Jin Gu
Department of Automation, Tsinghua University
Email: jgu@tsinghua.edu.cn
Phone: (010) 62794294-866

Chapter 2 Bayesian Networks: Representation

2021 Fall

Jin Gu (古槿)

Outlines

- Conditional independence
- Conditional parameterization
- Naïve Bayes model
- Bayesian networks
  - BNs and local independences
  - *I*-map and factorization
  - *d*-separation
  - From distribution to BNs

Textbook References

- Textbook 1
  - Chapter 3.1
  - Chapter 3.2
  - Chapter 3.3.1, 3.3.4, 3.4.1
- Textbook 2
  - Chapter 10.1
  - Chapter 10.2.1 ~ 10.2.3

These chapters are the ***minimal*** readings!

Decision with Probability

- When the variables you need to consider are very large, human brain will struggle to get an "optimal" decision.
- Why?

The parameters increase **exponentially** with the number of variables!

For binary variables: $\sim 2^n$ parameters

Representing Joint Distributions

- Random variables: $X_1, \ldots, X_n$
- P is a joint distribution over $X_1, \ldots, X_n$

If $X_1, \ldots, X_n$ binary, need 2^n-1 parameters to describe P

Can we **represent *P* more compactly**?

- Key: Exploit independence properties

Independent Random Variables

- Two variables X and Y are independent if
  - $P(X = x|Y = y) = P(X = x)$ for all values x, y
  - Equivalently, knowing Y does not change predictions of X

- If X and Y are independent then:
  - $P(X, Y) = P(X|Y)P(Y) = P(X)P(Y)$

- If $X_1, \dots, X_n$ are independent then:
  - $P(X_1, \dots, X_n) = P(X_1) \dots P(X_n)$
  - $O(n)$ parameters are needed
  - All 2^n probabilistic states are implicitly defined

This independent assumption is too strong to model complex problems!

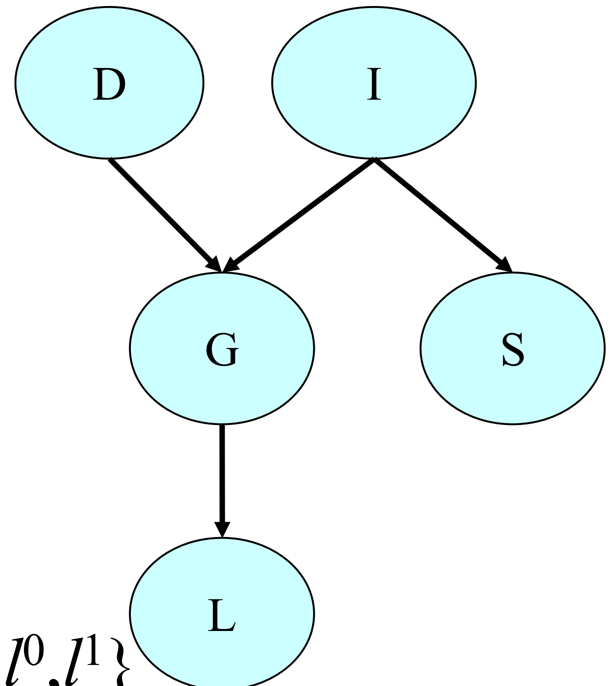
Conditional Independence

- Two variables X and Y are *conditionally independent* given Z, if:
  - $P(X = x|Y = y, Z = z) = P(X = x|Z = z)$ for all values x, y, z
  - Equivalently, if we know Z, then knowing Y does not change predictions of X
  - Notation: $Ind(X; Y \mid Z)$ or $(X \perp Y \mid Z)$

Can *conditional independence* reduce the required parameters?

Student Example

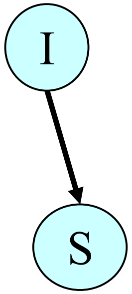
A concrete example to show that **independences can reduce the required parameters** for representing a distribution



- D = Course difficulty, Val(D)= $\{d^0,d^1\}$
- I = Intelligence, Val(I) = $\{i^0,i^1\}$
- S = Score on SAT, Val(S) = $\{s^0,s^1\}$
- G = Course grade, Val(G) = $\{g^0,g^1,g^2\}$
- L = Recommendation letter, Val(L) = $\{l^0,l^1\}$

- Assume that G and S are independent given I

Conditional Parameterization



- S = Score on SAT, Val(S) = $\{s^0, s^1\}$
- I = Intelligence, Val(I) = $\{i^0, i^1\}$

$P(I, S)$

I	***S***	***P(I,S)***
i^0	s^0	0.665
i^0	s^1	0.035
i^1	s^0	0.06
i^1	s^1	0.24

Joint parameterization

3 parameters

=

$P(I)$

I	
i^0	i^1
0.7	0.3

$P(S|I)$

	S	
I	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

Conditional parameterization

3 parameters

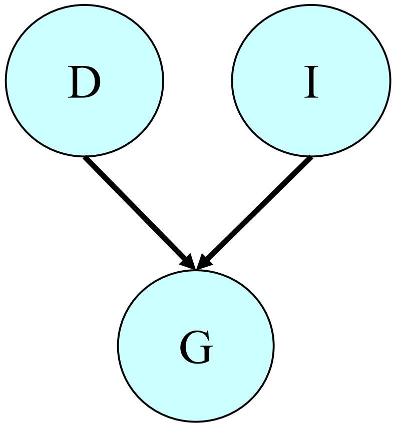
Alternative **conditional parameterization**: $P(S)$ and $P(I|S)$

