Causal Inference
CS 477-677

Statistical Models Of A Directed Acyclic Graph

Ilya Shpitser

Outline

Graphs And Probability

- Probability is great for reasoning consistently about uncertainty (unlike rule-based expert systems).
- However, probability is
  - Hard to think about because it's tables of numbers, or hard to visualize density functions.
  - Hard to reason about efficiently.

Graphs And Probability

- Probability is great for reasoning consistently about uncertainty (unlike rule-based expert systems).
- However, probability is
  - Hard to think about because it's tables of numbers, or hard to visualize density functions.
  - Hard to reason about efficiently.
- Humans are bad at tables of numbers or high dimensional curves.
- Humans are great at pictures!
- Graphical models use graphs to represent independence/irrelevance in probability distributions.
- Independence will help with efficiency, too!

Graphs And Probability

- Probability is great for reasoning consistently about uncertainty (unlike rule-based expert systems).
- However, probability is
  - Hard to think about because it's tables of numbers, or hard to visualize density functions.
  - Hard to reason about efficiently.
- Humans are bad at tables of numbers or high dimensional curves.
- Humans are great at pictures!
- Graphical models use graphs to represent independence/irrelevance in probability distributions.
- Independence will help with efficiency, too!
- Will discuss directed graph models in this class.

History: Undirected Models (Ising)

- Ernst Ising developed first undirected graph model for spin states in metals (1910).
- Iron itoms have two states: "up" and "down."
- Magnetized iron has most states pointing in the same direction.
- Atoms want to be like their neighbors.
- Ising's idea:

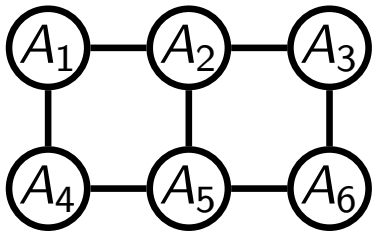
History: Undirected Models (Ising)

- Ernst Ising developed first undirected graph model for spin states in metals (1910).
- Iron itoms have two states: "up" and "down."
- Magnetized iron has most states pointing in the same direction.
- Atoms want to be like their neighbors.
- Ising's idea:
  - Vertices are atoms, edges connect neighbors. (2D/3D lattices are much harder).

  - Corresponding probability model: Atom state only depends on immediate neighbors.

History: Undirected Models (Ising)

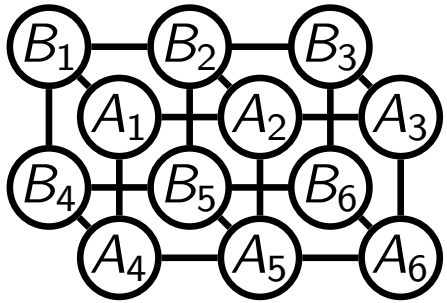
- Ernst Ising developed first undirected graph model for spin states in metals (1910).
- Iron itoms have two states: "up" and "down."
- Magnetized iron has most states pointing in the same direction.
- Atoms want to be like their neighbors.
- Ising's idea:
  - Vertices are atoms, edges connect neighbors. (2D/3D lattices are much harder).



  - Corresponding probability model: Atom state only depends on immediate neighbors.

History: Undirected Models (Ising)

- Ernst Ising developed first undirected graph model for spin states in metals (1910).
- Iron itoms have two states: "up" and "down."
- Magnetized iron has most states pointing in the same direction.
- Atoms want to be like their neighbors.
- Ising's idea:
  - Vertices are atoms, edges connect neighbors. (2D/3D lattices are much harder).



  - Corresponding probability model: Atom state only depends on immediate neighbors.

Markov Random Fields

- Undirected models are called Markov random fields (MRFs) today.
- Give probability of any configuration in terms of "local factors."

Markov Random Fields

- Undirected models are called Markov random fields (MRFs) today.
- Give probability of any configuration in terms of "local factors."
- For Ising and a graph $\mathcal{G}$,

$$p(a_1, a_2, \ldots, a_k) = \frac{1}{Z} \prod_{A_i \text{ a node in } \mathcal{G}} \phi(a_i) \prod_{(A_i, A_j) \text{ an edge in } \mathcal{G}} \phi(a_i, a_j).$$

- ϕ are not probabilities but map values of A_i to numbers, represent propensity of neighbors to have same value.
- Z normalizes so product is 1. Related to amount of energy in system for Ising.

Markov Random Fields

- Undirected models are called Markov random fields (MRFs) today.
- Give probability of any configuration in terms of "local factors."
- For Ising and a graph $\mathcal{G}$,

$$p(a_1, a_2, \ldots, a_k) = \frac{1}{Z} \prod_{A_i \text{ a node in } \mathcal{G}} \phi(a_i) \prod_{(A_i, A_j) \text{ an edge in } \mathcal{G}} \phi(a_i, a_j).$$

- ϕ are not probabilities but map values of A_i to numbers, represent propensity of neighbors to have same value.
- Z normalizes so product is 1. Related to amount of energy in system for Ising.
- High energy: atoms are disordered, p close to uniform.
- Low energy: atoms arrange as all "up" or all "down".
- If this happens quickly as energy is lowered, known as a phase transition.
- Model motivated originally by study of phase transitions.

Markov Random Fields

Main ideas for graphical models already present in Ising's model.

