Probabilistic Graphical Models

Bayesian Learning of parameters

Structure Learning

MLE limitations

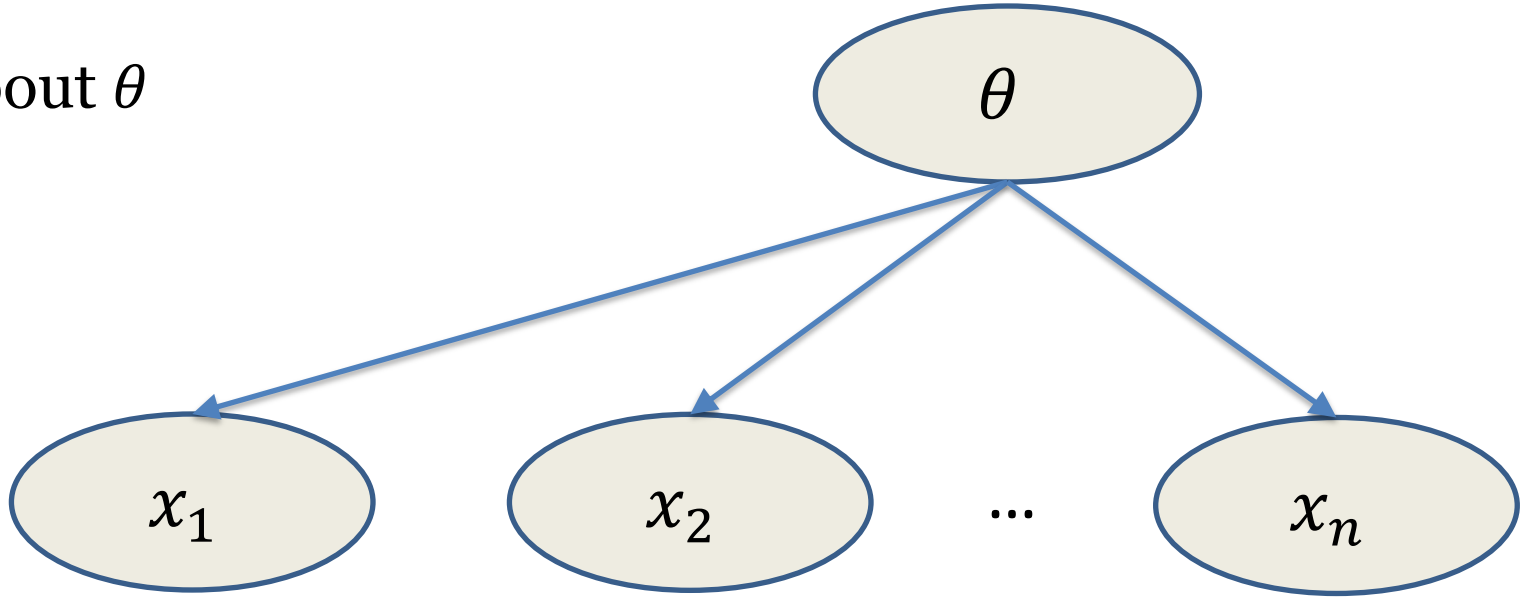
- Two teams play 10 times, and the first wins 7 of the 10 matches
  $\Rightarrow$ Probability of first team winning = 0.7
- A coin is tossed 10 times, and comes out 'heads' 7 of the 10 tosses
  $\Rightarrow$ Probability of heads = 0.7
- A coin is tossed 10000 times, and comes out 'heads' 7000 of the 10000 tosses
  $\Rightarrow$ Probability of heads = 0.7
- Before the first game, you cannot have an opinion on which team will win

Bayesian Inference

- Given a fixed θ, tosses are independent

- If θ is unknown, tosses are not marginally independent

each toss tells us something about θ



Bayesian Inference

- Given a fixed θ, tosses are independent

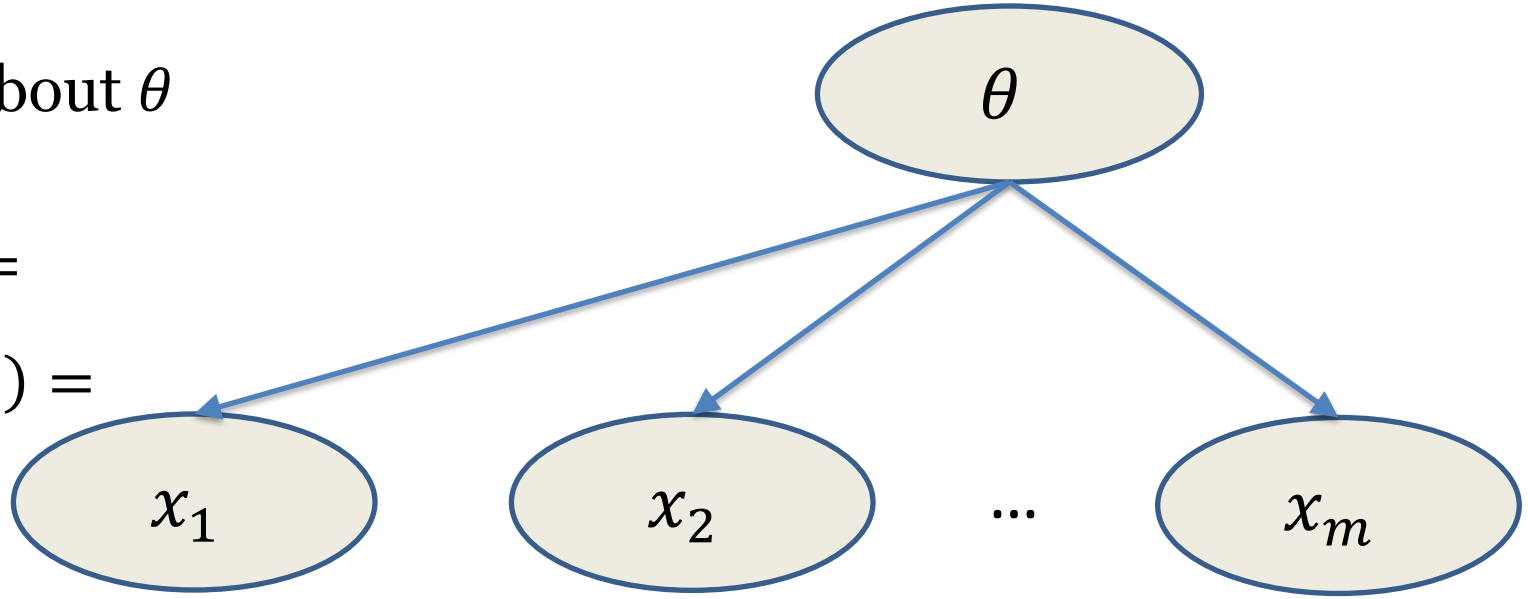
- If θ is unknown, tosses are not marginally independent

each toss tells us something about θ

$$P(\mathrm{x}[1], \ldots, x[m], \theta) =$$

$$\mathrm{P}(\mathrm{x}[1], \ldots, x[m], |\theta)P(\theta) =$$

$$P(\theta)\prod_{i}^{m} P(x[i]|\theta)$$

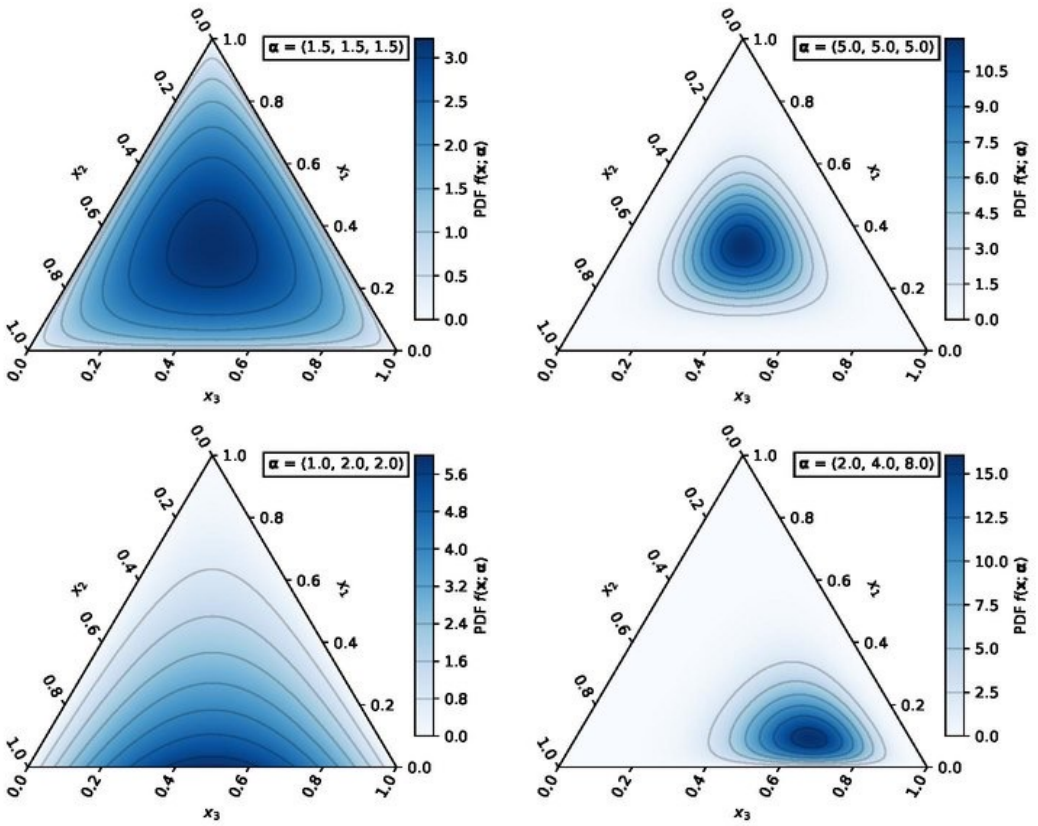


Bayesian Inference for Multinomial

Dirichlet distribution

$$f(\theta_1, \dots, \theta_k \mid \alpha_1, \dots, a_k) = \begin{cases} \frac{1}{B(\alpha)} \prod_{i=1}^{K} \theta_i^{a_i - 1}, \theta_i \in [0,1] \\ 0, \qquad otherwise \end{cases}$$

where $B(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\alpha_0)}, \alpha_0 = \sum_{i=1}^{K} \alpha_i$



Bayesian Inference for Multinomial

$$P(D \mid \theta) = \prod_{i=1}^{k} \theta_i^{M_i}$$

$$P(\theta) \propto \prod_{i=1}^{k} \theta_i^{a_i}$$

$$P(D|\theta)P(\theta) \propto \prod_{i=1}^{k} \theta_i^{a_i + M_i}$$

Update only uses sufficient statistics

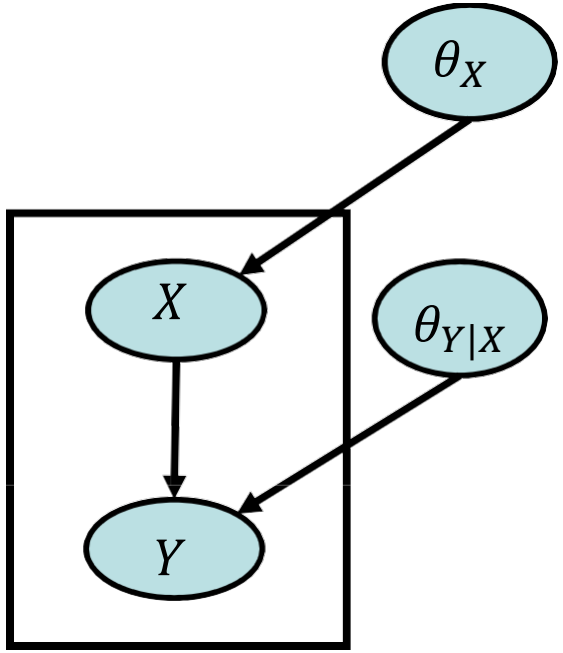
Bayesian Estimation for BNs

- Instances are independent given the parameters - (X[m'],Y[m']) are d-separated from (X [m], Y[m]) given θ

- Parameters for individual variables are independent a priori $P(\theta) = \prod P\left(\theta_{X_i} | P_a(X_i)\right)$

- Posteriors for θ are also independent given the data:
- $P\left(\theta_x, \theta_{y|x} \middle| D\right) = P\left(\theta_x | D\right) P\left(\theta_{y|x} \middle| D\right)$

  As in MLE, we can solve each estimation problem separately



Bayesian Estimation for BNs

- Instances are independent given the parameters - (X[m'],Y[m']) are d-separated from (X [m], Y[m]) given θ

- Parameters for individual variables are independent a priori $P(\theta) = \prod P\left(\theta_{X_i} | P_a(X_i)\right)$

- Posteriors for θ are also independent given the data:
- $P\left(\theta_x, \theta_{Y|X} \middle| D\right) = P\left(\theta_x \middle| D\right) P\left(\theta_{Y|X} \middle| D\right)$

  As in MLE, we can solve each estimation problem separately

- Posteriors of θ can be computed independently

– For multinomial $\theta_{X|\boldsymbol{u}}$ if prior is Dirichlet($a_{x^1|\boldsymbol{u}}, \ldots, a_{x^k|\boldsymbol{u}}$)

– posterior is Dirichlet($a_{x^1|\boldsymbol{u}} + M[x^1, \boldsymbol{u}], \ldots, a_{x^k|\boldsymbol{u}} + M[x^k, \boldsymbol{u}]$)

- Instances are independent given the parameters - (X[m'],Y[m']) are d-separated from (X [m], Y[m]) given θ

- Parameters for individual variables are independent a priori $P(\theta) = \prod P\left(\theta_{X_i} | P_a(X_i)\right)$

- Posteriors for θ are also independent given the data:
- $P\left(\theta_x, \theta_{Y|X} \middle| D\right) = P\left(\theta_x \middle| D\right) P\left(\theta_{Y|X} \middle| D\right)$

  As in MLE, we can solve each estimation problem separately

θ_X

X

$\theta_{Y|X}$

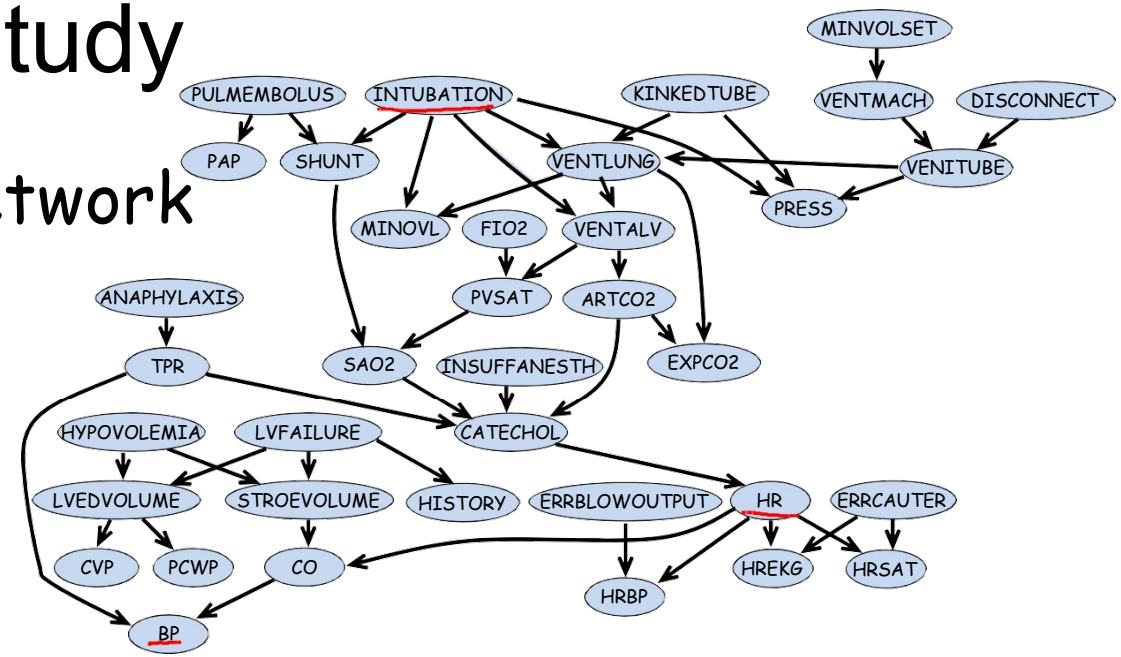
Y

Equivalent Sample size

- We need hyperparameter $\alpha_{x|\boldsymbol{u}}$ for each node X, value x, and parent assignment $\boldsymbol{u}$
  - Prior network with parameters Θ_o
  - Equivalent sample size parameter a
  - $\alpha_{x|\boldsymbol{u}} = \alpha P(x, \boldsymbol{u}|\Theta_0)$

Case Study

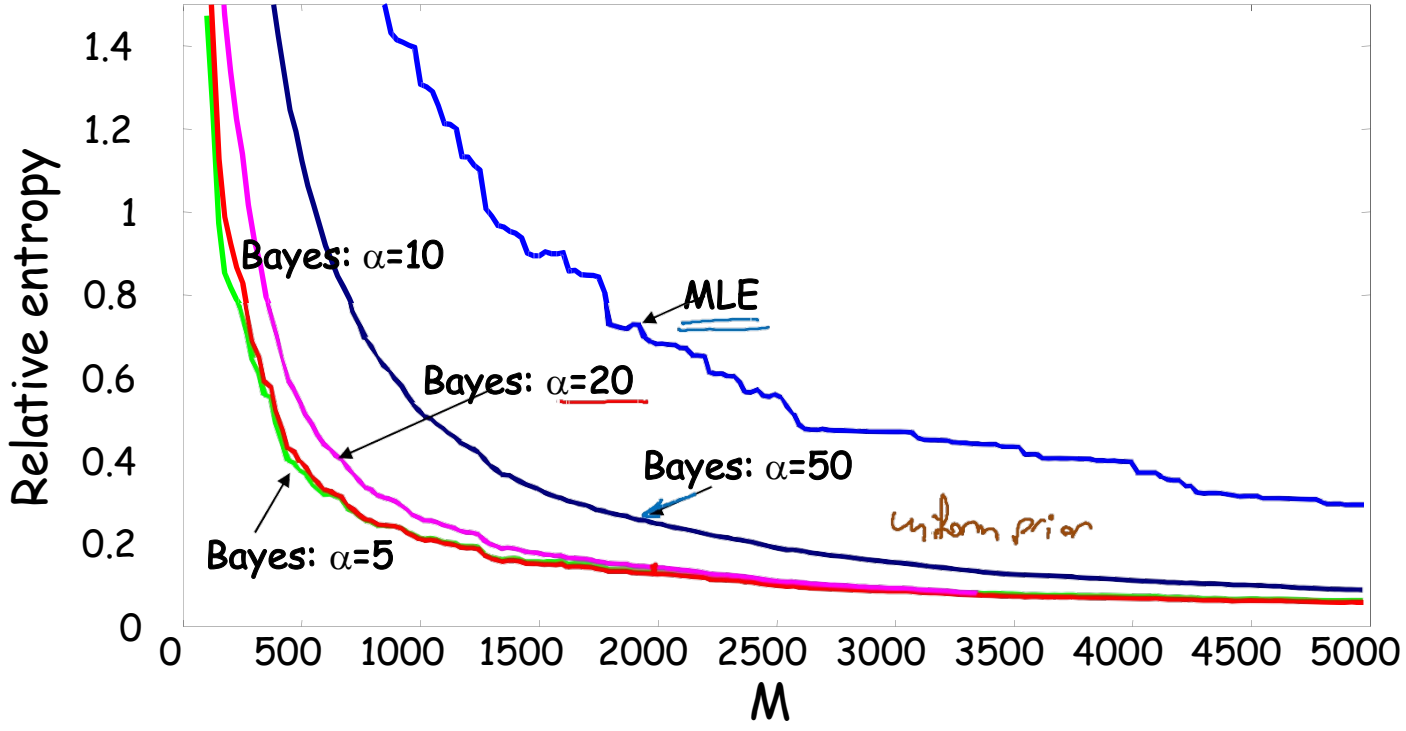
- ICU-Alarm network
  - 37 variables
  - 504 params



- Experiment
  - Sample instances from network
  - Relearn parameters

Daphne Koller

Case Study: ICU Alarm Network



Daphne Koller

Summary

- In Bayesian networks, if parameters are independent a priori, then also independent in the posterior

- For multinomial BNs, estimation uses sufficient statistics $M[x, \boldsymbol{u}]$

$$\hat{\theta}_{x|u} = \frac{M[x, \boldsymbol{u}]}{M[\boldsymbol{u}]}$$

MLE

$$E(x|\boldsymbol{u}, D) = \frac{\alpha_{x,u} + M[x, \boldsymbol{u}]}{\alpha_u + M[\boldsymbol{u}]}$$

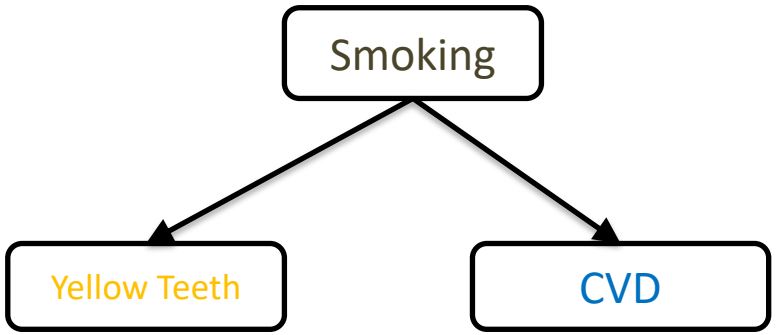
Bayesian (Dirichlet)

- Bayesian methods require choice of prior
  - can be elicited as prior network and equivalent sample size

Daphne Koller

What if you do not know the graph

Graph G captures the qualitative causal relations



JPD J encodes the quantitative probabilistic properties

		CVD		
Yellow Teeth	Smoking	Y	N	
Y	Y	0.17	0.06	0.13
N	Y	0.06	0.02	0.08
Y	N	0.02	0.06	0.08
N	N	0.15	0.46	0.61
		0.4	0.6	1

Markov Condition (MC):

Every variable is **independent** of its non-descendants in the graph given its parents.

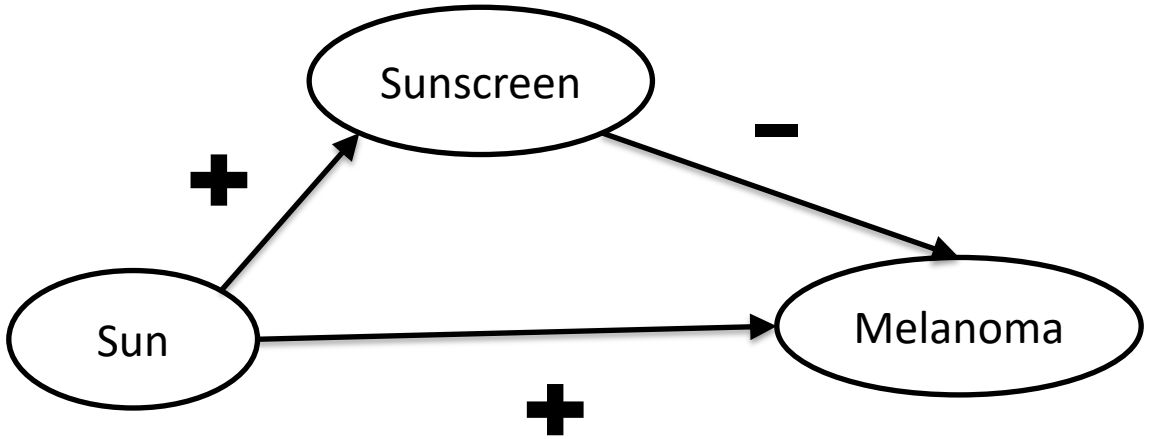
Faithfulness

Faithfulness Condition:
Independences stem **only** from the structure, **not the parameterization** of the distribution.
We say that the graph and the distribution are **faithful to each other**.

MC $\quad DSep(A, B|\boldsymbol{Z})$ in $G \Rightarrow \mathrm{A} \perp\!\!\!\perp \mathrm{B}|\mathbf{Z}$ in J

MC+FAITHFULNESS $\quad DSep(A, B|\boldsymbol{Z})$ in $G \Leftrightarrow \mathrm{A} \perp\!\!\!\perp \mathrm{B}|\mathbf{Z}$ in J

Faithfulness



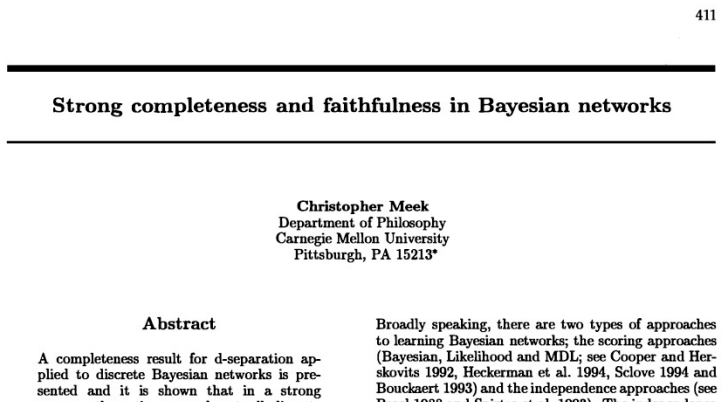
The parameters do not cancel each other out!

Faithfulness

Is it realistic?

Assume you are given a graph and you select the parameters of the conditional probability tables randomly following a Dirichlet distribution. The probability you get a non-faithful BN is zero (Lebesque measure is zero).

[Meek. C. UAI 1995]



411

Strong completeness and faithfulness in Bayesian networks

Christopher Meek
Department of Philosophy
Carnegie Mellon University
Pittsburgh, PA 15213*

Abstract

A completeness result for d-separation applied to discrete Bayesian networks is presented and it is shown that in a strong

Broadly speaking, there are two types of approaches to learning Bayesian networks; the scoring approaches (Bayesian, Likelihood and MDL; see Cooper and Herskovits 1992, Heckerman et al. 1994, Sclove 1994 and Bouckaert 1993) and the independence approaches (see

Faithfulness

Is it realistic?

Probable causes of non-faithfulness:

- Too low associations are not detectable for finite samples.
- Too high correlations (determinism or close-to-determinism).
- Natural selection may be biasing towards creating non-faithful distributions in systems in nature (e.g.. cells)!
- Not all joint probability distributions have a faithful representation.

The probability of getting an almost non-faithful distribution is non-zero.

Markov Condition + Faithfulness

$X \longleftarrow Y$ $\qquad$ $X \longrightarrow Y$

The edge is a d-connecting path that can not be broken given any other variables.

A useful implication of the Markov Condition

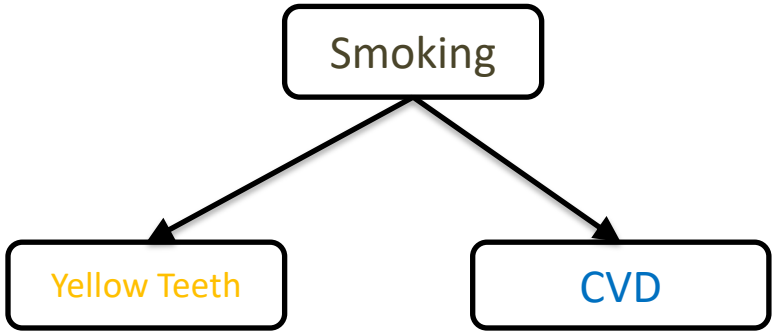
If X, Y are adjacent in the graph, then $\nexists\ \boldsymbol{Z}$ s.t. $(X, Y \perp\!\!\!\perp \boldsymbol{Z})$.

If $\exists\ \boldsymbol{Z}$ s.t. $(X, Y \perp\!\!\!\perp \boldsymbol{Z})$, X and Y are NOT adjacent in the graph.

An edge denotes unique information (given all other variables)

Bayesian Networks (BNs)

Graph $\mathcal{G}$



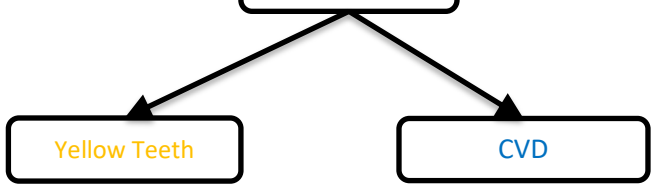
JPD$(\boldsymbol{V})$: $\mathcal{J}$

		CVD		
Yellow Teeth	Smoking	Y	N	
Y	Y	0.17	0.06	0.13
N	Y	0.06	0.02	0.08
Y	N	0.02	0.06	0.08
N	N	0.15	0.46	0.61
		0.4	0.6	1

Markov Condition +
Faithfulness =
Independence ↔ D-separation

Reverse-engineering the graph

What you want



P(Smoking)
0.3

Smoking	P(YT)
Y	0.85
N	0.10

Smoking	P(CVD)
Y	0.75
N	0.15

What you have

Sample (Person)	Smoking	CVD	Yellow Teeth
1	Yes	Yes	No
2	No	No	No
3	Yes	Yes	Yes
4	No	No	Yes
5	Yes	No	No
6	No	Yes	Yes
⋮			
52	No	Yes	No

Hypothesis Testing

- Identify the research question
- Writing the statistical hypotheses in terms of parameters of interest.
- Collect data and calculate a statistic
- Find the distribution of the statistic under the null hypothesis
- Find the p-value (probability that the result we got or a more extreme one happens just by chance given that the null hypothesis is true).
- Decide if the p-value is small or large
- Reject if p-value is lower than the significance threshold a.

Testing (In)Dependencies

Hypothesis Testing

- Identify the research question Is smoking independent from CVD?
- Writing the statistical hypotheses in terms of parameters of interest.
  P(smoking, CVD) = P(smoking)P(CVD)
- Collect data and calculate a statistic
- Find the distribution of the statistic under the null hypothesis
- Find the p-value (probability that the result we got or a more extreme one happens just by chance given that the null hypothesis is true).
- Decide if the p-value is small or large
- Reject if p-value is lower than the significance threshold a.

Example: Independence

- You have a population of 520 people
  - 160/520 smoke.
  - 210/520 have CVD.

		CVD		
		Y	N	Total
Smoking	Y	120	40	160
	N	90	270	360
	Total	210	310	520

Contingency table

Example: Independence

Null Hypothesis ($\mathbf{H}_0$) : Smoking is independent of CVD
Alternative Hypothesis ($\mathbf{H}_1$) : Smoking is dependent of CVD

Mathematically:

$$\mathbf{H}_0 = \forall\, i,j \;\; p_{ij} = p_{i.} \times p_{.j}$$
$$\mathbf{H}_1 = \exists\, i,j: \;\; p_{ij} \neq p_{i.} \times p_{.j}$$

	CVD=0	CVD=1	
S=0	p_{00}	p_{01}	$p_{0.}$
S=1	p_{10}	p_{11}	$p_{1.}$
	$p_{.0}$	$p_{.1}$	1

Reminder: Independence:

$$\forall x, y \; P(\mathrm{Y} = \mathrm{y}, \mathrm{X} = \mathrm{x}) = P(Y = \mathrm{y})\mathrm{P}(\mathrm{X} = \mathrm{x})$$

$$p_{ij} = P(X = i, \mathrm{Y} = j)$$
$$p_{i.} = P(X = i)$$
$$p_{.j} = P(Y = j)$$

Dependence

		CVD		
		Y	N	Total
Smoking	Y	120	40	160
	N	90	270	360
	Total	210	310	520

Contingency table

		CVD		
		Y	N	Total
Smoking	Y	.2308	.0769	.3077
	N	.1731	.5192	.6923
	Total	.4038	.5962	1

Joint Probability Distribution
$P(CVD, Smoking)$

Dependence

		CVD		
		Y	N	Total
Smoking	Y	120	40	160
	N	90	270	360
	Total	210	310	520

Contingency table

		CVD		
		Y	N	Total
Smoking	Y	.75	.25	1
	N	.25	.75	1

Conditional Probability Distribution
$P(CVD|Smoking)$

P(Smoking)

		CVD		
		Y	N	Total
Smoking	Y	.2308	.0769	.3077
	N	.1731	.5192	.6923
	Total	.4038	.5962	1

Joint Probability Distribution
$P(CVD, Smoking)$

P(Smoking|CVD=yes)

		CVD	
		Y	
Smoking	Y	.5714	.1290
	N	.4286	.8710
	Total	1	1

Conditional Probability Distribution
$P(Smoking|CVD)$

P(Smoking)≠P(Smoking|CVD=yes)

Test statistic: Expected counts

		CVD		
		Y	N	Total
Smoking	Y	.2308	.0769	.3077
	N	.1731	.5192	.6923
	Total	.4038	.5962	1

in your data

		CVD		
		Y	N	Total
Smoking	Y			.3077
	N			.6923
	Total	.4038	.5962	1

If Smoking and CVD were independent?

Are Smoking and CVD independent?

		CVD		
		Y	N	Total
Smoking	Y	.2308	.0769	.3077
	N	.1731	.5192	.6923
	Total	.4038	.5962	1

in your data

		CVD		
		Y	N	Total
Smoking	Y			.3077
	N			.6923
	Total	.4038	.5962	1

If Smoking and CVD were independent?

$$P(Smoking = Yes, CVD = Yes) = P(Smoking = Yes) * P(CVD = Yes)$$

Are Smoking and CVD independent?