Conditional Parameterization

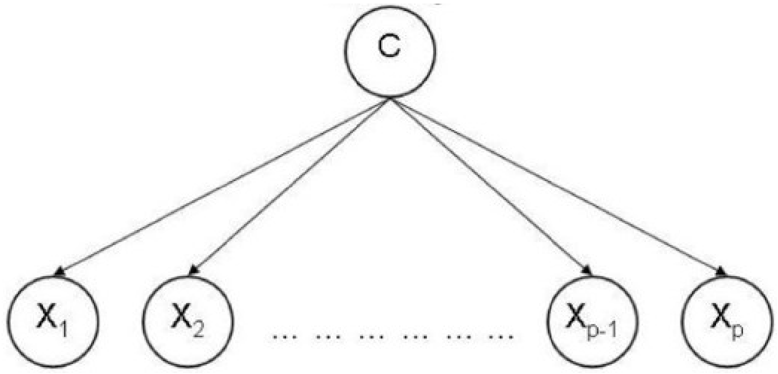
- I = Intelligence, Val(I) = $\{i^0, i^1\}$
- D = Difficulty, Val(D) = $\{d^0, d^1\}$
- G = Grade, Val(G) = $\{g^0, g^1, g^2\}$
- We know I and D are independent

Independences can reduce the required parameters

- Joint parameterization
  - $2 \cdot 2 \cdot 3 - 1 = 12 - 1 = 11$ independent parameters
- Conditional parameterization has
  - $P(D, I, G) = P(D)P(I)P(G|D, I)$
    - $P(D) \sim 1$ independent parameter
    - $P(I) \sim 1$ independent parameter
    - $P(G|D, I) \sim 4*(3-1) = 8$ independent parameters
  - 10 independent parameters



Naïve Bayes Model



- Class variable C, $Val(C) = \{c_1, \ldots, c_k\}$
- Evidence variables $X_1, \ldots, X_n$
- Naïve Bayes assumption: evidence variables are conditionally independent given C

$$P(C, X_1, \ldots, X_n) = P(C)\prod_{i=1}^{n} P(X_i \mid C)$$

- Applications in medical diagnosis, text classification
- Used as a classifier (i.e. $k=2$):

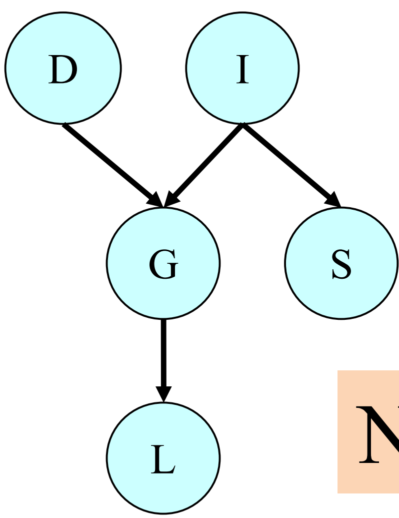
$$\frac{P(C=c_1 \mid x_1, \ldots, x_n)}{P(C=c_2 \mid x_1, \ldots, x_n)} = \frac{P(C=c_1)}{P(C=c_2)}\prod_{i=1}^{n}\frac{P(x_i \mid C=c_1)}{P(x_i \mid C=c_2)}$$

- Problem: Double counting correlated evidence. For example, fever and headache are highly correlated..

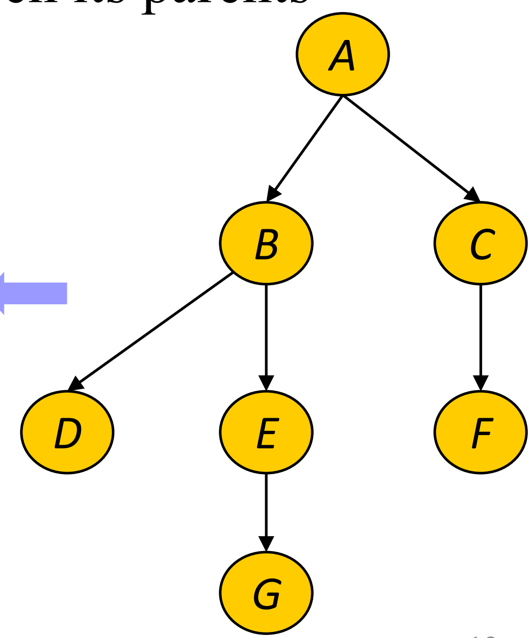
Bayesian Networks (Intuitive)

Can we find a simple graph model to equally or partially represent the probability with the same independences?

- Directed acyclic graph (DAG) G
  - Nodes $X_1, \dots, X_n$ represent random variables
- G encodes local independence assumptions
  - X_i is independent of its non-descendants given its parents
  - Formally: $(X_i \perp NonDesc(X_i) \mid Pa(X_i))$



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



Need formal proofs!!

