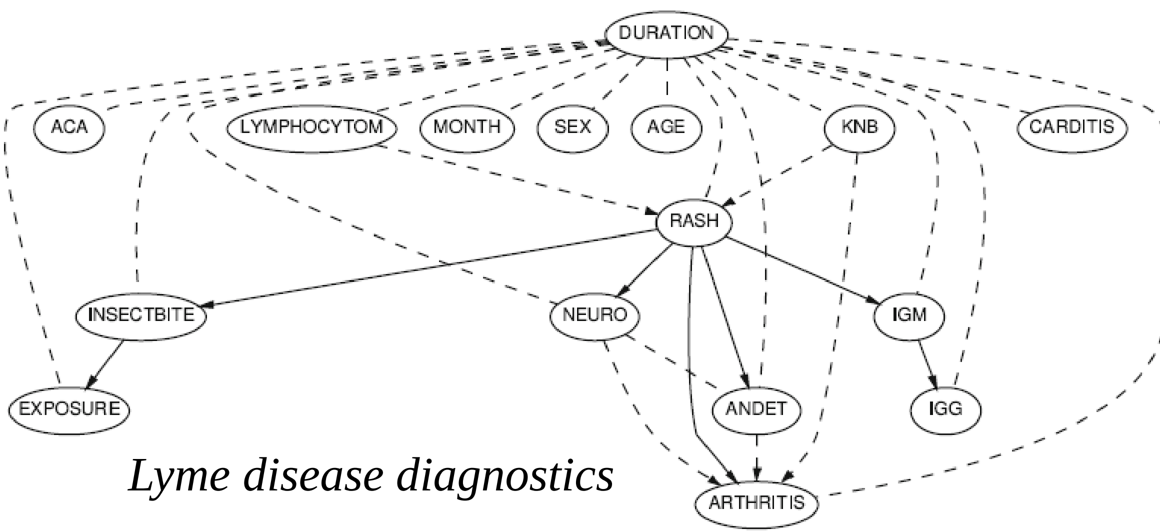
Universität Potsdam
Institut für Informatik
Lehrstuhl Maschinelles Lernen

Graphical Models

Niels Landwehr

Overview: Graphical Models

- Graphical models: tool for modelling domain with several random variables.
- For example medical domains: joint distribution over attributes of patients, symptoms, and diseases.
- Can be used to answer any probabilistic query.



Lyme disease diagnostics

Attribute name	Description
Exposure	Exposure to ticks, e.g., patient visited a forest
Duration	Duration of the disease
Month	Month the patient reported to a doctor
Rash	Whether the patient developed rash
IgM, IgG	Serological tests
Neuro	Neurological symptoms
ACA, KNB, Carditis, Lymphocytom, Andet	Various other symptoms

Agenda

- Graphical models: syntax and semantics.
- Inference in graphical models (exact, approximate)
- Graphical models in machine learning.

Agenda

- Graphical models: syntax and semantics.
- Inference in graphical models (exact, approximate)
- Graphical models in machine learning.

Recap: Random Variables, Distributions

- Random variables: $X, Y, Z, \ldots$
  - discrete random variables: distributions defined by probabilities for possible values.
  - continuous random variables: distributions defined by densities.
- Joint distribution $p(X, Y)$
- Conditional distribution $p(X \mid Y) = \dfrac{p(X, Y)}{p(Y)}$
- Product rule:

$$p(X, Y) = p(X \mid Y)\, p(Y) \quad \text{discrete or continuous}$$

- Sum rule:

$$p(x) = \sum_y p(x, y) \quad \text{discrete random variables}$$

$$p(x) = \int_{-\infty}^{\infty} p(x, y)\, dy \quad \text{continuous random variables}$$

Independence of Random Variables

- Independence (discrete or continuous)

  X,Y independent if and only if $p(X,Y) = p(X)p(Y)$

  X,Y independent if and only if $p(X \mid Y) = p(X)$

  X,Y independent if and only if $p(Y \mid X) = p(Y)$

- Conditional independence (discrete or continuous)