		CVD		
		Y	N	Total
Smoking	Y	.2308	.0769	.3077
	N	.1731	.5192	.6923
	Total	.4038	.5962	1

in your data

		CVD		
		Y	N	Total
Smoking	Y			.3077
	N			.6923
	Total	.4038	.5962	1

If Smoking and CVD were independent?

$$P(Smoking = Yes, CVD = Yes) = P(Smoking = Yes) * P(CVD = Yes) = 0.4038 * 0.3077$$

Are Smoking and CVD independent?

		CVD		
		Y	N	Total
Smoking	Y	.2308	.0769	.3077
	N	.1731	.5192	.6923
	Total	.4038	.5962	1

in your sample

		CVD		
		Y	N	Total
Smoking	Y	.1242	.1835	.3077
	N	.2796	.4127	.6923
	Total	.4038	.5962	1

If Smoking and CVD were independent?

Are Smoking and CVD independent?

		CVD	
		Y	N
Smoking	Y	120	40
	N	90	270

counts in your data

		CVD	
		Y	N
Smoking	Y	65	95
	N	145	215

Expected counts If Smoking and CVD were independent

$$P(Smoking = Yes, CVD = Yes) * \# \ samples = .1242*520$$

Summarize the differences

- n_{ij}: Counts in your data (# observations in cell i,j)
- e_{ij}: Expected counts under H_0

$$X^2 = \sum \frac{(observed - expected)^2}{expected} = \sum_{i,j} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

What is the probability of observing a value *t* at least as extreme as the one you observed in your data?

p-value: $P\left(X^2 > x_{obs}^2 \middle| \mathrm{H}_0\right)$

The chi-square distribution

- In order to determine if the χ^2 statistic we calculated is considered unusually high or not we need to first describe its distribution.

$$X^2 = \sum_{i=1}^{k} \frac{(N_i - np_i^o)^2}{np_i^0}$$

Under the null, when $n \to \infty, X^2 \sim \chi^2$ *with k-1 degrees of freedom.*

- The chi-square distribution has just one parameter called *degrees of freedom (df)*, which influences the shape, center, and spread of the distribution.

Chi square Distribution

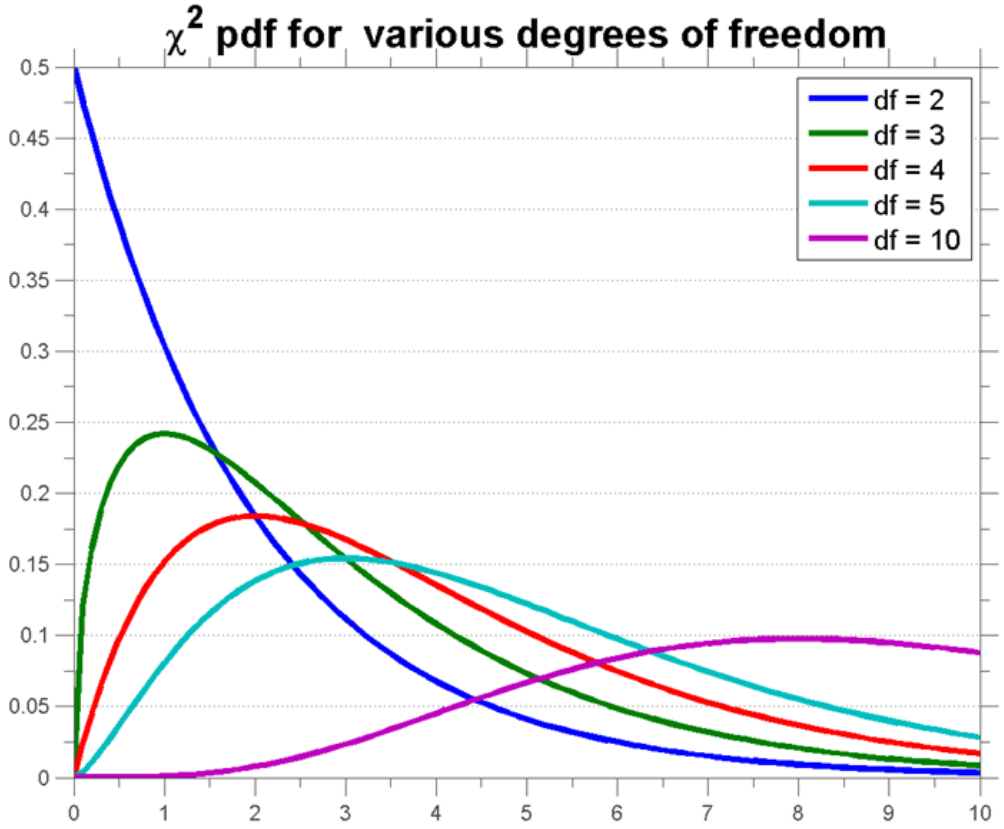
$$P(X^2 = t|H_0) \sim \frac{t^{\frac{df-2}{2}} e^{-\frac{t}{2}}}{2^{\frac{df}{2}} \Gamma\left(\frac{df}{2}\right)},$$

where df are the degrees of freedom, i.e. the number of parameters that are free to vary

For testing X ⫫ Y

$df = (\#\text{ possible values of } X - 1) \times$
$(\#\text{ of possible values of } Y - 1)$

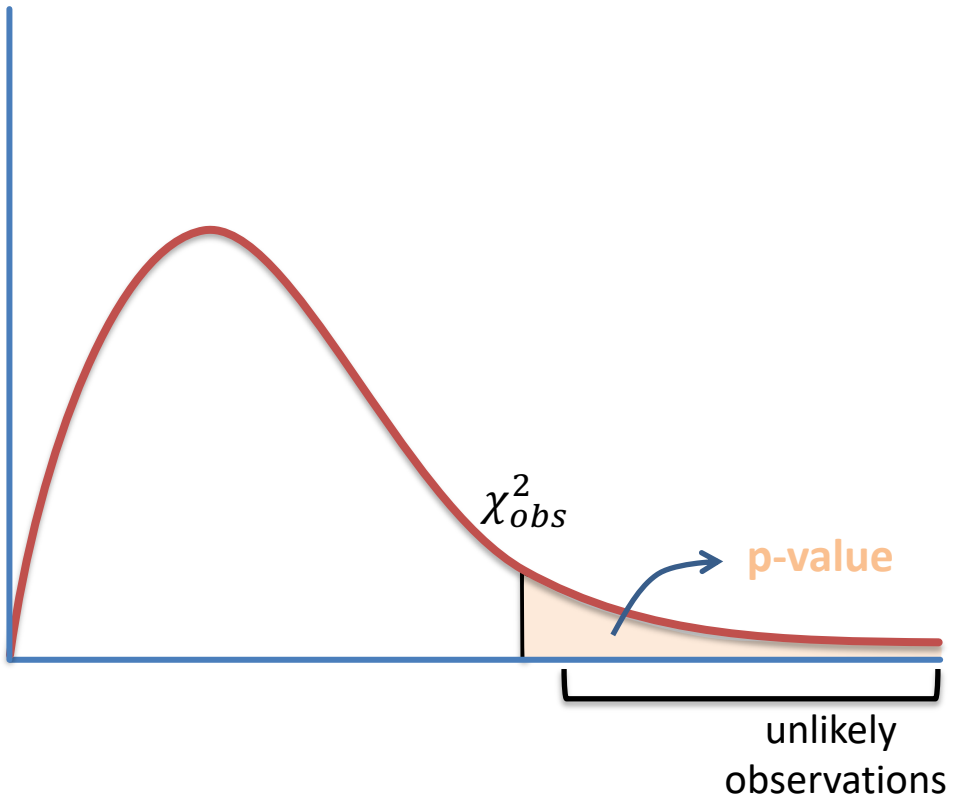
in our example $df = (2-1)\times(2-1) = 1$



Make a Decision

Check in the pdf
If the p-value is less than a significance threshold α, reject the null hypothesis.

p-value: $P\left(X^2 > \chi^2_{obs} \middle| \mathrm{H}_0\right)$

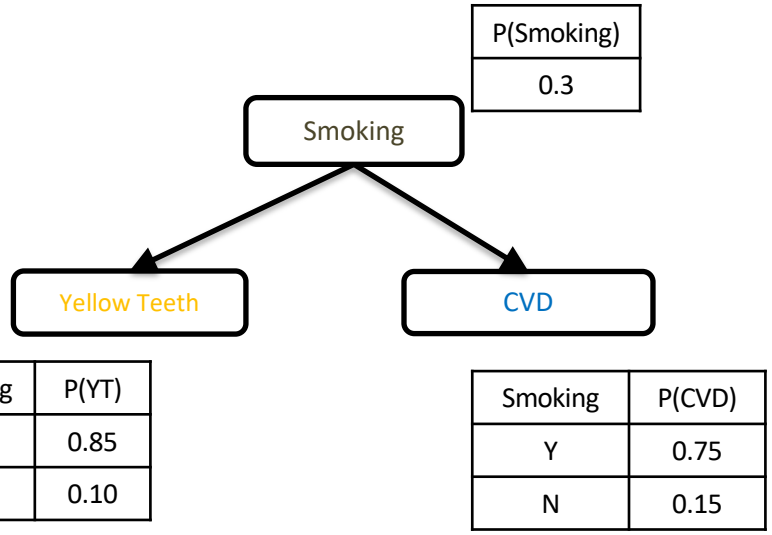


Now you can decide if you will reject H_o or not.

You can decide if X and Y are independent (given **Z**)

Reverse-engineering the graph

What you want



Can we find the graph where the only d-separation is CVD and Yellow teeth given smoking?

What you have

You can use tests of conditional independence to identify the set of conditional independencies:

Here you only have one independence:

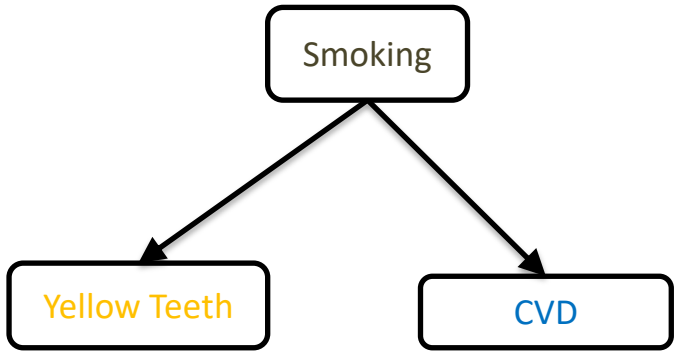
CVD $\perp\!\!\!\perp$ Yellow Teeth | Smoking

And the rest are dependencies:

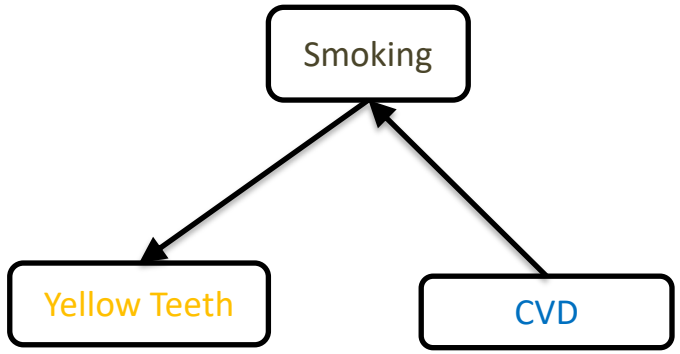
Smoking $\not\perp\!\!\!\perp$ Yellow Teeth | Ø
Smoking $\not\perp\!\!\!\perp$ Yellow Teeth | CVD

Smoking $\not\perp\!\!\!\perp$ CVD | Ø
Smoking $\not\perp\!\!\!\perp$ CVD | Yellow Teeth

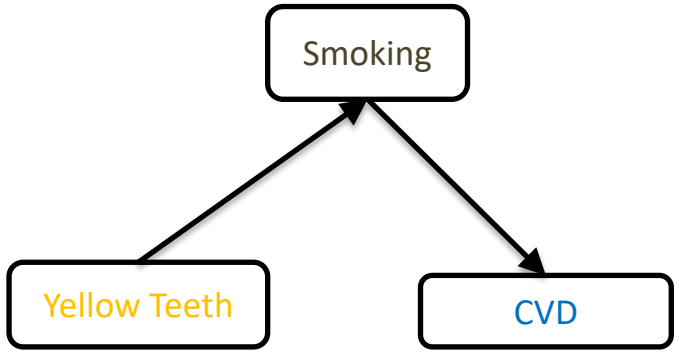
Markov Equivalence



$\text{CVD} \perp\!\!\!\perp \text{Yellow Teeth} \mid \text{Smoking}$



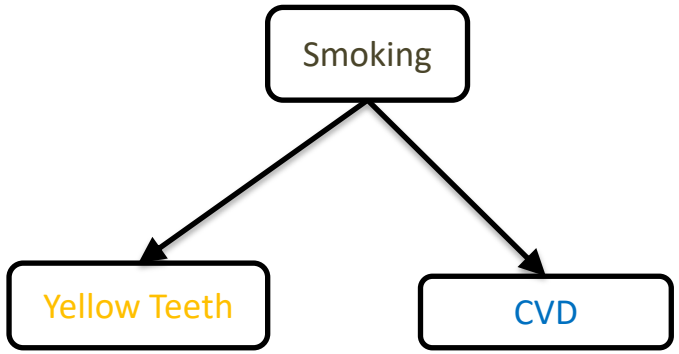
$\text{CVD} \perp\!\!\!\perp \text{Yellow Teeth} \mid \text{Smoking}$



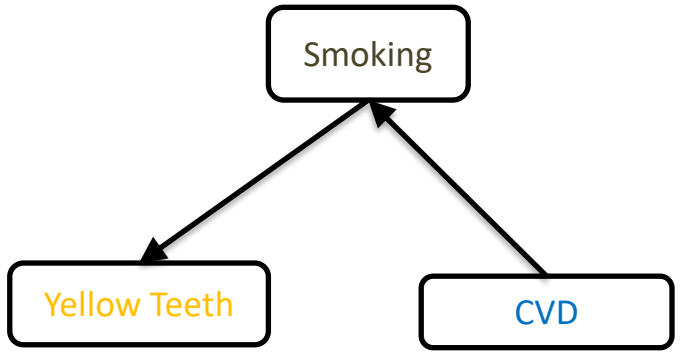
$\text{CVD} \perp\!\!\!\perp \text{Yellow Teeth} \mid \text{Smoking}$

Markov Condition entails the same conditional independence for all three graphs.

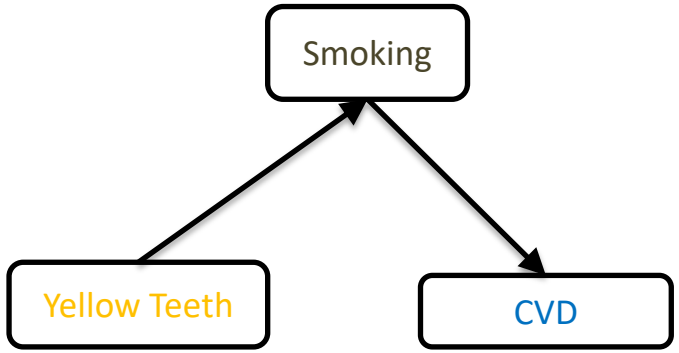
Markov Equivalence



$\text{CVD} \perp\!\!\!\perp \text{Yellow Teeth} \mid \text{Smoking}$



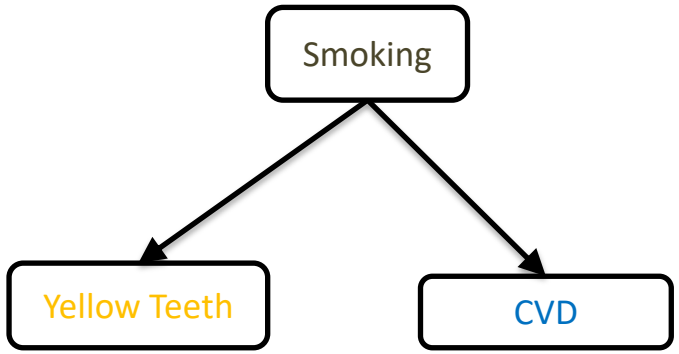
$\text{CVD} \perp\!\!\!\perp \text{Yellow Teeth} \mid \text{Smoking}$



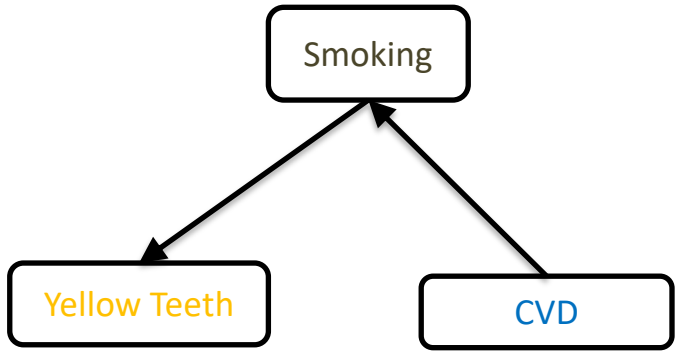
$\text{CVD} \perp\!\!\!\perp \text{Yellow Teeth} \mid \text{Smoking}$

- The graphs are called **Markov Equivalent.**
- All Markov equivalent graphs denote a **Markov equivalence class (MEC).**
- **We use [G] to denote the MEC of G.**

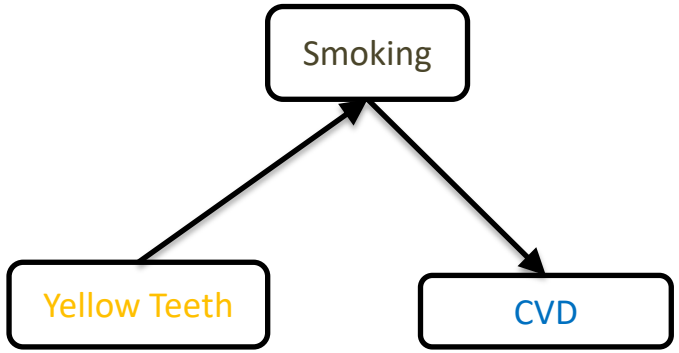
Markov Equivalence



CVD $\perp\!\!\!\perp$ Yellow Teeth | Smoking



CVD $\perp\!\!\!\perp$ Yellow Teeth | Smoking

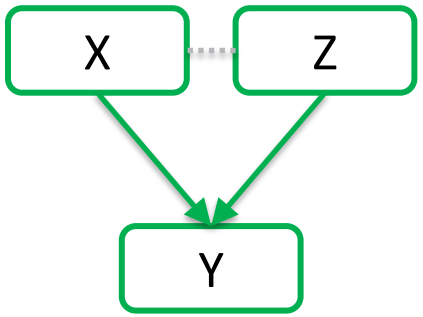


CVD $\perp\!\!\!\perp$ Yellow Teeth | Smoking

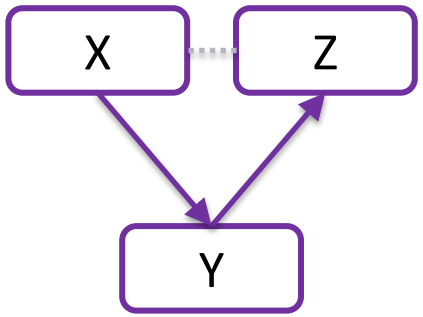
Markov Equivalent Graphs share

- the same skeleton (adjacencies).
- the same unshielded colliders

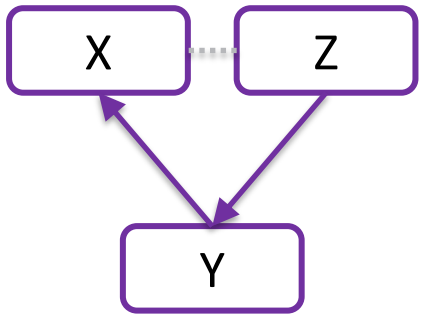
Reminder: (non) colliders



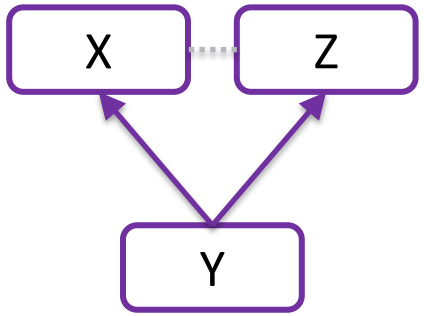
collider



non- collider



non- collider



non- collider

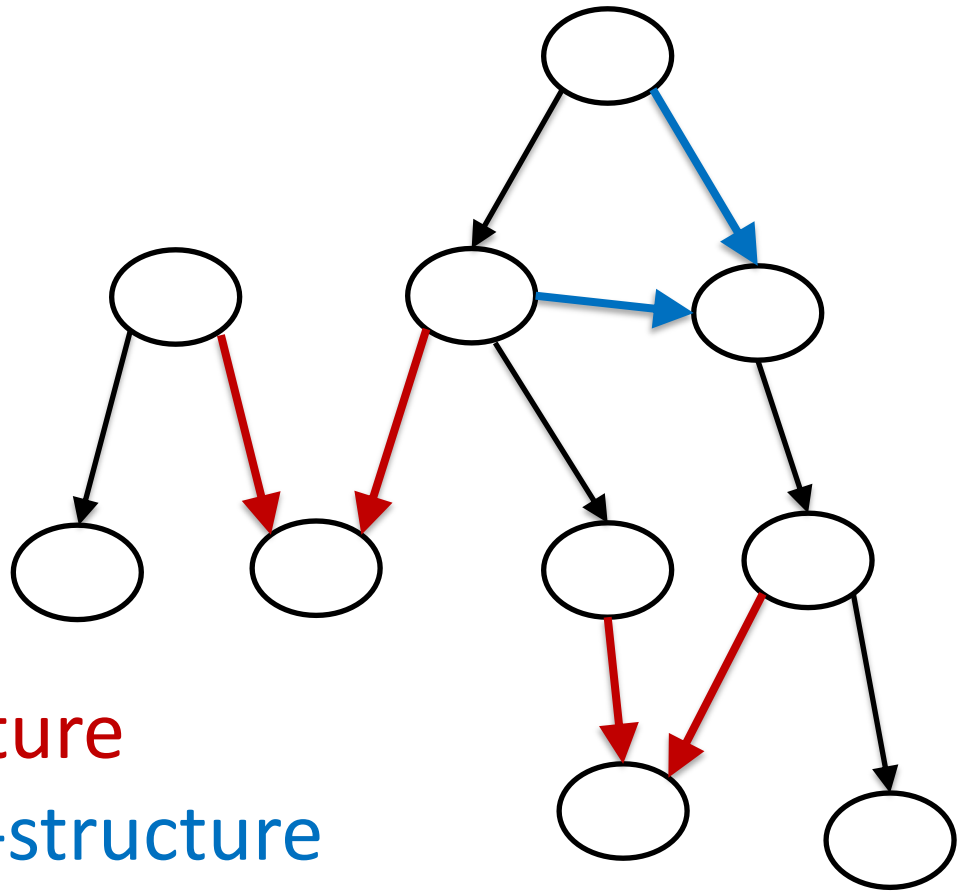
For a triple X-Y-Z:

If both edges are into Y, the triplet (and Y) is a collider.
Otherwise the triplet (and Y) is a non-collider.

The term is used to denote both the triplet and the middle node!

Characterization of the Markov Equivalence Class

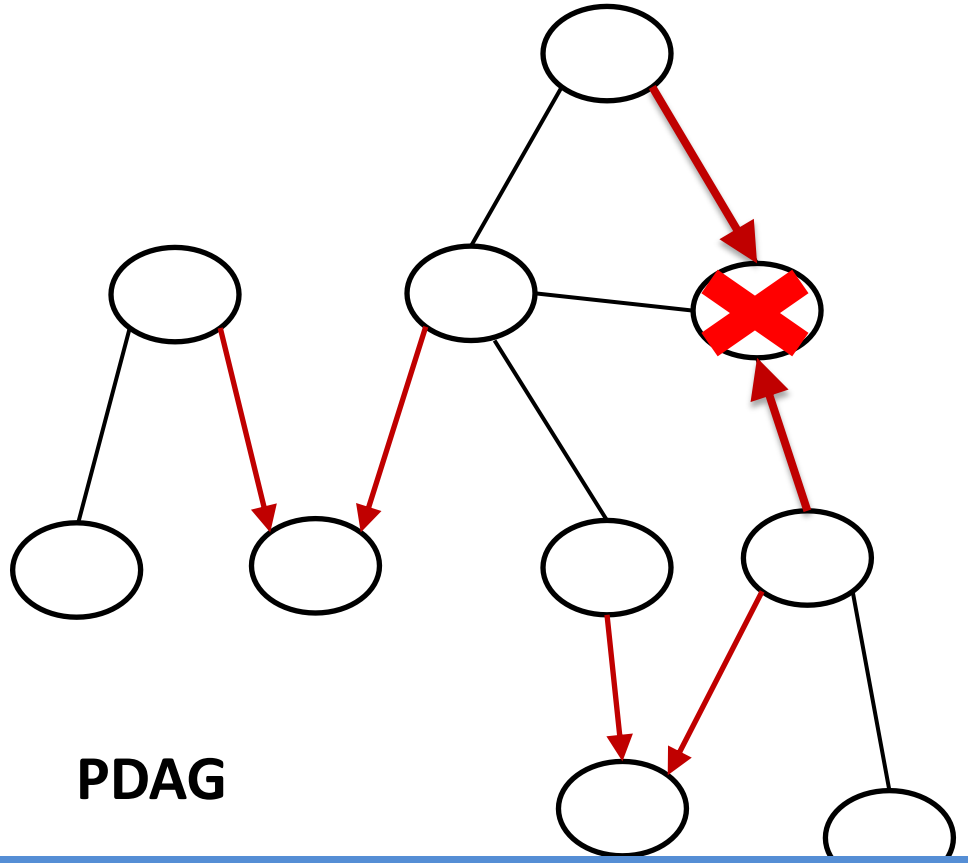
- Unshielded collider: A collider (X-Y-Z) where the endpoints (X, Z) are NOT adjacent.
- AKA v-structure.



- v-structure
- not a v-structure

Pattern DAGs

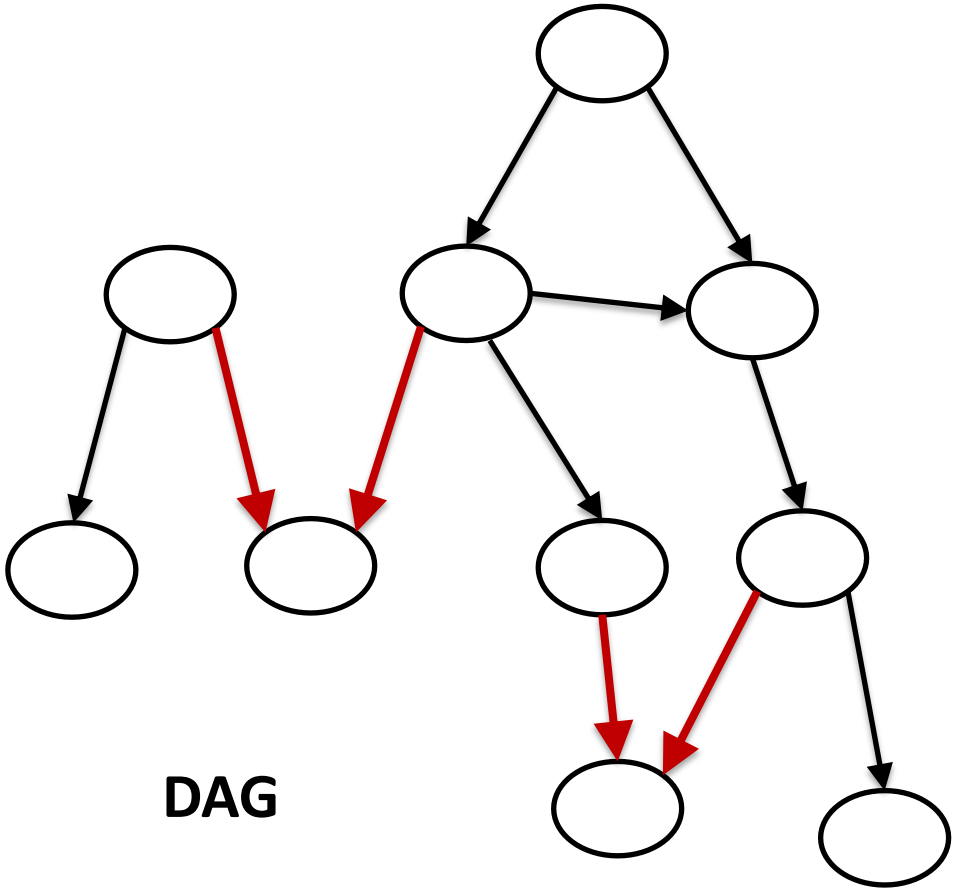
- Represents a **class** of Markov Equivalent DAGs.
- Has the **same edges** as every DAG in the class.
- Has only **orientations** (arrows) shared by **all the DAGs** in the class.
- Orient the PDAG as a DAG **without** creating a new collider or directed cycle!

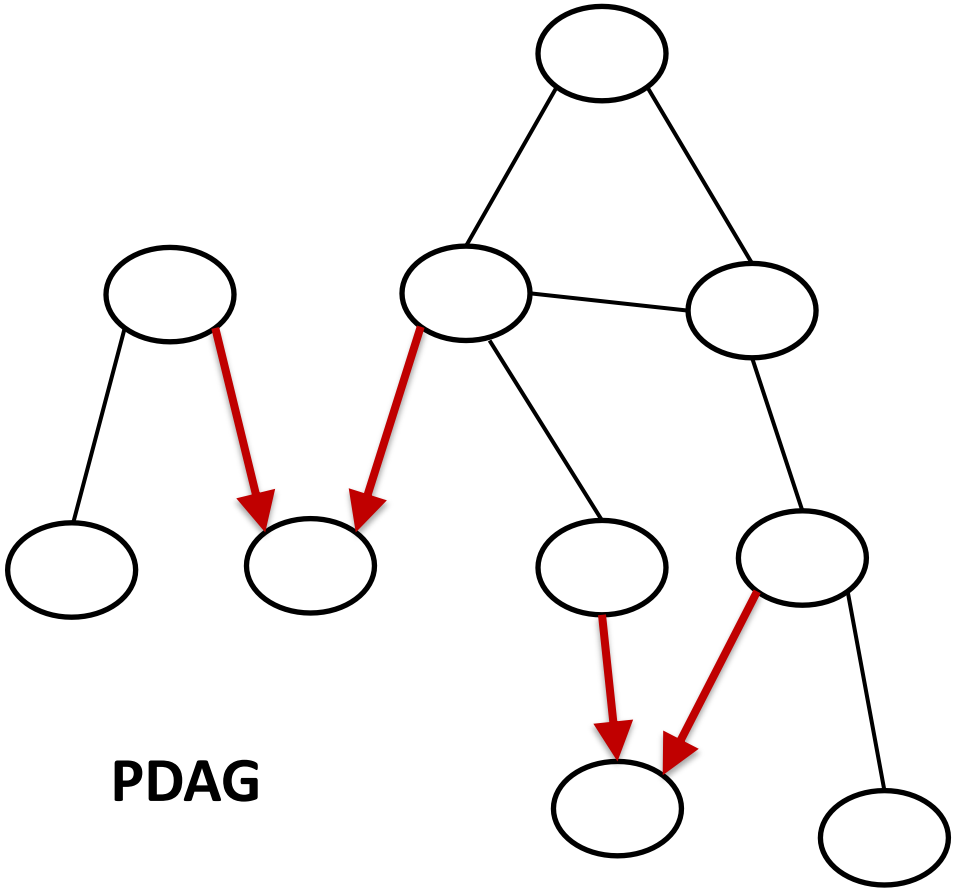


PDAG

Not all configurations are possible!

Pattern DAGs

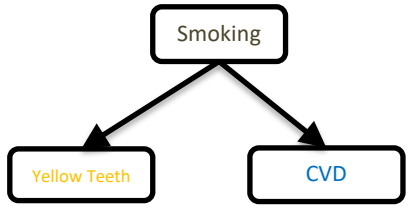




- You can still “read” all conditional independencies entailed by the Markov Condition in the graph using d-separation.

Reverse-engineering

Bayesian Network describing your variables



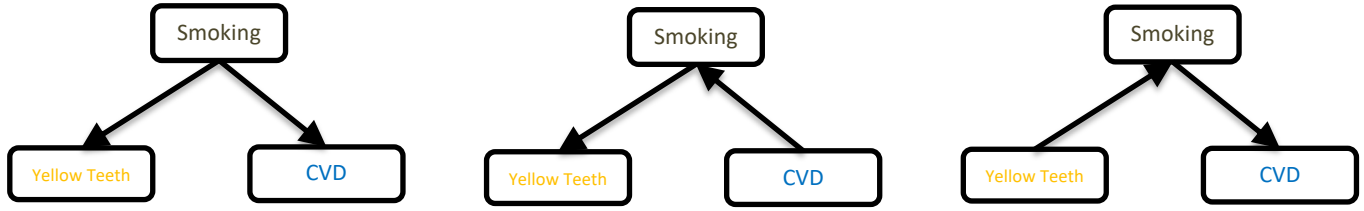
Independencies entailed by the CMC

CVD $\perp\!\!\!\perp$ Yellow Teeth | Smoking

For a causal structure, Causal Markov Condition entails a (possibly empty) set of conditional independencies.