Independency Mappings (*I*-Maps)

- *I*-Maps (Independence Maps)
  - Let P be a distribution over $\boldsymbol{X}$
  - Let $I(P)$ be the independencies in P
  - A Bayesian network is an *I*-map of P if $I(G) \subseteq I(P)$

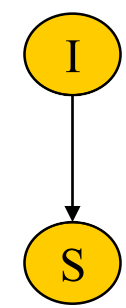
$I(G)=\{I \perp S\}$

I	***S***	***P(I,S)***
i^0	s^0	0.25
i^0	s^1	0.25
i^1	s^0	0.25
i^1	s^1	0.25

$I(P)=\{I \perp S\}$

I	***S***	***P(I,S)***
i^0	s^0	0.4
i^0	s^1	0.3
i^1	s^0	0.2
i^1	s^1	0.1

$I(P)=\varnothing$



$I(G)=\varnothing$

Factorization Theorem ***

If we define the independences in G as $X_i \perp NonDesc(X_i) \mid Pa(X_i)$

- G is an I-Map of P → $P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P(X_i \mid Pa(X_i))$

G is a given graph. If G is an I-Map of P, P can be factorized according to G.

- $P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P(X_i \mid Pa(X_i))$ → G is an I-Map of P

G is a given graph. If P can be factorized according to G, G is an I-Map of P.

If we define the independences in G as $X_i \perp NonDesc(X_i) | Pa(X_i)$

Proof: *I*-Map to Factorization

- If G is an *I*-Map of P, then

$$P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P(X_i \mid Pa(X_i))$$

Proof:

- wlog. $X_1, \ldots, X_n$ is an ordering consistent with G
- By chain rule: $P(X_1, \cdots, X_n) = \prod_{i=1}^{n} P(X_i | X_1, \cdots, X_{i-1})$
- From the ordering assumption
  - $Pa(X_i) \subseteq \{X_1, \cdots, X_{i-1}\}$ & $\{X_1, \cdots, X_{i-1}\} - Pa(X_i) \subseteq ND(X_i)$
- Since G is an *I*-Map → $(X_i \perp NonDesc(X_i) | Pa(X_i)) \subseteq I(P)$

$$\Downarrow$$

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

If we define the independences in G as $X_i \perp NonDesc(X_i) | Pa(X_i)$

Proof: Factorization Implies *I*-Map

- $P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P(X_i \mid Pa(X_i))$ → G is an *I*-Map of P
- Need to prove $(X_i \perp NonDesc(X_i) | Pa(X_i)) \subseteq I(P)$ or that $P(X_i \mid NonDesc(X_i)) = P(X_i \mid Pa(X_i))$

Proof:

- wlog. $X_1, \ldots, X_n$ is an ordering consistent with G

$$
\begin{aligned}
P(X_i \mid NonDesc(X_i)) &= \frac{P(X_i, NonDesc(X_i))}{P(NonDesc(X_i))} \\
&= \frac{\prod_{k=1}^{i} P(X_k \mid Pa(X_k))}{\prod_{k=1}^{i-1} P(X_k \mid Pa(X_k))} \\
&= P(X_i \mid Pa(X_i))
\end{aligned}
$$

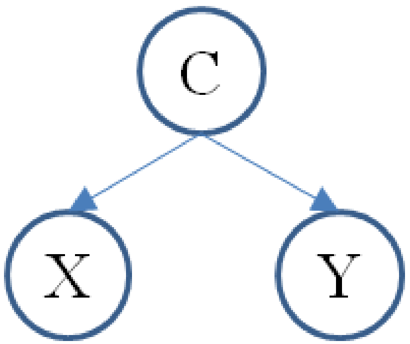
Formal Bayesian Network Definition

- A Bayesian network is a pair $\{P, G\}$
  - P factorizes over G
  - P is specified as set of conditional probability dependences (CPDs) associated with G's nodes

- Parameters
  - Joint distribution: $\sim 2^n$
  - Bayesian network (bounded in-degree k): $\sim n2^k$

How to Use BN Factorization?

- A simple example in paper assignment #1
  - $\mathrm{P}(X,Y,C)=\mathrm{P}(C)\mathrm{P}(X|C)\mathrm{P}(Y|C)$



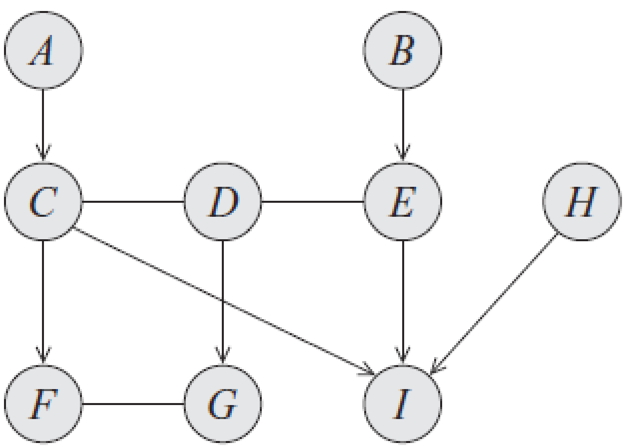
- We need to calculate $\mathrm{P}(X=x|Y=y)$
  - $\mathrm{P}(x|y) = \sum_c \mathrm{P}(x|y, C = c)\mathrm{P}(C|y)$
    - $\mathrm{P}(x|y, c) = \mathrm{P}(x|c)$
    - $\mathrm{P}(c|y) = \frac{\mathrm{P}(y|c)\mathrm{P}(c)}{\mathrm{P}(y)}$
      - $\mathrm{P}(y) = \sum_c \mathrm{P}(y|c)\mathrm{P}(c)$