- A graph $\mathcal{G}(\vec{V})$ encodes independences of a probability distribution $p(\vec{V})$.
- $\vec{V}$: vertices are variables.
- If a vertex A is "blocked" from reaching a vertex B by a set of vertices $\vec{C}$ in $\mathcal{G}$, then $A \perp\!\!\!\perp B \mid \vec{C}$ holds in the distribution.
- For an MRF: "blocked" means any path from A to B intersects $\vec{C}$.
- This relationship is called a Markov property (there are many).
- Graph gives a factorization: $p(\vec{V})$ as a product of small pieces.
- Factorization and Markov properties equivalent views of model.

Markov Random Fields

Main ideas for graphical models already present in Ising's model.

- A graph $\mathcal{G}(\vec{V})$ encodes independences of a probability distribution $p(\vec{V})$.
- $\vec{V}$: vertices are variables.
- If a vertex A is "blocked" from reaching a vertex B by a set of vertices $\vec{C}$ in $\mathcal{G}$, then $A \perp\!\!\!\perp B \mid \vec{C}$ holds in the distribution.
- For an MRF: "blocked" means any path from A to B intersects $\vec{C}$.
- This relationship is called a Markov property (there are many).
- Graph gives a factorization: $p(\vec{V})$ as a product of small pieces.
- Factorization and Markov properties equivalent views of model.
- Ising example

On the board

Markov Random Fields Today

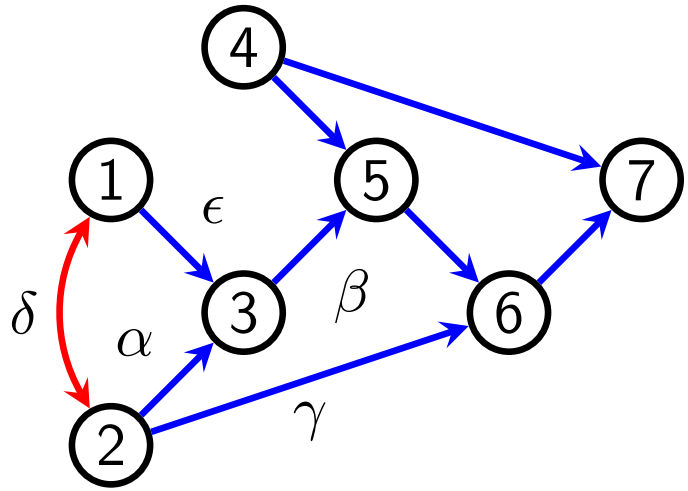
- Computer vision (vertices are pixels or features).
- Social network analysis.
- Association models for genomics data.
- Physics materials models, of course.
- Spacial statistics.
- Used in machine learning for efficient inference (junction trees, factor graphs, etc.)

History: Directed Models (Wright)

- Sewall Wright developed the first directed models for pedigree analysis for animals (1920s).
- Animals inherit traits from parents, probabilistically (see Mendel's Laws).

History: Directed Models (Wright)

- Sewall Wright developed the first directed models for pedigree analysis for animals (1920s).
- Animals inherit traits from parents, probabilistically (see Mendel's Laws).
- Question, formalize "degree of inbreeding":



- How inbred are x_5 and x_6?

Wright's Path Analysis

- Wright's model related variable Y and all X such that $X \rightarrow Y$ exists in graph $\mathcal{G}$, using a linear regression:

$$Y = w_0 + \sum_{X_i} w_i \cdot X_i + \epsilon_Y$$

- Coefficient w_i associated with $X_i \rightarrow Y$ in the graph.
- Allow correlations between noise terms: $\text{cov}(\epsilon_{Y_i}, \epsilon_{Y_j}) \neq 0$ (associated with $Y_i \leftrightarrow Y_j$).

Wright's Path Analysis

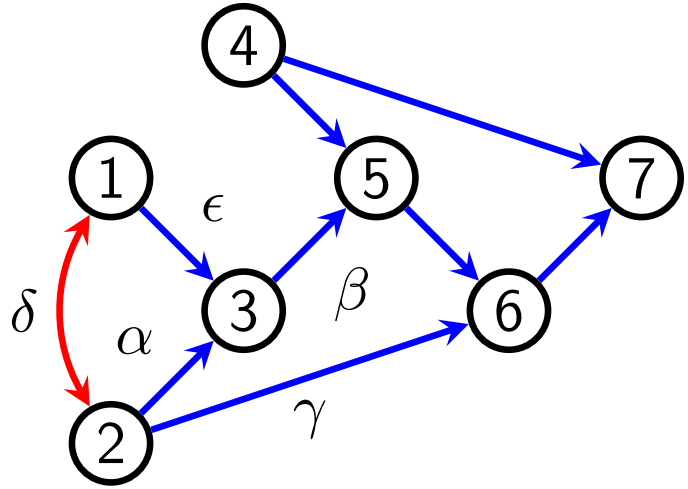
- Wright's model related variable Y and all X such that $X \rightarrow Y$ exists in graph $\mathcal{G}$, using a linear regression:

$$Y = w_0 + \sum_{X_i} w_i \cdot X_i + \epsilon_Y$$

- Coefficient w_i associated with $X_i \rightarrow Y$ in the graph.
- Allow correlations between noise terms: $\text{cov}(\epsilon_{Y_i}, \epsilon_{Y_j}) \neq 0$ (associated with $Y_i \leftrightarrow Y_j$).
- Inbreeding coefficient of X_5, X_6 as a measure of dependence in model.
- Add contributions from relevant paths in graph. Tracing relevant paths:
  - Can trace back then forward, but not forward and then back.
  - Pass through each variable only once.
  - At most one bidirected edge per path.
- Contribution of a path: multiply all coefficients along path.

