Probabilistic and Bayesian Machine Learning

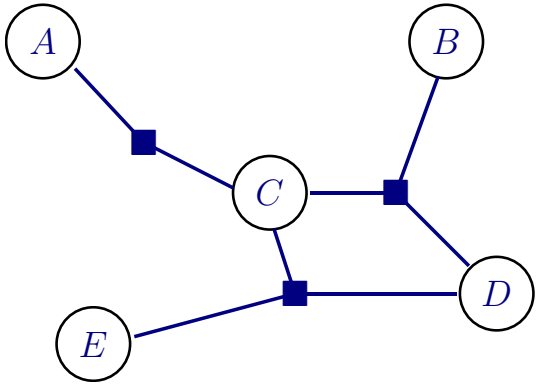
Lecture 2: Graphical Models

Yee Whye Teh
ywteh@gatsby.ucl.ac.uk

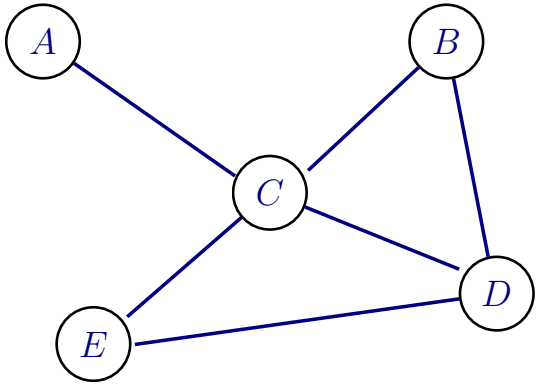
Gatsby Computational Neuroscience Unit
University College London