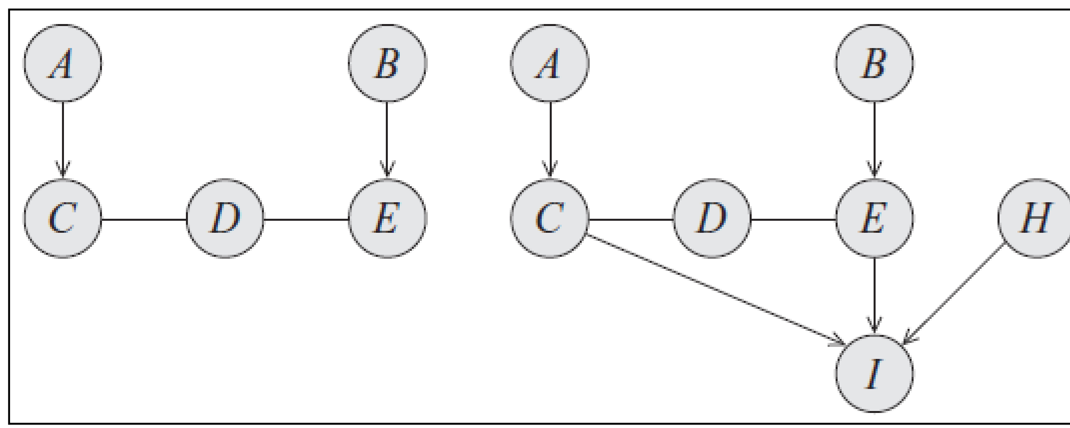
Try to transform the probability to the form of "local" probability as P(*child* | *Parents*)

How to Use BN Factorization?

Independence Inference??

- G encodes local independence assumptions
  - X_i is independent of its non-descendants given its parents (from the BN definition)
  - Formally: $(X_i \perp NonDesc(X_i) \mid Pa(X_i))$

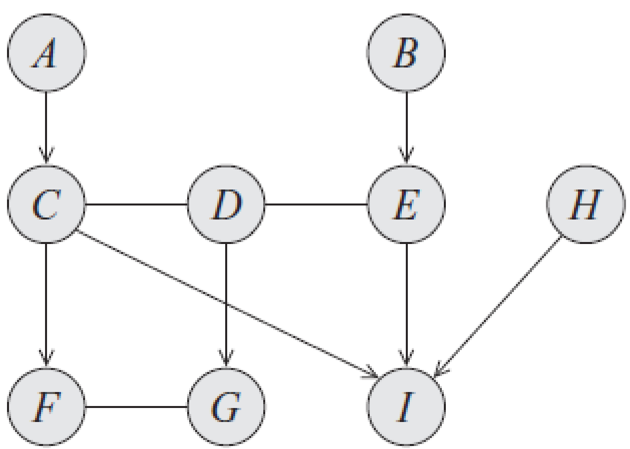


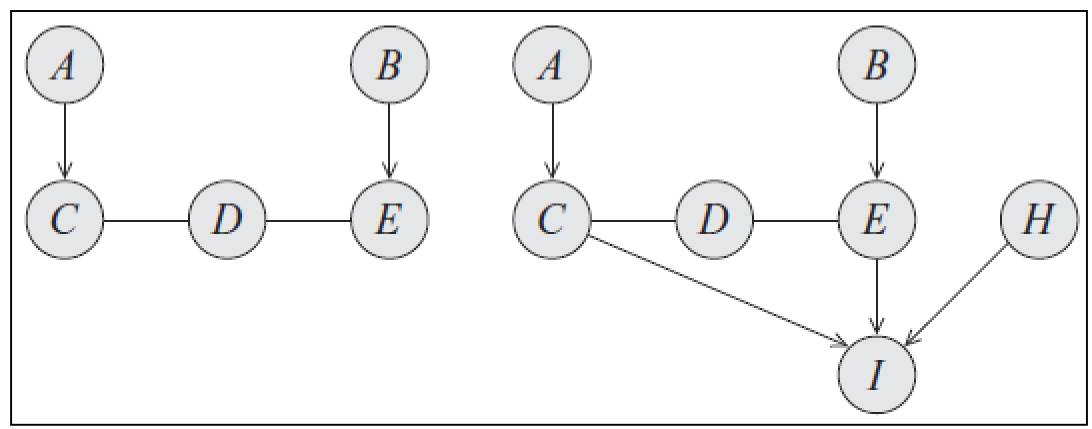


If the variables in ***upward closure*** are given, ...?

d-Separation in BNs

- G encodes local independence assumptions
  - X_i is independent of its non-descendants given its parents (from the BN definition)
  - Formally: $(X_i \perp NonDesc(X_i) \mid Pa(X_i))$





If the variables in ***upward closure*** are given, ...?

d-Separation in BNs

- G encodes local independence assumptions
  - X_i is independent of its non-descendants given its parents (from the BN definition)
  - Formally: $(X_i \perp NonDesc(X_i) \mid Pa(X_i))$

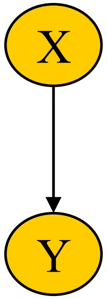
Does G encode other independence assumptions that hold in every distribution P that factorizes over G?