  X,Y independent given Z if and only if $p(X,Y \mid Z) = p(X \mid Z)p(Y \mid Z)$

  X,Y independent given Z if and only if $p(Y \mid X,Z) = p(Y \mid Z)$

  X,Y independent given Z if and only if $p(X \mid Y,Z) = p(X \mid Z)$

... simply application of the notion of independence
to the conditional joint distribution p(X,Y|Z)

Graphical Models: Idea/Goal

- Goal: Model for the joint distribution $p(X_1, \ldots, X_N)$ of a set of random variables $X_1, \ldots, X_N$
- Given $p(X_1, \ldots, X_N)$, we can compute…
  - All marginal distributions (by sum rule)

$$p(X_{i_1}, \ldots, X_{i_m}), \quad \{i_1, \ldots, i_m\} \subseteq \{1, \ldots, N\}$$

  - All conditional distributions (from marginal distributions)

$$p(X_{i_1}, \ldots, X_{i_m} \mid X_{i_{m+1}}, \ldots, X_{i_{m+k}}), \quad \{i_1, \ldots, i_{m+k}\} \subseteq \{1, \ldots, N\}$$

- Enough to answer all probabilistic queries („inference problems") over the random variables $X_1, \ldots, X_N$

Graphical Models: Idea/Goal

- Graphical models: combination of probability theory and graph theory.
- Compact, intuitive modeling of $p(X_1,\ldots,X_N)$.
  - Graph structure represents structure of the distributions (dependencies between variables $X_1,\ldots,X_N$).
  - Insight into structure of the model; easy to inject prior knowledge.
  - Efficient algorithms for inference that exploit the graph structure.
- Many machine learning methods can be represented as graphical models.
- Tasks such as finding the MAP model or computing Bayes-optimal predictions can be formulated as inference problems in graphical models.

Graphical Models: Example

- Example: „Alarm“ scenario
  - Our house in Los Angeles has an alarm system.
  - We are on holidays. Our neighbor calls in case he hears the alarm going off. In case of a burglary we would like to return.
  - Unfortunately, our neigbor is not always at home.
  - Unfortunately, the alarm can also be triggered by earthquakes.
- 5 binary random variables

B Burglary – burglary has taken place

E Earthquake – earthquake has taken place

A Alarm – alarm is triggered

N NeighborCalls – neighbor calls us

R RadioReport – Report about an earthquake on the radio

Graphical Models: Example

- Random variables have a joint distribution $p(B,E,A,N,R)$. How to specify? Which dependencies hold?
- Example for inference problem: neighbor has called ($N=1$), how likely that there was a burglary ($B=1$)?
  - Depends on several factors
    - How likely is a burglary a priori?
    - How likely is an earthquake a priori?
    - How likely that alarm is triggered?
    - …

(Naive) inference: $p(B=1 \mid N=1) = \dfrac{p(B=1, N=1)}{p(N=1)}$

$$= \frac{\sum_E \sum_A \sum_R p(B=1, E, A, N=1, R)}{\sum_B \sum_E \sum_A \sum_R p(B, E, A, N=1, R)}$$

Graphical Models: Example

- How do we model $p(B,E,A,N,R)$?
  - 1. Attempt: complete table of probabilities

2^N rows:

B	E	A	N	R	$P(B,E,A,N,R)$
0	0	0	0	0	0.6
1	0	0	0	0	0.005
0	1	0	0	0	0.01
…	…	…	…	…	…

+Any distribution $p(B,E,A,N,R)$ can be represented

- Exponential number of parameters

- Difficult to specify for humans

  - 2. Attempt: everything is independent

$$p(B,E,A,N,R) = p(B)\,p(E)\,p(A)\,p(N)\,p(R)$$

+ linear number of parameters

- too restrictive, independence assumption does not allow any meaningful inference

Graphical Models: Example

- Graphical model: selective independence assumptions, motivated by prior knowledge.
- Choose variable ordering: e.g. $B<E<A<N<R$
- Product rule:

$$\begin{aligned} p(B,E,A,N,R) &= p(B,E,A,N)\,p(R \mid B,E,A,N) \\ &= p(B,E,A)\,p(N \mid B,E,A)\,p(R \mid B,E,A,N) \\ &= p(B,E)\,p(A \mid B,E)\,p(N \mid B,E,A)\,p(R \mid B,E,A,N) \\ &= p(B)\,p(E \mid B)\,p(A \mid B,E)\,p(N \mid B,E,A)\,p(R \mid B,E,A,N) \end{aligned}$$

Factors describe distribution of one random variable as a function of other random variables.