Reverse-engineering

Causal Bayesian Network describing your variables



Independencies entailed by the MC

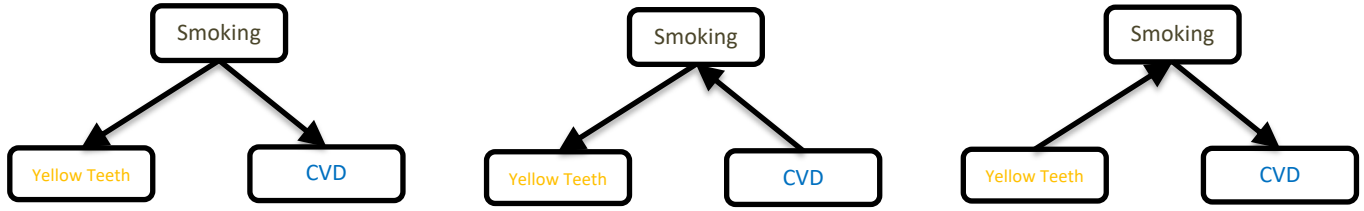
CVD $\perp\!\!\!\perp$ Yellow Teeth | Smoking

CVD $\perp\!\!\!\perp$ Yellow Teeth | Smoking

CVD $\perp\!\!\!\perp$ Yellow Teeth | Smoking

The same set of conditional independencies is entailed by all Markov equivalent networks.

Reverse-engineering

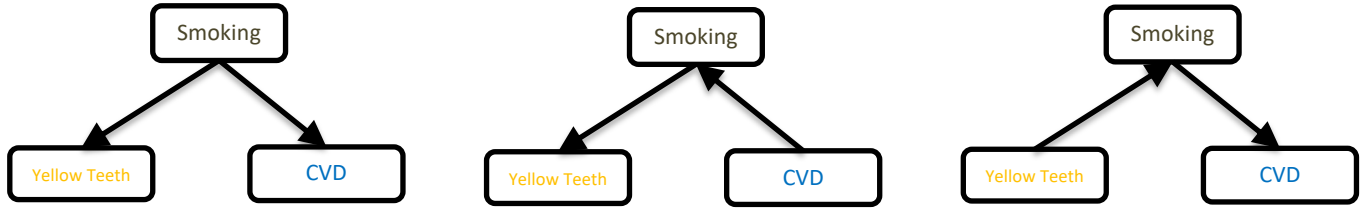


Causal Bayesian Network describing your variables			
Independencies entailed by the MC	**CVD $\perp\!\!\!\perp$ Yellow Teeth | Smoking**	**CVD $\perp\!\!\!\perp$ Yellow Teeth | Smoking**	**CVD $\perp\!\!\!\perp$ Yellow Teeth | Smoking**
Under Faithfulness	CVD $\not\perp\!\!\!\perp$ Yellow Teeth | $\emptyset$ Smoking $\not\perp\!\!\!\perp$ Yellow Teeth | $\emptyset$ Smoking $\not\perp\!\!\!\perp$ Yellow Teeth | CVD Smoking $\not\perp\!\!\!\perp$ CVD | $\emptyset$ Smoking $\not\perp\!\!\!\perp$ CVD | Yellow Teeth	CVD $\not\perp\!\!\!\perp$ Yellow Teeth | $\emptyset$ Smoking $\not\perp\!\!\!\perp$ Yellow Teeth | $\emptyset$ Smoking $\not\perp\!\!\!\perp$ Yellow Teeth | CVD Smoking $\not\perp\!\!\!\perp$ CVD | $\emptyset$ Smoking $\not\perp\!\!\!\perp$ CVD | Yellow Teeth	CVD $\not\perp\!\!\!\perp$ Yellow Teeth | $\emptyset$ Smoking $\not\perp\!\!\!\perp$ Yellow Teeth | $\emptyset$ Smoking $\not\perp\!\!\!\perp$ Yellow Teeth | CVD Smoking $\not\perp\!\!\!\perp$ CVD | $\emptyset$ Smoking $\not\perp\!\!\!\perp$ CVD | Yellow Teeth

If you also assume faithfulness, all remaining relationships are conditional dependencies.

Reverse-engineering

Causal Bayesian Network describing your variables



Independencies entailed by the MC

$\text{CVD} \perp\!\!\!\perp \text{Yellow Teeth} \mid \text{Smoking}$

$\text{CVD} \perp\!\!\!\perp \text{Yellow Teeth} \mid \text{Smoking}$

$\text{CVD} \perp\!\!\!\perp \text{Yellow Teeth} \mid \text{Smoking}$

Under Faithfulness

$\text{CVD} \not\perp\!\!\!\perp \text{Yellow Teeth} \mid \emptyset$

$\text{Smoking} \not\perp\!\!\!\perp \text{Yellow Teeth} \mid \emptyset$
$\text{Smoking} \not\perp\!\!\!\perp \text{Yellow Teeth} \mid \text{CVD}$

$\text{Smoking} \not\perp\!\!\!\perp \text{CVD} \mid \emptyset$
$\text{Smoking} \not\perp\!\!\!\perp \text{CVD} \mid \text{Yellow Teeth}$

$\text{CVD} \not\perp\!\!\!\perp \text{Yellow Teeth} \mid \emptyset$

$\text{Smoking} \not\perp\!\!\!\perp \text{Yellow Teeth} \mid \emptyset$
$\text{Smoking} \not\perp\!\!\!\perp \text{Yellow Teeth} \mid \text{CVD}$

$\text{Smoking} \not\perp\!\!\!\perp \text{CVD} \mid \emptyset$
$\text{Smoking} \not\perp\!\!\!\perp \text{CVD} \mid \text{Yellow Teeth}$

$\text{CVD} \not\perp\!\!\!\perp \text{Yellow Teeth} \mid \emptyset$

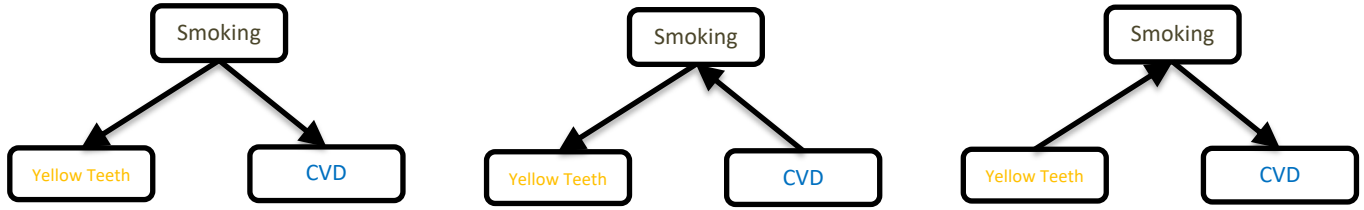
$\text{Smoking} \not\perp\!\!\!\perp \text{Yellow Teeth} \mid \emptyset$
$\text{Smoking} \not\perp\!\!\!\perp \text{Yellow Teeth} \mid \text{CVD}$

$\text{Smoking} \not\perp\!\!\!\perp \text{CVD} \mid \emptyset$
$\text{Smoking} \not\perp\!\!\!\perp \text{CVD} \mid \text{Yellow Teeth}$

Query the data to get the conditional (in) dependencies

Reverse-engineering the PDAG

Causal Bayesian Network describing your variables



Identify all DAGs that entail these (and only these) conditional independencies.

Independencies entailed by the MC	**CVD $\perp\!\!\!\perp$ Yellow Teeth | Smoking**	**CVD $\perp\!\!\!\perp$ Yellow Teeth | Smoking**	**CVD $\perp\!\!\!\perp$ Yellow Teeth | Smoking**
Under Faithfulness	CVD $\not\perp\!\!\!\perp$ Yellow Teeth | Ø Smoking $\not\perp\!\!\!\perp$ Yellow Teeth | Ø Smoking $\not\perp\!\!\!\perp$ Yellow Teeth | CVD Smoking $\not\perp\!\!\!\perp$ CVD | Ø Smoking $\not\perp\!\!\!\perp$ CVD | Yellow Teeth	CVD $\not\perp\!\!\!\perp$ Yellow Teeth | Ø Smoking $\not\perp\!\!\!\perp$ Yellow Teeth | Ø Smoking $\not\perp\!\!\!\perp$ Yellow Teeth | CVD Smoking $\not\perp\!\!\!\perp$ CVD | Ø Smoking $\not\perp\!\!\!\perp$ CVD | Yellow Teeth	CVD $\not\perp\!\!\!\perp$ Yellow Teeth | Ø Smoking $\not\perp\!\!\!\perp$ Yellow Teeth | Ø Smoking $\not\perp\!\!\!\perp$ Yellow Teeth | CVD Smoking $\not\perp\!\!\!\perp$ CVD | Ø Smoking $\not\perp\!\!\!\perp$ CVD | Yellow Teeth

Query the data to get the conditional (in) dependencies

Brute force: generate all possible DAGs and test using d-separation.

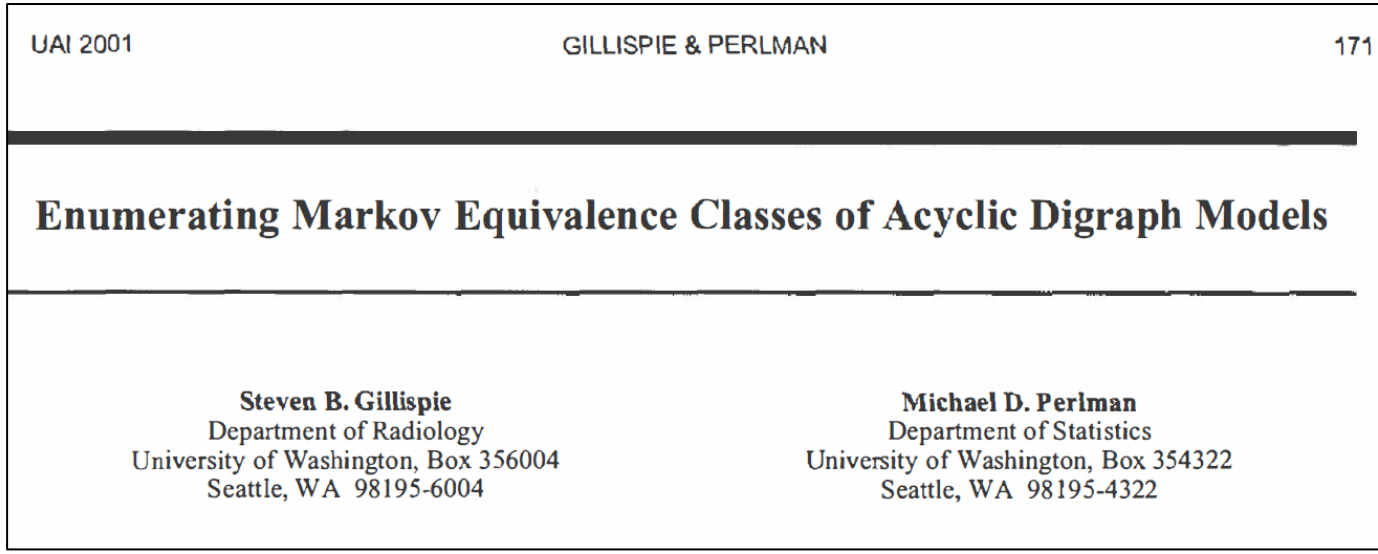
Learning Bayesian Networks is NP-complete

How many possible DAGs?

# variables	# Possible DAGs
2	3
3	25
4	543
5	29,281
10	$O(10^{18})$

$$G(n) = \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} 2^{k(n-k)} G(n-k)$$

[Gillespie and Perlman 2001, 2002]



UAI 2001 GILLISPIE & PERLMAN 171

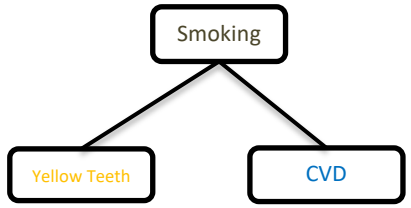
Enumerating Markov Equivalence Classes of Acyclic Digraph Models

Steven B. Gillispie
Department of Radiology
University of Washington, Box 356004
Seattle, WA 98195-6004

Michael D. Perlman
Department of Statistics
University of Washington, Box 354322
Seattle, WA 98195-4322

Reverse-engineering

Causal PDAG describing your variables



Identify all PDAGs that entail these (and only these) conditional independencies.

Independencies entailed by the MC

CVD $\perp\!\!\!\perp$ Yellow Teeth | Smoking

Under Faithfulness

CVD $\not\perp\!\!\!\perp$ Yellow Teeth | Ø

Smoking $\not\perp\!\!\!\perp$ Yellow Teeth | Ø
Smoking $\not\perp\!\!\!\perp$ Yellow Teeth | CVD

Smoking $\not\perp\!\!\!\perp$ CVD | Ø
Smoking $\not\perp\!\!\!\perp$ CVD | Yellow Teeth

Query the data to get the conditional (in) dependencies

Brute force: generate all possible PDAGs and test using d-separation.

Still NP-Complete

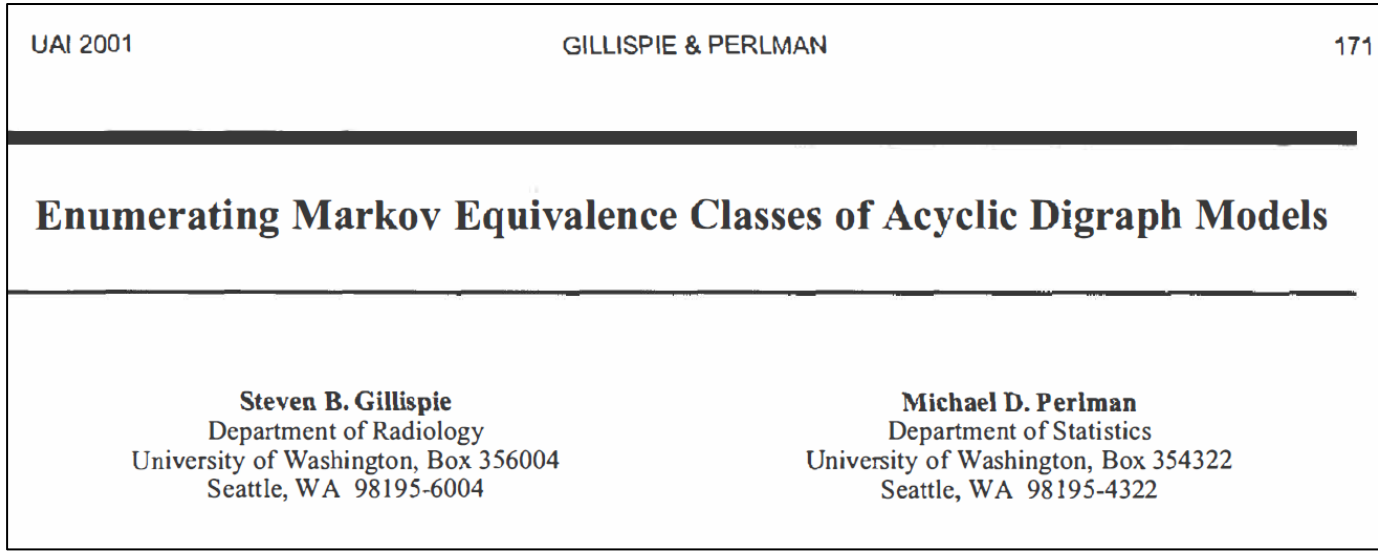
How many possible PDAGs?

# variables	# Possible DAGs	# Possible PDAGs
2	3	2
3	25	11
4	543	185
5	29,281	8,782
10	$O(10^{18})$	1,118,902,054,495,975,141

$$G(n) = \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} 2^{k(n-k)} G(n-k)$$

$$G'(n) \sim 0.267 \times G(n)$$

[Gillespie and Perlman 2001, 2002]



UAI 2001 GILLISPIE & PERLMAN 171

Enumerating Markov Equivalence Classes of Acyclic Digraph Models

Steven B. Gillispie
Department of Radiology
University of Washington, Box 356004
Seattle, WA 98195-6004

Michael D. Perlman
Department of Statistics
University of Washington, Box 354322
Seattle, WA 98195-4322

Learning BNs : Constraint-based approach

Good news:

You can identify all invariant characteristics of a Markov equivalence class of causal Bayesian networks that faithfully represent the conditional independencies in your data.

Bad news:

There are too many possible networks (DAGs/PDAGs).

There may not be a faithful representation.

You need:

A search strategy.

A test of conditional independence suitable for your data.

Reminder : Markov Condition + Faithfulness

$X \longleftarrow Y$ $\qquad$ $X \longrightarrow Y$

The edge is a d-connecting path that can not be broken given any other variables.

A useful implication of the Markov Condition

If X, Y are adjacent in the graph, then $\nexists$ $\boldsymbol{Z}$ s.t. $X \perp\!\!\!\perp Y \mid \boldsymbol{Z}$.

If $\exists$ $\boldsymbol{Z}$ s.t. $X \perp\!\!\!\perp Y \mid \boldsymbol{Z}$, X and Y are NOT adjacent in the graph.

You find a conditional independence $X \perp\!\!\!\perp Y \mid \mathbf{Z}$ **if and only if**

X and Y are not adjacent in the DAG.

Learning the skeleton of a BN

Search strategy:

Identify the skeleton of your PDAG:

Begin with the full graph.

For each pair of adjacent variables look for a set of observed variabl

If you find succeed, remove X-Y.

Until no more edges can be removed.

For each pair of variables:
2^{N-2} tests of independence.

Assume you have 20 variables. You may need to condition on 18 variables, which means 2^{18} possible configurations of the conditioning set.

You need a MANY samples
For finite sizes, very low power, tests that cannot be performed.

Learning the skeleton of a BN

Search strategy:

Identify the skeleton of your PDAG:

Begin with the full graph.

For each pair of adjacent variables look for a set of observed variables **Z** such that $X \perp\!\!\!\perp Y \mid \mathbf{Z}$.

If you find succeed, remove X-Y.

Until no more edges can be removed.

Theorem (Spirtes and Glymour, 1993): If S_G is the skeleton of the true DAG and $S_{G'}$ has a superset of edges, then the separating set of X , Y is a subset of the neighbors of X or Y in $S_{G'}$.

- You do not know the neighbors of each node.
- You begin with the full graph, so at each step of the algorithm you each variable is adjacent to a superset of its real neighbors.
- As you remove edges, the neighbor sets are reduced.
- You only have to check the adjacent nodes of X or Y at the current step of the algorithm.
- For a sparse graph, this really speeds up the skeleton search.
- Worst-case complexity is still exponential.

Learning the skeleton of a BN: PC algorithm

Search strategy:

- Identify the skeleton of your PDAG:
  - Begin with the full graph.
  - For k=0:number of variables-2 (or until k greater than the size of any neighborhood)
    - For each pair of adjacent variables X, Y,
      - Look within Adjacencies(X)\Y or Adjacencies(Y)\X for a set of k observed variables **Z** such that $X \perp\!\!\!\perp Y \mid \mathbf{Z}$.
      - If you succeed, remove X-Y

Essentially three loops: conditioning set size, pairs, conditioning sets

Learning the skeleton of a BN: PC algorithm

Search strategy:

Identify the skeleton of your PDAG:

Begin with the full graph.

For k=0:number of variables-2 (or until k greater than the size of any neighborhood)

For each pair of adjacent variables X, Y,

Look within Adjacencies(X)\Y or Adjacencies(Y)\X for a set of k observed variables **Z** such that $X \perp\!\!\!\perp Y \mid \mathbf{Z}$.

If you succeed, remove X-Y

Essentially three loops: conditioning set size, pairs, conditioning sets

How do you pick which edges/neighbors to try first?

- Naïve choice: lexicographic order
- Smart choice: (**HEURISTIC 3**, Causation, Prediction and Search, 1993):

You want to remove edges (X, Y) and you are looking for conditioning sets within Adjacencies(X)\Y.

-Start from the pair (X, Y) with the weakest pairwise association.

(weakest pairwise association more likely corresponds to non-adjacent variables)

-Start from the neighbor with the highest pairwise association with X (or Y).

(variables strongly associated with X are more probable to be neighbors/mediators on the path from X to Y)

Learning the skeleton of a BN

- Search strategy:
  - Identify the skeleton of your PDAG:
    - Begin with the full graph.
    - For k=0:number of variables -2
    - Using heuristic 3
      - For each pair of adjacent variables X, Y,
        - look within Adjacencies(X)\Y or Adjacencies(Y)\X for a set of k observed variables **Z** such that $X \perp\!\!\!\perp Y \mid \mathbf{Z}$.
        - If you succeed, remove X-Y.

You have identified the skeleton of your graph!

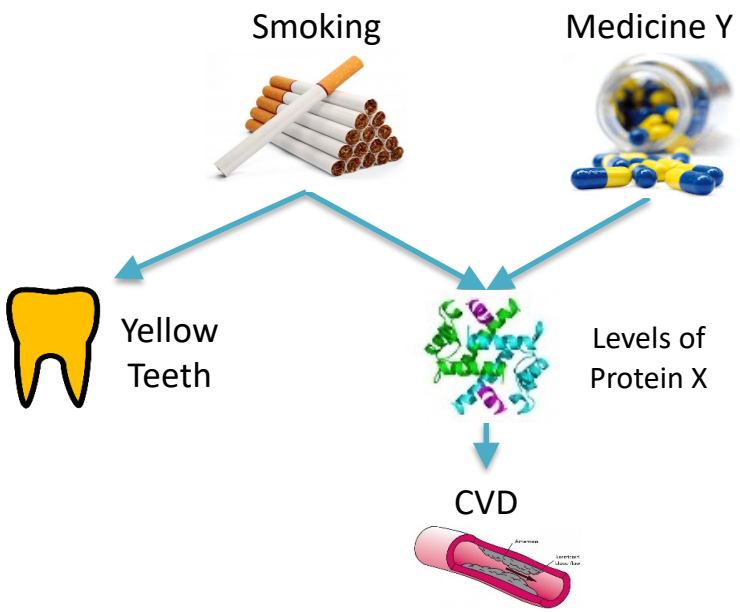
This is the skeleton identification step of the PC algorithm, introduced in 1993 by Peter Spirtes and Clark Glymour.

PC Algorithm – an example

Dataset measuring your variables.

Variables / Samples	Yellow Teeth	Smoking	CVD	Medicine Y	Levels of Protein X
1					
2					
3					
4					
5					
6					
7					
8					
9					
⋮					
999					
1000					

TRUE, **UNKNOWN** causal DAG

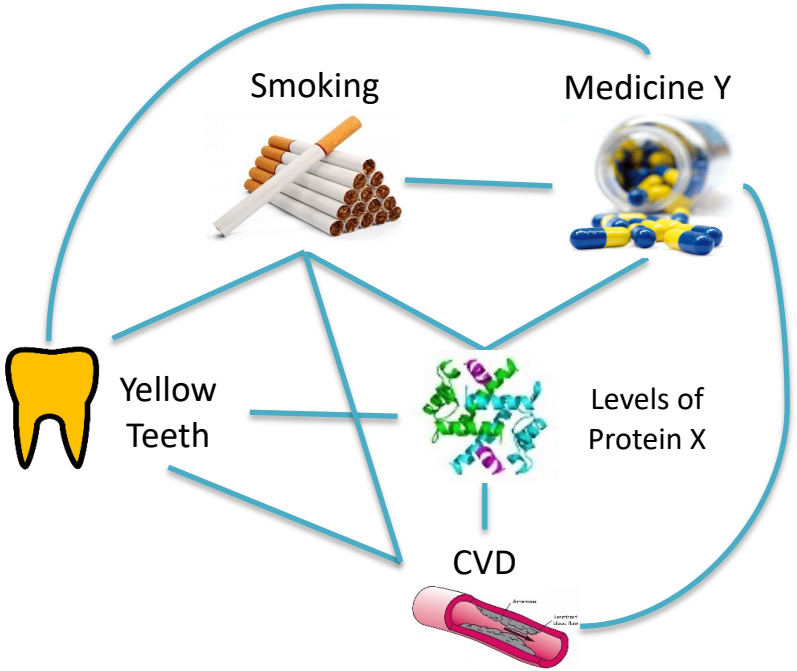


Let's see an example of the PC algorithm skeleton identification step.
Assuming:

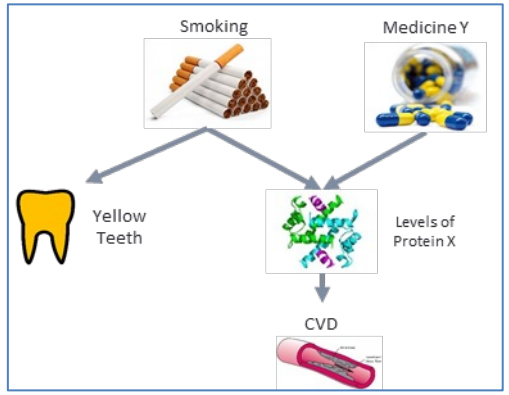
1. You have a data-set of measuring Yellow Teeth, Smoking, Medicine Y, Levels of Protein X and CVD in a sample of people.
2. MC and Faithfulness hold for your distribution and the causal DAG.
3. Your threshold for statistical significance is 0.05

PC Algorithm – an example

1. Begin with the full graph.

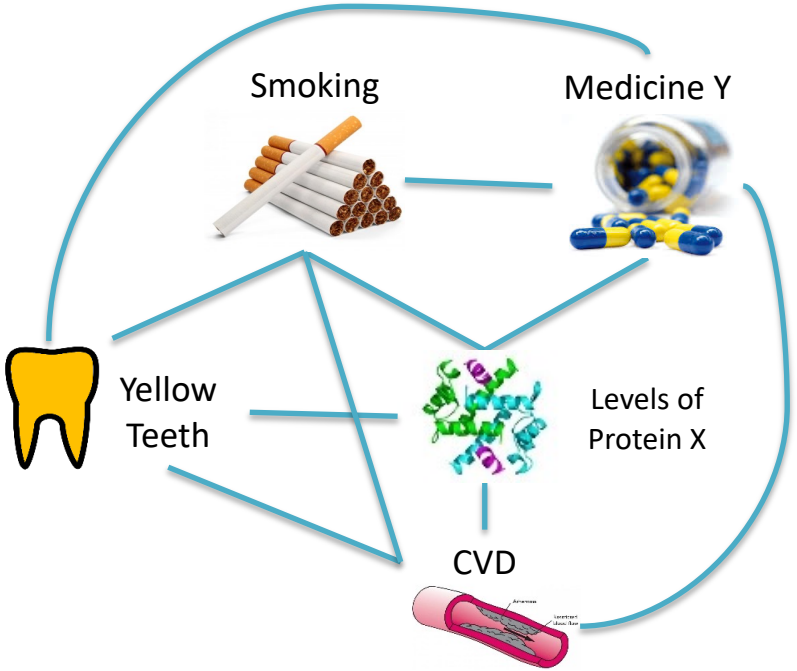


True (unknown) graph

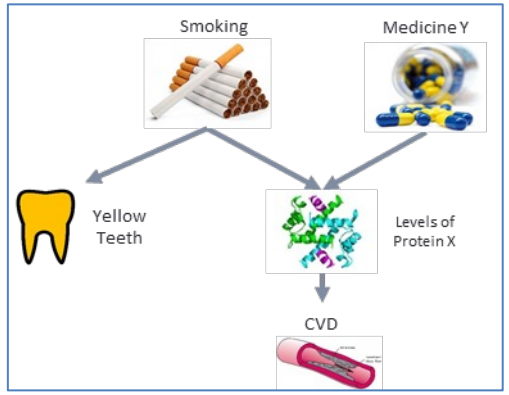


PC Algorithm – an example

2. k=0

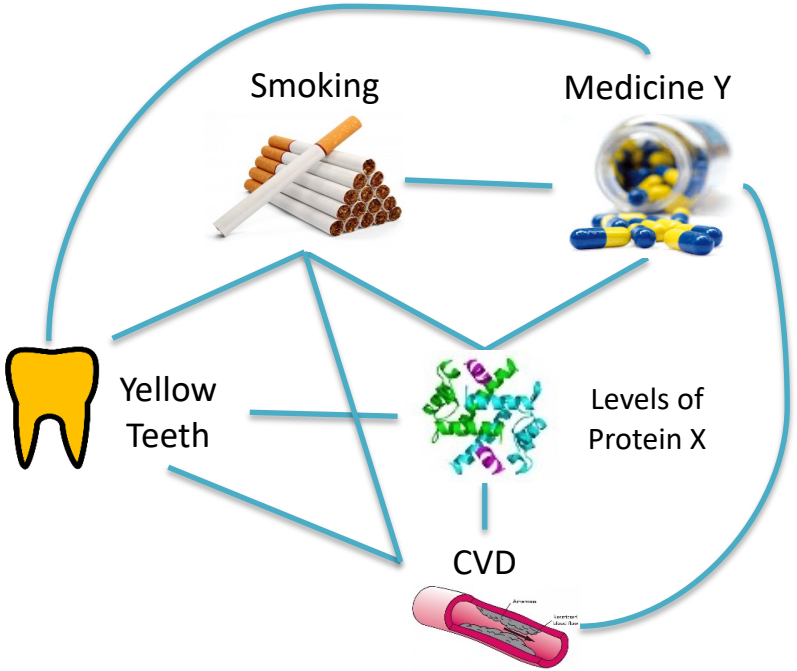


True (unknown) graph



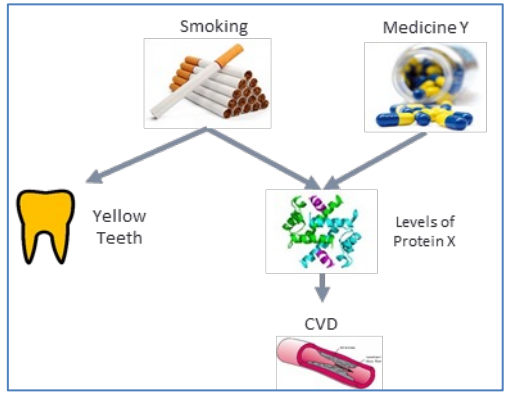
PC Algorithm – an example

2. k=0



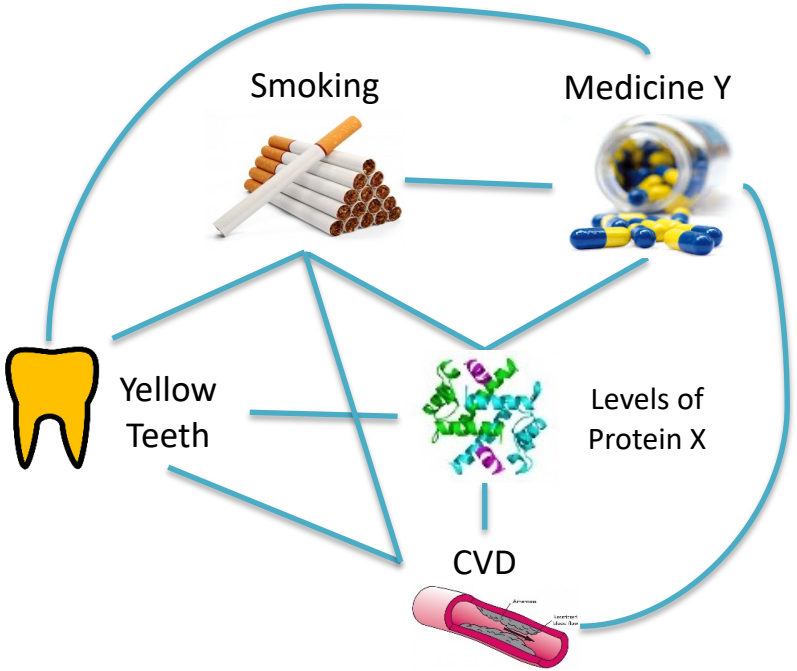
Tests attempted	p-value
Yellow Teeth, Smoking	0.00002

True (unknown) graph



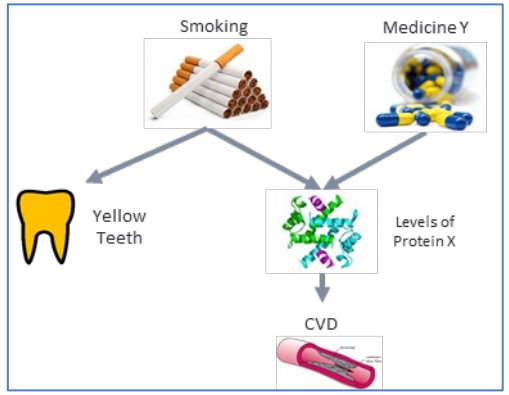
PC Algorithm – an example

2. k=0



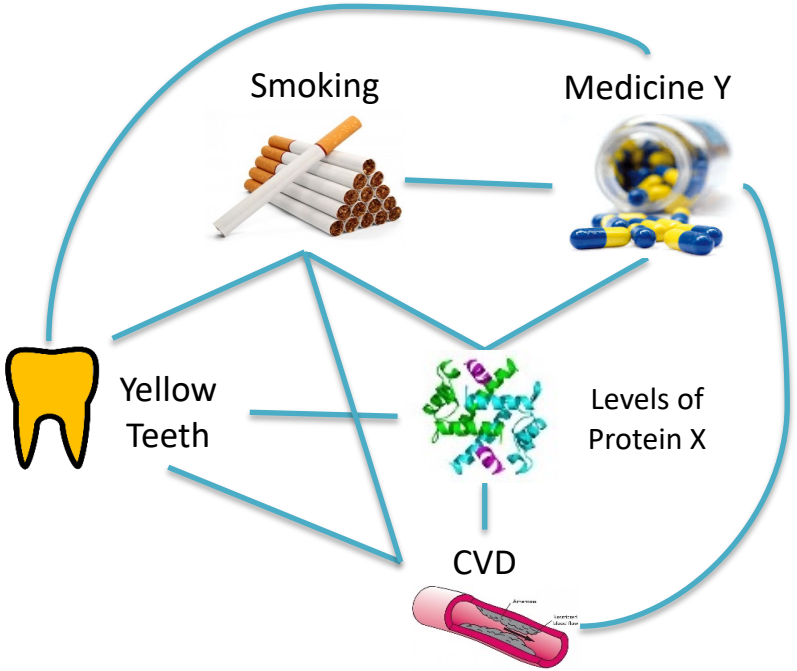
Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
Yellow Teeth, CVD	0.00384