http://www.gatsby.ucl.ac.uk/∼ywteh/teaching/probmodels

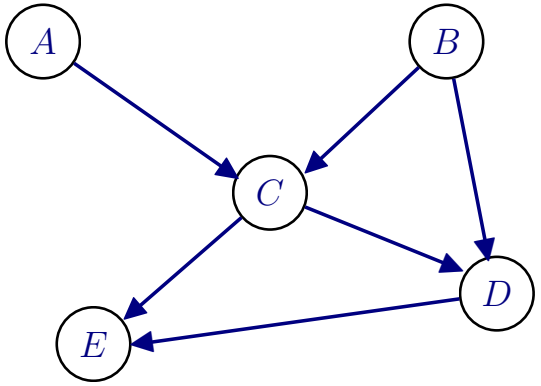
Graphical Models



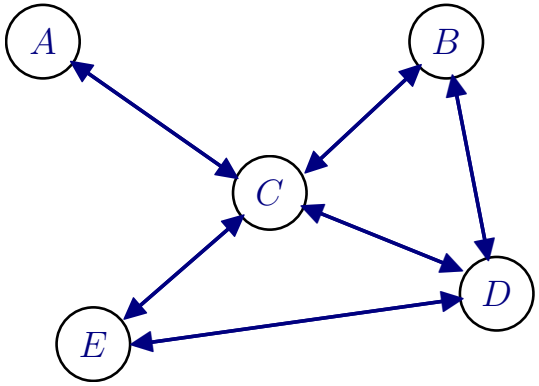
factor graph



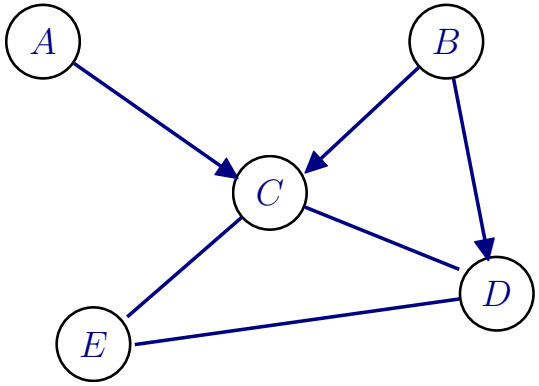
undirected graph



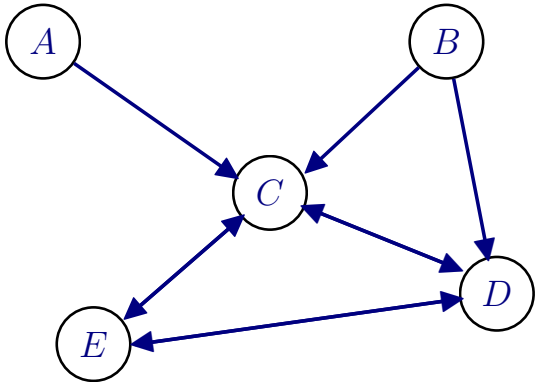
directed graph



bidirected graph



chain graph



mixed graph

Nodes in graph correspond to random variables.

Edges in graph correspond to statistical dependencies between the corresponding variables.

(Absence of edges correspond to conditional independencies between variables).

Why Do We Need Graphical Models?

- Graphs are an intuitive way of representing and visualising the relationships between many variables. (Examples: family trees, electric circuit diagrams, neural networks)

- Graphical models are a precise language to describe these relationships between variables.

- Graphical models allow us to build complex probabilistic models out of simpler building blocks.

- A graph allows us to abstract out the conditional independence relationships between the variables from the details of their parametric forms. Thus we can ask questions like: "Is A dependent on B given that we know the value of C ?" just by looking at the graph.

- Graphical models allow us to define general-purpose message-passing algorithms that implement Bayesian inference efficiently. Thus we can answer queries like "What is $P(A|C = c)$?" without enumerating all settings of all variables in the model.

Conditional Independence

Conditional Independence:

$$X \perp\!\!\!\perp Y|V \quad \Leftrightarrow \quad P(X|Y,V) = P(X|V)$$

when $P(Y,V) > 0$. Also

$$X \perp\!\!\!\perp Y|V \quad \Leftrightarrow \quad P(X,Y|V) = P(X|V)P(Y|V)$$

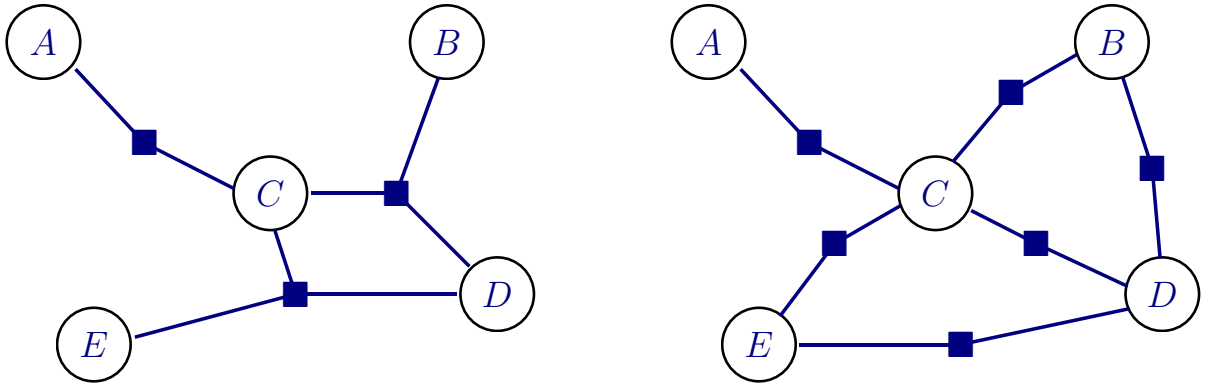
In general we can think of conditional independence between **sets of variables**:

$$\mathcal{X} \perp\!\!\!\perp \mathcal{Y}|\mathcal{V} \quad \Leftrightarrow \quad \{X \perp\!\!\!\perp Y|\mathcal{V}, \ \ \forall X \in \mathcal{X} \ \text{ and } \ \forall Y \in \mathcal{Y}\}$$

Marginal Independence:

$$X \perp\!\!\!\perp Y \quad \Leftrightarrow \quad X \perp\!\!\!\perp Y|\emptyset \quad \Leftrightarrow \quad P(X,Y) = P(X)P(Y)$$

Factor Graphs



(a) (b)

The circles in a factor graph represent random variables.
The filled squares represent factors in the joint distribution.