Can we simplify these factors?
Which dependencies really hold in our domain?

Graphical Models: Example

- Decomposition into factors according to product rule:

$$p(B,E,A,N,R) = p(B)\,p(E \mid B)\,p(A \mid B,E)\,p(N \mid B,E,A)\,p(R \mid B,E,A,N)$$

- Conditional independence assumptions (remove variables from conditional expression)

$p(E \mid B) = p(E)$ *Earthquake does not depend on burglary*

$p(A \mid B,E) = p(A \mid B,E)$ *Alarm does depend on burglary and earthquake*

$p(N \mid B,E,A) = p(N \mid A)$ *Whether neighbor calls only depends on alarm*

$p(R \mid B,E,A,N) = p(R \mid E)$ *Report on the radio only depends on earthquake*

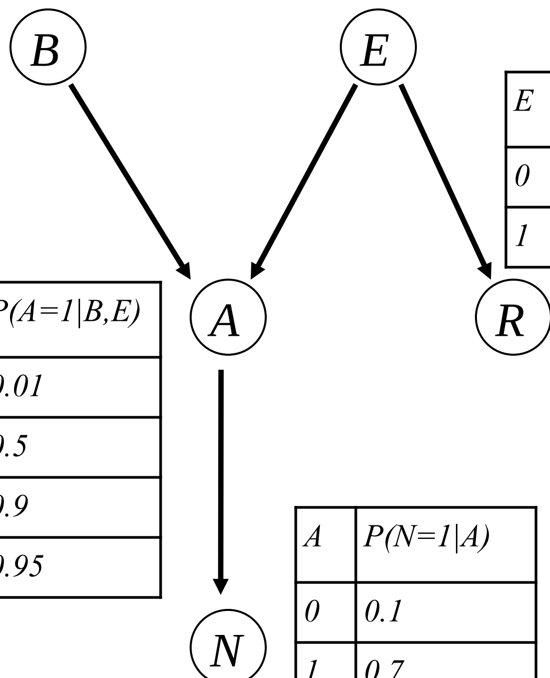
- Arriving at simplified form of joint distribution:

$$p(B,E,A,N,R) = p(B)\,p(E)\,p(A \mid E,B)\,p(N \mid A)\,p(R \mid E)$$

Simplified factors

Graphical Models: Example

- Graphical model for „Alarm“ scenario



P(B=1)
0.1

P(E=1)
0.2

E	P(R=1|E)
0	0.01
1	0.5

B	E	P(A=1|B,E)
0	0	0.01
0	1	0.5
1	0	0.9
1	1	0.95

A	P(N=1|A)
0	0.1
1	0.7

Distribution modeled:

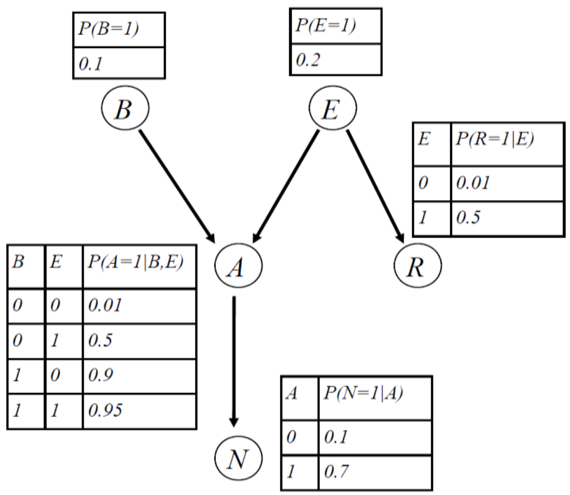
$$p(B,E,A,N,R) = p(B)\,p(E)\,p(A \mid E,B)\,p(N \mid A)\,p(R \mid E)$$