⬇

Devise a procedure to find all independencies in G

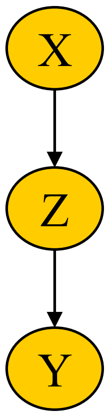
Not Separated: Direct Connection

- X and Y directly connected in G → no Z exists for which $Ind(X;Y \mid Z)$ holds in any factorizing distribution
  - Example: deterministic function



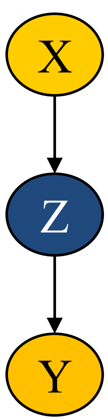
Not Separated: Indirect Connection

Active



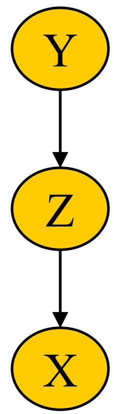
Case 1:

Indirect causal effect



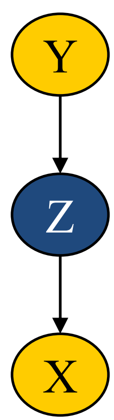
Blocked

Active



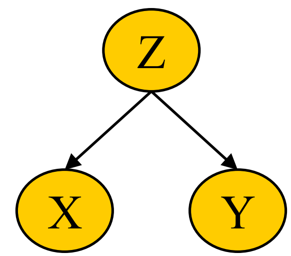
Case 2:

Indirect evidential effect



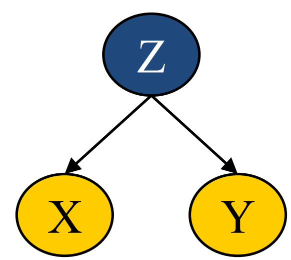
Blocked

Active



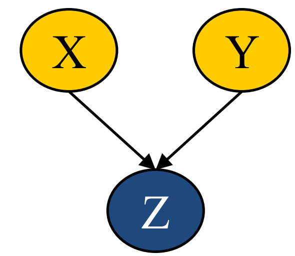
Case 3:

Common cause



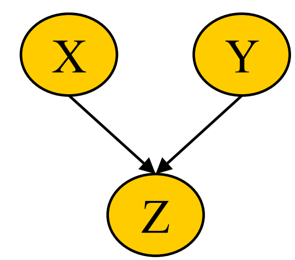
Blocked

Active



Case 4:

Common effect



Blocked

Not Separated: the General Case

- Let G be a Bayesian network structure
- Let $X_1 \leftrightarrow \ldots \leftrightarrow X_n$ be a trail in G
- Let $\boldsymbol{E}$ be a subset of evidence nodes in G

⇓

The trail $X_1 \leftrightarrow \ldots \leftrightarrow X_n$ is active given evidence E if:

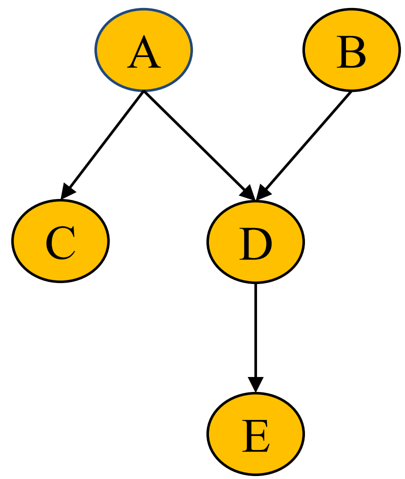
- For every V-structure $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$, X_i or one of its descendants is observed
- No other nodes along the trail are in $\boldsymbol{E}$

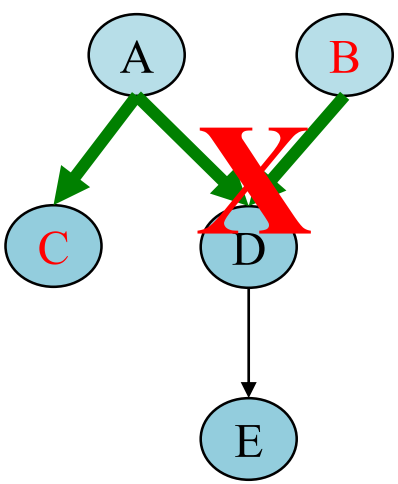
d-Separation

- $\boldsymbol{X}$ and $\boldsymbol{Y}$ are *d-separated* in G given $\boldsymbol{Z}$, denoted $d\text{-}sep_G(\boldsymbol{X}; \boldsymbol{Y}|\boldsymbol{Z})$ if there is no active trail between any node $X \in \boldsymbol{X}$ and any node $Y \in \boldsymbol{Y}$ in G