True (unknown) graph



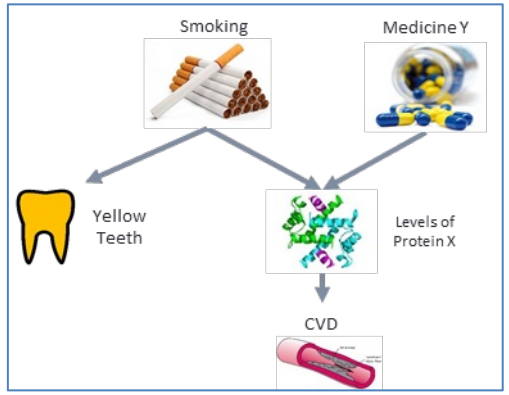
PC Algorithm – an example

2. k=0



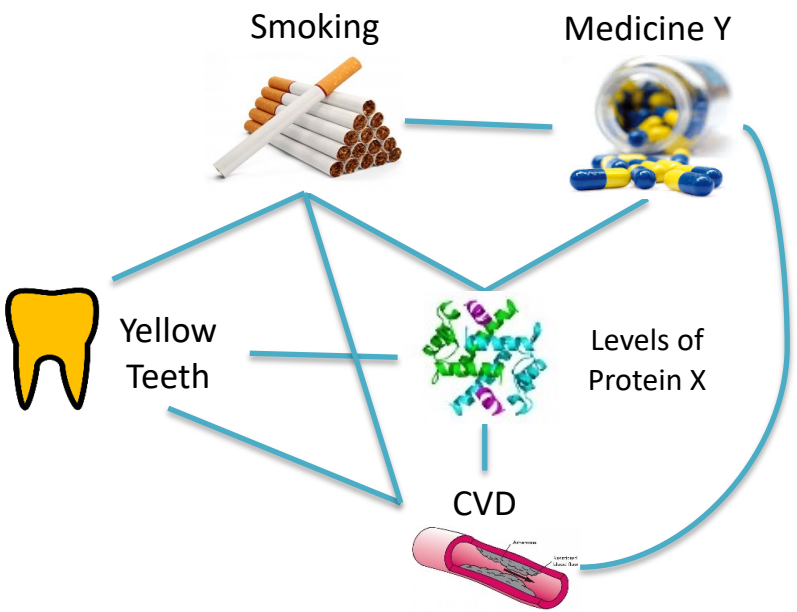
Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
Yellow Teeth, CVD	0.00384
Yellow Teeth, Medicine Y	0.54501

True (unknown) graph



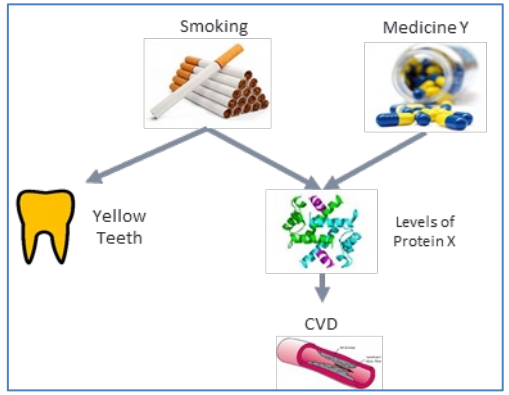
PC Algorithm – an example

2. k=0



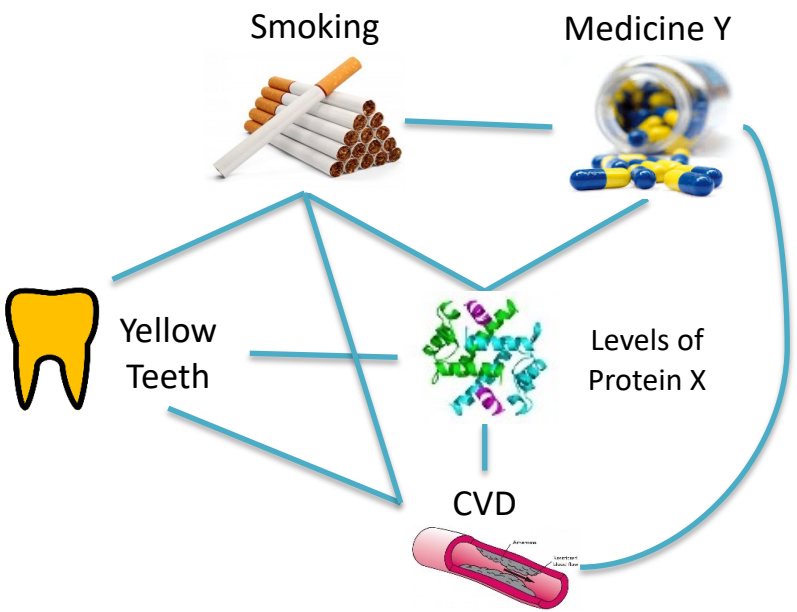
Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
Yellow Teeth, CVD	0.00384
Yellow Teeth, Medicine Y	0.54501
Yellow Teeth, Protein X	0.00056

True (unknown) graph



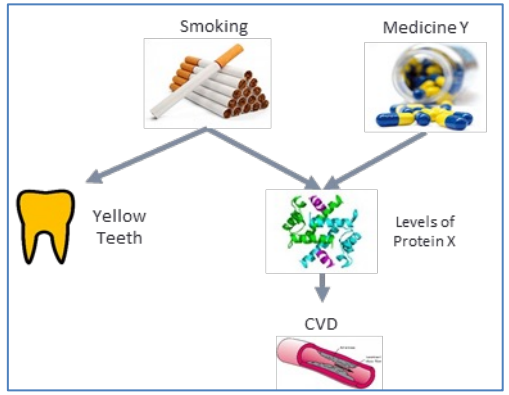
PC Algorithm – an example

2. k=0



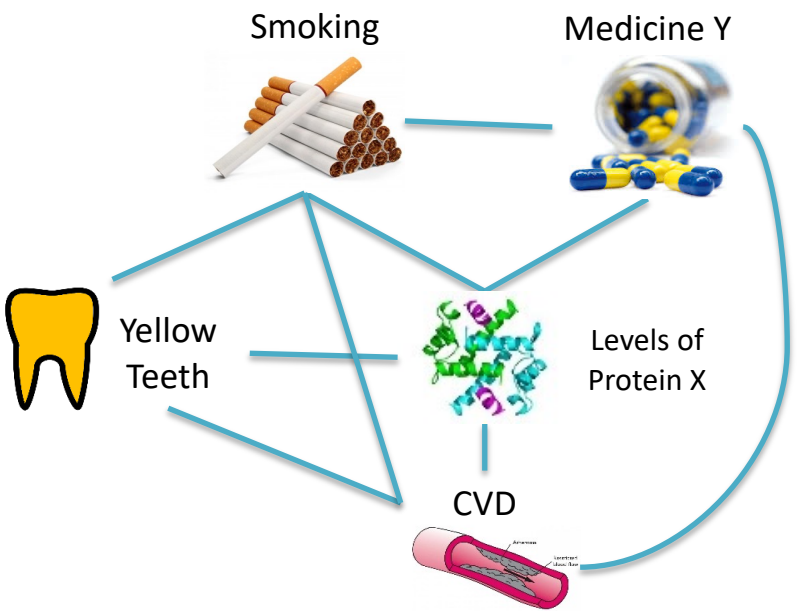
Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
Yellow Teeth, CVD	0.00384
Yellow Teeth, Medicine Y	0.54501
Yellow Teeth, Protein X	0.00056
Smoking, CVD	0.00015

True (unknown) graph



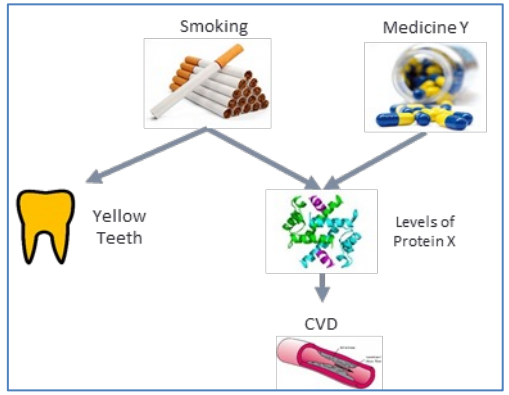
PC Algorithm – an example

2. k=0



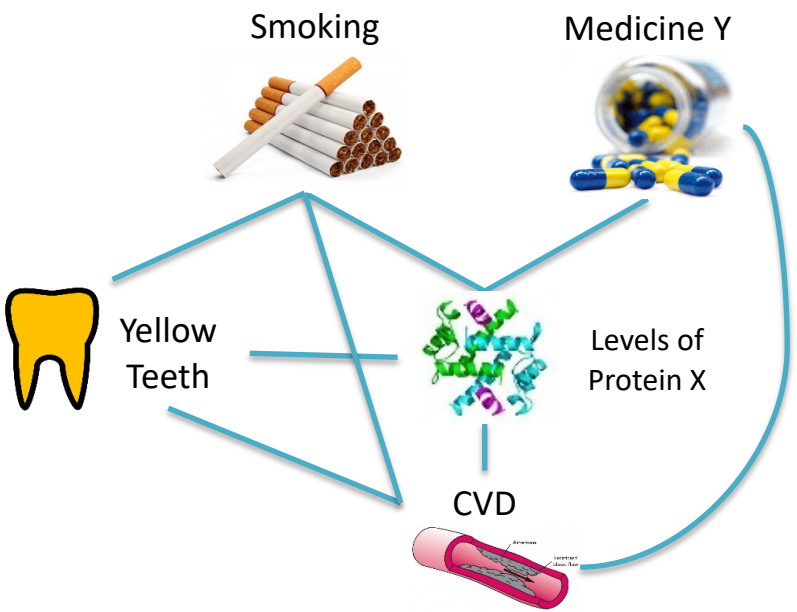
Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
Yellow Teeth, CVD	0.00384
Yellow Teeth, Medicine Y	0.54501
Yellow Teeth, Protein X	0.00056
Smoking, CVD	0.00015
Smoking, Medicine Y	0.36458

True (unknown) graph



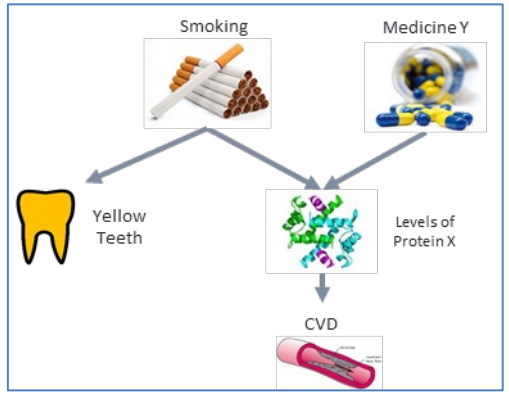
PC Algorithm – an example

2. k=0



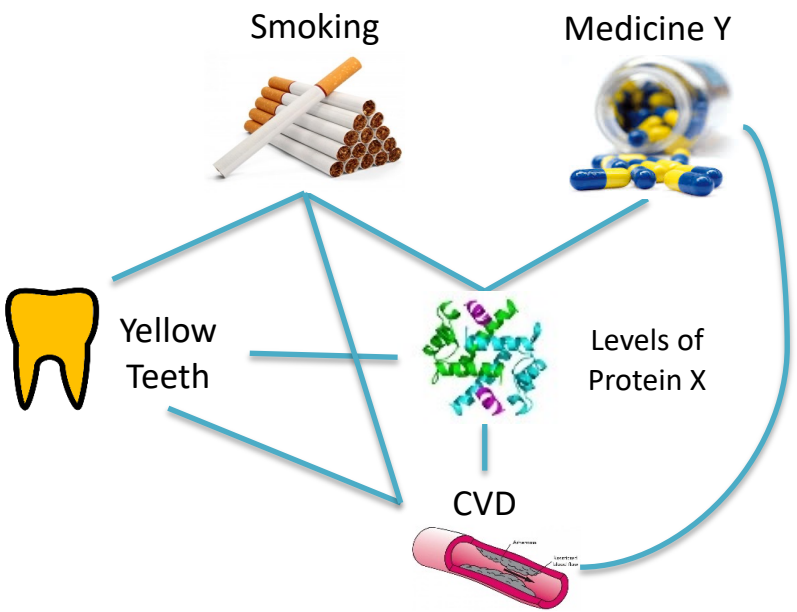
Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
Yellow Teeth, CVD	0.00384
Yellow Teeth, Medicine Y	0.54501
Yellow Teeth, Protein X	0.00056
Smoking, CVD	0.00015
Smoking, Medicine Y	0.36458
Smoking, Protein X	0.00003

True (unknown) graph



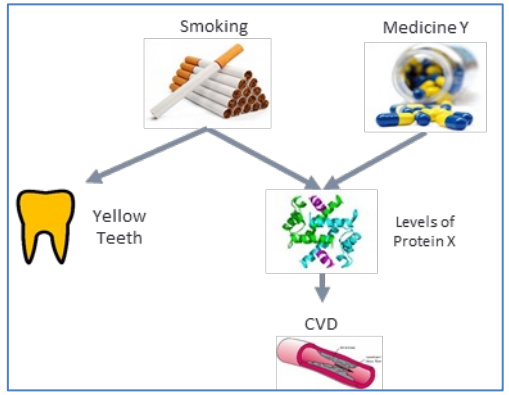
PC Algorithm – an example

2. k=0



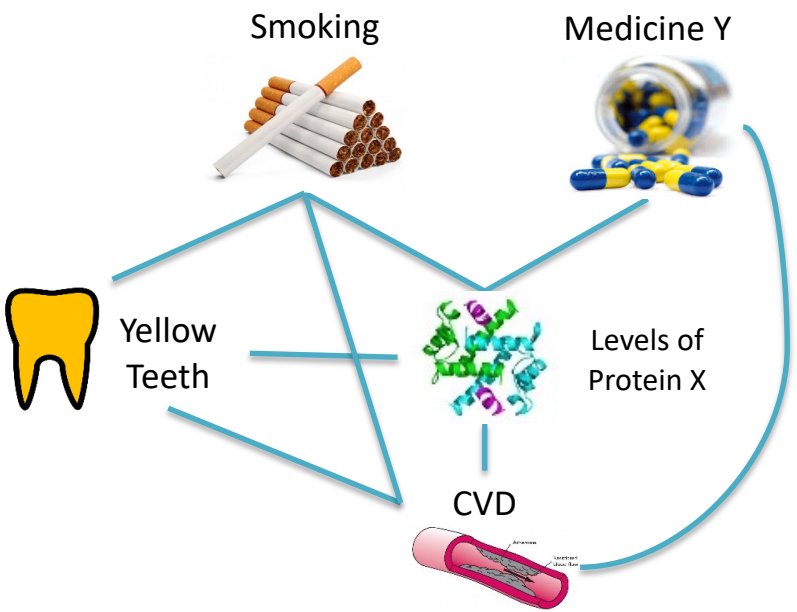
Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
Yellow Teeth, CVD	0.00384
Yellow Teeth, Medicine Y	0.54501
Yellow Teeth, Protein X	0.00056
Smoking, CVD	0.00015
Smoking, Medicine Y	0.36458
Smoking, Protein X	0.00003
CVD, Medicine Y	0.00012

True (unknown) graph



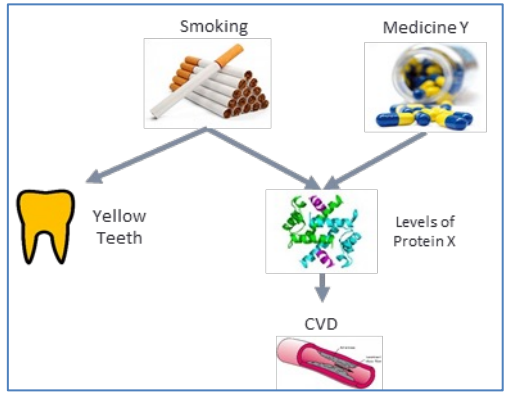
PC Algorithm – an example

2. k=0



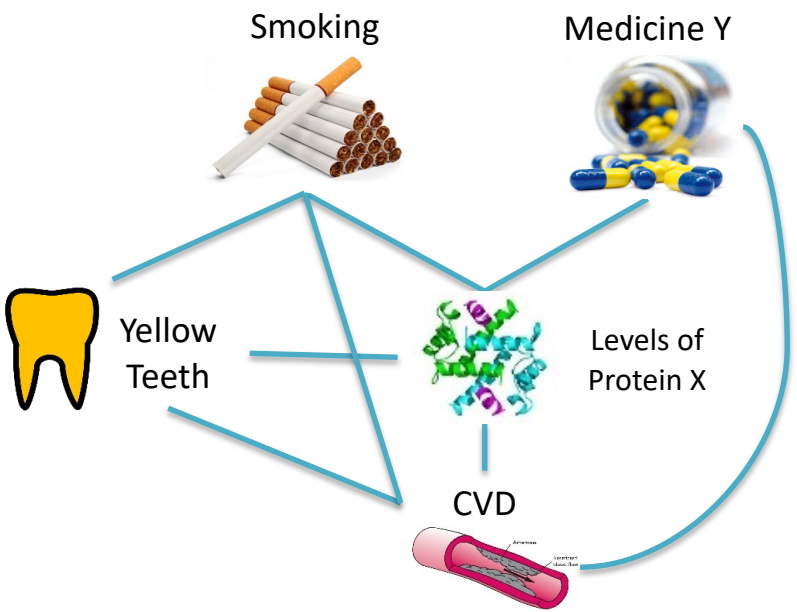
Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
Yellow Teeth, CVD	0.00384
Yellow Teeth, Medicine Y	0.54501
Yellow Teeth, Protein X	0.00056
Smoking, CVD	0.00015
Smoking, Medicine Y	0.36458
Smoking, Protein X	0.00003
CVD, Medicine Y	0.00012
CVD, Protein X	0.00024

True (unknown) graph



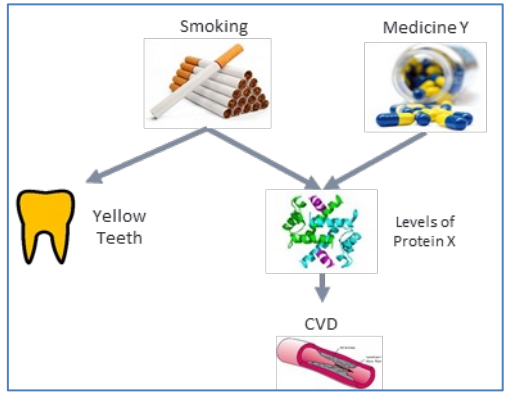
PC Algorithm – an example

2. k=0



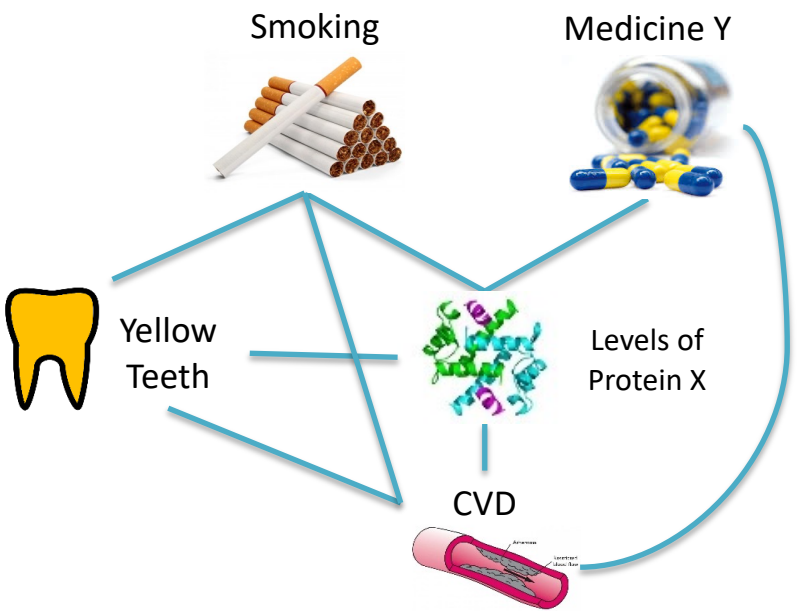
Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
Yellow Teeth, CVD	0.00384
Yellow Teeth, Medicine Y	0.54501
Yellow Teeth, Protein X	0.00056
Smoking, CVD	0.00015
Smoking, Medicine Y	0.36458
Smoking, Protein X	0.00003
CVD, Medicine Y	0.00012
CVD, Protein X	0.00024
Medicine Y, Protein X	0.00007

True (unknown) graph



PC Algorithm – an example

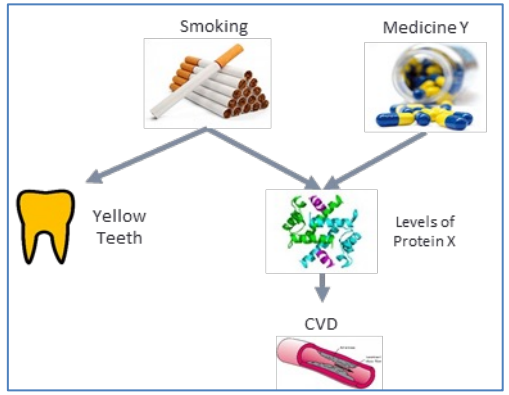
2. k=1



Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
Yellow Teeth, CVD	0.00384
Yellow Teeth, Medicine Y	0.54501
Yellow Teeth, Protein X	0.00056
Smoking, CVD	0.00015
Smoking, Medicine Y	0.36458
Smoking, Protein X	0.00003
CVD, Medicine Y	0.00012
CVD, Protein X	0.00024
Medicine Y, Protein X	0.00007

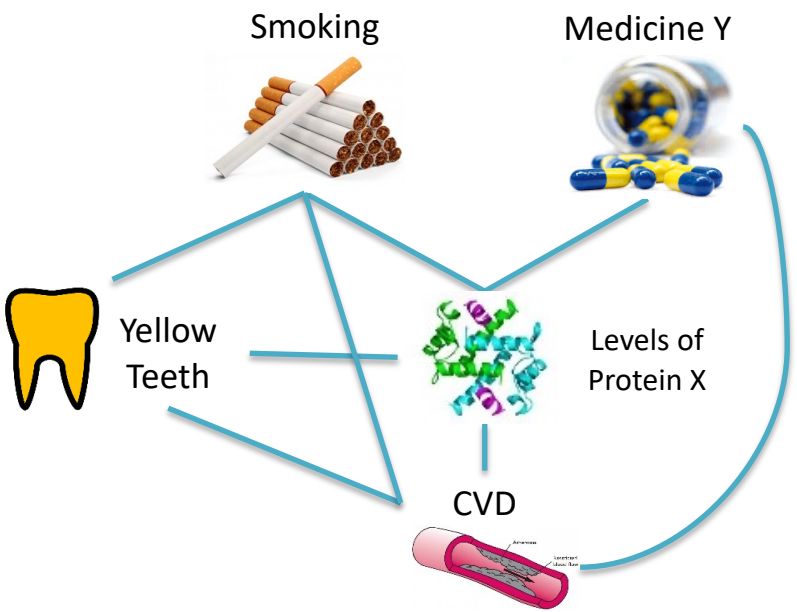
You want to identify the least correlated variables Since all variables are binary, you can check the p-values (in decreasing order)

True (unknown) graph



PC Algorithm – an example

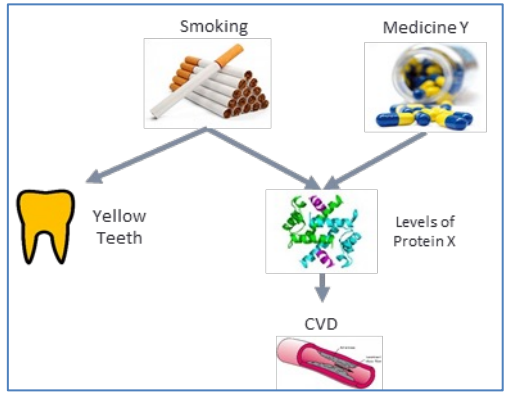
2. k=1



Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
Yellow Teeth, CVD	0.00384
Yellow Teeth, Medicine Y	0.54501
Yellow Teeth, Protein X	0.00056
Smoking, CVD	0.00015
Smoking, Medicine Y	0.36458
Smoking, Protein X	0.00003
CVD, Medicine Y	0.00012
CVD, Protein X	0.00024
Medicine Y, Protein X	0.00007

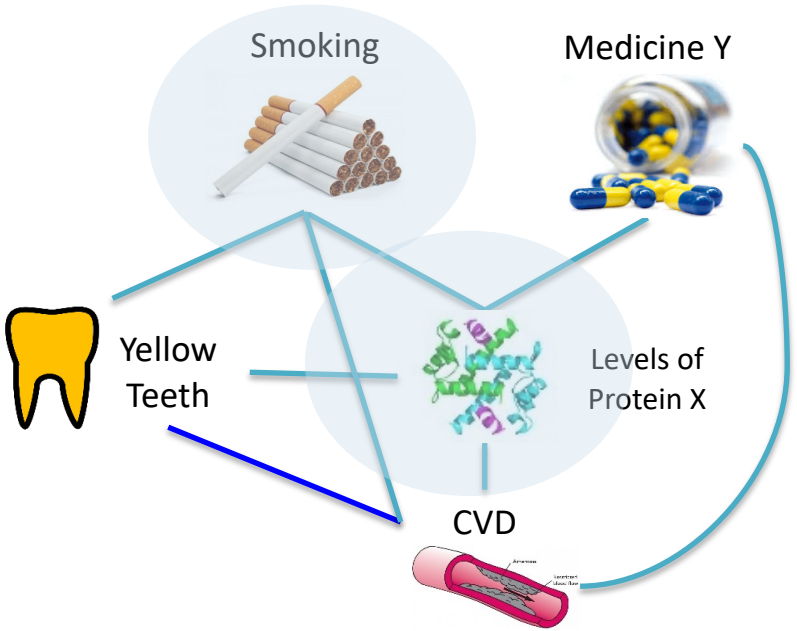
Yellow Teeth, CVD are the least associated variables

True (unknown) graph



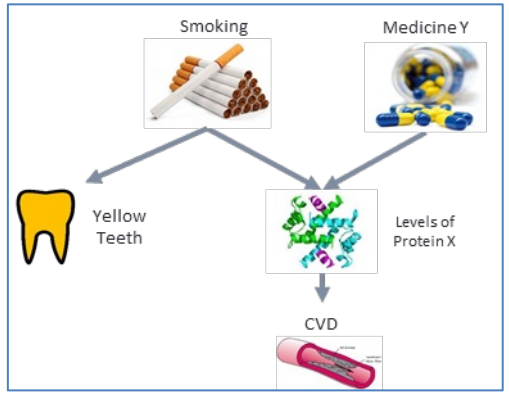
PC Algorithm – an example

2. k=1



Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
Yellow Teeth, CVD	0.00384
Yellow Teeth, Medicine Y	0.54501
Yellow Teeth, Protein X	0.00056
Smoking, CVD	0.00015
Smoking, Medicine Y	0.36458
Smoking, Protein X	0.00003
CVD, Medicine Y	0.00012
CVD, Protein X	0.00024
Medicine Y, Protein X	0.00007

True (unknown) graph

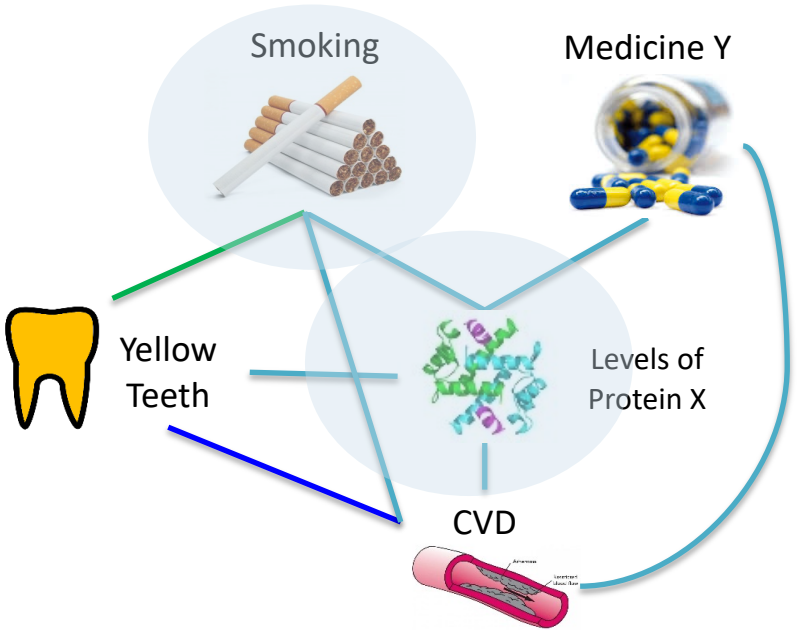


Yellow Teeth, CVD are the least associated variables

Adjacent(Yellow Teeth)\CVD = {Smoking, Protein X}

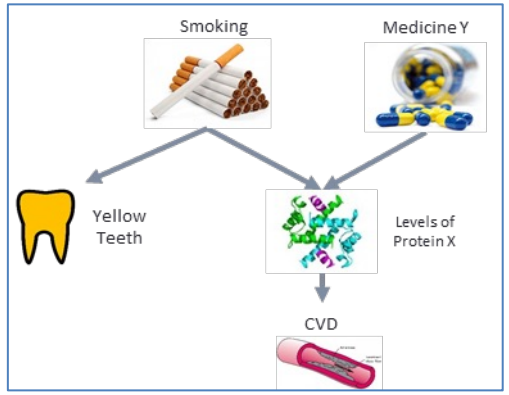
PC Algorithm – an example

2. k=1



Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
Yellow Teeth, CVD	0.00384
Yellow Teeth, Medicine Y	0.54501
Yellow Teeth, Protein X	0.00056
Smoking, CVD	0.00015
Smoking, Medicine Y	0.36458
Smoking, Protein X	0.00003
CVD, Medicine Y	0.00012
CVD, Protein X	0.00024
Medicine Y, Protein X	0.00007

True (unknown) graph



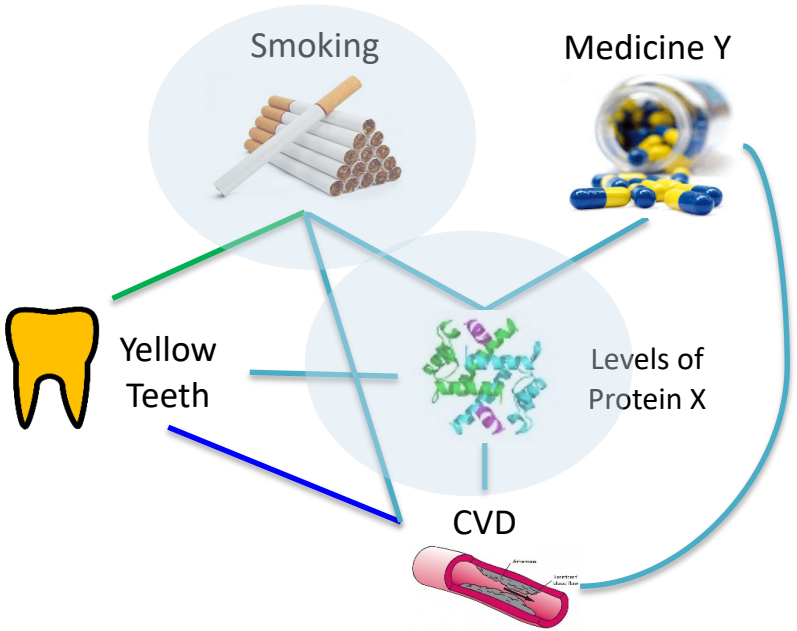
Yellow Teeth, CVD are the least associated variables

Adjacent(Yellow Teeth)\CVD = {Smoking, Protein X}

Yellow Teeth, Smoking are the most associated variables

PC Algorithm – an example

2. k=1



Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
Yellow Teeth, CVD	0.00384
Yellow Teeth, Medicine Y	0.54501
Yellow Teeth, Protein X	0.00056
Smoking, CVD	0.00015
Smoking, Medicine Y	0.36458
Smoking, Protein X	0.00003
CVD, Medicine Y	0.00012
CVD, Protein X	0.00024
Medicine Y, Protein X	0.00007