Graphical model:

- There is one node for each random variable
- For each factor of the form $p(X \mid X_1,\ldots,X_k)$ there is a directed edge from the X_i to X in the graph
- Model is parameterized with conditional distributions $p(X \mid X_1,\ldots,X_k)$

Graphical Models: Example

- Graphical model for „Alarm“ scenario



  - Number of parameters: $O(N2^K)$, K= max. number of parents of a node.
  - Here 1+1+2+2+4 instead of 2^5-1 parameters for full table.
- These directed graphical models are also called **Bayesian networks**.

Directed Graphical Models: Definition

- Given a set of random variables $\{X_1, \ldots, X_N\}$
- A directed graphical model over the random variables $\{X_1, \ldots, X_N\}$ is a directed graph with
  - Node set $X_1, \ldots, X_N$
  - There are no directed cycles $X_{i_1} \to X_{i_2} \to \ldots \to X_{i_k} \to X_{i_1}$
  - Nodes are associated with parameterized conditional distributions $p(X_i \mid pa(X_i))$, where $pa(X_i) = \{X_j \mid X_j \to X_i\}$ denotes the set of parent nodes of a node.

- The graphical model represents a joint distribution over $X_1, \ldots, X_N$ by

$$p(X_1, \ldots, X_N) = \prod_{i=1}^{N} p(X_i \mid pa(X_i))$$

Directed Graphical Models: Definition

- Why does the graph have to be acyclic?
  - Theorem from graph theory:

$$G \text{ is acyclic} \Leftrightarrow \text{ there is an ordering } \leq_G \text{ of the nodes such that all directed edges respect the ordering } (N \rightarrow N' \;\Rightarrow\; N \leq_G N')$$

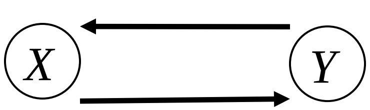
  - For such an ordering, we can factorize

$$p(X_1,\ldots,X_N) = \prod_{i=1}^{N} p(X_i \mid pa(X_i))$$

Before X_i in variable ordering

according to product rule + conditional independence assumptions (variables sorted according to $\leq_G$)

- Counterexample (not a graphical model)



$$p(X,Y) \neq p(X \mid Y)\, p(Y \mid X)$$

Graphical Models: Independence

- The graph structure of a graphical model implies (conditional) independencies between random variables.
- Notation: for variables X, Y, Z we write

$$X \perp Y \mid Z \Leftrightarrow p(X \mid Y, Z) = p(X \mid Z)$$

$$\text{"}X \text{ independent of } Y \text{ given } Z\text{"}$$

- Extension to disjoint sets A, B, C of random variables:

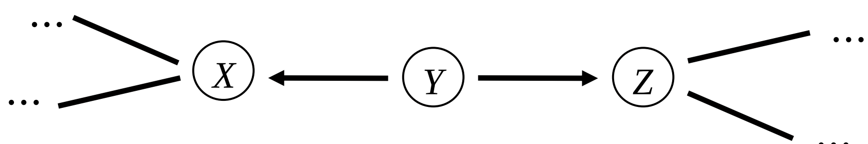
$$\mathrm{A} \perp \mathrm{B} \mid C \Leftrightarrow p(A \mid B, C) = p(A \mid C)$$

Graphical Models: Independence

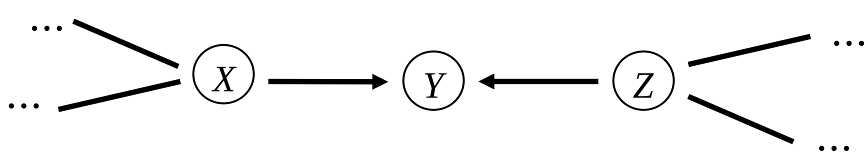
- Which independence assumptions of the form $A \perp B \mid C$ are modeled by the graph structure?
  - Can be checked by sum/product rule starting from the modeled joint distribution (lots of work!)
  - For graphical models, independence assumptions can be read off the graph structure → much easier.
  - „D-separation“: Set of simple rules from which all independence assumptions encoded in the graph can be derived.
  - „D“ in „D-separation“ stands for „Directed“, because we are talking about directed graphical models (similar mechanism exists for „undirected“ models, which we do not cover).