(a) $P(A, B, C, D, E) = \frac{1}{Z} f_1(A, C) f_2(B, C, D) f_3(C, D, E)$

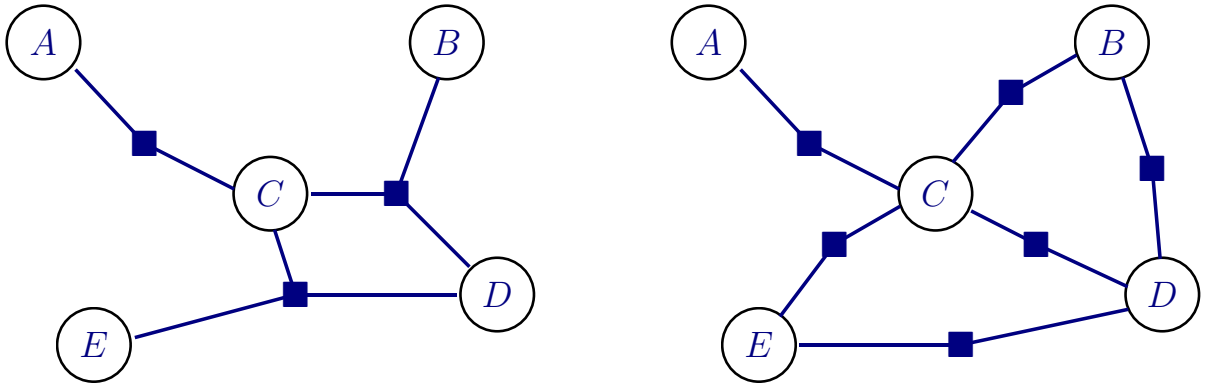
(b) $P(A, B, C, D, E) = \frac{1}{Z} f_1(A, C) f_2(B, C) f_3(C, D) f_4(B, D) f_5(C, E) f_6(D, E)$

The f_j are non-negative functions of their arguments, and Z is a normalization constant, e.g. in (a):

$$Z = \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} \sum_{c \in \mathcal{C}} \sum_{d \in \mathcal{D}} \sum_{e \in \mathcal{E}} f_1(a, c) f_2(b, c, d) f_3(c, d, e)$$

where $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, $\mathcal{D}$ and $\mathcal{E}$ are domains for the corresponding variables.

Factor Graphs



(a) (b)

In a **factor graph**, the probability distribution over variables factorizes:

$$P(\mathbf{X}) = \frac{1}{Z} \prod_j f_j(\mathbf{X}_{C_j})$$

where $\mathbf{X} = (X_1, \ldots, X_K)$, $\mathbf{X}_S = (X_i \,:\, i \in S)$, j indexes the factors f_j, C_j are the indices of variables adjacent to factor j, and Z is a normalization constant.
Factors describe compatibility between adjacent variables.

Two nodes are neighbours if they share a common factor.
Fact: $X \perp\!\!\!\perp Y | \mathcal{V}$ if every path between X and Y contains some node $V \in \mathcal{V}$.
Corollary: Given the neighbours of X, the variable X is conditionally independent of all other variables: $X \perp\!\!\!\perp Y | \,\mathrm{ne}(X), \quad \forall Y \notin \{X \cup \mathrm{ne}(X)\}$.