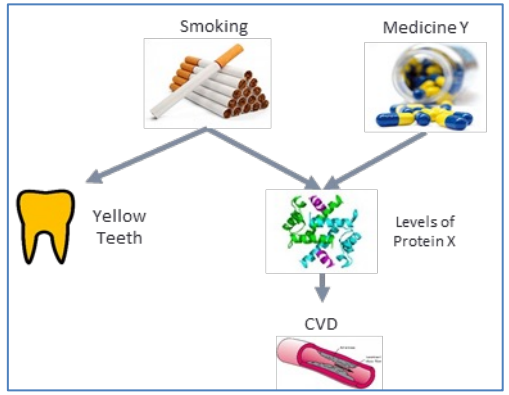
Tests attempted	p-value
Yellow Teeth, CVD| Smoking	0.78961

Yellow Teeth, CVD are the least associated variables

Adjacent(Yellow Teeth)\CVD = {Smoking, Protein X}

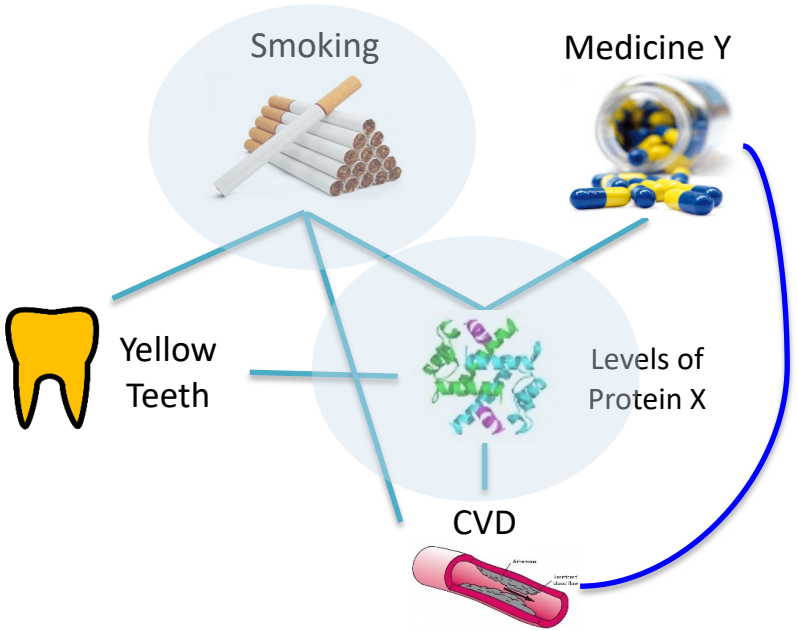
Yellow Teeth, Smoking are the most associated variables

True (unknown) graph



PC Algorithm – an example

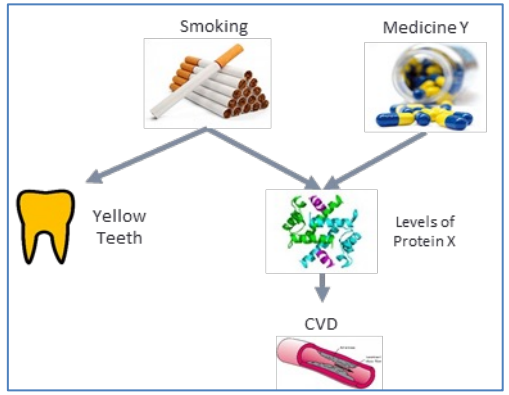
2. k=1



Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
~~Yellow Teeth, CVD~~	~~0.00384~~
Yellow Teeth, Medicine Y	0.54501
Yellow Teeth, Protein X	0.00056
Smoking, CVD	0.00035
Smoking, Medicine Y	0.36458
Smoking, Protein X	0.00003
CVD, Medicine Y	0.00062
CVD, Protein X	0.00014
Medicine Y, Protein X	0.00007

Tests attempted	p-value
Yellow Teeth, CVD| Smoking	0.78961

True (unknown) graph



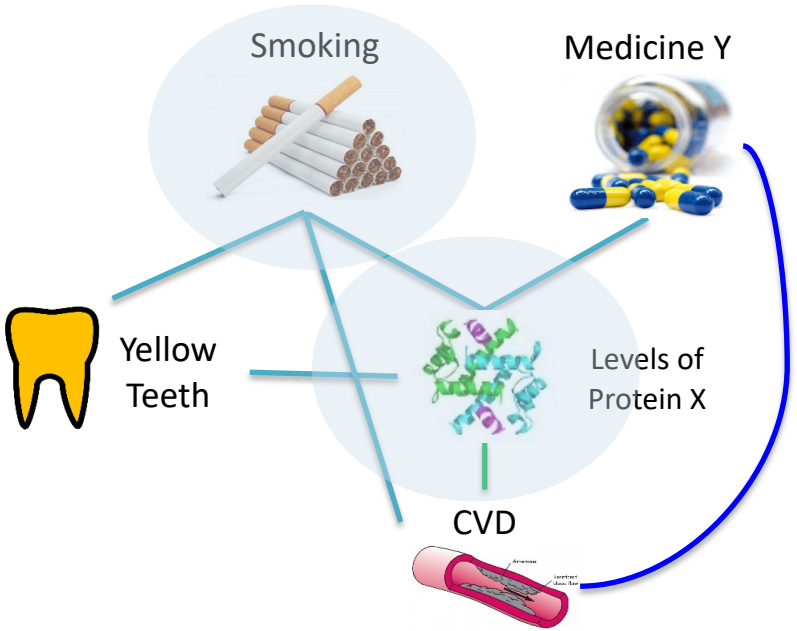
CVD, Medicine Y are the least associated variables

Adjacent(CVD)\Medicine Y = {Smoking, Protein X}

CVD, Protein X are the most associated variables

PC Algorithm – an example

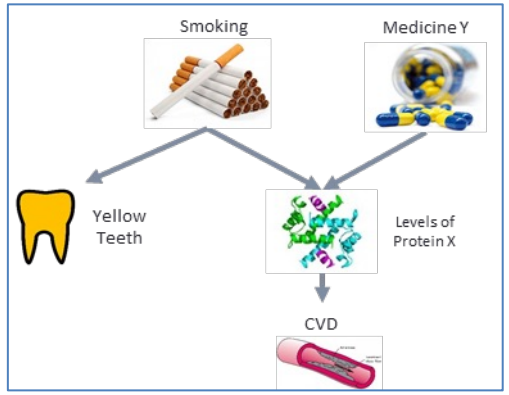
2. k=1



Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
~~Yellow Teeth, CVD~~	~~0.00384~~
Yellow Teeth, Medicine Y	0.54501
Yellow Teeth, Protein X	0.00056
Smoking, CVD	0.00035
Smoking, Medicine Y	0.36458
Smoking, Protein X	0.00003
CVD, Medicine Y	0.00062
CVD, Protein X	0.00014
Medicine Y, Protein X	0.00007

Tests attempted	p-value
Yellow Teeth, CVD| Smoking	0.78961
CVD, Medicine Y | Protein X	0.15092

True (unknown) graph



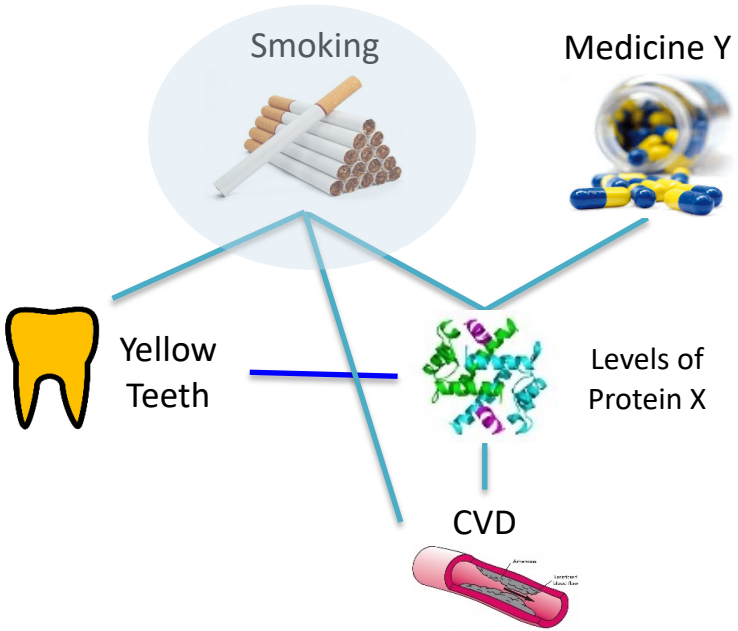
CVD, Medicine Y are the least associated variables

Adjacent(CVD)\Medicine Y = {Smoking, Protein X}

CVD, Protein X are the most associated variables

PC Algorithm – an example

2. k=1



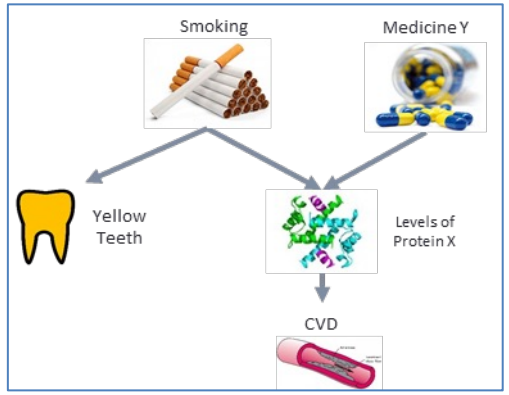
Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
~~Yellow Teeth, CVD~~	~~0.00384~~
Yellow Teeth, Medicine Y	0.54501
Yellow Teeth, Protein X	0.00056
Smoking, CVD	0.00035
Smoking, Medicine Y	0.36458
Smoking, Protein X	0.00003
~~CVD, Medicine Y~~	~~0.00062~~
CVD, Protein X	0.00014
Medicine Y, Protein X	0.00007

Tests attempted	p-value
Yellow Teeth, CVD| Smoking	0.78961
CVD, Medicine Y | Protein X	0.15092

Yellow Teeth, Protein X are the least associated variables

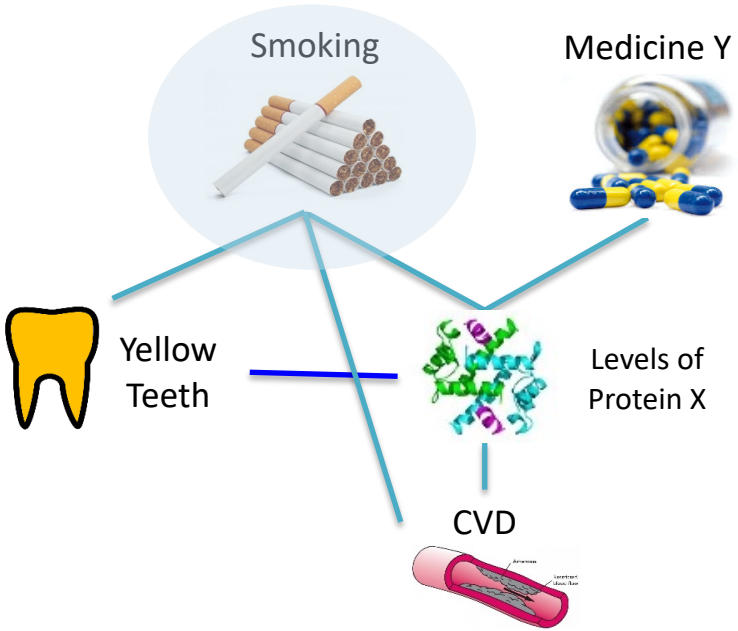
Adjacent(Yellow Teeth)\Protein X = {Smoking}

True (unknown) graph



PC Algorithm – an example

2. k=1



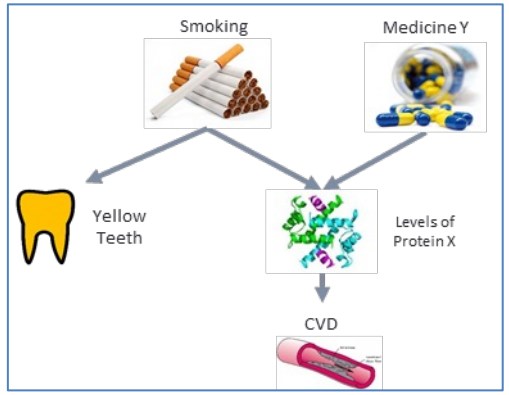
Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
~~Yellow Teeth, CVD~~	~~0.00384~~
Yellow Teeth, Medicine Y	0.54501
Yellow Teeth, Protein X	0.00056
Smoking, CVD	0.00035
Smoking, Medicine Y	0.36458
Smoking, Protein X	0.00003
~~CVD, Medicine Y~~	~~0.00062~~
CVD, Protein X	0.00014
Medicine Y, Protein X	0.00007

Tests attempted	p-value
Yellow Teeth, CVD| Smoking	0.78961
CVD, Medicine Y | Protein X	0.15092
Yellow Teeth, Protein X | Smoking	0.23567

Yellow Teeth, Protein X are the least associated variables

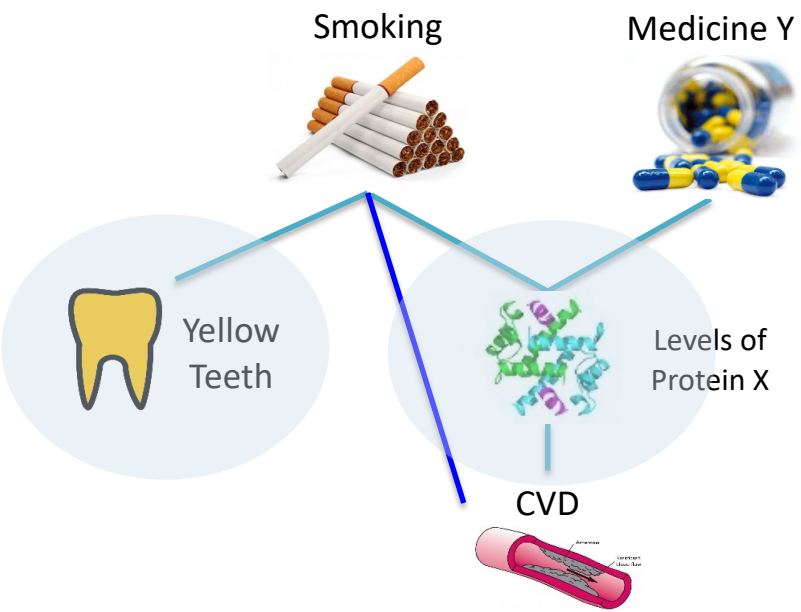
Adjacent(Yellow Teeth)\Protein X = {Smoking}

True (unknown) graph



PC Algorithm – an example

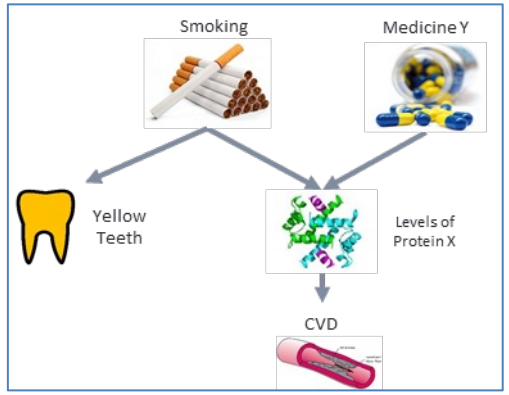
2. k=1



Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
~~Yellow Teeth, CVD~~	~~0.00384~~
Yellow Teeth, Medicine Y	0.54501
~~Yellow Teeth, Protein X~~	~~0.00056~~
Smoking, CVD	0.00035
Smoking, Medicine Y	0.36458
Smoking, Protein X	0.00003
~~CVD, Medicine Y~~	~~0.00062~~
CVD, Protein X	0.00014
Medicine Y, Protein X	0.00007

Tests attempted	p-value
Yellow Teeth, CVD| Smoking	0.78961
CVD, Medicine Y | Protein X	0.15092
Yellow Teeth, Protein X | Smoking	0.23567

True (unknown) graph



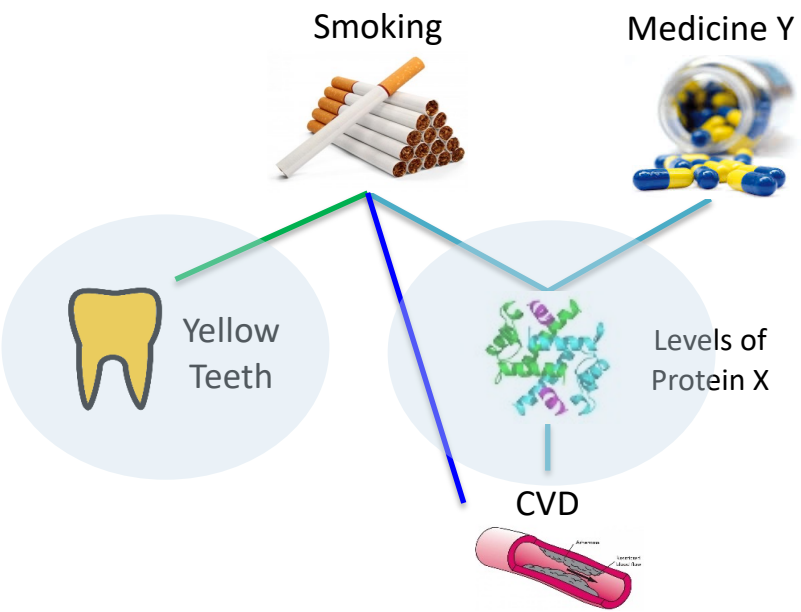
Smoking, CVD are the least associated variables

Adjacent(Smoking)\CVD= {Yellow Teeth, Protein X}

Smoking, Yellow Teeth are the most associated variables

PC Algorithm – an example

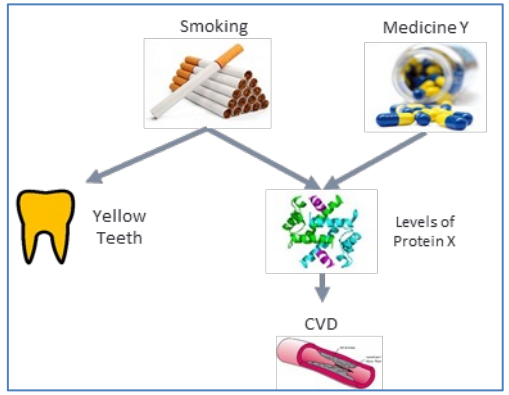
2. k=1



Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
~~Yellow Teeth, CVD~~	~~0.00384~~
Yellow Teeth, Medicine Y	0.54501
~~Yellow Teeth, Protein X~~	~~0.00056~~
Smoking, CVD	0.00035
Smoking, Medicine Y	0.36458
Smoking, Protein X	0.00003
~~CVD, Medicine Y~~	~~0.00062~~
CVD, Protein X	0.00014
Medicine Y, Protein X	0.00007

Tests attempted	p-value
Yellow Teeth, CVD| Smoking	0.78961
CVD, Medicine Y | Protein X	0.15092
Yellow Teeth, Protein X | Smoking	0.23567
Smoking, CVD |Yellow Teeth	0.00345

True (unknown) graph



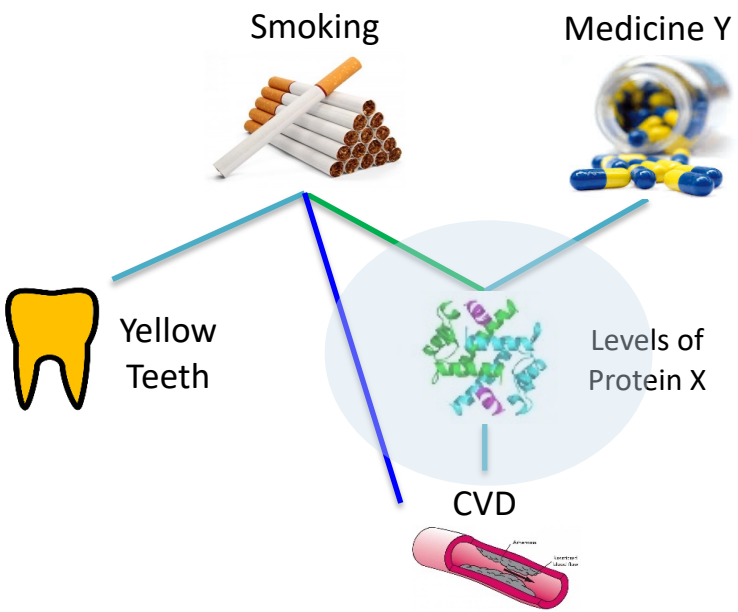
Smoking, CVD are the least associated variables

Adjacent(Smoking)\CVD= {Yellow Teeth, Protein X}

Smoking, Yellow Teeth are the most associated variables

PC Algorithm – an example

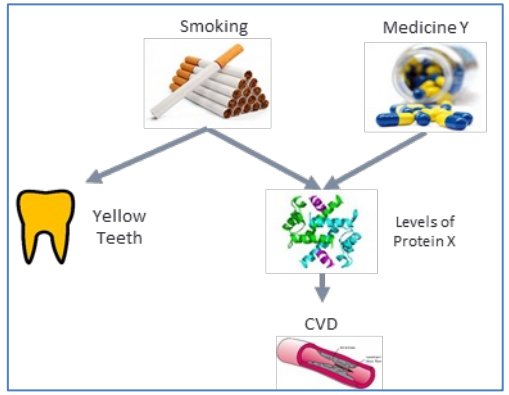
2. k=1



Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
~~Yellow Teeth, CVD~~	~~0.00384~~
Yellow Teeth, Medicine Y	0.54501
~~Yellow Teeth, Protein X~~	~~0.00056~~
Smoking, CVD	0.00035
Smoking, Medicine Y	0.36458
Smoking, Protein X	0.00003
~~CVD, Medicine Y~~	~~0.00062~~
CVD, Protein X	0.00014
Medicine Y, Protein X	0.00007

Tests attempted	p-value
Yellow Teeth, CVD| Smoking	0.78961
CVD, Medicine Y | Protein X	0.15092
Yellow Teeth, Protein X | Smoking	0.23567
Smoking, CVD |Yellow Teeth	0.00345
Smoking, CVD | Protein X	0.12365

True (unknown) graph



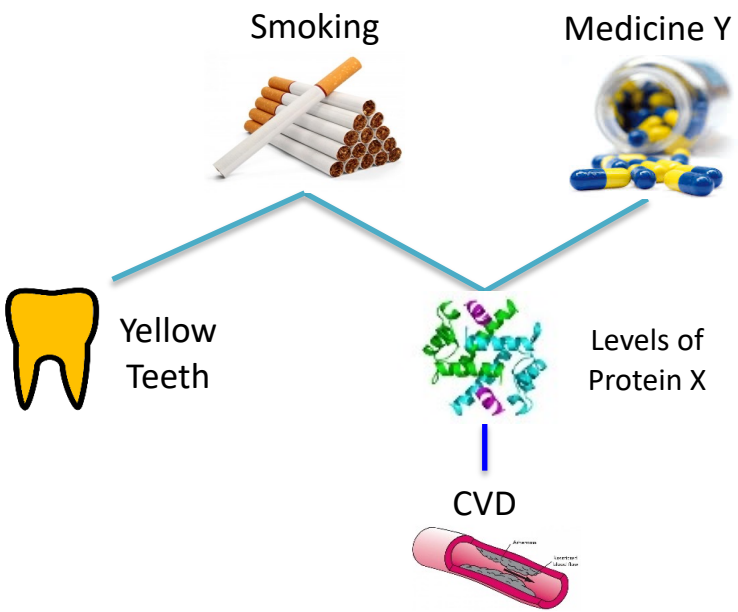
Smoking, CVD are the least associated variables

Adjacent(Smoking)\CVD= {Yellow Teeth, Protein X}

Smoking, Protein X are the next most associated variables

PC Algorithm – an example

2. k=1



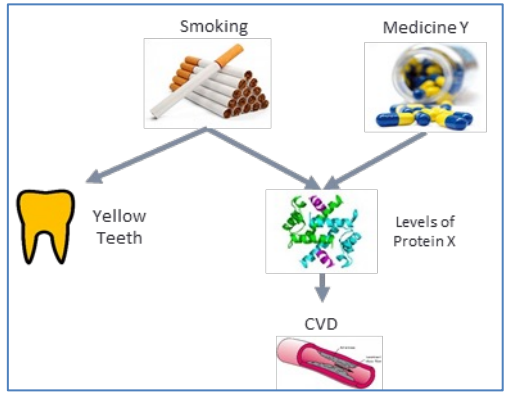
Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
~~Yellow Teeth, CVD~~	~~0.00384~~
Yellow Teeth, Medicine Y	0.54501
~~Yellow Teeth, Protein X~~	~~0.00056~~
~~Smoking, CVD~~	~~0.00035~~
Smoking, Medicine Y	0.36458
Smoking, Protein X	0.00003
~~CVD, Medicine Y~~	~~0.00062~~
CVD, Protein X	0.00014
Medicine Y, Protein X	0.00007

Tests attempted	p-value
Yellow Teeth, CVD| Smoking	0.78961
CVD, Medicine Y | Protein X	0.15092
Yellow Teeth, Protein X | Smoking	0.23567
Smoking, CVD |Yellow Teeth	0.00345
Smoking, CVD | Protein X	0.12365

CVD, Protein X are the least associated variables

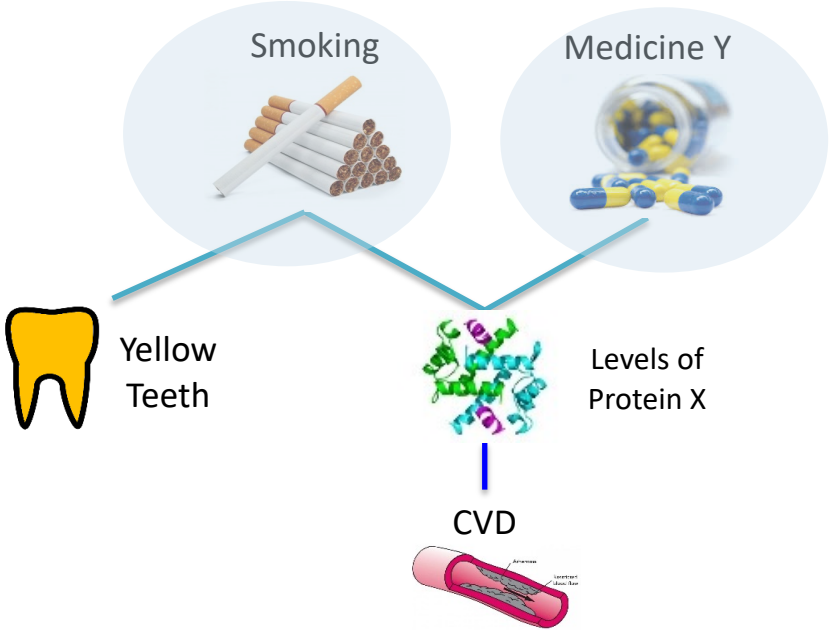
Adjacent(CVD)\Protein X = {}

True (unknown) graph



PC Algorithm – an example

2. k=1



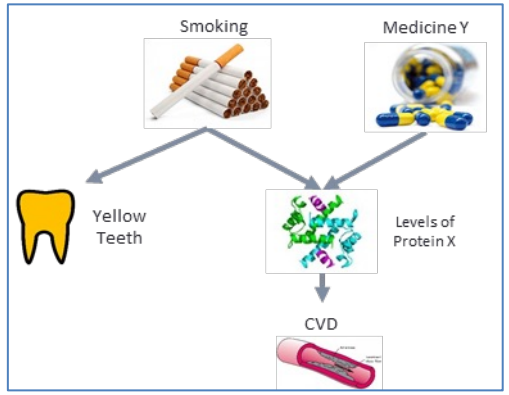
Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
~~Yellow Teeth, CVD~~	~~0.00384~~
Yellow Teeth, Medicine Y	0.54501
~~Yellow Teeth, Protein X~~	~~0.00056~~
~~Smoking, CVD~~	~~0.00035~~
Smoking, Medicine Y	0.36458
Smoking, Protein X	0.00003
~~CVD, Medicine Y~~	~~0.00062~~
CVD, Protein X	0.00014
Medicine Y, Protein X	0.00007

Tests attempted	p-value
Yellow Teeth, CVD| Smoking	0.78961
CVD, Medicine Y | Protein X	0.15092
Yellow Teeth, Protein X | Smoking	0.23567
Smoking, CVD |Yellow Teeth	0.00345
Smoking, CVD | Protein X	0.12365

CVD, Protein X are the least associated variables

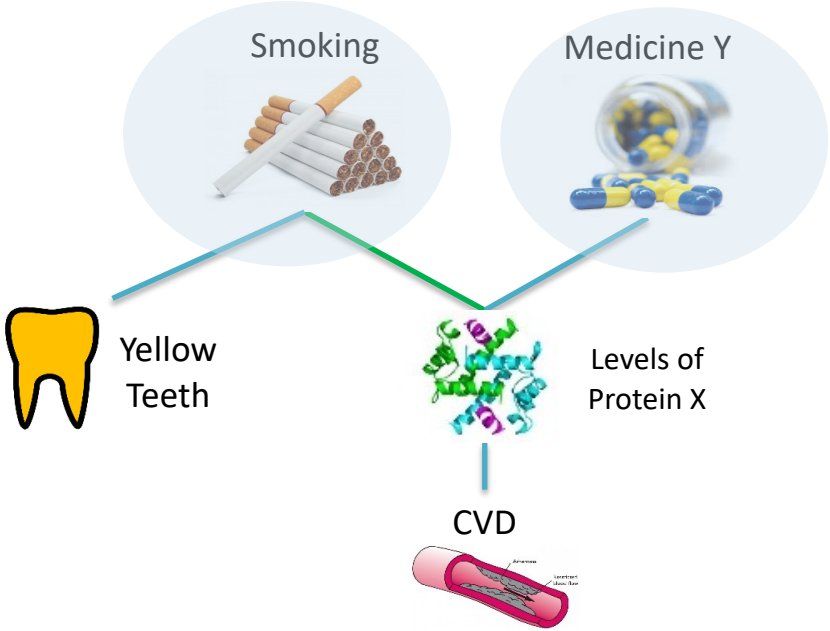
Adjacent(Protein X)\CVD= {Smoking, Medicine Y}

True (unknown) graph



PC Algorithm – an example

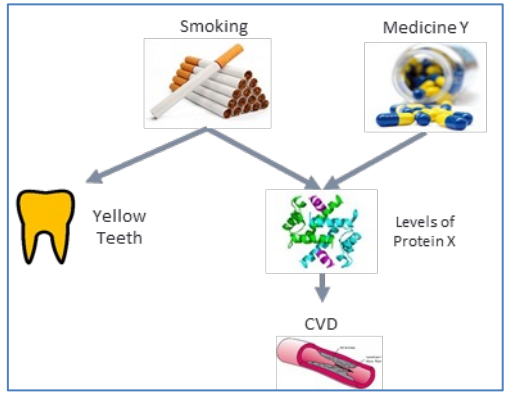
2. k=1



Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
~~Yellow Teeth, CVD~~	~~0.00384~~
Yellow Teeth, Medicine Y	0.54501
~~Yellow Teeth, Protein X~~	~~0.00056~~
~~Smoking, CVD~~	~~0.00035~~
Smoking, Medicine Y	0.36458
Smoking, Protein X	0.00003
~~CVD, Medicine Y~~	~~0.00062~~
CVD, Protein X	0.00014
Medicine Y, Protein X	0.00007

Tests attempted	p-value
Yellow Teeth, CVD| Smoking	0.78961
CVD, Medicine Y | Protein X	0.15092
Yellow Teeth, Protein X | Smoking	0.23567
Smoking, CVD |Yellow Teeth	0.00345
Smoking, CVD | Protein X	0.12365
CVD, Protein X | Smoking	0.00045

True (unknown) graph



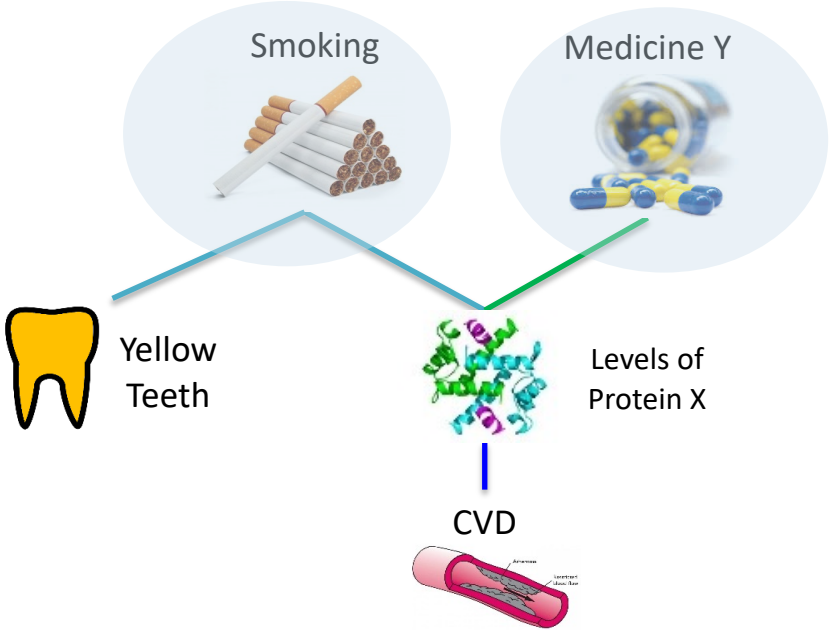
CVD, Protein X are the least associated variables