Graphical Models: Independence

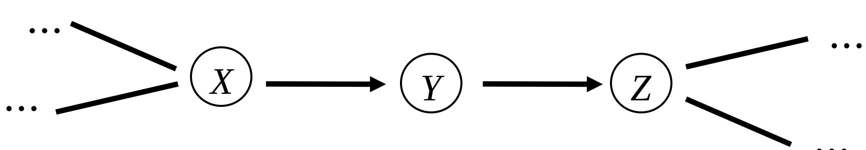
- D-separation: Which indepence assumptions $A \perp B \mid C$ are modeled by the graph structure?
- Idea: can be checked by looking at pathes connecting random variables.
- Notation:



Path between X and Z has a **diverging connection** at Y („tail to tail").

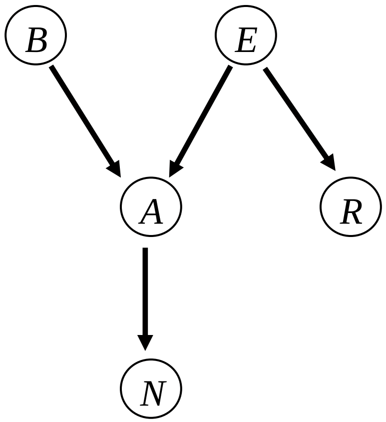


Path between X and Z has a **converging connection** at Y („head to head").



Path between X and Z has a **serial connection** at Y („head to tail").

Diverging Connections



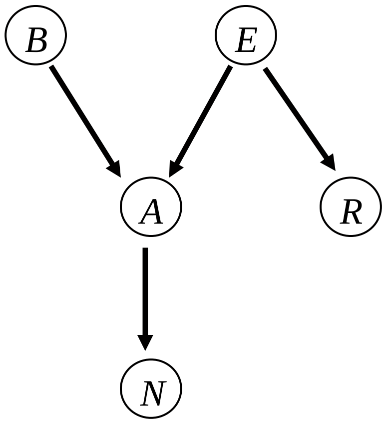
Joint distribution:

$$p(B, E, A, N, R) =$$
$$p(B)\,p(E)\,p(A \mid E, B)\,p(N \mid A)\,p(R \mid E)$$

B=„*Burglary*" *N*=„*Neighbor calls*"
E=„*Earthquake*" *R*=„*Radio report*"
A=„*Alarm*"

- Looking at path $A \leftarrow E \rightarrow R$. Does $A \perp R \mid \emptyset$ hold?
  -

Diverging Connections



Joint distribution:

$$p(B,E,A,N,R) =$$
$$p(B)\,p(E)\,p(A\,|\,E,B)\,p(N\,|\,A)\,p(R\,|\,E)$$

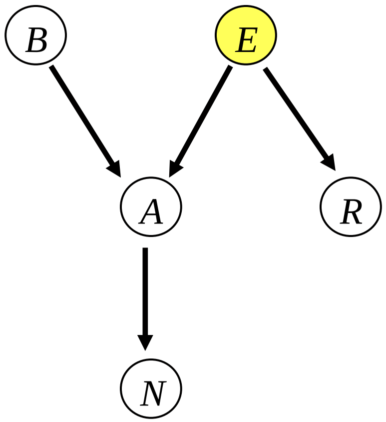
B=„Burglary“ *N=„Neighbor calls“*
E=„Earthquake“ *R=„Radio report“*
A=„Alarm“

- Looking at path $A \leftarrow E \rightarrow R$. Does $A \perp R\,|\,\varnothing$ hold?
  - No, $p(A\,|\,R) \neq p(A)$ [Can be derived from joint distribution]
  - Intuitively:

    Radio report $\Rightarrow$ probably earthquake $\Rightarrow$ probably alarm

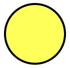
    $p(A=1\,|\,R=1) > p(A=1\,|\,R=0)$
  - Variable R influences variable A through the diverging connection $R \leftarrow E \rightarrow A$

Diverging Connections



Joint distribution:

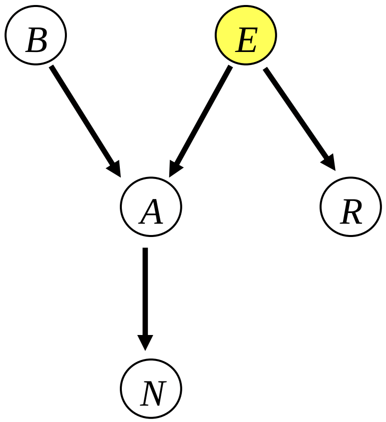
$$p(B,E,A,N,R) = p(B)\,p(E)\,p(A\,|\,E,B)\,p(N\,|\,A)\,p(R\,|\,E)$$



observed variable

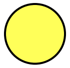
- Looking at path $A \leftarrow E \rightarrow R$. Does $\mathrm{A} \perp R\,|\,E$ hold?
  -

Diverging Connections



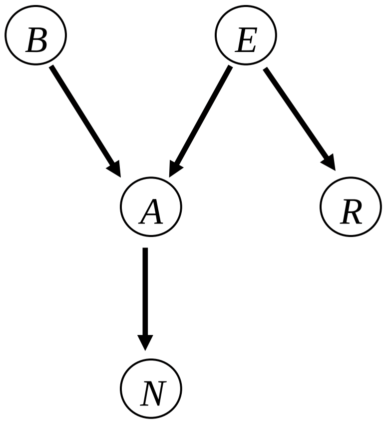
Joint distribution:

$$p(B,E,A,N,R) = p(B)\,p(E)\,p(A \mid E,B)\,p(N \mid A)\,p(R \mid E)$$

 observed variable

- Looking at path $A \leftarrow E \rightarrow R$. Does $A \perp R \mid E$ hold?
  - Yes, $p(A \mid R,E) = p(A \mid E)$ [Can be derived from joint distribution]
  - Intuitively:
    If we allready know that an earthquake has occured the probably for alarm is not increased or decreased because of radio report.
  - The diverging path $A \leftarrow E \rightarrow R$ is *blocked* by the observation of E.

Serial Connections



Joint distribution:

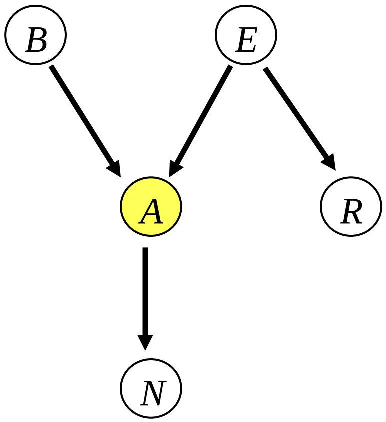
$$p(B,E,A,N,R) = p(B)p(E)p(A|E,B)p(N|A)p(R|E)$$

- Looking at path $N \leftarrow A \leftarrow B$. Does $\mathrm{B} \perp \mathrm{N} | \varnothing$ hold?
  - No, $p(B|N) \neq p(B)$ [Can be derived from joint distribution]
  - Intuitively:

    Neighbor calls $\Rightarrow$ probably alarm $\Rightarrow$ probably burglary

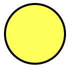
    $p(B=1|N=1) > p(B=1|N=0)$
  - Variable N influences variable B through the serial connection $\mathrm{N} \leftarrow A \leftarrow B$