Path Analysis Example

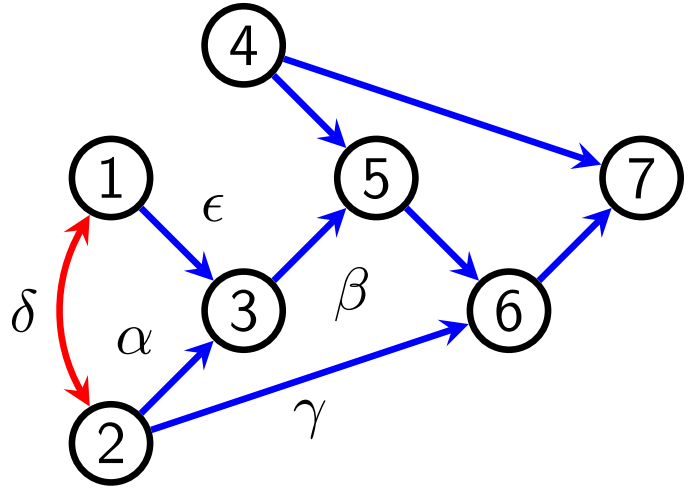
- How inbred are x_5 and x_6?



- Can trace back then forward, but not forward and then back.
- Pass through each variable only once.
- At most one bidirected edge per path.

Path Analysis Example

- How inbred are x_5 and x_6?



- Can trace back then forward, but not forward and then back.
- Pass through each variable only once.
- At most one bidirected edge per path.
- Relevant paths: $5 \leftarrow 3 \leftarrow 1 \leftrightarrow 2 \rightarrow 6$, $5 \leftarrow 3 \leftarrow 2 \rightarrow 6$.

$$\text{Inbreeding coefficient } f = \beta \cdot \epsilon \cdot \delta \cdot \gamma + \beta \cdot \alpha \cdot \gamma.$$

Structural Equation Models (SEMs)

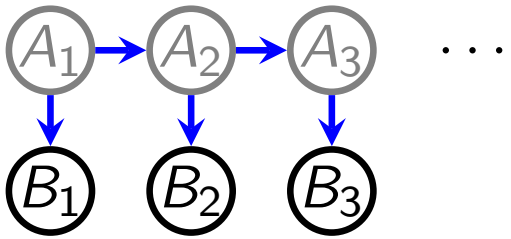
- For Wright, $\rightarrow$ were causal: parent traits cause child trains due to how gamete reproduction works.
- Model easy to fit without $\leftrightarrow$ (will discuss later).
- With $\leftrightarrow$, can use iterative methods (Drton et al).
- Path analysis came to be heavily used in social sciences and economics, due to (Haavelmo, 1943), and computer programs to do iterative fitting (1960s).

Structural Equation Models (SEMs)

- For Wright, $\rightarrow$ were causal: parent traits cause child trains due to how gamete reproduction works.
- Model easy to fit without $\leftrightarrow$ (will discuss later).
- With $\leftrightarrow$, can use iterative methods (Drton et al).
- Path analysis came to be heavily used in social sciences and economics, due to (Haavelmo, 1943), and computer programs to do iterative fitting (1960s).
- Applied to lots of areas, many not causal anymore.
- Loose terminology, eventually causal interpetation was abandoned.
- Still used today in the original form. Will talk about non-parametric version later.

Hidden Markov Models

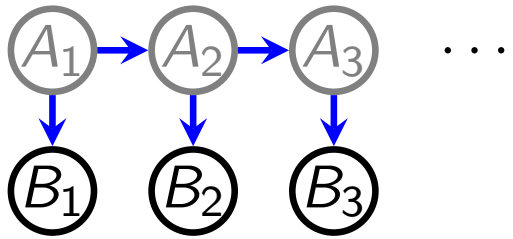
- Developed for speech processing in the 1960s (Stratonovich, Baum, etc.).
- Discrete hidden state of known complexity evolves in discrete time, we see a noisy version:



- Model $p(A_i \mid A_{i-1})$, and $p(B_i \mid A_i)$ (tables of numbers).

Hidden Markov Models

- Developed for speech processing in the 1960s (Stratonovich, Baum, etc.).
- Discrete hidden state of known complexity evolves in discrete time, we see a noisy version:

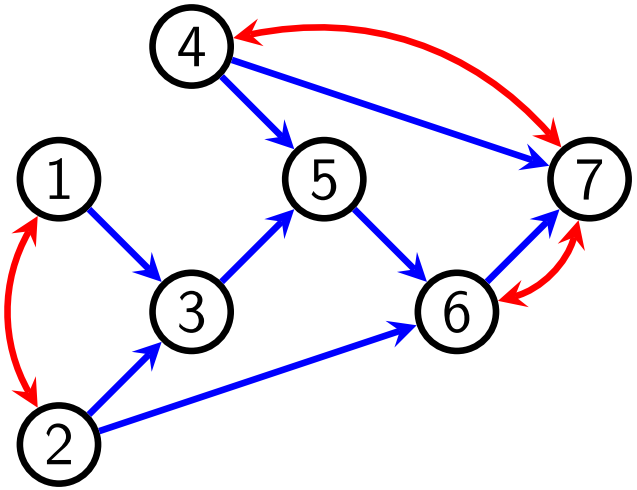


- Model $p(A_i \mid A_{i-1})$, and $p(B_i \mid A_i)$ (tables of numbers).
- Example: $A_1, A_2, \ldots$ are the words said, $B_1, B_2, \ldots$ is the speech sound wave.
- Interested in probable value assignments to $A_1, A_2, \ldots$ conditional on $b_1, b_2, \ldots$.
- Efficient algorithms (Viterbi) for this.

