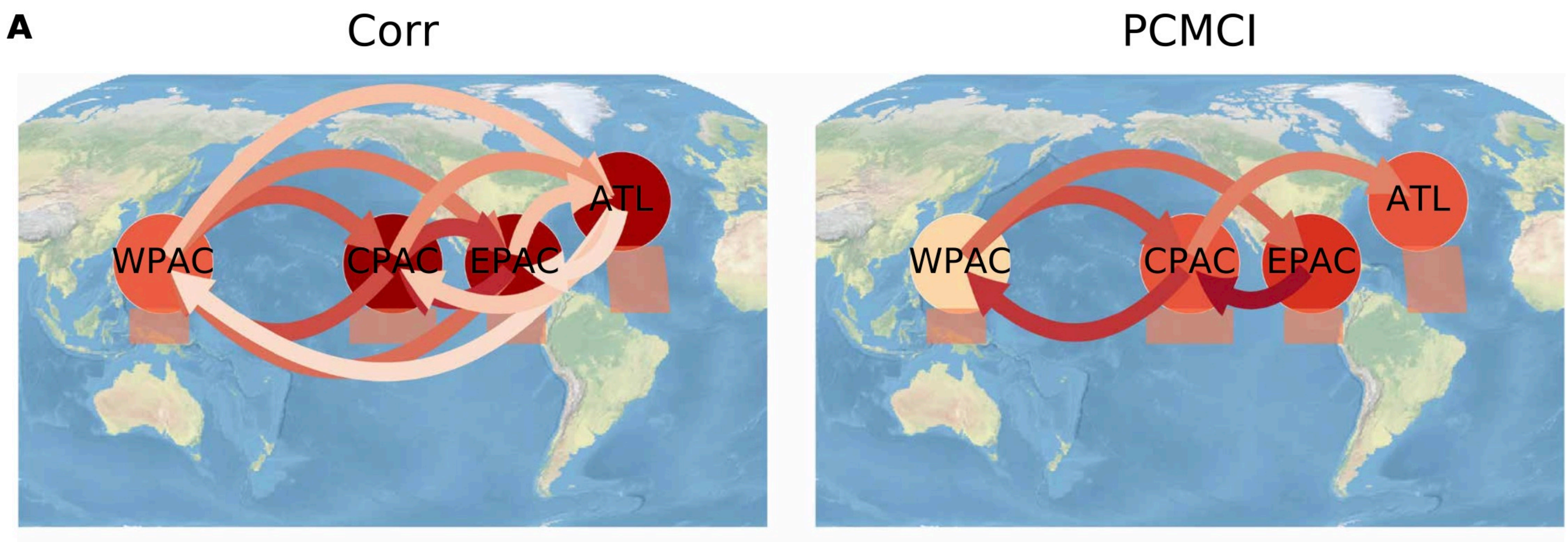
The PC-algorithm

Issues:

- CI tests can be hard $\longrightarrow$ false positives/negatives
- Runtime can be $\sim$exp(#nodes)
- Many assumptions that can be hard to justify
- Many extensions have been made