Serial Connections



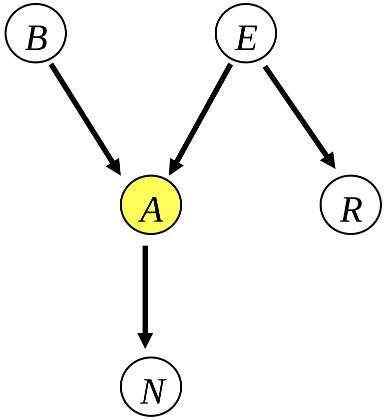
Joint distribution:

$$p(B,E,A,N,R) = p(B)\,p(E)\,p(A \mid E,B)\,p(N \mid A)\,p(R \mid E)$$

 observed variable

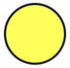
- Looking at path $N \leftarrow A \leftarrow B$. Does $\mathrm{B} \perp \mathrm{N} \mid A$ hold?
    -

Serial Connections



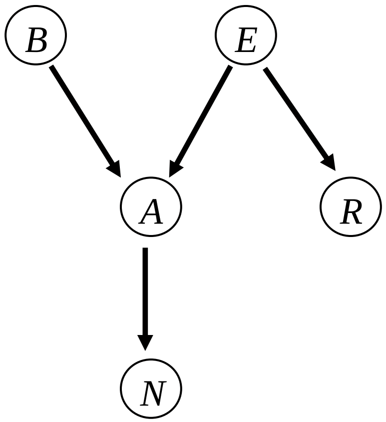
Joint distribution:

$$p(B, E, A, N, R) = p(B)\,p(E)\,p(A \mid E, B)\,p(N \mid A)\,p(R \mid E)$$


observed variable

- Looking at path $N \leftarrow A \leftarrow B$. Does $\mathrm{B} \perp \mathrm{N} \mid A$ hold?
  - Yes, $p(B \mid N, A) = p(B \mid A)$ [Can be derived from joint distribution]
  - Intuitively:
    If we already know that alarm was triggered, the probably for burglary does not increase or decrease because the neighbor calls.
  - The serial connection $N \leftarrow A \leftarrow B$ is *blocked* by the observation of A.

Converging Connections

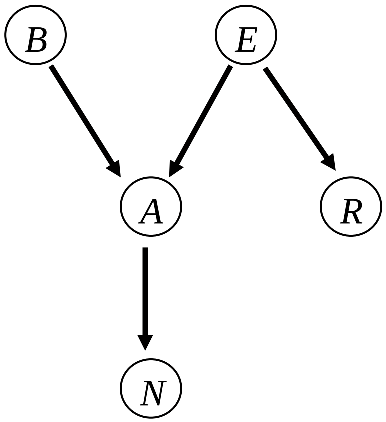


Joint distribution:

$$p(B,E,A,N,R) = p(B)\,p(E)\,p(A \mid E,B)\,p(N \mid A)\,p(R \mid E)$$

- Looking at path $B \rightarrow A \leftarrow E$. Does $\mathrm{B} \perp \mathrm{E} \mid \varnothing$ hold?
    -

Converging Connections



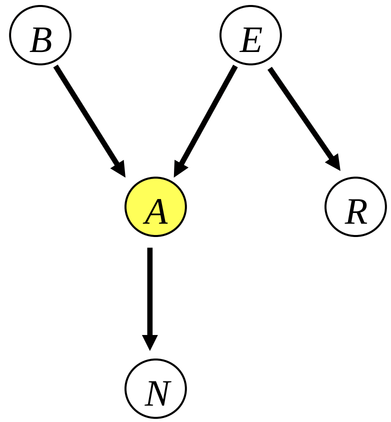
Joint distribution:

$$p(B,E,A,N,R) = p(B)\,p(E)\,p(A \mid E,B)\,p(N \mid A)\,p(R \mid E)$$

- Looking at path $B \longrightarrow A \longleftarrow E$. Does $\mathrm{B} \perp \mathrm{E} \mid \varnothing$ hold?
  - Yes, $p(B \mid E) = p(B)$ [Can be derived from joint distribution]
  - Intuitively:

    Burglaries are not more/less frequent on days with earthquakes
  - The converging path $\mathrm{B} \rightarrow A \leftarrow E$ is blocked if A is **not** observed