The Rise of Bayesian Networks

- Other special cases exist (Kalman filters, 1960s).
- Fusion began in 1980s after Pearl's *Probabilistic Reasoning in Intelligent Systems* book.
- Directed models are very heavily used in ML, statistics, public health, social sciences.
- Many applications.
- Unlike MRFs, models could be either statistical or causal.
- Will talk about statistical models today.

Graph Terms



- $\mathrm{pa}_{\mathcal{G}}(.)$: parents, $\mathrm{ch}_{\mathcal{G}}(.)$: children, $\mathrm{an}_{\mathcal{G}}(.)$: ancestors, $\mathrm{de}_{\mathcal{G}}(.)$: descendants, $\mathrm{sb}_{\mathcal{G}}(.)$: siblings.
- $1 \rightarrow 3$: $1 \in \mathrm{pa}_{\mathcal{G}}(3)$, $3 \in \mathrm{ch}_{\mathcal{G}}(1)$.
- $1 \rightarrow 3 \rightarrow 5 \rightarrow 6$: $1 \in \mathrm{an}_{\mathcal{G}}(6)$, $6 \in \mathrm{de}_{\mathcal{G}}(1)$.
- $1 \leftrightarrow 2$: $1 \in \mathrm{sb}_{\mathcal{G}}(2)$.
- Districts: connected sets via $\leftrightarrow$: $\{1, 2\}$, $\{4, 6, 7\}$, $\{3\}$, $\{5\}$.
- $6 \in \mathrm{dis}_{\mathcal{G}}(4)$.

Bayesian Networks

- Bayesian networks link a probability distribution and a directed acyclic graph.
- Directed: only $\rightarrow$ edges.
- Acyclic: if $X \in \mathrm{de}_{\mathcal{G}}(Y)$, $Y \notin \mathrm{ch}_{\mathcal{G}}(X)$.
- (Slightly imprecise term, but entrenched now).

Bayesian Networks

- Bayesian networks link a probability distribution and a directed acyclic graph.
- Directed: only $\rightarrow$ edges.
- Acyclic: if $X \in \mathrm{de}_{\mathcal{G}}(Y)$, $Y \notin \mathrm{ch}_{\mathcal{G}}(X)$.
- (Slightly imprecise term, but entrenched now).
- Big point of confusion:

  In a Bayesian network graph, $\rightarrow$ are not causal!

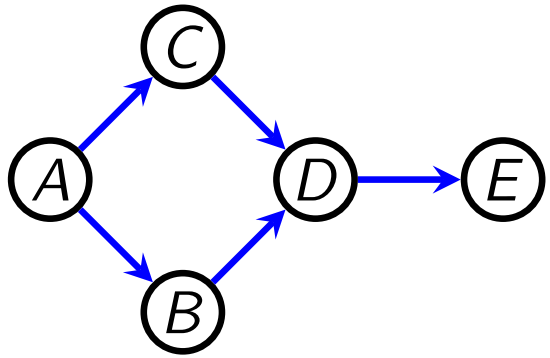
Bayesian Networks

- Bayesian networks link a probability distribution and a directed acyclic graph.
- Directed: only $\rightarrow$ edges.
- Acyclic: if $X \in \mathrm{de}_{\mathcal{G}}(Y)$, $Y \notin \mathrm{ch}_{\mathcal{G}}(X)$.
- (Slightly imprecise term, but entrenched now).
- Big point of confusion:

  In a Bayesian network graph, $\rightarrow$ are not causal!
- Three definitions (all involve the graph):
  - Factorization (probability distribution as a set of small factors).
  - Local Markov property (a small set of independence constraints).
  - Global Markov property (all independence constraints in the model).

Factorization

- Factorization has a term for a conditional distribution of a variable given its parents.



$$p(A, B, C, D, E) = p(E \mid D)p(D \mid B, C)p(C \mid A)p(B \mid A)p(A).$$

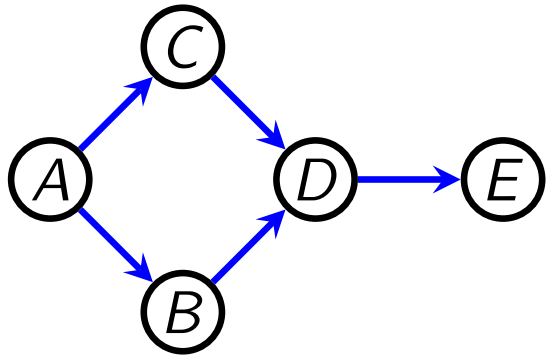
- Exactly what we did for SEMs and HMMs!
- If the graph has few edges, need few parameters.
- In a binary model need $2^5 - 1 = 31$ to specify LHS, but only $2^1 + 2^2 + 2^1 + 2^1 + 1 = 11$ to specify RHS.