Adjacent(Protein X)\CVD= {Smoking, Medicine Y}

Protein X, Smoking are the most associated variables

PC Algorithm – an example

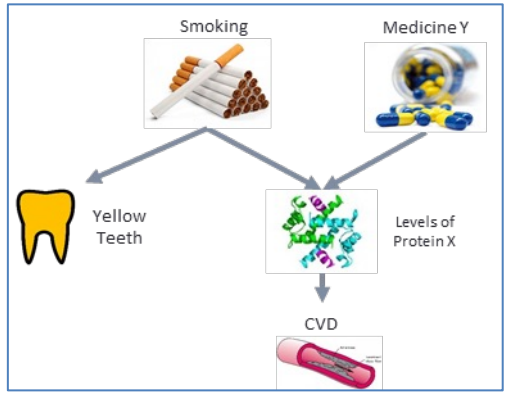
2. k=1



Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
~~Yellow Teeth, CVD~~	~~0.00384~~
Yellow Teeth, Medicine Y	0.54501
~~Yellow Teeth, Protein X~~	~~0.00056~~
~~Smoking, CVD~~	~~0.00035~~
Smoking, Medicine Y	0.36458
Smoking, Protein X	0.00003
~~CVD, Medicine Y~~	~~0.00062~~
~~CVD, Protein X~~	~~0.00014~~
Medicine Y, Protein X	0.00007

Tests attempted	p-value
Yellow Teeth, CVD| Smoking	0.78961
CVD, Medicine Y | Protein X	0.15092
Yellow Teeth, Protein X | Smoking	0.23567
Smoking, CVD |Yellow Teeth	0.00345
Smoking, CVD | Protein X	0.12365
CVD, Protein X | Smoking	0.00045
CVD, Protein X | Medicine Y	0.00389

True (unknown) graph



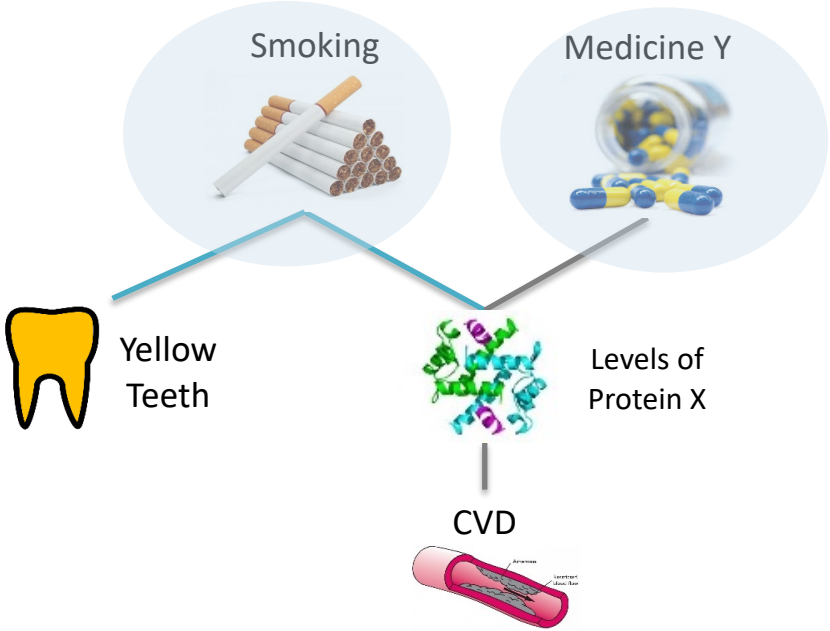
CVD, Protein X are the least associated variables

Adjacent(Protein X)\CVD= {Smoking, Medicine Y}

Protein X, Medicine Y are the next most associated variables

PC Algorithm – an example

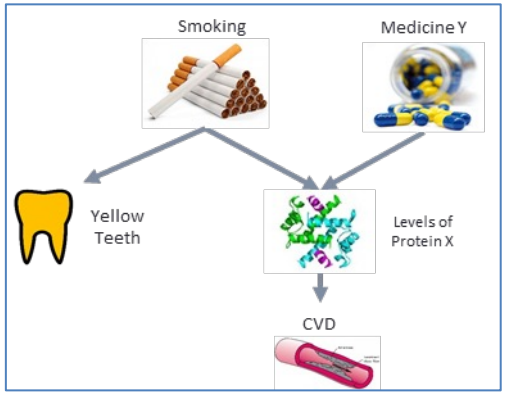
2. k=1



Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
~~Yellow Teeth, CVD~~	~~0.00384~~
Yellow Teeth, Medicine Y	0.54501
~~Yellow Teeth, Protein X~~	~~0.00056~~
~~Smoking, CVD~~	~~0.00035~~
Smoking, Medicine Y	0.36458
Smoking, Protein X	0.00003
~~CVD, Medicine Y~~	~~0.00062~~
~~CVD, Protein X~~	~~0.00014~~
Medicine Y, Protein X	0.00007

Tests attempted	p-value
Yellow Teeth, CVD| Smoking	0.78961
CVD, Medicine Y | Protein X	0.15092
Yellow Teeth, Protein X | Smoking	0.23567
Smoking, CVD |Yellow Teeth	0.00345
Smoking, CVD | Protein X	0.12365
CVD, Protein X | Smoking	0.00045
CVD, Protein X | Medicine Y	0.00389
Medicine Y, Protein X|CVD	0.00972
Smoking, Protein X | Yellow Teeth	0.00126
Smoking, Protein X|CVD	0.00438
Yellow Teeth, Smoking|Protein X	0.00072

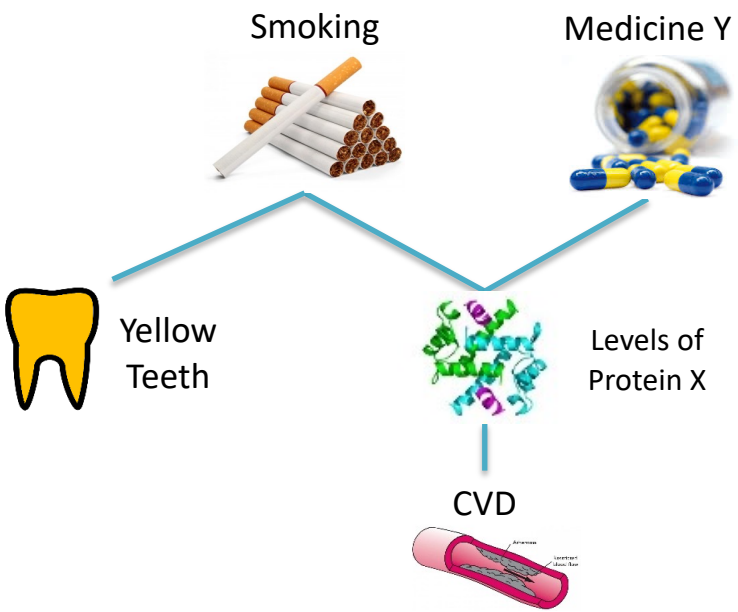
True (unknown) graph



...

PC Algorithm – an example

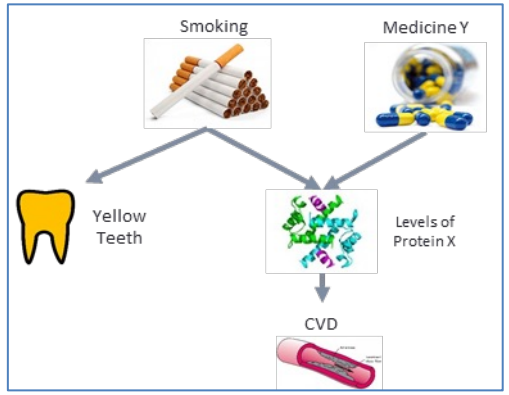
2. k=2



Tests attempted	p-value
~~Yellow Teeth, Smoking~~	~~0.00002~~
~~Yellow Teeth, CVD~~	~~0.00384~~
Yellow Teeth, Medicine Y	0.54501
~~Yellow Teeth, Protein X~~	~~0.00056~~
~~Smoking, CVD~~	~~0.00035~~
Smoking, Medicine Y	0.36458
~~Smoking, Protein X~~	~~0.00003~~
~~CVD, Medicine Y~~	~~0.00062~~
~~CVD, Protein X~~	~~0.00014~~
~~Medicine Y, Protein X~~	~~0.00007~~

Tests attempted	p-value
Yellow Teeth, CVD| Smoking	0.78961
CVD, Medicine Y | Protein X	0.15092
Yellow Teeth, Protein X | Smoking	0.23567
Smoking, CVD |Yellow Teeth	0.00345
Smoking, CVD | Protein X	0.12365
CVD, Protein X | Smoking	0.00045
CVD, Protein X | Medicine Y	0.00389
Medicine Y, Protein X|CVD	0.00972
Smoking, Protein X | Yellow Teeth	0.00126
Smoking, Protein X|CVD	0.00438
Yellow Teeth, Smoking|Protein X	0.00072

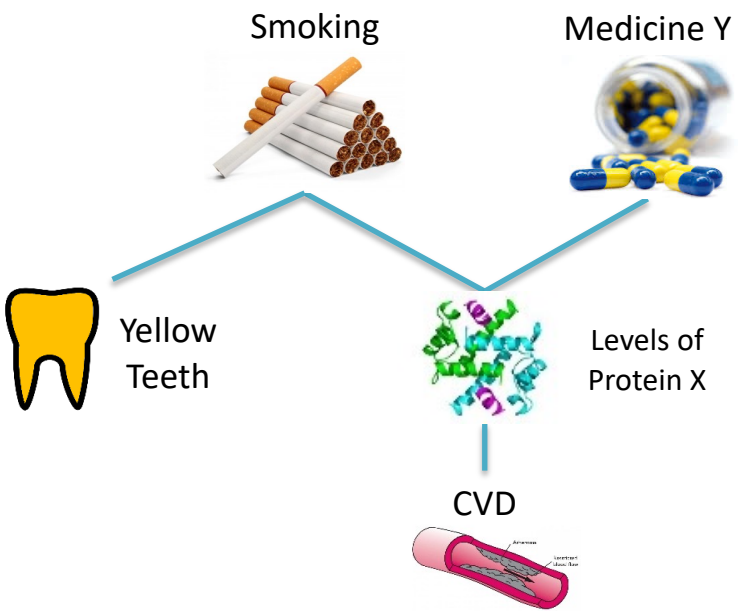
True (unknown) graph



Only Protein X has two neighbors.

PC Algorithm – an example

2. k=2



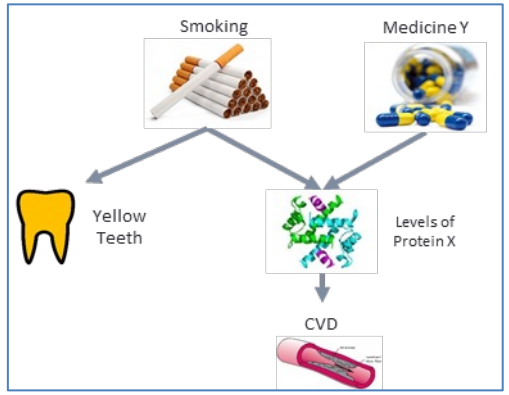
Tests attempted	p-value
~~Yellow Teeth, Smoking~~	~~0.00002~~
~~Yellow Teeth, CVD~~	~~0.00384~~
Yellow Teeth, Medicine Y	0.54501
~~Yellow Teeth, Protein X~~	~~0.00056~~
~~Smoking, CVD~~	~~0.00035~~
Smoking, Medicine Y	0.36458
~~Smoking, Protein X~~	~~0.00003~~
~~CVD, Medicine Y~~	~~0.00062~~
~~CVD, Protein X~~	~~0.00014~~
~~Medicine Y, Protein X~~	~~0.00007~~

Tests attempted	p-value
Yellow Teeth, CVD| Smoking	0.78961
CVD, Medicine Y | Protein X	0.15092
Yellow Teeth, Protein X | Smoking	0.23567
Smoking, CVD |Yellow Teeth	0.00345
Smoking, CVD | Protein X	0.12365
CVD, Protein X | Smoking	0.00045
CVD, Protein X | Medicine Y	0.00389
Medicine Y, Protein X|CVD	0.00972
Smoking, Protein X | Yellow Teeth	0.00126
Smoking, Protein X|CVD	0.00438
Yellow Teeth, Smoking|Protein X	0.00072

Only Protein X has two neighbors

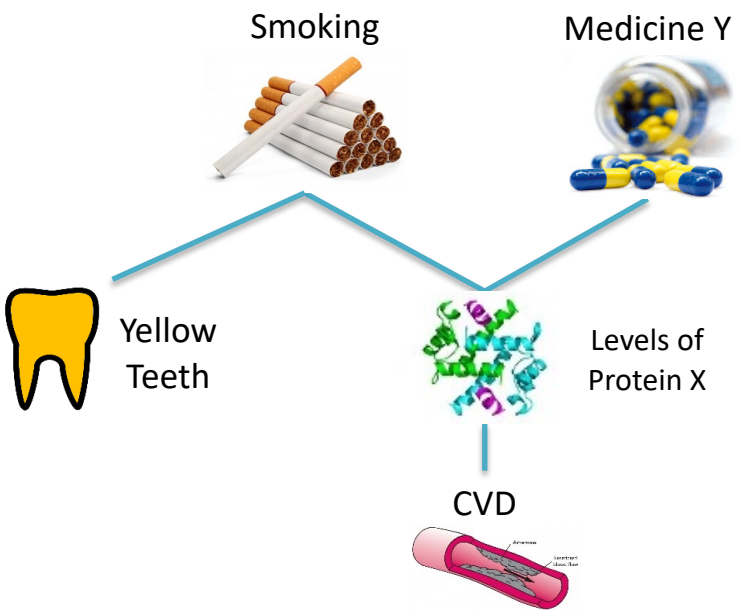
Tests attempted	p-value
CVD, Protein X| Smoking, Medicine Y	0.02356
Smoking, Protein X|CVD, Medicine Y	0.00498
Medicine Y, Protein X| Smoking, CVD	0.00074

True (unknown) graph



PC Algorithm – an example

2. k=3



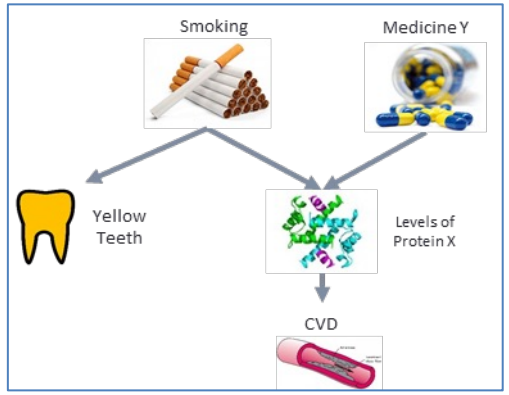
Tests attempted	p-value
~~Yellow Teeth, Smoking~~	~~0.00002~~
~~Yellow Teeth, CVD~~	~~0.00384~~
Yellow Teeth, Medicine Y	0.54501
~~Yellow Teeth, Protein X~~	~~0.00056~~
~~Smoking, CVD~~	~~0.00035~~
Smoking, Medicine Y	0.36458
~~Smoking, Protein X~~	~~0.00003~~
~~CVD, Medicine Y~~	~~0.00062~~
~~CVD, Protein X~~	~~0.00014~~
~~Medicine Y, Protein X~~	~~0.00007~~

Tests attempted	p-value
Yellow Teeth, CVD| Smoking	0.78961
CVD, Medicine Y | Protein X	0.15092
Yellow Teeth, Protein X | Smoking	0.23567
Smoking, CVD |Yellow Teeth	0.00345
Smoking, CVD | Protein X	0.12365
CVD, Protein X | Smoking	0.00045
CVD, Protein X | Medicine Y	0.00389
Medicine Y, Protein X|CVD	0.00972
Smoking, Protein X | Yellow Teeth	0.00126
Smoking, Protein X|CVD	0.00438
Yellow Teeth, Smoking|Protein X	0.00072

No variable has four neighbors.

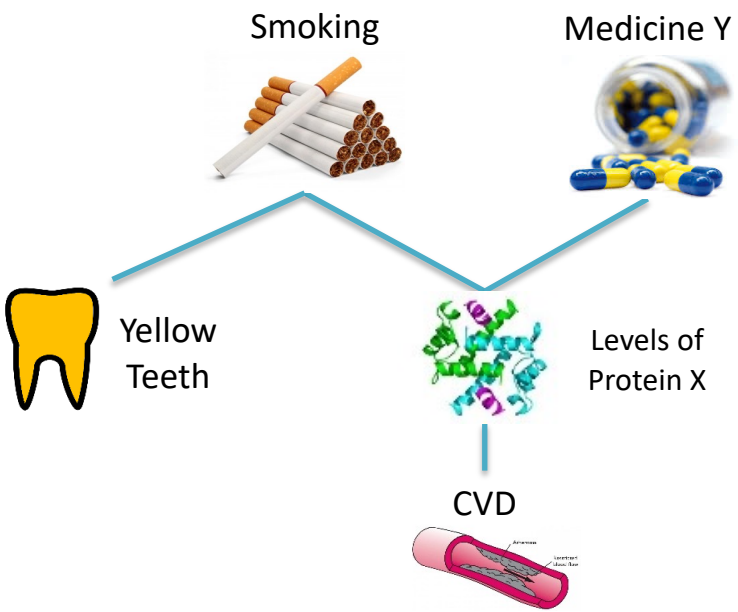
Tests attempted	p-value
CVD, Protein X| Smoking, Medicine Y	0.02356
Smoking, Protein X|CVD, Medicine Y	0.00498
Medicine Y, Protein X| Smoking, CVD	0.00074

True (unknown) graph



PC Algorithm – an example

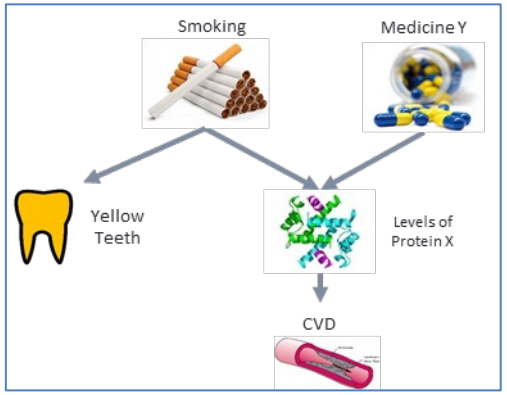
No more edges can be removed.



Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
Yellow Teeth, CVD	0.00384
Yellow Teeth, Medicine Y	0.54501
Yellow Teeth, Protein X	0.00056
Smoking, CVD	0.00035
Smoking, Medicine Y	0.36458
Smoking, Protein X	0.00003
CVD, Medicine Y	0.00062
CVD, Protein X	0.00014
Medicine Y, Protein X	0.00007

Tests attempted	p-value
Yellow Teeth, CVD| Smoking	0.78961
CVD, Medicine Y | Protein X	0.15092
Yellow Teeth, Protein X | Smoking	0.23567
Smoking, CVD |Yellow Teeth	0.00345
Smoking, CVD | Protein X	0.12365
CVD, Protein X | Smoking	0.00045
CVD, Protein X | Medicine Y	0.00389
Medicine Y, Protein X|CVD	0.00972
Smoking, Protein X | Yellow Teeth	0.00126
Smoking, Protein X|CVD	0.00438
Yellow Teeth, Smoking|Protein X	0.00072

True (unknown) graph



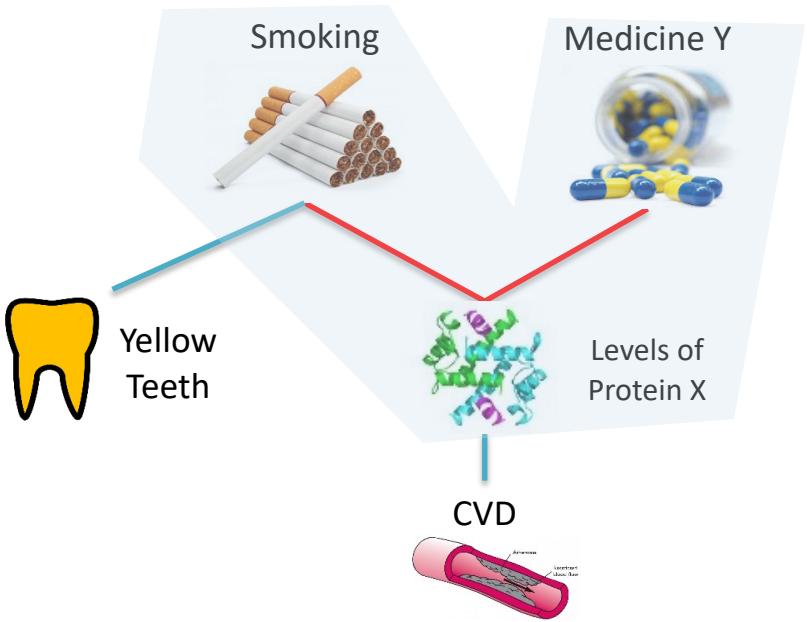
Tests attempted	p-value
CVD, Protein X| Smoking, Medicine Y	0.02356
Smoking, Protein X|CVD, Medicine Y	0.00498
Medicine Y, Protein X| Smoking, CVD	0.00074

You have (correctly) identified the skeleton of your graph

For causal discovery, you also want to identify some edge directions!

PC Algorithm – an example

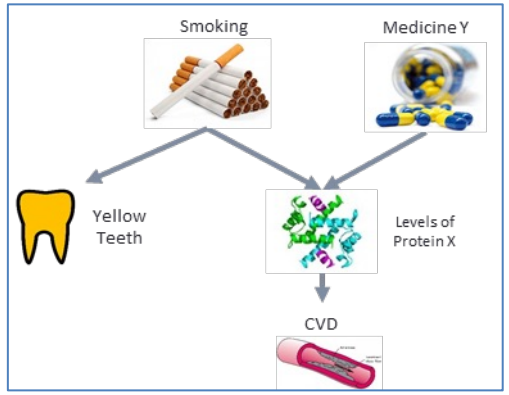
Smoking and Medicine Y are independent given the empty set.



Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
Yellow Teeth, CVD	0.00384
Yellow Teeth, Medicine Y	0.54501
Yellow Teeth, Protein X	0.00056
Smoking, CVD	0.00035
Smoking, Medicine Y	**0.36458**
Smoking, Protein X	0.00003
CVD, Medicine Y	0.00062
CVD, Protein X	0.00014
Medicine Y, Protein X	0.00007

Tests attempted	p-value
Yellow Teeth, CVD| Smoking	0.78961
CVD, Medicine Y | Protein X	0.15092
Yellow Teeth, Protein X | Smoking	0.23567
Smoking, CVD |Yellow Teeth	0.00345
Smoking, CVD | Protein X	0.12365
CVD, Protein X | Smoking	0.00045
CVD, Protein X | Medicine Y	0.00389
Medicine Y, Protein X|CVD	0.00972
Smoking, Protein X | Yellow Teeth	0.00126
Smoking, Protein X|CVD	0.00438
Yellow Teeth, Smoking|Protein X	0.00072

True (unknown) graph

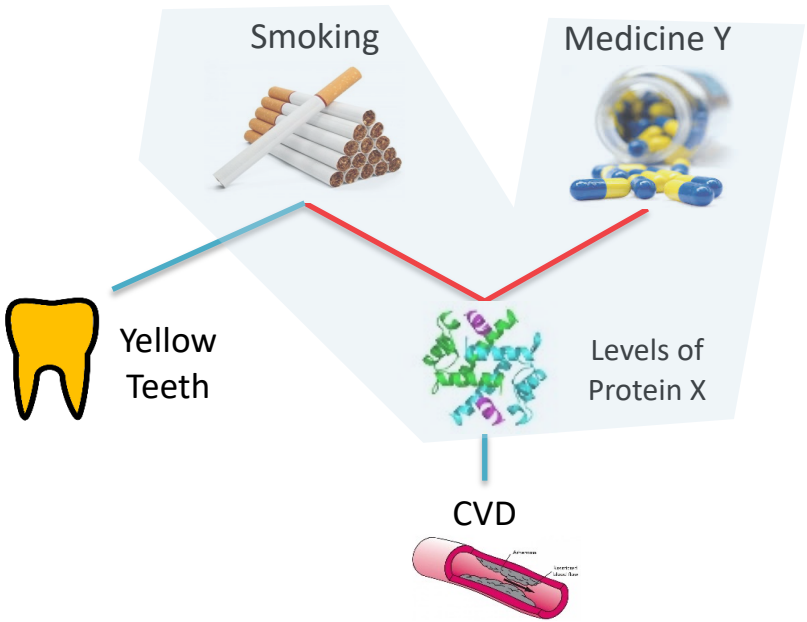


If Levels of Protein X was a non-collider on the path Smoking – Protein X – Medicine Y the path would be d-connecting given the empty set

Tests attempted	p-value
CVD, Protein X| Smoking, Medicine Y	0.02356
Smoking, Protein X|CVD, Medicine Y	0.00498
Medicine Y, Protein X| Smoking, CVD	0.00074

PC Algorithm – an example

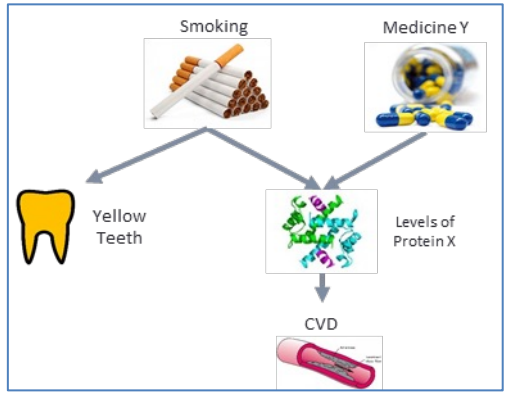
Smoking and Medicine Y are independent given the empty set.



Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
Yellow Teeth, CVD	0.00384
Yellow Teeth, Medicine Y	0.54501
Yellow Teeth, Protein X	0.00056
Smoking, CVD	0.00035
Smoking, Medicine Y	**0.36458**
Smoking, Protein X	0.00003
CVD, Medicine Y	0.00062
CVD, Protein X	0.00014
Medicine Y, Protein X	0.00007

Tests attempted	p-value
Yellow Teeth, CVD| Smoking	0.78961
CVD, Medicine Y | Protein X	0.15092
Yellow Teeth, Protein X | Smoking	0.23567
Smoking, CVD |Yellow Teeth	0.00345
Smoking, CVD | Protein X	0.12365
CVD, Protein X | Smoking	0.00045
CVD, Protein X | Medicine Y	0.00389
Medicine Y, Protein X|CVD	0.00972
Smoking, Protein X | Yellow Teeth	0.00126
Smoking, Protein X|CVD	0.00438
Yellow Teeth, Smoking|Protein X	0.00072

True (unknown) graph



Tests attempted	p-value
CVD, Protein X| Smoking, Medicine Y	0.02356
Smoking, Protein X|CVD, Medicine Y	0.00498
Medicine Y, Protein X| Smoking, CVD	0.00074

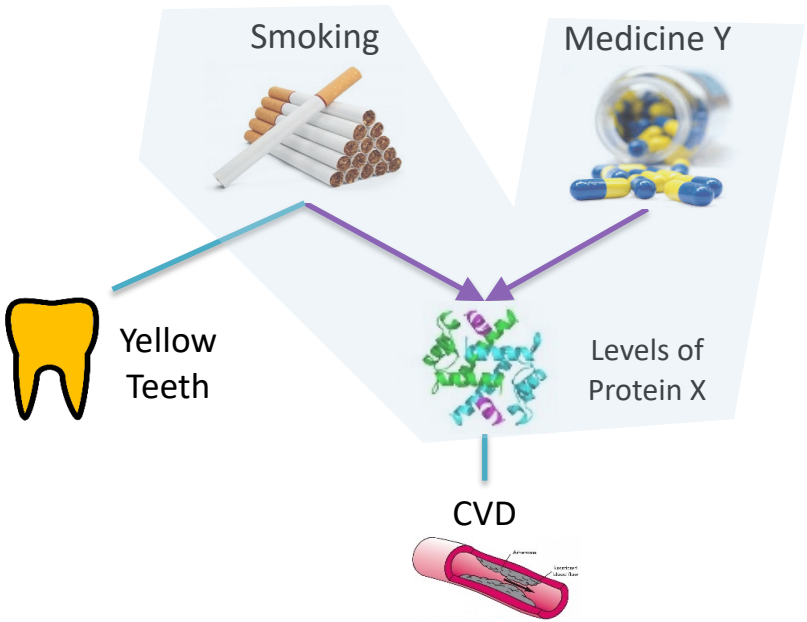
You would expect a dependence

$$\text{Smoking} \not\perp\!\!\!\perp \text{Medicine Y} \mid \emptyset$$

(i.e. a p-value < 0.05)

PC Algorithm – an example

Smoking and Medicine Y are independent given the empty set.



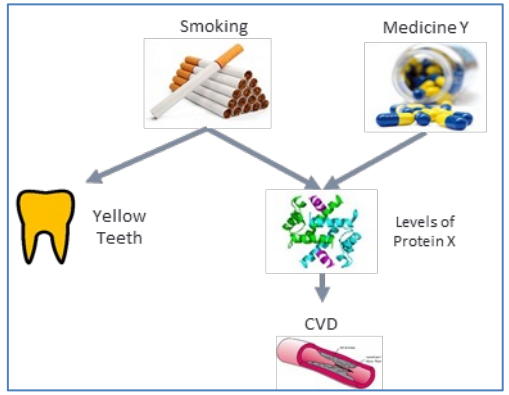
Thus, the triple must be a **collider**!

Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
Yellow Teeth, CVD	0.00384
Yellow Teeth, Medicine Y	0.54501
Yellow Teeth, Protein X	0.00056
Smoking, CVD	0.00035
Smoking, Medicine Y	**0.36458**
Smoking, Protein X	0.00003
CVD, Medicine Y	0.00062
CVD, Protein X	0.00014
Medicine Y, Protein X	0.00007

Tests attempted	p-value
Yellow Teeth, CVD| Smoking	0.78961
CVD, Medicine Y | Protein X	0.15092
Yellow Teeth, Protein X | Smoking	0.23567
Smoking, CVD |Yellow Teeth	0.00345
Smoking, CVD | Protein X	0.12365
CVD, Protein X | Smoking	0.00045
CVD, Protein X | Medicine Y	0.00389
Medicine Y, Protein X|CVD	0.00972
Smoking, Protein X | Yellow Teeth	0.00126
Smoking, Protein X|CVD	0.00438
Yellow Teeth, Smoking|Protein X	0.00072

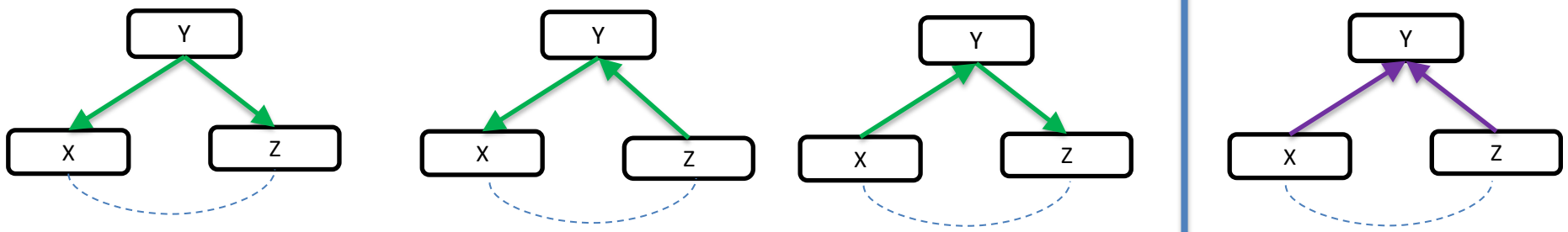
Tests attempted	p-value
CVD, Protein X| Smoking, Medicine Y	0.02356
Smoking, Protein X|CVD, Medicine Y	0.00498
Medicine Y, Protein X| Smoking, CVD	0.00074

True (unknown) graph



Unshielded colliders in BNs

Causal Bayesian Network describing your variables



Independencies entailed by the CMC

$X \perp\!\!\!\perp Z \mid \{Y \ldots\}$

$X \perp\!\!\!\perp Z \mid \{Y \ldots\}$

$X \perp\!\!\!\perp Z \mid \{Y \ldots\}$

$X \perp\!\!\!\perp Z \mid \{\ldots\}$

$X \not\perp\!\!\!\perp Z \mid \{Y \ldots\}$

- - - - - - - - - Could be connected by another path, but not an edge

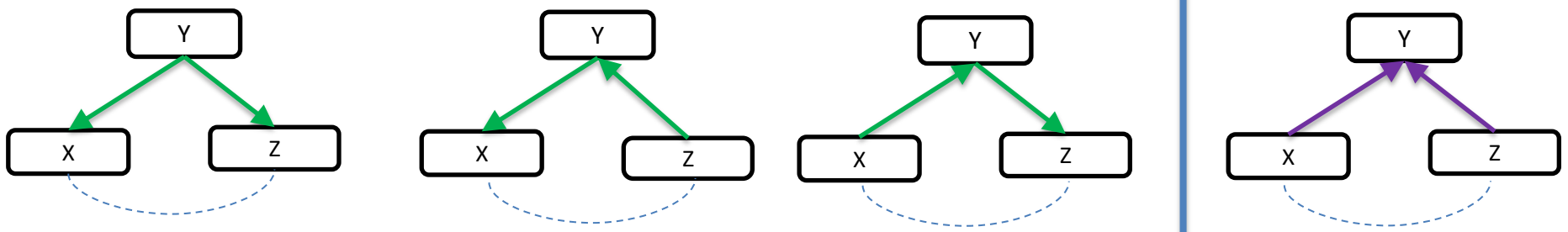
You need Y to d-separate X and Z

You DON'T need Y to d-separate X and Z

In fact, conditioning on Y would make X, Z dependent

Unshielded colliders in BNs

Causal Bayesian Network describing your variables



Independencies entailed by the CMC

$X \perp\!\!\!\perp Z \mid \{Y \ldots\}$

$X \perp\!\!\!\perp Z \mid \{Y \ldots\}$

$X \perp\!\!\!\perp Z \mid \{Y \ldots\}$

$X \perp\!\!\!\perp Z \mid \{\ldots\}$

$X \not\perp\!\!\!\perp Z \mid \{Y \ldots\}$

- - - - - - - - - Could be connected by another path, but not an edge

You need Y to d-separate X and Z

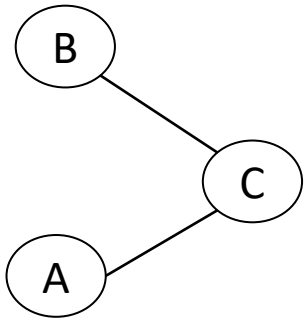
You DON'T need Y to d-separate X and Z

In fact, conditioning on Y would make X, Z dependent

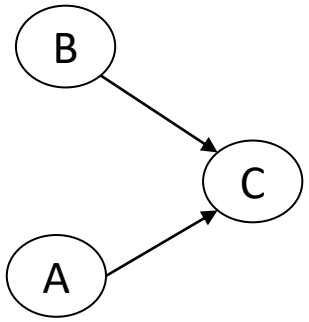
if X-Z-Y form an unshielded triplet, you can distinguish whether the triplet is a collider or a non-collider.