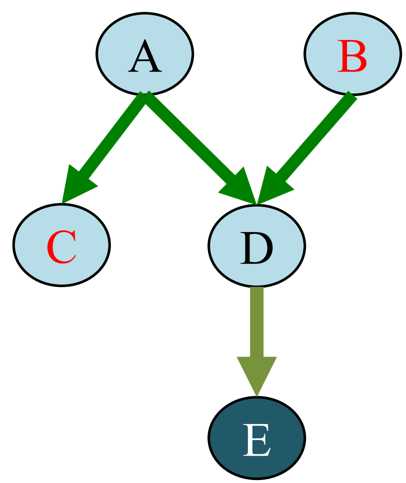
- Get all independences from d-separation
  - $I(G) = \{(\boldsymbol{X} \perp \boldsymbol{Y}|\boldsymbol{Z}) : d - sep_G(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{Z})\}$

d-Separation Examples

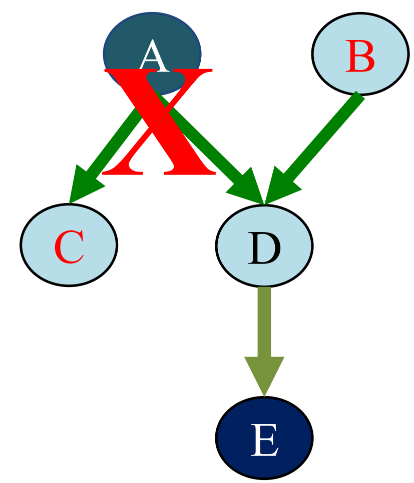




d-seq(B,C)=yes



d-sep(B,C|E)=no



d-sep(B,C|A,E)=yes

Forward vs Backward in BNs

- According to above examples, we should intuitively know that the *backward* process (given results, infer causes or learn parameters) is much harder than the *forward* process

- Why?
  - The child nodes will activate the trails (probabilistic dependences) between their parents and these dependences can propagate upward in the graphs

Algorithm for *d*-Separation

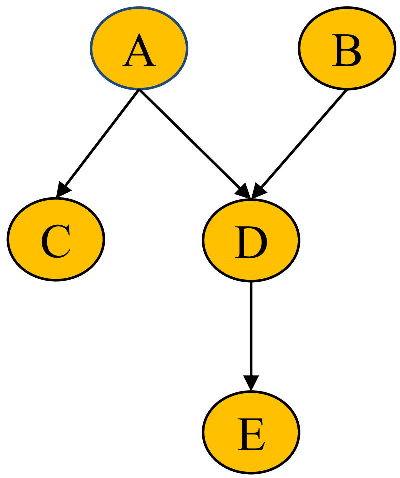
Procedure Reachable (
  $\mathcal{G}$, // Bayesian network graph
  X, // Source variable
  $\boldsymbol{Z}$ // Observations
)
  // Phase I: Insert all ancestors of $\boldsymbol{Z}$ into $\boldsymbol{A}$
  $\boldsymbol{L} \leftarrow \boldsymbol{Z}$ // Nodes to be visited
  $\boldsymbol{A} \leftarrow \emptyset$ // Ancestors of $\boldsymbol{Z}$
  while $\boldsymbol{L} \neq \emptyset$
    Select some Y from $\boldsymbol{L}$
    $\boldsymbol{L} \leftarrow \boldsymbol{L} - \{Y\}$
    if $Y \notin \boldsymbol{A}$ **then**
      $\boldsymbol{L} \leftarrow \boldsymbol{L} \cup \mathrm{Pa}_Y$ // Y's parents need to be visited
    $\boldsymbol{A} \leftarrow \boldsymbol{A} \cup \{Y\}$ // Y is ancestor of evidence