Local Markov Property

- Graph implies a small list of independences that imply the rest.
- Every X is independent of non-parental non-descendants, conditional on parents.

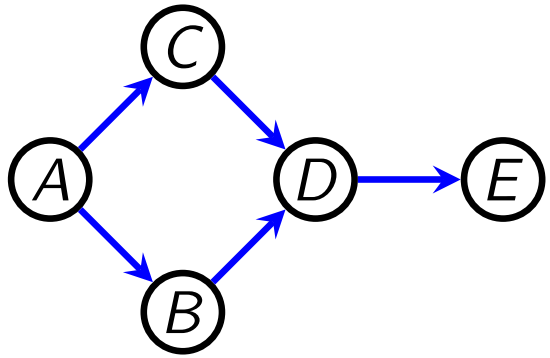
Local Markov Property

- Graph implies a small list of independences that imply the rest.
- Every X is independent of non-parental non-descendants, conditional on parents.



Local Markov Property

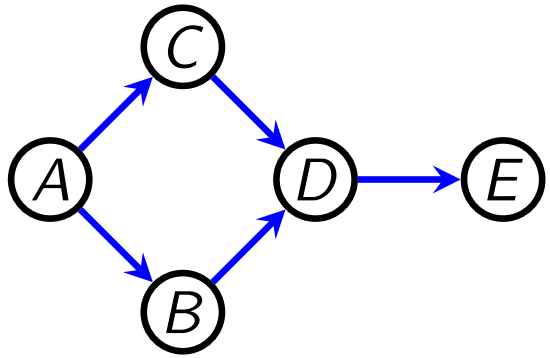
- Graph implies a small list of independences that imply the rest.
- Every X is independent of non-parental non-descendants, conditional on parents.



- $(C \perp\!\!\!\perp B \mid A)$, $(D \perp\!\!\!\perp A \mid B, C)$, $(E \perp\!\!\!\perp A, B, C \mid D)$.

Local Markov Property

- Graph implies a small list of independences that imply the rest.
- Every X is independent of non-parental non-descendants, conditional on parents.



- $(C \perp\!\!\!\perp B \mid A)$, $(D \perp\!\!\!\perp A \mid B, C)$, $(E \perp\!\!\!\perp A, B, C \mid D)$.
- Often expressed causal intuition: for any X only "direct causes" matter to specify it.
- Why is this not entirely right?

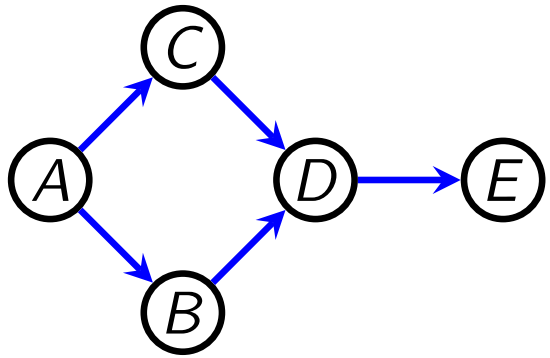
Observational Equivalence

- Consider the following two DAGs:

- Local Markov property gives same independence: $(A \perp\!\!\!\perp C \mid B)$.
- In fact, the only independence in this model.
- If one graph is causal, the other isn't...
- These graphs are called observationally equivalent.
- This is going to create a lot of problems for us later.

Global Markov Property

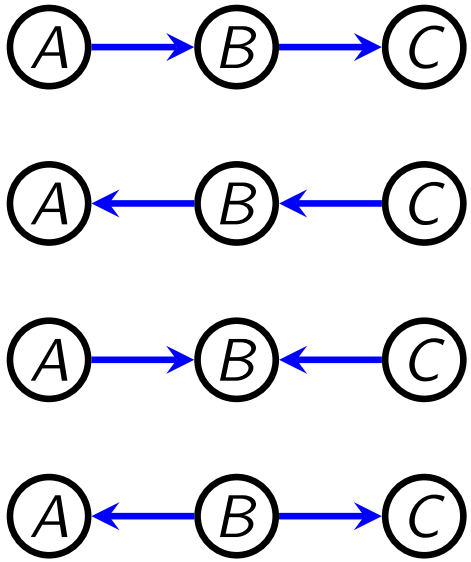
- Local Markov property specifies only a small set of $\perp\!\!\!\perp$.
- Can we read independences directly from graph, e.g. is A independent of E given C, D?



- In MRFs, to check if $A \perp\!\!\!\perp B \mid \vec{C}$ checked if $\vec{C}$ blocked any path from A to B.
- Will do the same here, but defining "blocked" will be harder.

Triplets On A Path

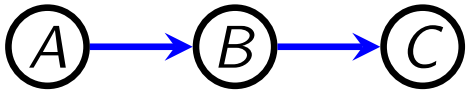
- In a directed graph, we have one of 4 possibilities for 3 consecutive vertices in a path:



- First two are directed, the third is called a collider, and the fourth is called a split or a fork.
- Need to think about how influence behaves along these.

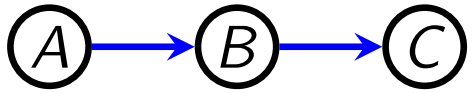
Directed Triplet

- Will have two cases, B is unobserved, and B is observed (conditioned on).