Orientation rules

R0.

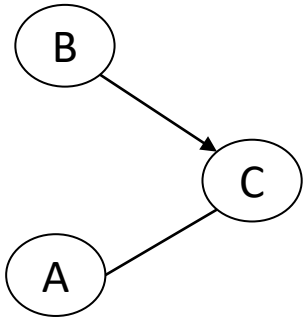


If C is NOT in the d-separating set of A, B

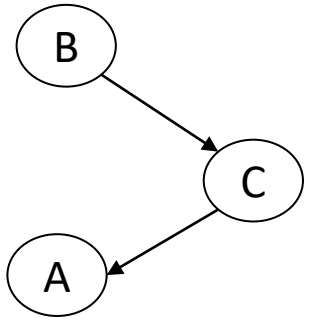


Orient Unshielded Colliders

R1.



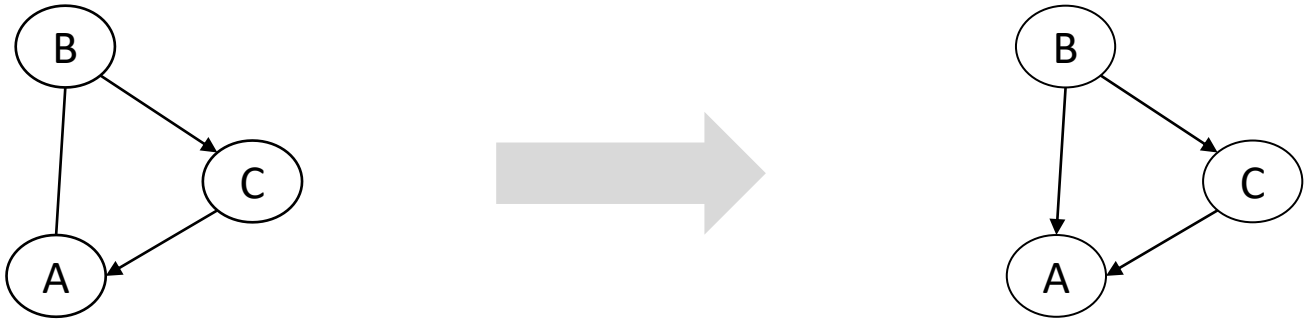
If C is in the d-separating set of A, B



Away from collider

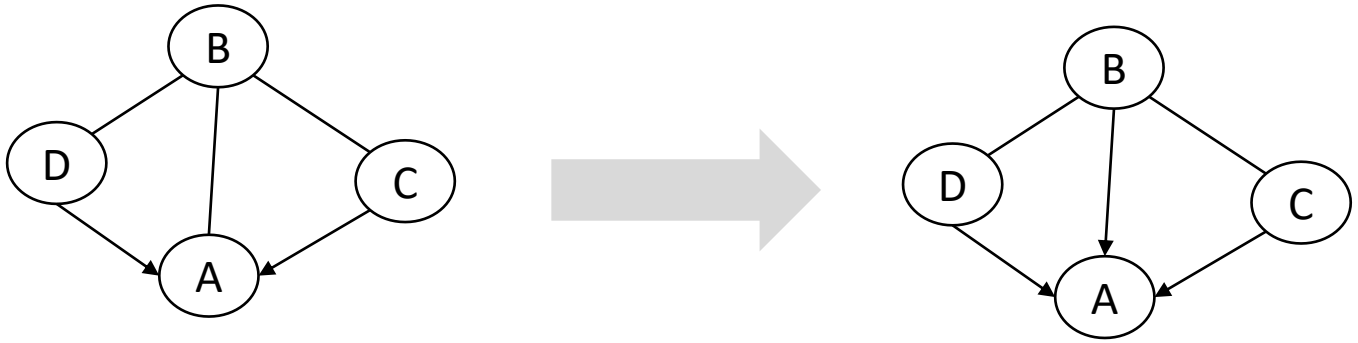
Orientation rules

R2.



Away from cycles

R3.



Double triangle

The PC algorithm

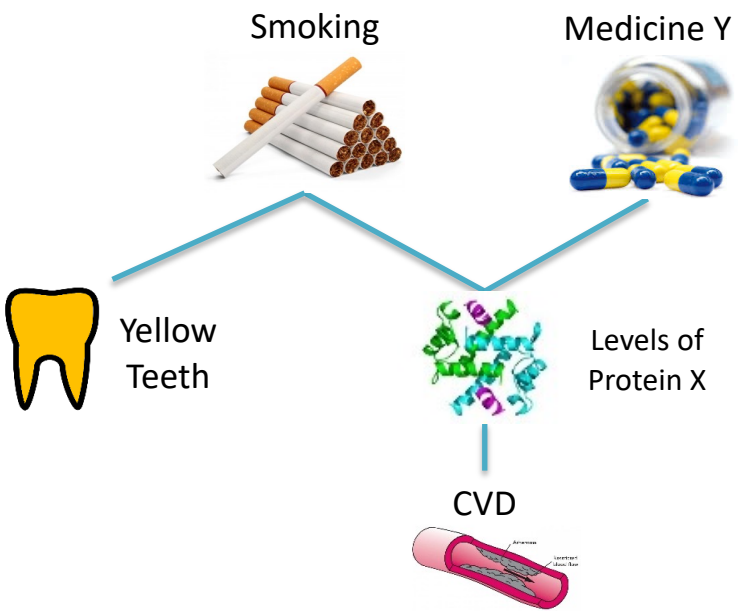
Search strategy:

- Identify the skeleton of your PDAG:
  - Begin with the full graph.
  - For k=0:number of variables -2
  - Using heuristic 3
    - For each pair of adjacent variables X, Y,
      - look within Adjacencies(X)\Y or Adjacencies(Y)\X for a set of k observed variables **Z** such that $X \perp\!\!\!\perp Y \mid \mathbf{Z}$.
      - If you succeed, remove X-Y.
- Orient all invariant edges of the Markov Equivalence class
  - Apply R0
  - While no more rules are applicable, apply R1-R3

Rules R0-R3 are complete (Meek, 1995)

PC Algorithm – an example

Apply orientation rules

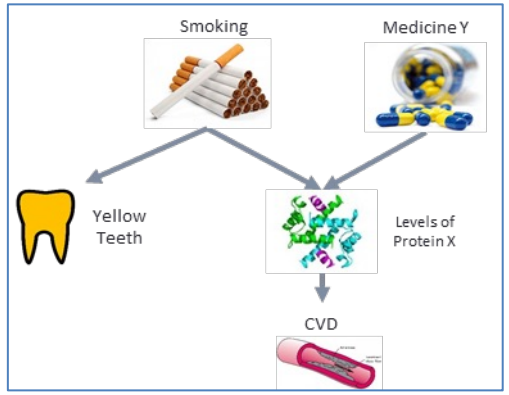


Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
Yellow Teeth, CVD	0.00384
Yellow Teeth, Medicine Y	0.54501
Yellow Teeth, Protein X	0.00056
Smoking, CVD	0.00035
Smoking, Medicine Y	0.36458
Smoking, Protein X	0.00003
CVD, Medicine Y	0.00062
CVD, Protein X	0.00014
Medicine Y, Protein X	0.00007

Tests attempted	p-value
Yellow Teeth, CVD| Smoking	0.78961
CVD, Medicine Y | Protein X	0.15092
Yellow Teeth, Protein X | Smoking	0.23567
Smoking, CVD |Yellow Teeth	0.00345
Smoking, CVD | Protein X	0.12365
CVD, Protein X | Smoking	0.00045
CVD, Protein X | Medicine Y	0.00389

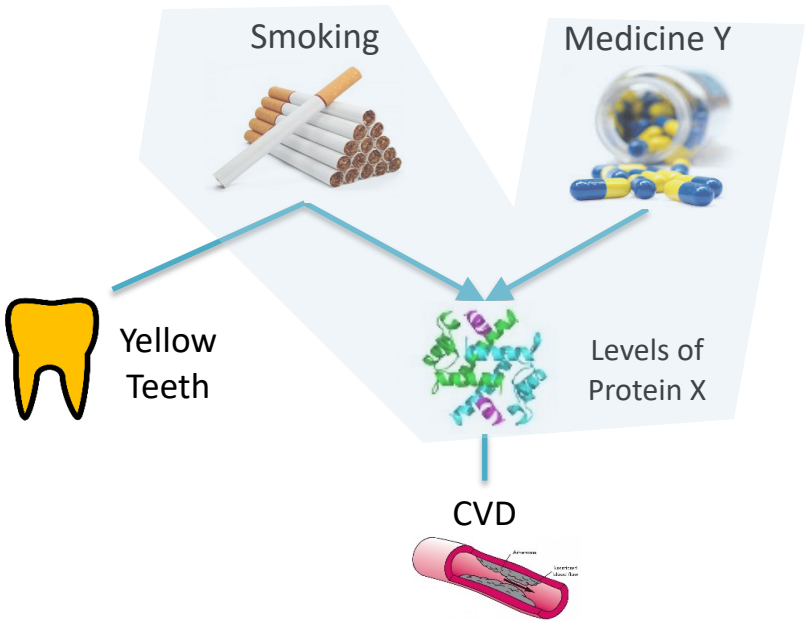
Tests attempted	p-value
CVD, Protein X| Smoking, Medicine Y	0.02356
Smoking, Protein X|CVD, Medicine Y	0.00498
Medicine Y, Protein X| Smoking, CVD	0.00074

True (unknown) graph



PC Algorithm – an example

Apply orientation rules

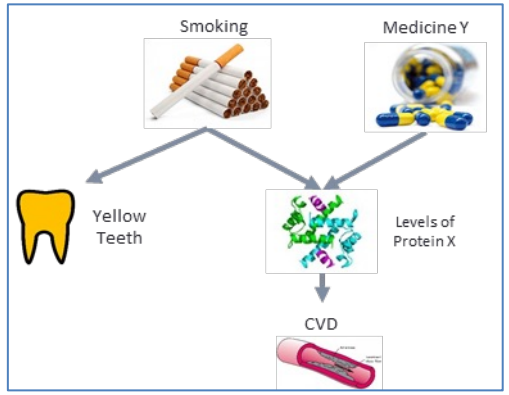


Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
Yellow Teeth, CVD	0.00384
Yellow Teeth, Medicine Y	0.54501
Yellow Teeth, Protein X	0.00056
Smoking, CVD	0.00035
Smoking, Medicine Y	**0.36458**
Smoking, Protein X	0.00003
CVD, Medicine Y	0.00062
CVD, Protein X	0.00014
Medicine Y, Protein X	0.00007

Tests attempted	p-value
Yellow Teeth, CVD| Smoking	0.78961
CVD, Medicine Y | Protein X	0.15092
Yellow Teeth, Protein X | Smoking	0.23567
Smoking, CVD |Yellow Teeth	0.00345
Smoking, CVD | Protein X	0.12365
CVD, Protein X | Smoking	0.00045
CVD, Protein X | Medicine Y	0.00389

Tests attempted	p-value
CVD, Protein X| Smoking, Medicine Y	0.02356
Smoking, Protein X|CVD, Medicine Y	0.00498
Medicine Y, Protein X| Smoking, CVD	0.00074

True (unknown) graph

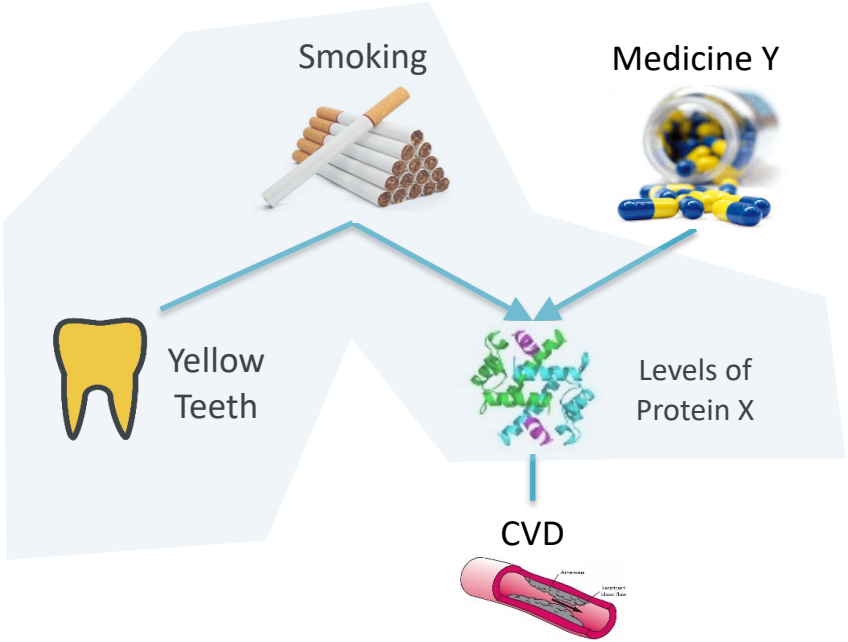


Orient unshielded colliders

Smoking-Protein X-Medicine Y is a collider

PC Algorithm – an example

Apply orientation rules

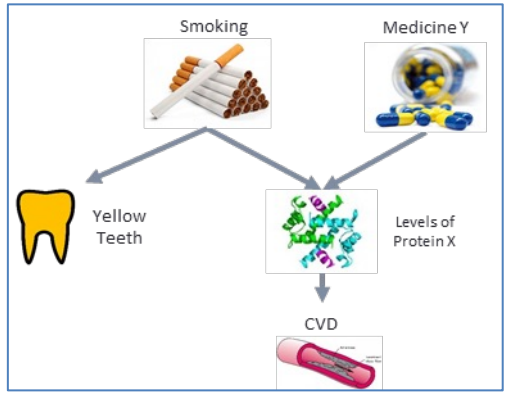


Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
Yellow Teeth, CVD	0.00384
Yellow Teeth, Medicine Y	0.54501
Yellow Teeth, Protein X	0.00056
Smoking, CVD	0.00035
Smoking, Medicine Y	**0.36458**
Smoking, Protein X	0.00003
CVD, Medicine Y	0.00062
CVD, Protein X	0.00014
Medicine Y, Protein X	0.00007

Tests attempted	p-value
Yellow Teeth, CVD| Smoking	0.78961
CVD, Medicine Y | Protein X	0.15092
Yellow Teeth, Protein X | Smoking	0.23567
Smoking, CVD |Yellow Teeth	0.00345
Smoking, CVD | Protein X	0.12365
CVD, Protein X | Smoking	0.00045
CVD, Protein X | Medicine Y	0.00389

Tests attempted	p-value
CVD, Protein X| Smoking, Medicine Y	0.02356
Smoking, Protein X|CVD, Medicine Y	0.00498
Medicine Y, Protein X| Smoking, CVD	0.00074

True (unknown) graph

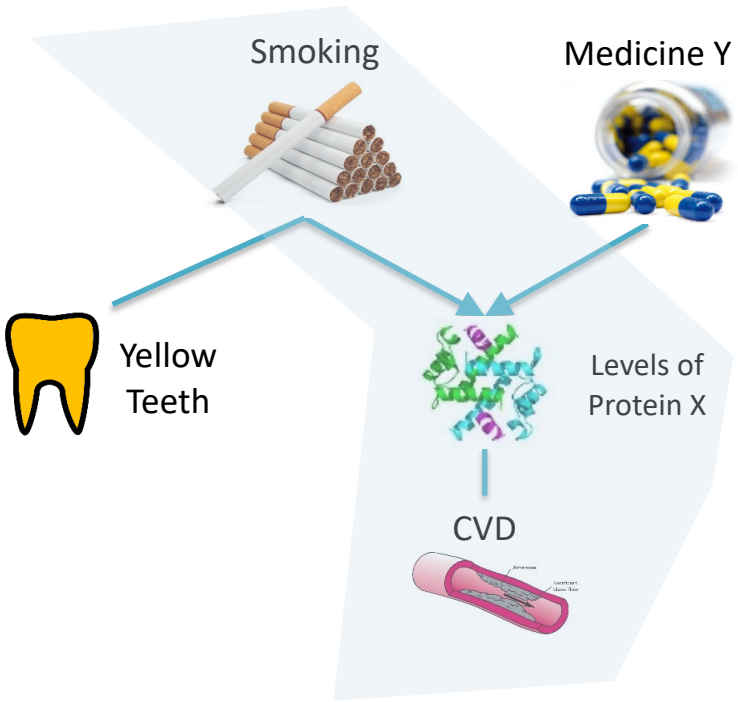


Orient unshielded colliders

Yellow Teeth-Smoking-Protein X is a non collider

PC Algorithm – an example

Apply orientation rules

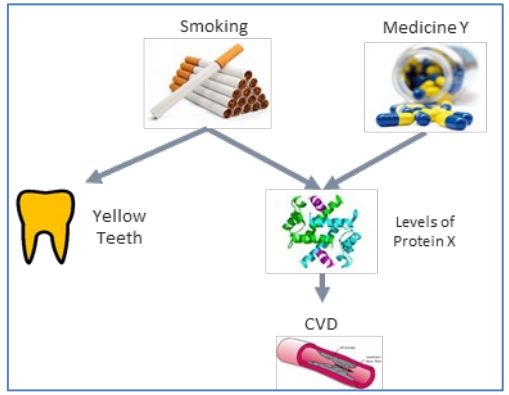


Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
Yellow Teeth, CVD	0.00384
Yellow Teeth, Medicine Y	0.54501
Yellow Teeth, Protein X	0.00056
Smoking, CVD	0.00035
Smoking, Medicine Y	**0.36458**
Smoking, Protein X	0.00003
CVD, Medicine Y	0.00062
CVD, Protein X	0.00014
Medicine Y, Protein X	0.00007

Tests attempted	p-value
Yellow Teeth, CVD| Smoking	0.78961
CVD, Medicine Y | Protein X	0.15092
Yellow Teeth, Protein X | Smoking	0.23567
Smoking, CVD |Yellow Teeth	0.00345
Smoking, CVD | Protein X	0.12365
CVD, Protein X | Smoking	0.00045
CVD, Protein X | Medicine Y	0.00389

Tests attempted	p-value
CVD, Protein X| Smoking, Medicine Y	0.02356
Smoking, Protein X|CVD, Medicine Y	0.00498
Medicine Y, Protein X| Smoking, CVD	0.00074

True (unknown) graph

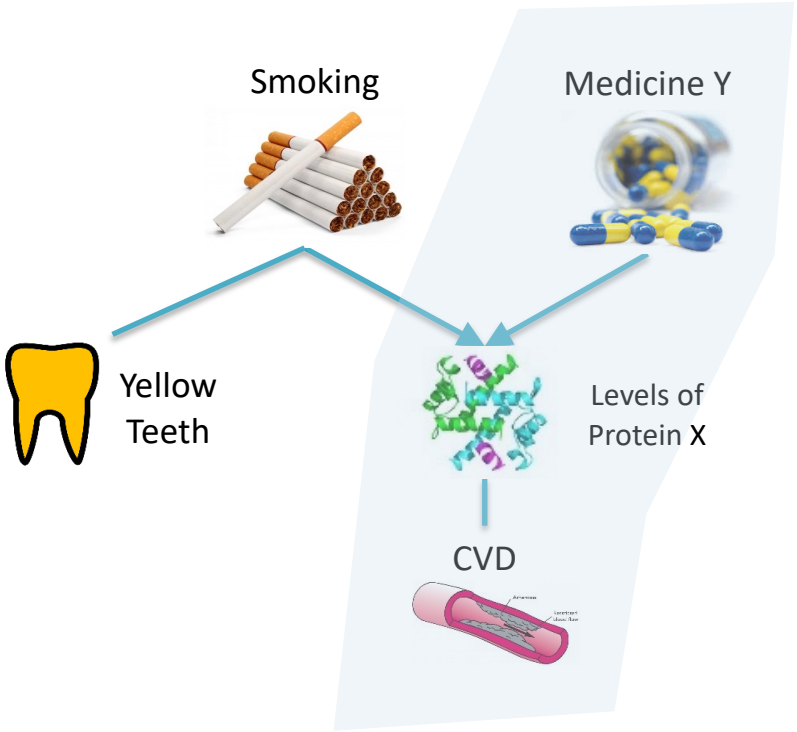


Orient unshielded colliders

Smoking-Protein X- CVD is a non collider

PC Algorithm – an example

Apply orientation rules

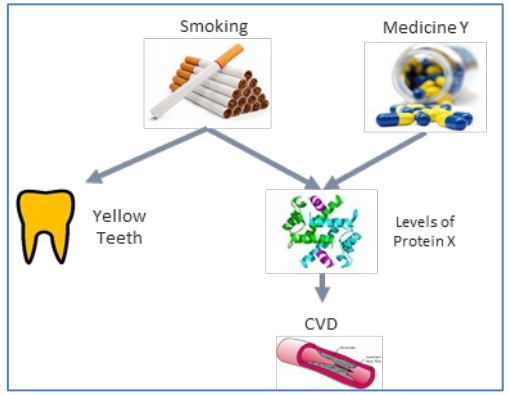


Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
Yellow Teeth, CVD	0.00384
Yellow Teeth, Medicine Y	0.54501
Yellow Teeth, Protein X	0.00056
Smoking, CVD	0.00035
Smoking, Medicine Y	**0.36458**
Smoking, Protein X	0.00003
CVD, Medicine Y	0.00062
CVD, Protein X	0.00014
Medicine Y, Protein X	0.00007

Tests attempted	p-value
Yellow Teeth, CVD| Smoking	0.78961
CVD, Medicine Y | Protein X	0.15092
Yellow Teeth, Protein X | Smoking	0.23567
Smoking, CVD |Yellow Teeth	0.00345
Smoking, CVD | Protein X	0.12365
CVD, Protein X | Smoking	0.00045
CVD, Protein X | Medicine Y	0.00389

Tests attempted	p-value
CVD, Protein X| Smoking, Medicine Y	0.02356
Smoking, Protein X|CVD, Medicine Y	0.00498
Medicine Y, Protein X| Smoking, CVD	0.00074

True (unknown) graph

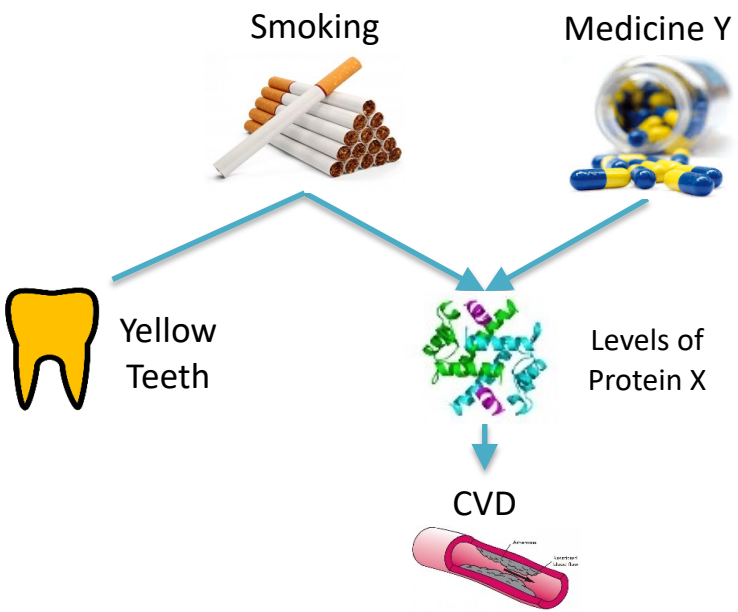


Orient unshielded colliders

Medicine Y -Protein X- CVD is a non collider

PC Algorithm – an example

Apply orientation rules



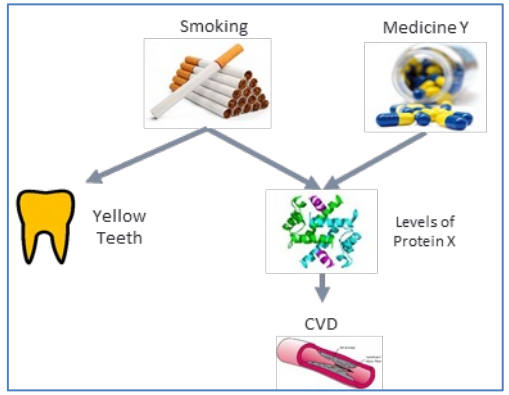
Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
Yellow Teeth, CVD	0.00384
Yellow Teeth, Medicine Y	0.54501
Yellow Teeth, Protein X	0.00056
Smoking, CVD	0.00035
Smoking, Medicine Y	0.36458
Smoking, Protein X	0.00003
CVD, Medicine Y	0.00062
CVD, Protein X	0.00014
Medicine Y, Protein X	0.00007

Tests attempted	p-value
Yellow Teeth, CVD| Smoking	0.78961
CVD, Medicine Y | Protein X	0.15092
Yellow Teeth, Protein X | Smoking	0.23567
Smoking, CVD |Yellow Teeth	0.00345
Smoking, CVD | Protein X	**0.12365**
CVD, Protein X | Smoking	0.00045
CVD, Protein X | Medicine Y	0.00389

Away from collider

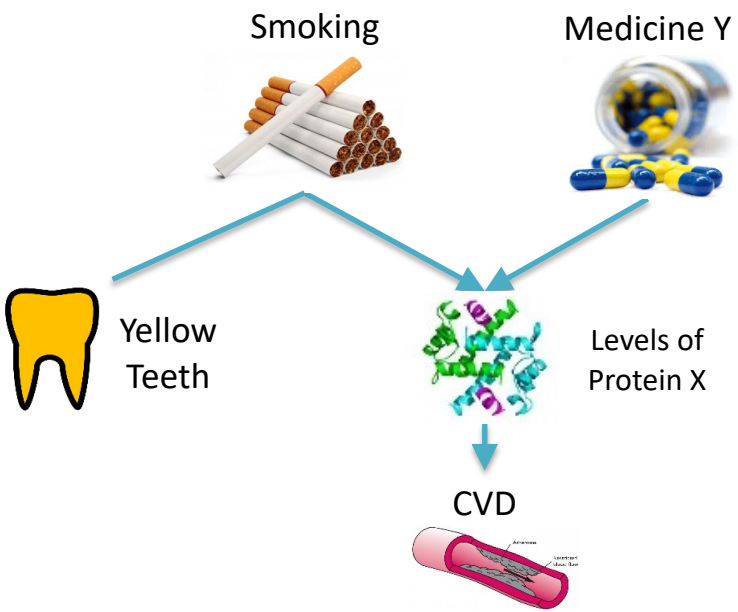
Tests attempted	p-value
CVD, Protein X| Smoking, Medicine Y	0.02356
Smoking, Protein X|CVD, Medicine Y	0.00498
Medicine Y, Protein X| Smoking, CVD	0.00074

True (unknown) graph



PC Algorithm – an example

Apply orientation rules



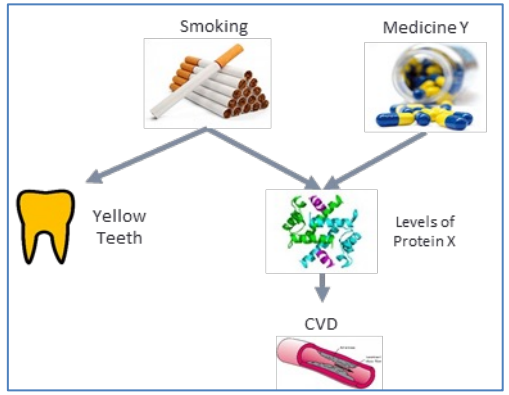
No more rules are applicable

Tests attempted	p-value
Yellow Teeth, Smoking	0.00002
Yellow Teeth, CVD	0.00384
Yellow Teeth, Medicine Y	0.54501
Yellow Teeth, Protein X	0.00056
Smoking, CVD	0.00035
Smoking, Medicine Y	0.36458
Smoking, Protein X	0.00003
CVD, Medicine Y	0.00062
CVD, Protein X	0.00014
Medicine Y, Protein X	0.00007

Tests attempted	p-value
Yellow Teeth, CVD| Smoking	0.78961
CVD, Medicine Y | Protein X	0.15092
Yellow Teeth, Protein X | Smoking	0.23567
Smoking, CVD |Yellow Teeth	0.00345
Smoking, CVD | Protein X	0.12365
CVD, Protein X | Smoking	0.00045
CVD, Protein X | Medicine Y	0.00389

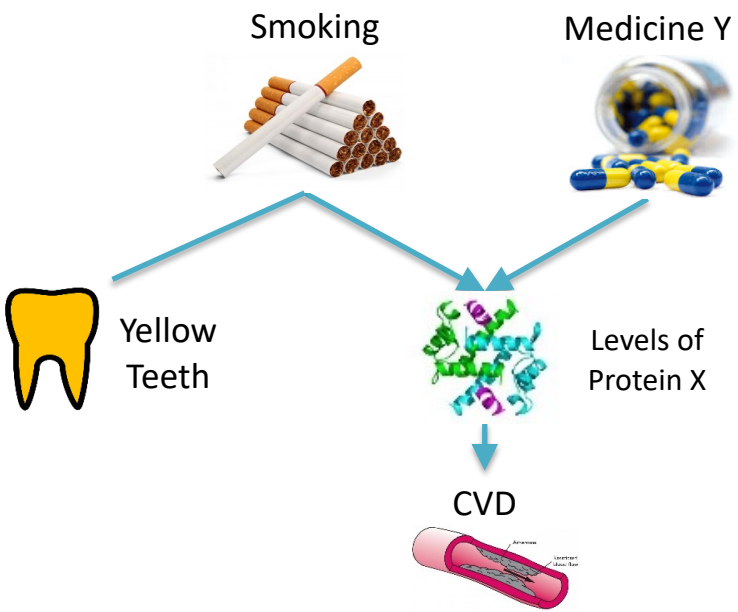
Tests attempted	p-value
CVD, Protein X| Smoking, Medicine Y	0.02356
Smoking, Protein X|CVD, Medicine Y	0.00498
Medicine Y, Protein X| Smoking, CVD	0.00074

True (unknown) graph

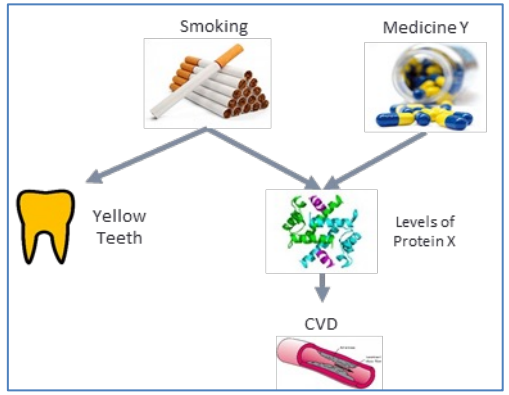


PC Algorithm output

PDAG returned by the PC algorithm



True (unknown) graph



PC algorithm

Introduced by Peter Spirtes and Clark Glymour in 1993.
One of the first algorithms to perform causal discovery from cross-sectional data.
Uses a complete set of orientation rules and therefore identifies the PDAG that faithfully represents the conditional independencies it identifies.

The PDAG is maximally informative, in the sense that every un-oriented edge has different orientations in different DAGs in the Markov Equivalence class.

Most current constraint-based algorithms are extensions/improvements of the PC algorithm.