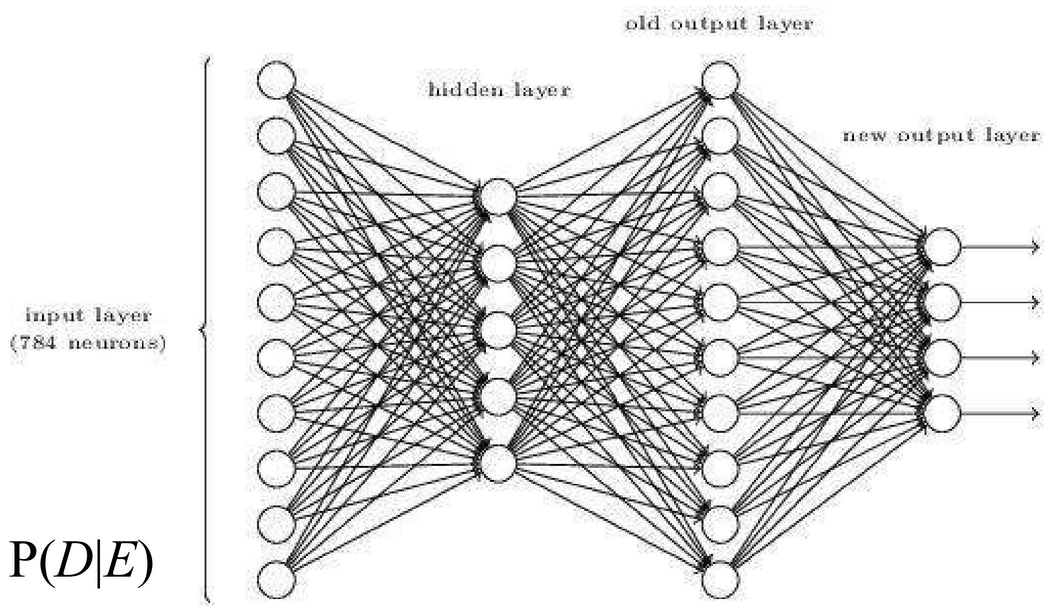
  // Phase II: traverse active trails starting from X
  $\boldsymbol{L} \leftarrow \{(X, \uparrow)\}$ // (Node,direction) to be visited
  $\boldsymbol{V} \leftarrow \emptyset$ // (Node,direction) marked as visited
  $\boldsymbol{R} \leftarrow \emptyset$ // Nodes reachable via active trail
  while $\boldsymbol{L} \neq \emptyset$
    Select some (Y, d) from $\boldsymbol{L}$
    $\boldsymbol{L} \leftarrow \boldsymbol{L} - \{(Y, d)\}$
    if $(Y, d) \notin \boldsymbol{V}$ **then**
      if $Y \notin \boldsymbol{Z}$ **then**
        $\boldsymbol{R} \leftarrow \boldsymbol{R} \cup \{Y\}$ // Y is reachable
      $\boldsymbol{V} \leftarrow \boldsymbol{V} \cup \{(Y, d)\}$ // Mark (Y, d) as visited
      if $d = \uparrow$ **and** $Y \notin \boldsymbol{Z}$ **then** // Trail up through Y active if Y not in $\boldsymbol{Z}$
        for each $Z \in \mathrm{Pa}_Y$
          $\boldsymbol{L} \leftarrow \boldsymbol{L} \cup \{(Z, \uparrow)\}$ // Y's parents to be visited from bottom
        for each $Z \in \mathrm{Ch}_Y$
          $\boldsymbol{L} \leftarrow \boldsymbol{L} \cup \{(Z, \downarrow)\}$ // Y's children to be visited from top
      else if $d = \downarrow$ **then** // Trails down through Y
        if $Y \notin \boldsymbol{Z}$ **then**
          // Downward trails to Y's children are active
          for each $Z \in \mathrm{Ch}_Y$
            $\boldsymbol{L} \leftarrow \boldsymbol{L} \cup \{(Z, \downarrow)\}$ // Y's children to be visited from top
        if $Y \in \boldsymbol{A}$ **then** // v-structure trails are active
          for each $Z \in \mathrm{Pa}_Y$
            $\boldsymbol{L} \leftarrow \boldsymbol{L} \cup \{(Z, \uparrow)\}$ // Y's parents to be visited from bottom
return $\boldsymbol{R}$

Aim: find all reachable nodes from ***X*** given ***Z***

Independences Ease Inferences

- Except less parameters, for inferences, you can only consider the *local dependent sub-graphs*



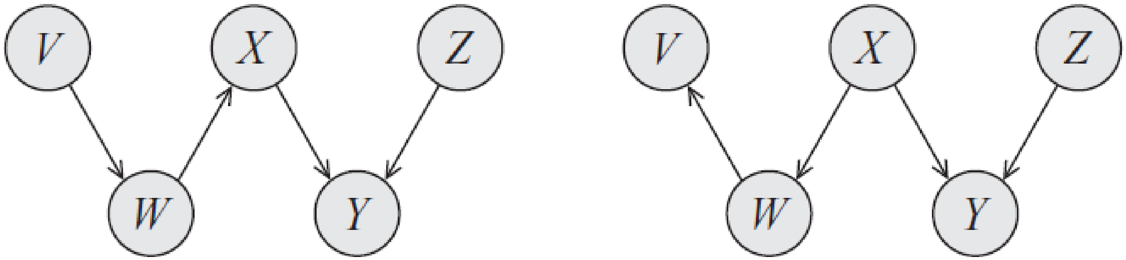


C is irrelevant for calculating P(*D*|*E*)

P(*A*), P(*B*), P(*D*|*A*,*B*), P(*E*|*D*) are needed

I-Equivalence

- Two graphs $G1$ and $G2$ are *I*-equivalent, if $I(G1) = I(G2)$. It means that the independences encoded by the two graphs should be the same.
- $G1$ and $G2$ have the same skeleton and the same set of immoralities (v-structures) if and only if they are I-equivalent.



From Distributions to BNs

- If P factorizes over G, we can derive a rich set of independence assertions that hold for P by simply examining G.
- Given a distribution P (a complex distribution hard to get the encoded independencies), to what extent can we construct a graph G **whose independencies** are a reasonable surrogate for the independencies in P?