Proving Conditional Independence

Assume the following factorization:

$$P(X, Y, V) = \frac{1}{Z} g_1(X, V) g_2(Y, V)$$

Let's show that the above implies the conditional independence:

$$X \perp\!\!\!\perp Y | V \Leftrightarrow P(X|Y, V) = P(X|V)$$

Summing over X,

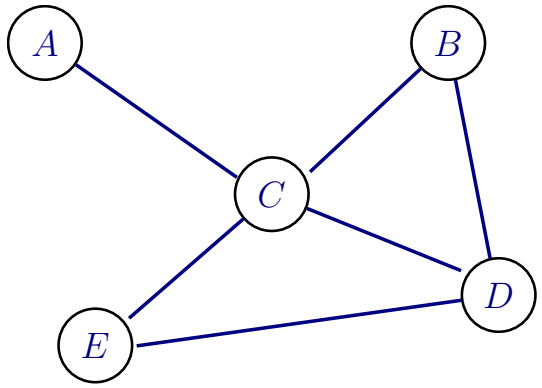
$$p(Y, V) = \frac{1}{Z} \left(\sum_X g_1(X, V) \right) g_2(Y, V)$$

So dividing $P(X, Y, V)$ by $P(Y, V)$ gives

$$P(X|Y, V) = \frac{P(X, Y, V)}{P(Y, V)} = \frac{g_1(X, V)}{\sum_{X'} g_1(X', V)}$$

Since the RHS does not depend on Y, it follows that X is independent from Y given V.

Undirected Graphical Models



In an **undirected graphical model**, the joint probability over all variables can be written in a factored form:

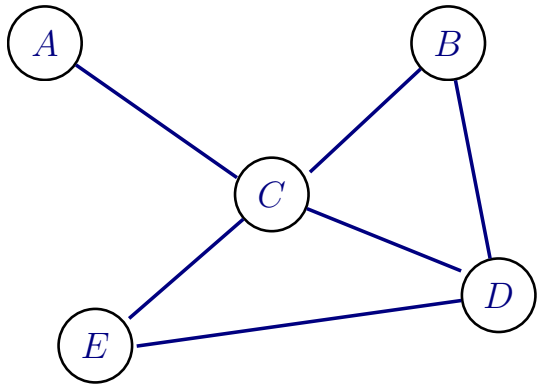
$$P(\mathbf{X}) = \frac{1}{Z} \prod_j f_j(\mathbf{X}_{C_j})$$

where C_j are the maximal cliques of the undirected graph.
(Cliques are fully connected subgraphs, maximal cliques are cliques not contained in other cliques).

Undirected graphical models are also called **Markov networks**.

Undirected graphical models are very similar to factor graphs (not quite equivalent).

Undirected Graphical Models



$$P(A, B, C, D, E) = \frac{1}{Z} g_1(A, C) g_2(B, C, D) g_3(C, D, E)$$

Fact: $X \perp\!\!\!\perp Y | \mathcal{V}$ if every path between X and Y contains some node $V \in \mathcal{V}$

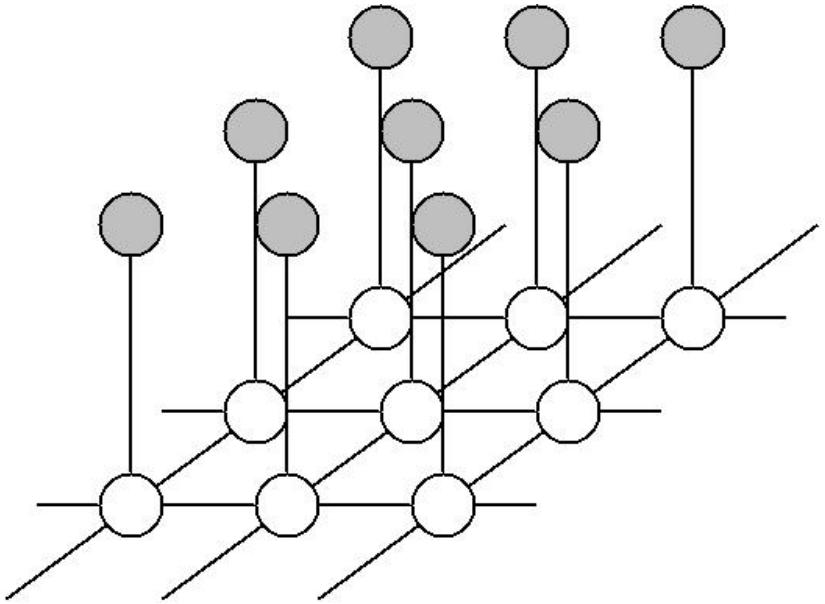
Corollary: Given the neighbours of X, the variable X is conditionally independent of all other variables: $X \perp\!\!\!\perp Y | \operatorname{ne}(X), \quad \forall Y \notin \{X \cup \operatorname{ne}(X)\}$