Unobserved case:  $A \not\perp\!\!\!\perp C$

Directed Triplet

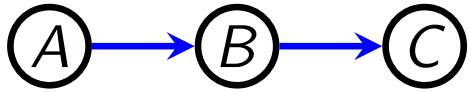
- Will have two cases, B is unobserved, and B is observed (conditioned on).

Unobserved case:  $A \not\perp\!\!\!\perp C$

- Wrong intuition: A influences B, and B influences C, therefore A influences C.
- Causal intuition: if B is a noisy version of A and C is a noisy version of B, then C is a noisy version of A. Example?
- Noncausal intuition: if arrows don't meet, edges act undirected.

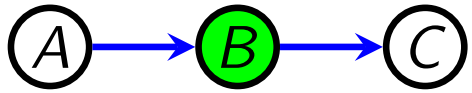
Directed Triplet

- Will have two cases, B is unobserved, and B is observed (conditioned on).

Unobserved case: 

$A \not\perp\!\!\!\perp C$

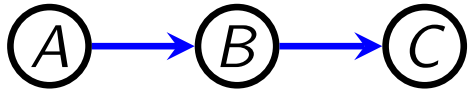
- Wrong intuition: A influences B, and B influences C, therefore A influences C.
- Causal intuition: if B is a noisy version of A and C is a noisy version of B, then C is a noisy version of A. Example?
- Noncausal intuition: if arrows don't meet, edges act undirected.

Observed case: 

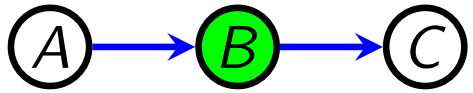
$A \perp\!\!\!\perp C \mid B$

Directed Triplet

- Will have two cases, B is unobserved, and B is observed (conditioned on).

Unobserved case:  $A \not\perp\!\!\!\perp C$

- Wrong intuition: A influences B, and B influences C, therefore A influences C.
- Causal intuition: if B is a noisy version of A and C is a noisy version of B, then C is a noisy version of A. Example?
- Noncausal intuition: if arrows don't meet, edges act undirected.

Observed case:  $A \perp\!\!\!\perp C \mid B$

- Acts like undirected path blocking.
- "Wiggles" in A translate into "wiggles" in C via "wiggles" B.
- If B is conditioned to a value, this stops happening. Example?

Split Triplet

- Will have two cases, as before.

Unobserved case: 

$A \not\perp\!\!\!\perp C$

Split Triplet

- Will have two cases, as before.

  Unobserved case: 

  $A \not\perp\!\!\!\perp C$

- Wrong intuition: A influences B, and B influences C, therefore A influences C.
- Causal intuition: if A and C share a common cause, they become dependent. Example?
- Noncausal intuition: if arrows don't meet, edges act undirected.

Split Triplet

- Will have two cases, as before.

Unobserved case: 

$A \not\perp\!\!\!\perp C$

- Wrong intuition: A influences B, and B influences C, therefore A influences C.
- Causal intuition: if A and C share a common cause, they become dependent. Example?
- Noncausal intuition: if arrows don't meet, edges act undirected.

Observed case: 

$A \perp\!\!\!\perp C \mid B$

Split Triplet

- Will have two cases, as before.

Unobserved case: 

$A \not\perp\!\!\!\perp C$

- Wrong intuition: A influences B, and B influences C, therefore A influences C.
- Causal intuition: if A and C share a common cause, they become dependent. Example?
- Noncausal intuition: if arrows don't meet, edges act undirected.

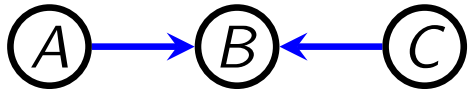
Observed case: 

$A \perp\!\!\!\perp C \mid B$

- Acts like undirected path blocking.
- "Wiggles" in A translate into "wiggles" in C via "wiggles" B.
- If B is conditioned to a value, this stops happening. Example?

Collider Triplet

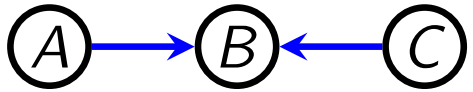
- Will have two cases, as before.

Unobserved case: 

$A \perp\!\!\!\perp C$

Collider Triplet

- Will have two cases, as before.

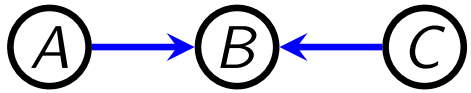
  Unobserved case: 

  $A \perp\!\!\!\perp C$

- Causal intuition: two independent causes of an effect. Example?

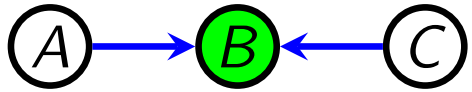
Collider Triplet

- Will have two cases, as before.

  Unobserved case: 

  $A \perp\!\!\!\perp C$

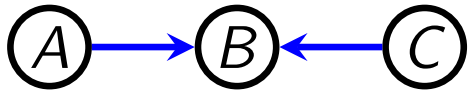
- Causal intuition: two independent causes of an effect. Example?