PC algorithm

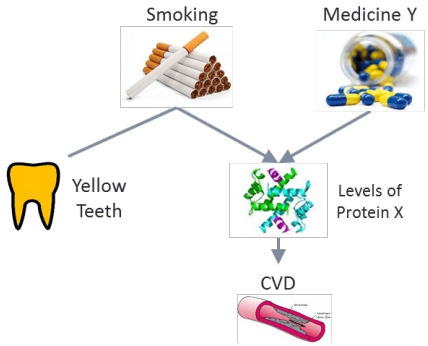
Remember the assumptions:

Causal Markov Condition.
Faithfulness.
Acyclicity.
Causal Sufficiency (no hidden confounders).

INPUT

Variables / Samples	Yellow Teeth	Smoking	CVD	Medicine Y	Levels of Protein X
1					
2					
3					
4					
5					
6					
7					
8					
9					
⋮					
999					
1000					

OUTPUT



PC Algorithm - Complexity

Suppose that the maximum number of parents for any variable in the graph is k.
Then the worst-case number of tests of conditional independence performed by PC is:

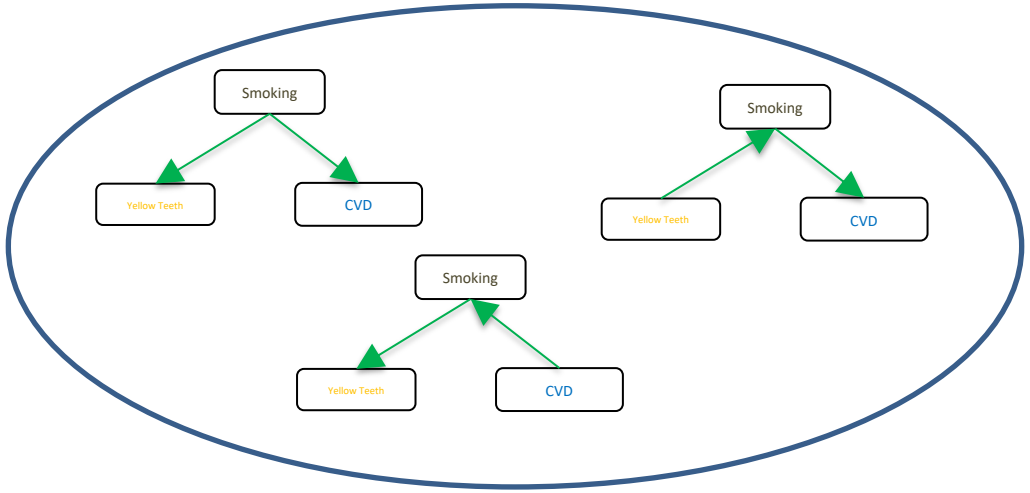
$$2\left(\frac{n}{2}\right)\sum_{i=0}^{k}\left(\frac{n-1}{i}\right)$$

which is bounded by

$$\frac{n^2(n-1)^{k-1}}{(k-1)!}$$

i.e., polynomial to the number of variables, exponential to the maximum number of parents.

Learning causal networks as a model selection problem



Sample (Person)	Smoking	CVD	Yellow Teeth
1	Yes	Yes	No
2	No	No	No
3	Yes	Yes	Yes
4	No	No	Yes
5	Yes	No	No
6	No	Yes	Yes
⋮			
52	No	Yes	No

Identify all DAGs that maximize the posterior probability of the graph given the data: P(G|D) (or some other data-fitting criterion in general)

Posterior probability of the graph

Probability of the data given the graph

Prior probability of the graph

$$P(G|D) = \frac{P(D|G) \times P(G)}{P(D)} = \frac{\boxed{P(D|G)} \times \boxed{P(G)}}{\boxed{P(D)}}$$

Normalization constant

Posterior probability of the graph

Probability of the data given the graph

Prior probability of the graph

$$P(G|D) = \frac{P(D|G) \times P(G)}{P(D)} = \frac{\boxed{P(D|G)} \times \boxed{P(G)}}{\boxed{P(D)}}$$

You can ignore it since it does not depend on the graph structure.

Scoring function

Find G: $argmax_G \; P(D|G) \times P(G)$

Scoring function

Uniform/based on prior knowledge/favoring sparsity

Find G: $argmax_G \; P(D|G) \times \boxed{P(G)}$

Scoring function

Uniform/based on prior knowledge/favoring sparsity

Find G: $argmax_G \boxed{P(D|G)} \times \boxed{P(G)}$

Average over all possible parameters (of the joint probability distribution).

$$\int_{\boldsymbol{\theta}} P(D|G,\theta)P(\theta)d\boldsymbol{\theta}$$

Scoring function

Uniform/based on prior knowledge/favoring sparsity

Find G: $argmax_G \boxed{P(D|G)} \times \boxed{P(G)}$

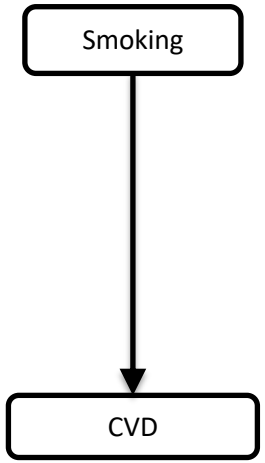
Average over all possible parameters (of the joint probability distribution).

$$\int_{\boldsymbol{\theta}} P(D|G,\theta)P(\theta)d\boldsymbol{\theta} = \int_{\boldsymbol{\theta}} P\left(D \middle| \boldsymbol{\theta}_{\boldsymbol{x}|\boldsymbol{pa}(\boldsymbol{x})}\right) f(\boldsymbol{\theta})d\boldsymbol{\theta}$$

The parameterization depends on the graphical structure.

Scoring function

$$P(D|G) = \int_{\boldsymbol{\theta}} P\left(D \middle| \boldsymbol{\theta}_{\boldsymbol{x}|\boldsymbol{pa}(\boldsymbol{x})}\right) f(\boldsymbol{\theta}) d\boldsymbol{\theta} =$$



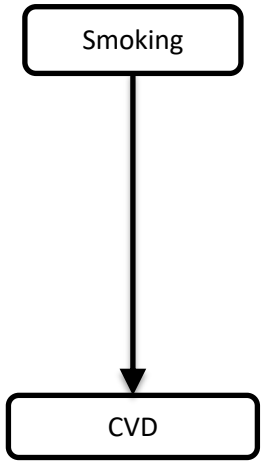
	P(Smoking)
Yes	θ_s
No	$1 - \theta_s$

	P(CVD)	
Smoking	Yes	No
Yes	$\theta_{C\|S}$	$1 - \theta_{C\|S}$
No	$\theta_{C\|NS}$	$1 - \theta_{C\|NS}$

Scoring function

$$P(D|G) = \int_{\boldsymbol{\theta}} P\left(D \middle| \boldsymbol{\theta}_{\boldsymbol{x}|\boldsymbol{pa}(\boldsymbol{x})}\right) f(\boldsymbol{\theta}) d\boldsymbol{\theta} =$$

$$\prod_{x} \int_{\theta_{x|pa(x)}} P\left(D \middle| \theta_{x|pa(x)}\right) f\left(\theta_{x|pa(x)}\right) d\theta_{x|pa(x)}$$



	P(Smoking)
Yes	θ_s
No	$1 - \theta_s$

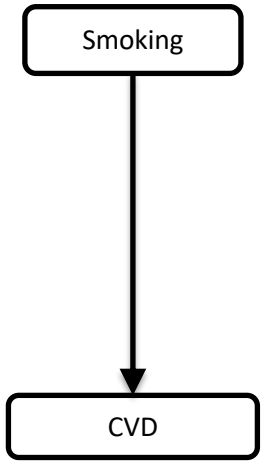
	P(CVD)	
Smoking	Yes	No
Yes	$\theta_{C\|S}$	$1 - \theta_{C\|S}$
No	$\theta_{C\|NS}$	$1 - \theta_{C\|NS}$

- Score is decomposable:
- It is a product of terms involving only a variable and its parents.

Scoring function

$$P(D|G) = \int_{\boldsymbol{\theta}} P\left(D\middle|G, \boldsymbol{\theta}_{\boldsymbol{x}|\boldsymbol{pa}(\boldsymbol{x})}\right) f(\boldsymbol{\theta}) d\boldsymbol{\theta} =$$

$$\prod_{x} \int_{\theta_{x|pa(x)}} P\left(D\middle|G, \theta_{x|pa(x)}\right) f\left(\theta_{x|pa(x)}\right) d\theta_{x|pa(x)}$$



	P(Smoking)
Yes	θ_s
No	$1 - \theta_s$

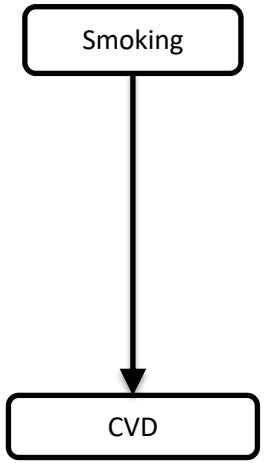
	P(CVD)	
Smoking	Yes	No
Yes	$\theta_{C\|S}$	$1 - \theta_{C\|S}$
No	$\theta_{C\|NS}$	$1 - \theta_{C\|NS}$

$$\int_{\theta_s} P(D|\,\theta_s)\ f(\theta_s) d\theta_s \int_{\theta_{c|ns}} P\left(D\middle|\,\theta_{c|s}\right)\ f\left(\theta_{c|s}\right) d\theta_{c|s} \int_{\theta_{c|ns}} P\left(D\middle|\,\theta_{c|ns}\right)\ f\left(\theta_{c|ns}\right) d\theta_{c|ns}$$

Scoring function

$$P(D|G) = \int_{\boldsymbol{\theta}} P\left(D \middle| G, \boldsymbol{\theta}_{\boldsymbol{x}|\boldsymbol{pa}(\boldsymbol{x})}\right) f(\boldsymbol{\theta}) d\boldsymbol{\theta} =$$

$$\prod_{x} \int_{\theta_{x|pa(x)}} P\left(D \middle| G, \theta_{x|pa(x)}\right) f\left(\theta_{x|pa(x)}\right) d\theta_{x|pa(x)}$$



	P(Smoking)
Yes	θ_s
No	$1 - \theta_s$

	P(CVD)	
Smoking	Yes	No
Yes	$\theta_{C\|S}$	$1 - \theta_{C\|S}$
No	$\theta_{C\|NS}$	$1 - \theta_{C\|NS}$

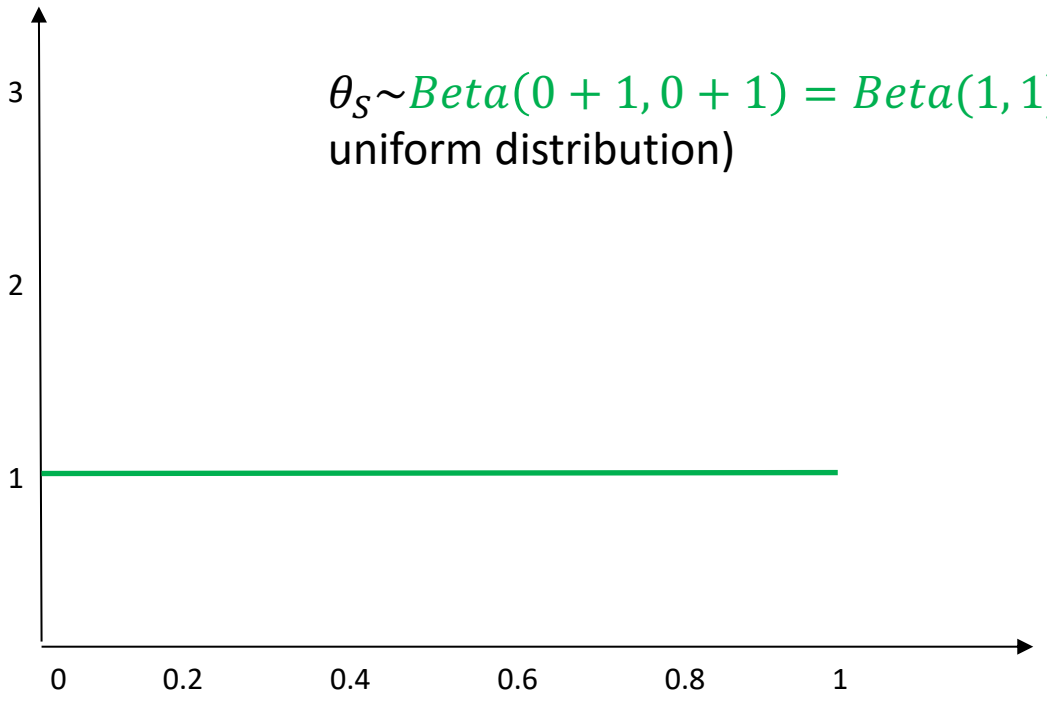
$$\boxed{\int_{\theta_s} P(D|\,\theta_s)\ f(\theta_s) d\theta_s} \int_{\theta_{c|ns}} P\left(D \middle|\, \theta_{c|s}\right)\ f\left(\theta_{c|s}\right) d\theta_{c|s} \int_{\theta_{c|ns}} P\left(D \middle|\, \theta_{c|ns}\right)\ f\left(\theta_{c|ns}\right) d\theta_{c|ns}$$

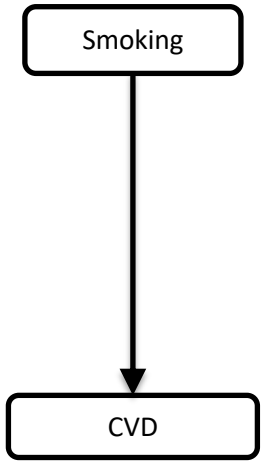
This score can be computed in closed form for some families of distributions that have conjugate priors

Scoring function

You have observed 0 smokers and 0 non smokers. (Prior)

$\theta_S \sim Beta(0+1, 0+1) = Beta(1,1)$ (the uniform distribution)





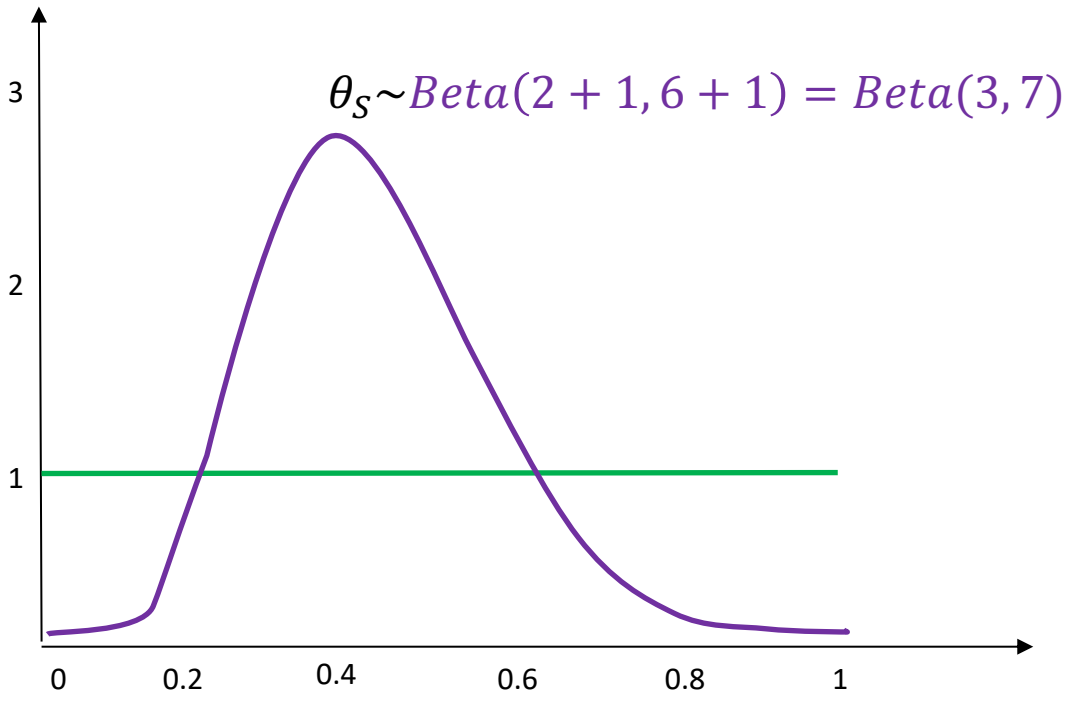
	P(Smoking)
Yes	θ_s
No	$1-\theta_s$

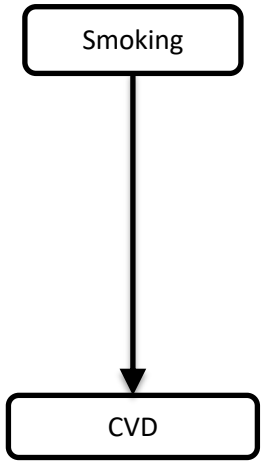
	P(CVD)	
Smoking	Yes	No
Yes	$\theta_{C\|S}$	$1-\theta_{C\|S}$
No	$\theta_{C\|NS}$	$1-\theta_{C\|NS}$

Reminder: Bayesian Statistics.

Scoring function

You then observe 2 smokers and 6 non-smokers. Bayesian Update :





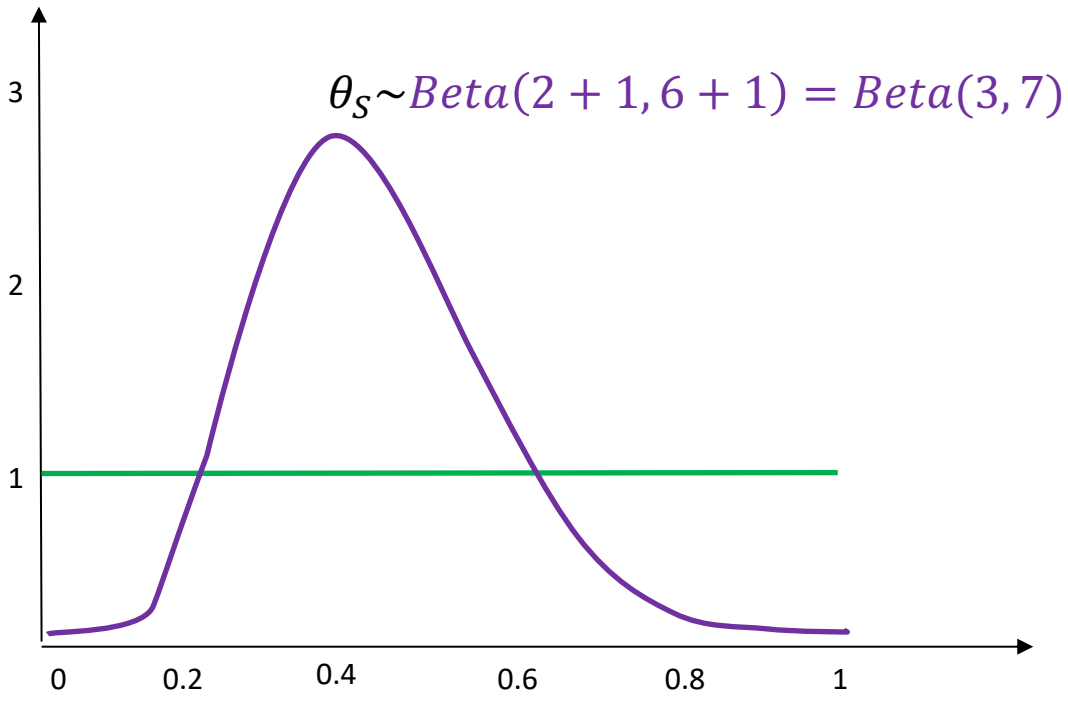
	P(Smoking)
Yes	θ_s
No	$1-\theta_s$

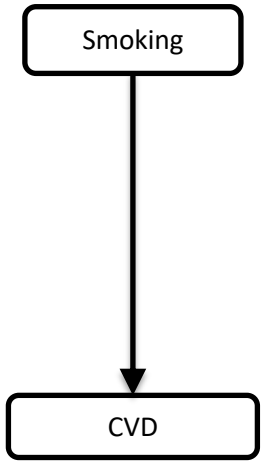
	P(CVD)	
Smoking	Yes	No
Yes	$\theta_{C\|S}$	$1-\theta_{C\|S}$
No	$\theta_{C\|NS}$	$1-\theta_{C\|NS}$

Reminder: Bayesian Statistics.

Scoring function

You then observe 2 smokers and 6 non-smokers. Bayesian Update:





	P(Smoking)
Yes	θ_s
No	$1-\theta_s$

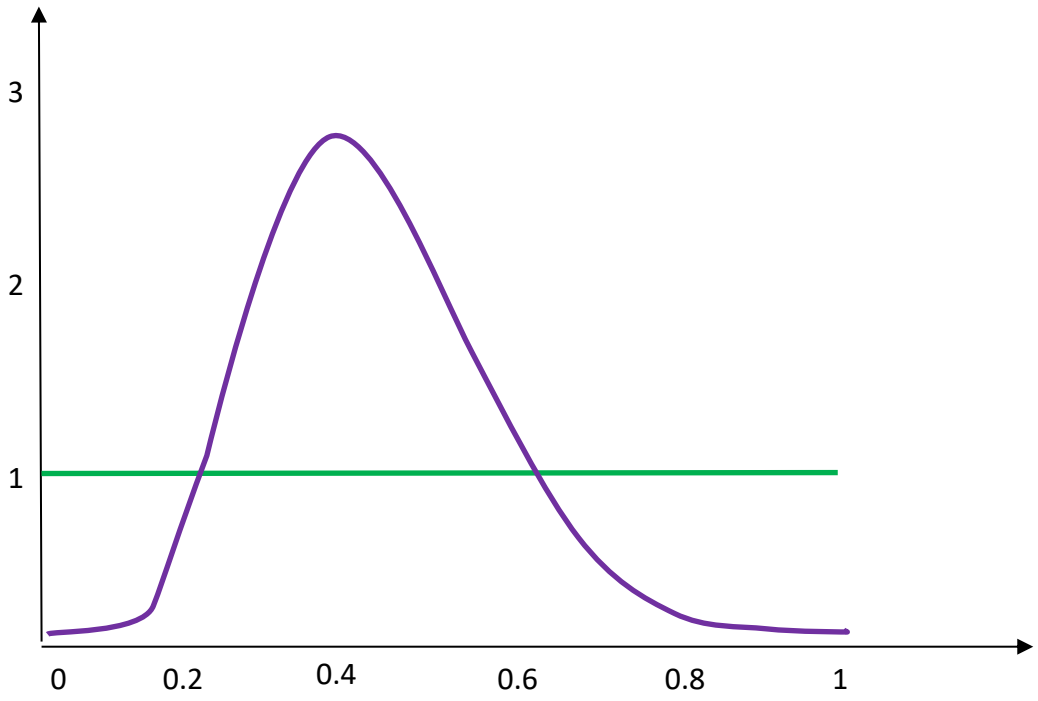
	P(CVD)	
Smoking	Yes	No
Yes	$\theta_{C\|S}$	$1-\theta_{C\|S}$
No	$\theta_{C\|NS}$	$1-\theta_{C\|NS}$

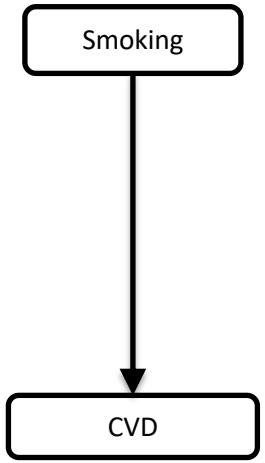
You now believe that the proportion of smokers to non smokers is close to 3:7

Bayesian Statistics.

Scoring function

You then observe 2 smokers and 6 non-smokers. Posterior:





	P(Smoking)
Yes	θ_s
No	$1-\theta_s$

	P(CVD)	
Smoking	Yes	No
Yes	$\theta_{C\|S}$	$1-\theta_{C\|S}$
No	$\theta_{C\|NS}$	$1-\theta_{C\|NS}$

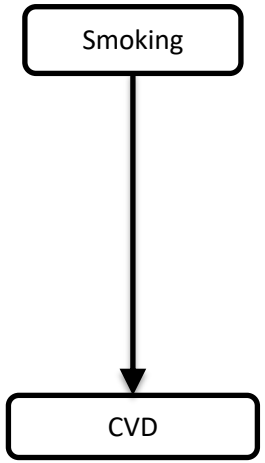
You now believe that the proportion of smokers to non smokers is close to 3:7

Bayesian Statistics.

Scoring function

$$\int_{\theta_S} P(D|\theta_s)\ f(\theta_s)d\theta_s = \int_{\theta_S} \prod_i (X_i|\theta_s)\, f(\theta_s)d\theta_s =$$

$$\frac{\Gamma(2)\Gamma(6)}{\Gamma(8)} = 0.0238$$



	P(Smoking)
Yes	θ_s
No	$1-\theta_s$

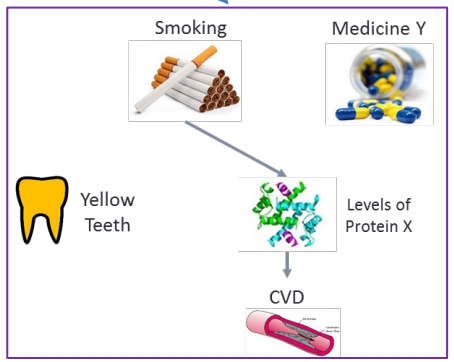
	P(CVD)	
Smoking	Yes	No
Yes	$\theta_{C\|S}$	$1-\theta_{C\|S}$
No	$\theta_{C\|NS}$	$1-\theta_{C\|NS}$

Computed in closed form!

Example Search Strategy (Greedy Search)

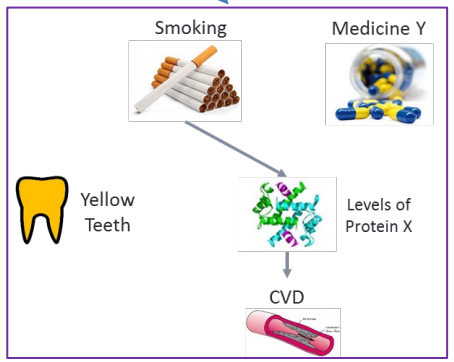
Initialize G as the empty/full/random graph and score.
Score all networks that can be produced by G with a single change: adding/removing/reversing an edge, ensuring G remains a DAG (no cycles). Keep the change that resulted in the highest-scoring network.
Until no single action improves the score.

Example Search Strategy (Greedy Search)



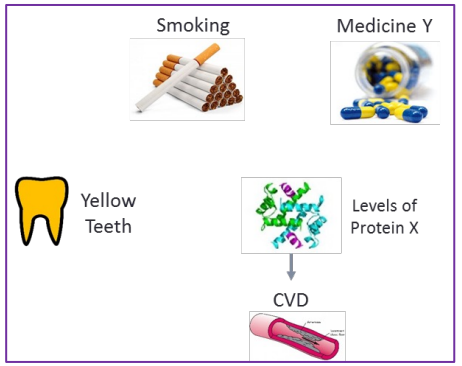
Score=-100

Example Search Strategy (Greedy Search)



Score=-100

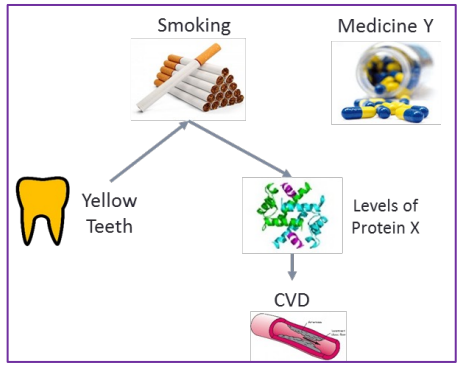
Remove Smoking → Protein X



Score=-104

…

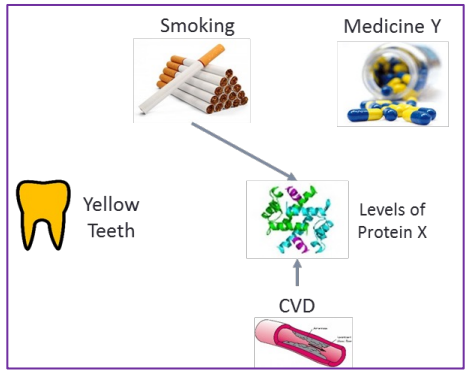
Add Yellow Teeth→ Smoking



Score=-90

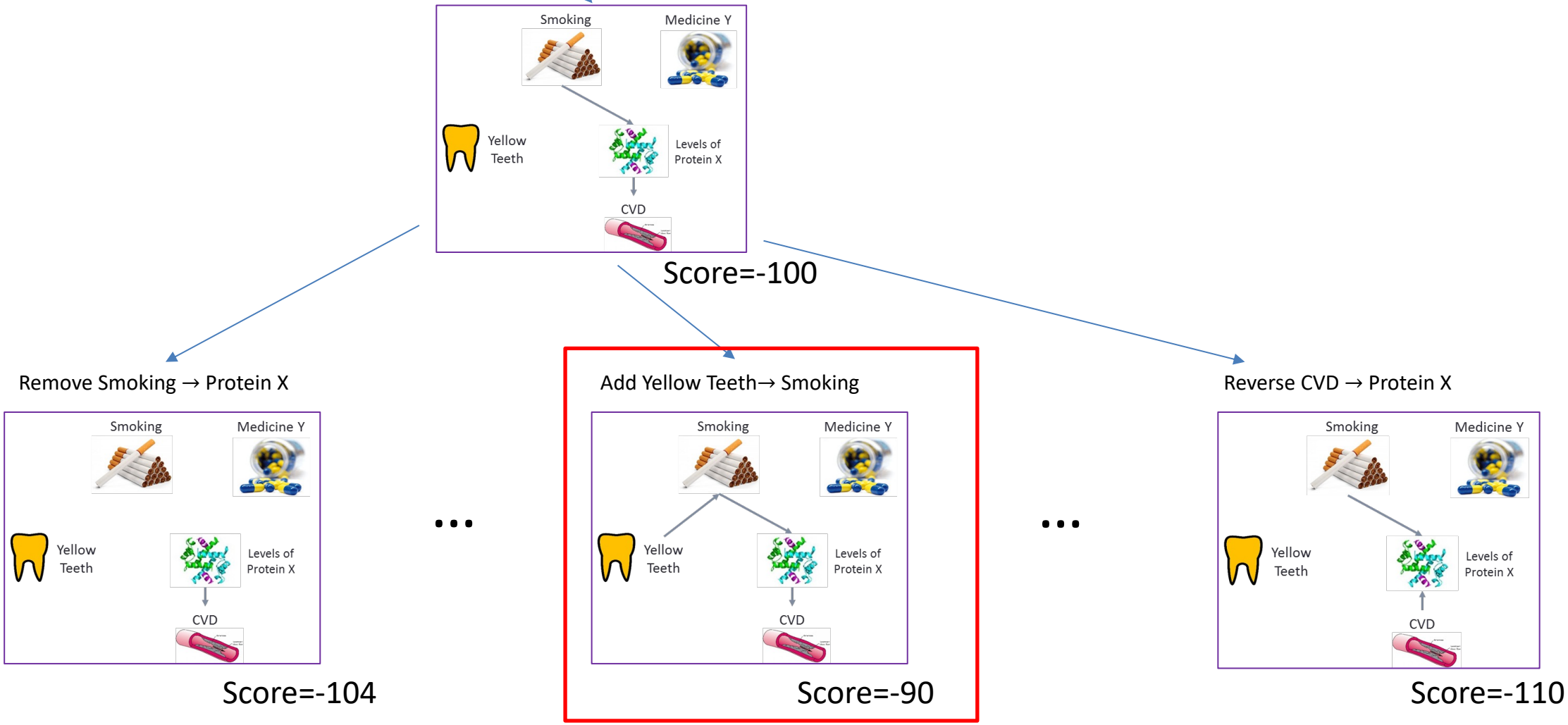
…

Reverse CVD → Protein X



Score=-110

Example Search Strategy (Greedy Search)



Example Search Strategy (Greedy Search)

Search-and-Score CBN learning

Other search strategies are possible.

e.g. BFS, DFS, Genetic algorithms, TABU search.

You can search in the space of PDAGs.

e.g. GES algorithm, (Chickering, 1996)

You may get stuck in local minima.

Avoid by random restarts, simulated annealing, stochastic greedy search.

Exact methods exist for actually scoring all possible networks (e.g. Koivisto and Sood, 2004)

Using dynamic programming & bounded number of parents per variable.

$O(n2^n)$ space + time complexity, not possible for more than ~20-40 variables.

Comparison

Constraint-Based

Easier to extend to different types of data (e.g., censored).

Easier to extend to networks with latent variables (next time).

More efficient in learning the skeleton of the network.

Search-and-score

Robust to small samples.

Easier to incorporate priors on the networks.

Better in identifying the edge orientations.

Exact methods also exist, limited to ~20-40 variables.

Study Material

P Spirtes, C Glymour, R Scheines. Causation, Prediction and Search, MIT press, 2001.

Cooper, Gregory F., and Edward Herskovits. "A Bayesian method for the induction of probabilistic networks from data." Machine learning 9.4 (1992): 309-347.

Carvallo, A.M. Scoring functions for learning Bayesian networks, INESC-ID Tec. Rep. 54/2009 (2009).

Tsamardinos, I., Brown, L. E. & Aliferis, C. F. The max-min hill-climbing Bayesian network structure learning algorithm. Mach. Learn. 65, 31–78 (2006).

Cooper, G. F. & Yoo, C. Causal Discovery from a Mixture of Experimental and Observational Data. *(UAI 1999)* **10,** 116–125 (1999).