$\mathcal{V}$ is a Markov blanket for X iff $X \perp\!\!\!\perp Y | \mathcal{V}$ for all $Y \notin \{X \cup \mathcal{V}\}$.

Markov boundary: minimal Markov blanket $\equiv \operatorname{ne}(X)$ for undirected graphs and factor graphs.

Examples of Undirected Graphical Models

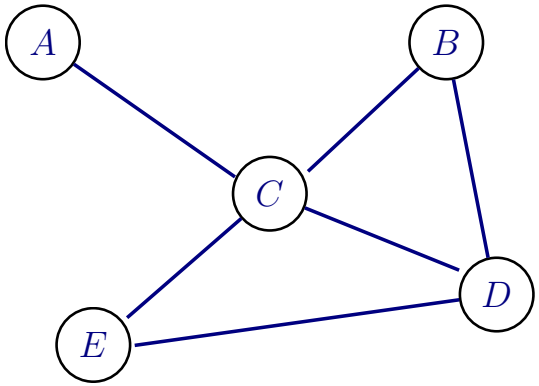
- Markov random fields (image processing, computer vision).



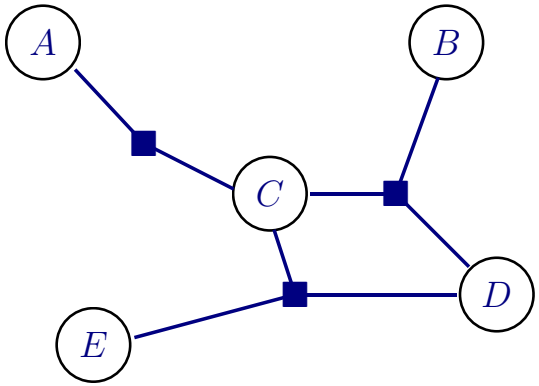
- Conditional random fields and maximum entropy models (text, natural language processing).

- Boltzmann machines (connectionist systems).

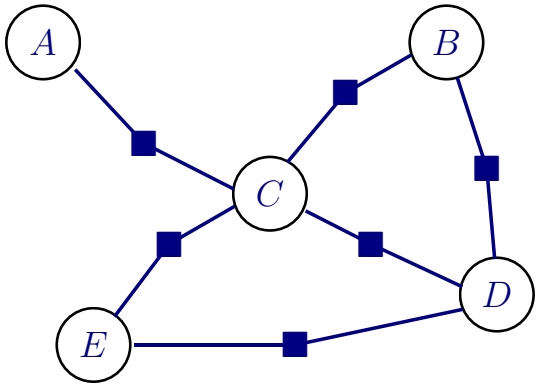
Comparing Undirected Graphs and Factor Graphs



(a)



(b)



(c)

All nodes in (a), (b), and (c) have exactly the same neighbours and therefore these three graphs represent exactly the same conditional independence relationships.