Application in climate science

Causal discovery in time series

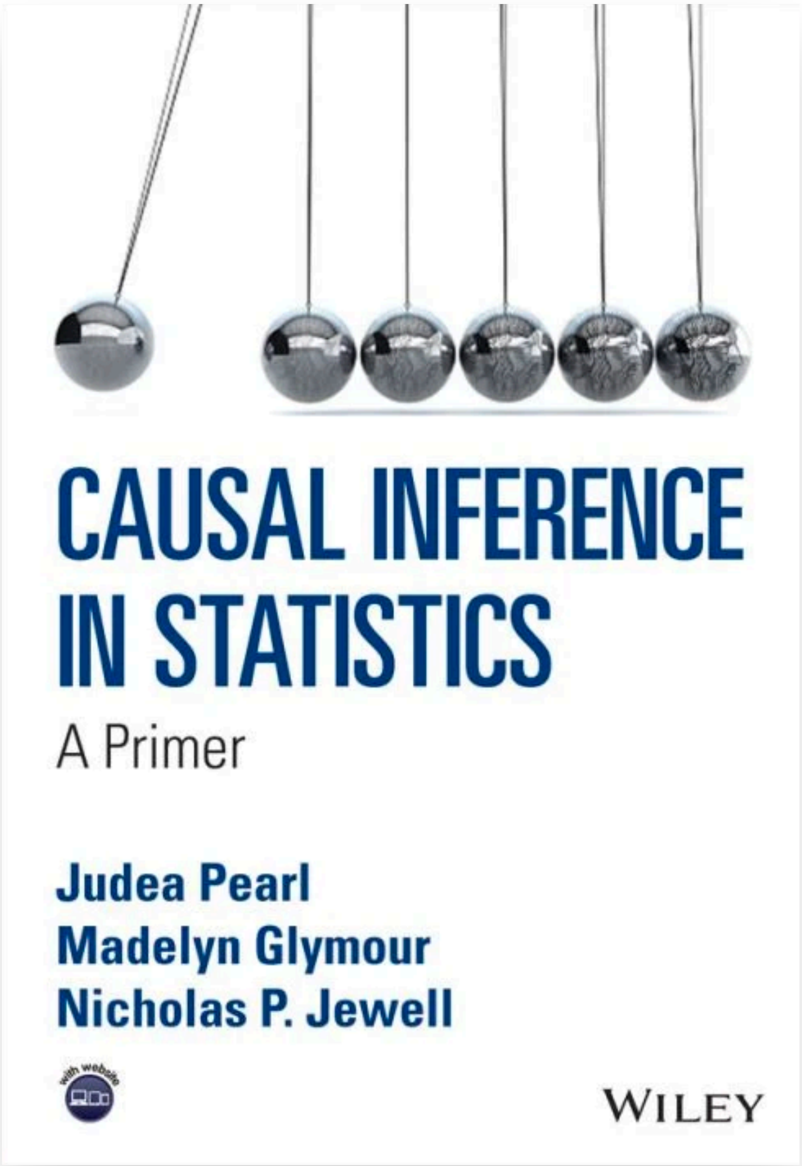


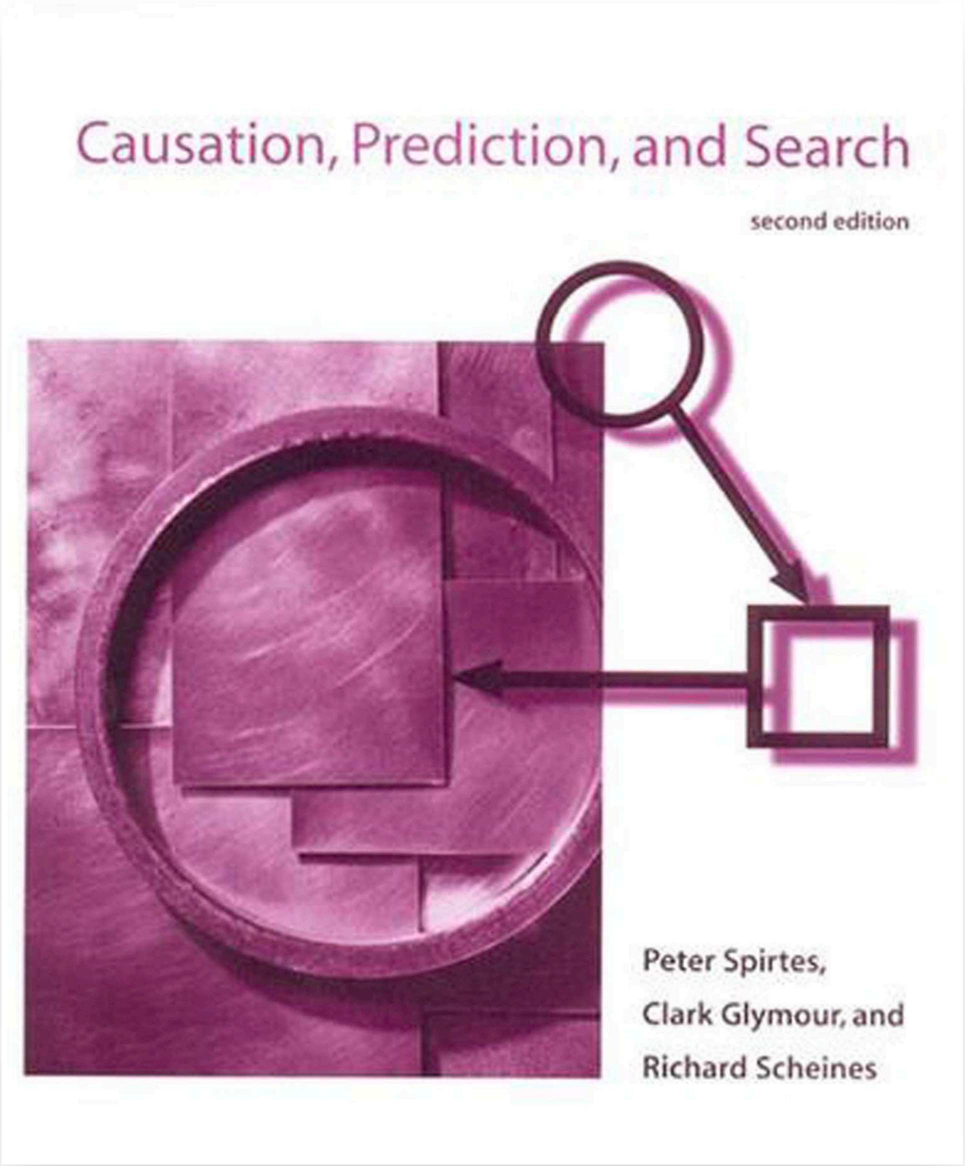
PC identifies possible causal links, then Momentary CI tests against time-lags.

Walker circulation: warm air travels westward from the east pacific (EPAC) over the central pacific (CPAC) to the west pacific (WPAC), where it becomes moist and rises before it travels back east. EPAC is also linked to the tropical atlantic.

[Jakob Runge et al. "Detecting and quantifying causal associations in large nonlinear time series datasets", 2019]

Resources


CAUSAL INFERENCE IN STATISTICS
A Primer
Judea Pearl
Madelyn Glymour
Nicholas P. Jewell
WILEY


Causation, Prediction, and Search
second edition
Peter Spirtes,
Clark Glymour, and
Richard Scheines

Assumptions

Assumption 1 (Causal Markov Condition):
Conditional independence in data reflects the absence of direct causal relationships (d-separation).

Assumption 2 (Faithfulness Condition):
Conditional independencies represent the true underlying causal structure.

Assumption 3 (Causal sufficiency):
We have measured all the common causes of the measured variables.