  Observed case: 

  $A \not\perp\!\!\!\perp C \mid B$

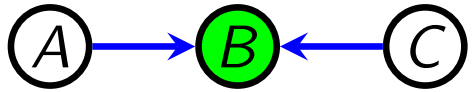
Collider Triplet

- Will have two cases, as before.

Unobserved case: 

$A \perp\!\!\!\perp C$

- Causal intuition: two independent causes of an effect. Example?

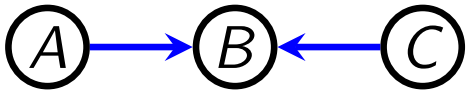
Observed case: 

$A \not\perp\!\!\!\perp C \mid B$

- Causal intuition: knowing a shared effect create dependence between causes (NBA example).
- Hard to think of a non-causal intuition, colliders often arise in causal systems. Can we think of a non-causal example?

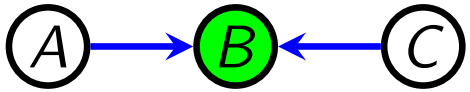
Collider Triplet

- Will have two cases, as before.

Unobserved case: 

$A \perp\!\!\!\perp C$

- Causal intuition: two independent causes of an effect. Example?

Observed case: 

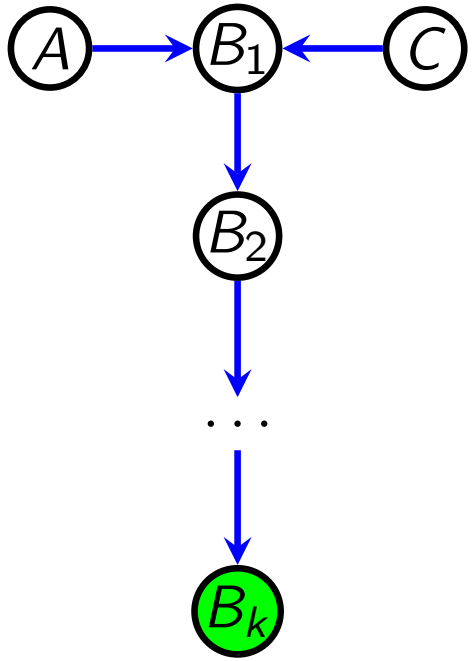
$A \not\perp\!\!\!\perp C \mid B$

- Causal intuition: knowing a shared effect create dependence between causes (NBA example).
- Hard to think of a non-causal intuition, colliders often arise in causal systems. Can we think of a non-causal example?
- Berkson's bias/paradox: if edges meet, dependence behaves in the opposite way from undirected.
- A can depend on B, and B on C, but A and C could be independent.
- Conditioning on B can create dependence, not just remove it.

Important Note On Colliders

- Do not have to condition on B, can condition on descendant:

Observed case:

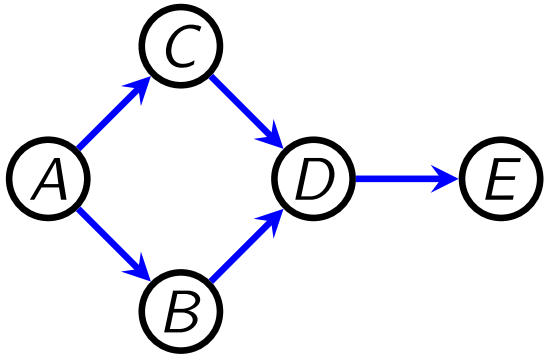


$A \not\perp\!\!\!\perp C \mid B_k$

- Can just extend the NBA example to observed consequences of being in the NBA.

From Path Blocking To d-separation

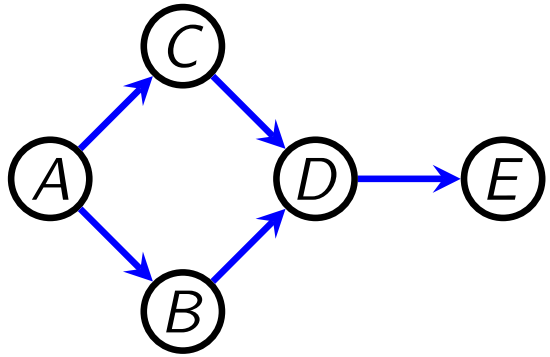
- Will say a path from A to B is blocked by $\vec{C}$ if there is a blocking triplet on the path.
- Intuition: dependence is water flow, paths are pipes. A single block is enough.
- A and B is said to be d-separated given $\vec{C}$ if all paths from A to B are blocked by $\vec{C}$ in this way. Examples:



- $A \perp\!\!\!\perp E \mid C$?
- $C \perp\!\!\!\perp B \mid A$?
- $C \perp\!\!\!\perp B \mid E, A$?

From Path Blocking To d-separation

- Will say a path from A to B is blocked by $\vec{C}$ if there is a blocking triplet on the path.
- Intuition: dependence is water flow, paths are pipes. A single block is enough.
- A and B is said to be d-separated given $\vec{C}$ if all paths from A to B are blocked by $\vec{C}$ in this way. Examples:



- $A \perp\!\!\!\perp E \mid C$? No: $A \rightarrow B \rightarrow D \rightarrow E$ not blocked.
- $C \perp\!\!\!\perp B \mid A$? Yes: $C \leftarrow A \rightarrow B$ and $C \rightarrow D \leftarrow B$ blocked.
- $C \perp\!\!\!\perp B \mid E, A$? No: $C \rightarrow D \leftarrow B$ open because of E.