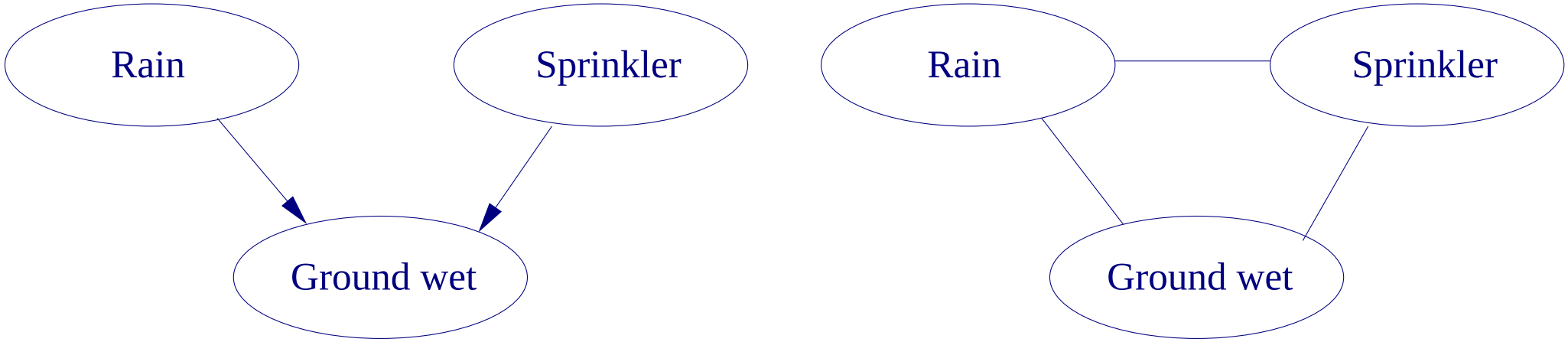
(c) also represents the fact that the probability factors into a product of pairwise functions.

Consider the case where each variables is discrete and can take on K possible values. Then the functions in (a) and (b) are tables with $\mathcal{O}(K^3)$ cells, whereas in (c) they are $\mathcal{O}(K^2)$.

Factor graphs have richer expressive power than undirected graphical models.

Problems with Undirected Graphs and Factor Graphs

In undirected and factor graphs, many useful independencies are unrepresented—two variables are connected merely because some other variable depends on them:

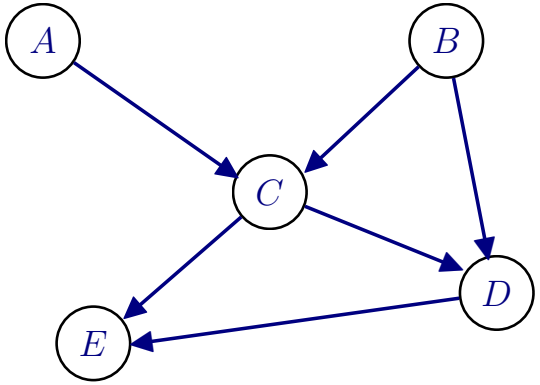


This highlights the difference between **marginal independence** and **conditional independence**.

R and S are marginally independent (i.e. given nothing), but they are conditionally dependent given G.

Explaining Away: Observing that the spinkler is on, explains away the fact that the ground was wet, therefore we don't need to believe that it rained.

Directed Acyclic Graphical Models



A **directed acyclic graphical (DAG) model** represents a factorization of the joint probability distribution in terms of conditionals:

$$P(A, B, C, D, E) = P(A)P(B)P(C|A, B)P(D|B, C)P(E|C, D)$$

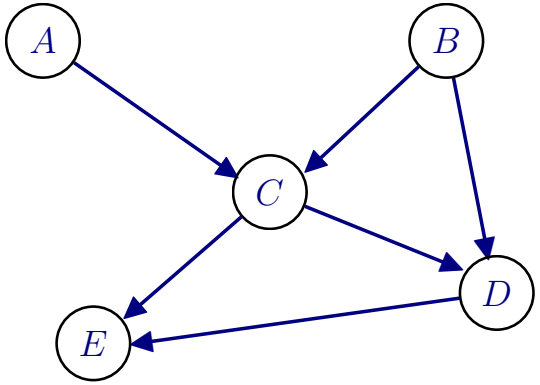
In general:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|X_{\mathsf{pa}(i)})$$

where $\mathsf{pa}(i)$ are the parents of node i.

DAG models are also known as **Bayesian networks** or **Bayes net**.