Minimal *I*-Maps

- A graph ***G*** is a minimal *I*-map for a set of independences ***I*** if it is an *I*-map for ***I***, and if the removal of even a single edge from ***G*** renders it not an *I*-map. **Removal of an edge means additional independences!**

Procedure Build-Minimal-I-Map (
  $X_1, \ldots, X_n$ // an ordering of random variables in $\mathcal{X}$
  $\mathcal{I}$ // Set of independencies
)
  Set $\mathcal{G}$ to an empty graph over $\mathcal{X}$
  for $i = 1, \ldots, n$
    $U \leftarrow \{X_1, \ldots, X_{i-1}\}$ // U is the current candidate for parents of X_i
    for $U' \subseteq \{X_1, \ldots, X_{i-1}\}$
      if $U' \subset U$ and $(X_i \perp \{X_1, \ldots, X_{i-1}\} - U' \mid U') \in \mathcal{I}$ **then**
        $U \leftarrow U'$
      // At this stage U is a minimal set satisfying $(X_i \perp \{X_1, \ldots, X_{i-1}\} - U \mid U)$
      // Now set U to be the parents of X_i
    for $X_j \in U$
      Add $X_j \rightarrow X_i$ to $\mathcal{G}$
  return $\mathcal{G}$

Note: different initial orderings may generate different networks.

Basic idea: for *i*-th variable X_i, find the minimal sets of parents of X_i from the previous variables.

Perfect Maps

- A graph ***G*** is a perfect map (P-map) for a set of independences ***I*** if I(G) = ***I*** or ***I***(***G***) = ***I***(***P***). ***P*** is a distribution.
- Enumerate all independences in ***G*** and ***P*** to see weather ***G*** is a P-map of ***P***.

- *How a find a graph which is a P-map of distribution? (Please read textbook 3.4.3)*

Summary

- Independences can reduce the required parameters to represent a distribution

- ***Factorization theorem*** establish a mapping from a distribution and a graph

- Minimal *I*-Maps provide a possible way to find a graph representation of a distribution

Further Thinking

- Less links ⇔ more independences

- With the similar losses, do you prefer:
  - the models with fewer links?
  - the models with more links?

- Why?
- *Random dropouts in deep learning*..
- *Densely connected CNNs*..

Further Thinking: Causality / Intervention

- If the directed edges mean "*causal* effects"

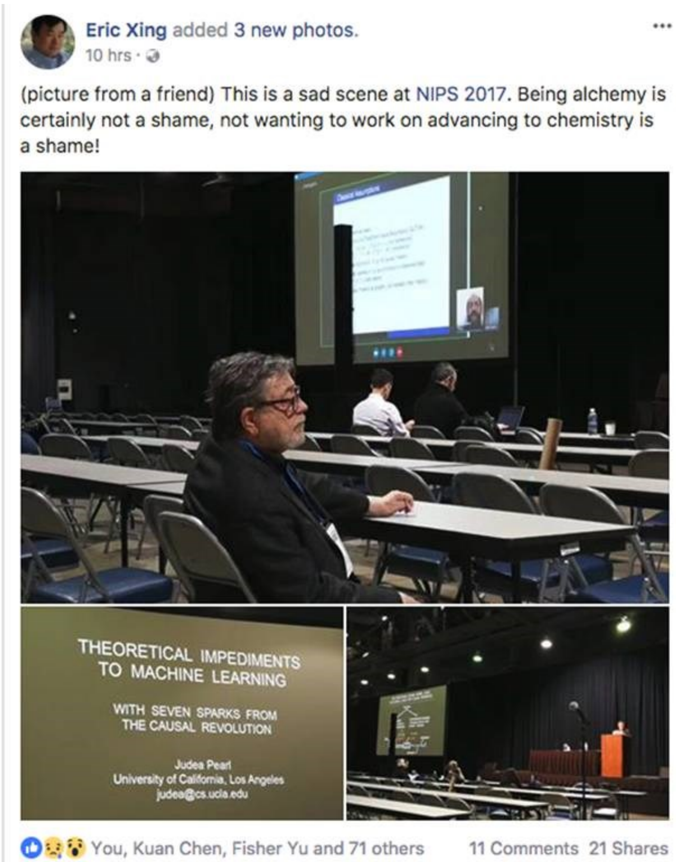
Theoretical Impediments to Machine Learning
With Seven Sparks from the Causal Revolution

Judea Pearl
University of California, Los Angeles
Computer Science Department
Los Angeles, CA, 90095-1596, USA
judea@cs.ucla.edu

January 15, 2018

因果 or 机理？

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?



For *causal edge*, *intervention* on child node can be regarded as "*delete the edge*"

- I-equivalence
  - If only consider dependence
  - X->Y ⇔ Y->X
- Causality
  - X: speed by a detector follows a conditional distribution $P(X|Y)$
  - Y: car speed follows a distribution $P(Y)$
  - Y->X is causal
  - If we *manually* set X=100, how about $P(Y|do(X=x))$?

Seven Pillars for Representing Causality

- Encoding causal assumptions – transparency and testability
- *Do*-calculus and the control of confounding
- The algorithmization of counterfactuals
- Mediation analysis and the assessment of direct and indirect effects
- External validity and sample selection bias
- Missing data
- Causal discovery

A Paradox for Drug Effect

- Does the drug take effect?

	Recovery	No recovery	Total
Drug	20	20	40
No drug	16	24	40
Total	36	44	80

- Please answer the question again!!

Females	Recovery	No recovery	Total
Drug	2	8	10
No drug	9	21	30
Total	11	29	40

Males	Recovery	No recovery	Total
Drug	18	12	30
No drug	7	3	10
Total	25	15	40