Global Markov Property

- Will write $(A \perp\!\!\!\perp_d B \mid \vec{C})_{\mathcal{G}}$ to denote "A is d-separated from B given $\vec{C}$ in $\mathcal{G}$."
- Extends to sets: $(\vec{A} \perp\!\!\!\perp_d \vec{B} \mid \vec{C})_{\mathcal{G}}$ if for all $A \in \vec{A}, B \in \vec{B}$, $(A \perp\!\!\!\perp_d B \mid \vec{C})_{\mathcal{G}}$.
- Global Markov property for a Bayesian network model with DAG $\mathcal{G}(\vec{V})$:

$$(\vec{A} \perp\!\!\!\perp_d \vec{B} \mid \vec{C})_{\mathcal{G}(\vec{V})} \Rightarrow (\vec{A} \perp\!\!\!\perp \vec{B} \mid \vec{C})_{p(\vec{V})}.$$

- One way implication! Could have extra independences in $p(\vec{V})$.
- Example (3-chain).

Global Markov Property

- Will write $(A \perp\!\!\!\perp_d B \mid \vec{C})_{\mathcal{G}}$ to denote "A is d-separated from B given $\vec{C}$ in $\mathcal{G}$."
- Extends to sets: $(\vec{A} \perp\!\!\!\perp_d \vec{B} \mid \vec{C})_{\mathcal{G}}$ if for all $A \in \vec{A}, B \in \vec{B}$, $(A \perp\!\!\!\perp_d B \mid \vec{C})_{\mathcal{G}}$.
- Global Markov property for a Bayesian network model with DAG $\mathcal{G}(\vec{V})$:

$$(\vec{A} \perp\!\!\!\perp_d \vec{B} \mid \vec{C})_{\mathcal{G}(\vec{V})} \Rightarrow (\vec{A} \perp\!\!\!\perp \vec{B} \mid \vec{C})_{p(\vec{V})}.$$

- One way implication! Could have extra independences in $p(\vec{V})$.
- Example (3-chain).
- Distributions where implication is two way are very special and useful.
- Will have much to say about them later.

Equivalent Definitions

Important theorem:

Theorem (Verma and Pearl)

Given a DAG $\mathcal{G}(\vec{V})$, a distribution $p(\vec{V})$ factorizes according to $\mathcal{G}$ if and only if it obeys the local Markov property according to $\mathcal{G}$ if and only if it obeys the global Markov property according to $\mathcal{G}$.

Equivalent Definitions

Important theorem:

Theorem (Verma and Pearl)

Given a DAG $\mathcal{G}(\vec{V})$, a distribution $p(\vec{V})$ factorizes according to $\mathcal{G}$ if and only if it obeys the local Markov property according to $\mathcal{G}$ if and only if it obeys the global Markov property according to $\mathcal{G}$.

- Why is this true?

Equivalent Definitions

Important theorem:

Theorem (Verma and Pearl)

Given a DAG $\mathcal{G}(\vec{V})$, a distribution $p(\vec{V})$ factorizes according to $\mathcal{G}$ if and only if it obeys the local Markov property according to $\mathcal{G}$ if and only if it obeys the global Markov property according to $\mathcal{G}$.

- Why is this true?
- Global $\Rightarrow$ local is easy (properties of d-separation).

Equivalent Definitions

Important theorem:

Theorem (Verma and Pearl)

Given a DAG $\mathcal{G}(\vec{V})$, a distribution $p(\vec{V})$ factorizes according to $\mathcal{G}$ if and only if it obeys the local Markov property according to $\mathcal{G}$ if and only if it obeys the global Markov property according to $\mathcal{G}$.

- Why is this true?
- Global $\Rightarrow$ local is easy (properties of d-separation).
- Local $\Rightarrow$ factorization is easy (chain rule using topological order, and use local).

Equivalent Definitions

Important theorem:

Theorem (Verma and Pearl)

Given a DAG $\mathcal{G}(\vec{V})$, a distribution $p(\vec{V})$ factorizes according to $\mathcal{G}$ if and only if it obeys the local Markov property according to $\mathcal{G}$ if and only if it obeys the global Markov property according to $\mathcal{G}$.

- Why is this true?
- Global $\Rightarrow$ local is easy (properties of d-separation).
- Local $\Rightarrow$ factorization is easy (chain rule using topological order, and use local).
- Factorization $\Rightarrow$ global is hard (easiest proof in Lauritzen's *Graphical Models* book).

Statistical Vs Causal Models

- A statistical model is formally a set of distributions. For example:

$$\left\{ p(\vec{V}) \middle| (\forall \vec{A} \dot{\cup} \vec{B} \dot{\cup} \vec{C} \in \vec{V})(\vec{A} \perp\!\!\!\perp_d \vec{B} \mid \vec{C})_{\mathcal{G}} \Rightarrow (\vec{A} \perp\!\!\!\perp \vec{B} \mid \vec{C})_{p(\vec{V})} \right\}$$

- This is talking about $p(\vec{V})$, the observed data distribution.
- Nothing about potential experiments.
- Nothing about causality.
- May use causal intuitions, but these are informal.
- Need to represent "directly causes" formally.
- Next: causal models, as sets of distributions on counterfactual and factual random variables.

Next time: Causal Models Of A DAG.