Conditional Independence in Directed Acyclic Graphs

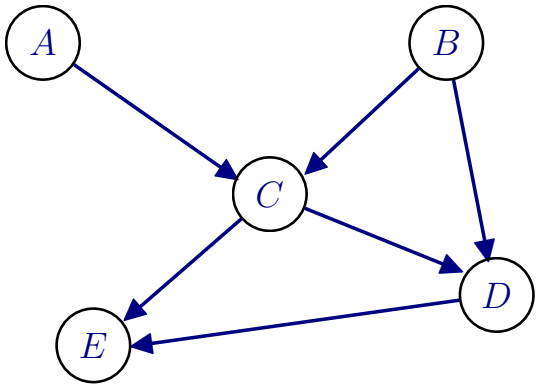


Reading conditional independence from DAGs is more complicated than in undirected graphs.

- $A \perp\!\!\!\perp E \mid \{B, C\}$: observed nodes block paths
- $A \not\perp\!\!\!\perp B \mid C$: observed node creates path by explaining away
- $A \not\perp\!\!\!\perp E \mid C$: created path extends to E via D
- $A \perp\!\!\!\perp E \mid \{C, D\}$: extra path blocked by observing D

So observing (i.e. conditioning on) nodes can both create and remove dependencies.

D-separation



Consider two nodes X, Y and a set of observed nodes $\mathcal{V}$. When is $X \perp\!\!\!\perp Y | \mathcal{V}$?

We consider every undirected path[1] between X and Y. Say that the path is **blocked** by $\mathcal{V}$ if there is a node W on the path such that either:

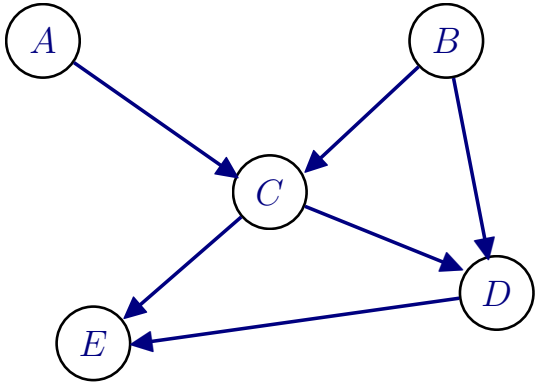
- W has convergent arrows ($\rightarrow W \leftarrow$) on the path and **neither W nor its descendents are in $\mathcal{V}$**. W is also called a collider node.
- or W does not have convergent arrows ($\rightarrow W \rightarrow$ or $\leftarrow W \rightarrow$) and $W \in \mathcal{V}$. This is similar to the undirected graph semantics.

If all paths are blocked, we say $\mathcal{V}$ **d-separates** X from Y (d for directed), and $X \perp\!\!\!\perp Y | \mathcal{V}$.

Markov boundary for X: $\{\text{parents}(X) \cup \text{children}(X) \cup \text{parents-of-children}(X)\}$.

[1] A path in the DAG ignoring the direction of edges.

The Bayes-ball algorithm



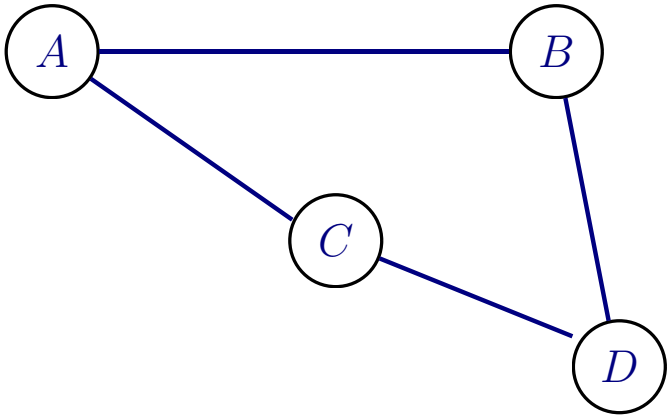
Game: can you get a ball from X to Y without being blocked by $\mathcal{V}$? If so, $X \not\perp\!\!\!\perp Y \,|\, \mathcal{V}$.

Rules: the ball can only be passed along or bounced back under the following conditions:

- If $W \notin \mathcal{V}$ and either $\rightarrow W \rightarrow$ or $\leftarrow W \leftarrow$, then the ball is passed along.
- If $W \notin \mathcal{V}$, then the ball is bounced from any child to any child (including the child it came from).
- If $W \in \mathcal{V}$, then the ball is bounced from any parent to any parent (including the parent it came from).

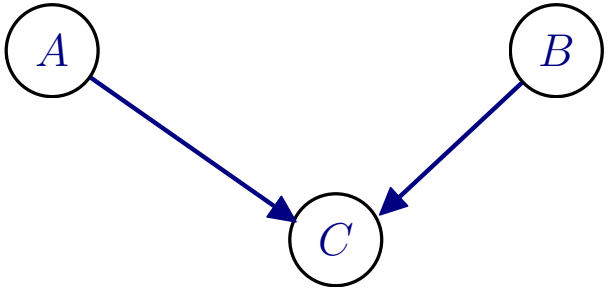
Note: if $W \in \mathcal{V}$ it blocks all balls from its children, and balls from parents are blocked from reaching its children.

Expressive Power of Directed and Undirected Graphs



No DAG can represent these and only these independencies

No matter how we direct the arrows there will always be two non-adjacent parents sharing a common child $\Longrightarrow$ dependence in DAG but independence in undirected graph.



No undirected or factor graph can represent these and only these independencies

Graphs, Conditional Independencies, and Families of Distributions

Corresponding to each (factor, undirected, directed acyclic) graph G is a set of conditional independency statements $\mathcal{C}_G = \{X_i \perp\!\!\!\perp Y_i | \mathcal{V}_i\}$.

Corresponding to G is also a parameterized family of distributions $\mathcal{P}_G$, e.g. for factor graph

$$\mathcal{P}_G = \{P(\mathbf{X}) \ : \ P(\mathbf{X}) = \frac{1}{Z}\prod_j f_j(\mathbf{X}_{C_j}), \text{ for some non-negative functions } f_j \ \}$$

Corresponding to any set of conditional independency statements $\mathcal{C}$ is also a family of distributions satisfying all statements in $\mathcal{C}$:

$$\mathcal{P}_{\mathcal{C}} = \{P(\mathbf{X}) \ : \ P(X_i, Y_i|\mathcal{V}_i) = P(X_i|\mathcal{V}_i)P(Y_i|\mathcal{V}_i) \text{ for all } X_i \perp\!\!\!\perp Y_i|\mathcal{V}_i \text{ in } \mathcal{C} \ \}$$

- Adding edges to graph $\Rightarrow$ removing conditional independency statements $\Rightarrow$ enlarging the family of distributions (converse true for removing edges).
- For directed graphs we have $\mathcal{P}_G = \mathcal{P}_{\mathcal{C}_G}$.
- For undirected graphs we also have $\mathcal{P}_G = \mathcal{P}_{\mathcal{C}_G}$ if we assume that all distributions are positive, i.e. $p(\mathbf{X}) > 0$ for all values $\mathbf{X}$ take on (Hammersley-Clifford Theorem).
- There are factor graphs for which $\mathcal{P}_G \neq \mathcal{P}_{\mathcal{C}_G}$.
- Factor graphs are more expressive than undirected graphs: for every undirected graph G_1 there is a factor graph G_2 with $\mathcal{P}_{G_1} = \mathcal{P}_{G_2}$ but not vice versa.

End Notes

Recent textbook on graphical models:

Koller and Friedman (2009):
Probabilistic Graphical Models: Principles and Techniques. MIT Press.

More theoretical textbooks on graphical models:

Cowell, Dawid, Lauritzen and Spiegelhalter (2007). Probabilistic Networks and Expert Systems. Springer.

Jensen and Graven-Nielsen (2007). Bayesian Networks and Decision Graphs. Springer.

End Notes