Converging Connections



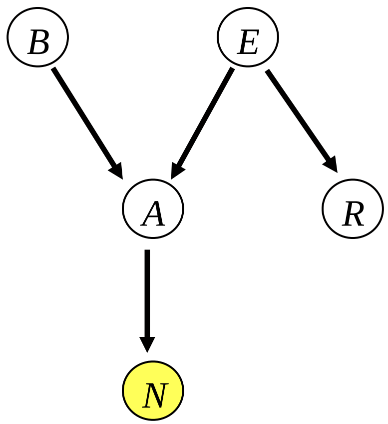
Joint distribution:

$$p(B,E,A,N,R) = p(B)p(E)p(A \mid E,B)p(N \mid A)p(R \mid E)$$

observed variable

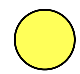
- Looking at path $B \rightarrow A \leftarrow E$. Does $\mathrm{B} \perp \mathrm{E} \mid A$ hold?
  - No, $p(B \mid E,A) \neq p(B \mid A)$ [Derive from joint distribution]
  - Intuitively:
    Alarm was triggered. If we observed an earthquake, this explains the alarm, thus probability for burglary is reduced ("explaining away" phenomenon).
  - The converging path $\mathrm{B} \rightarrow A \leftarrow E$ is *unblocked* by observation of A

Converging connections



Joint distribution:

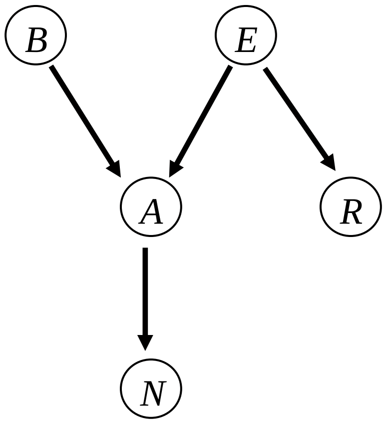
$$p(B,E,A,N,R) = p(B)\,p(E)\,p(A \mid E,B)\,p(N \mid A)\,p(R \mid E)$$

 observed variable

- Looking at path $B \rightarrow A \leftarrow E$. Does $\mathrm{B} \perp \mathrm{E} \mid N$ hold?
  - No, $p(B \mid N, A) \neq p(B \mid A)$ [Derive from joint distribution]
  - Intuitively:

    Neighbor calls is an indirect observation of alarm. Observation of an earthquake explains the alarm, probability for burglary is reduced ("explaining away").
  - The converging path $\mathrm{B} \rightarrow A \leftarrow E$ is *unblocked* by observing N.

Summary Pathes



Joint distribution:

$$p(B,E,A,N,R) = p(B)\,p(E)\,p(A \mid E,B)\,p(N \mid A)\,p(R \mid E)$$

- Summary: a path …-X-Y-Z-… is
  - Blocked at Y, if
    - Diverging connection, and Y is observed, or
    - Serial connection, and Y is observed, or
    - Converging connection, and neither Y nor one of its descendents $Y' \in Descendants(Y)$ is observed
    - $Descendants(Y)=\{Y' \mid \text{there is a directed path from } Y \text{ zu } Y'\}$
  - If the path it not blocked at Y, it is free at Y.

D-Separation: Are All Pathes Blocked?

- So far we have defined if a path is blocked at a particular node.
- A path is blocked overall, if it is blocked at one of its nodes:
  - Let X, X' be random variables, C a set of observed random variables, $X, X' \notin C$
  - A path $X - X_1 - \ldots - X_n - X'$ between X and X' is blocked given C if and only if there is a node X_i such that the path is blocked at the node X_i given C.

- **D-Separation**: are all pathes blocked?
  - Let X, Y be random variables, C a set of random variables with $X, Y \notin C$.
  - Definition: X and Y are d-separated given C if and only if every path from X to Y is blocked given C.

D-Separation: Correct and Complete

- Given a graphical model over random variables $\{X_1, \ldots, X_N\}$ with graph structure G.
- The graphical model defines a joint distribution by

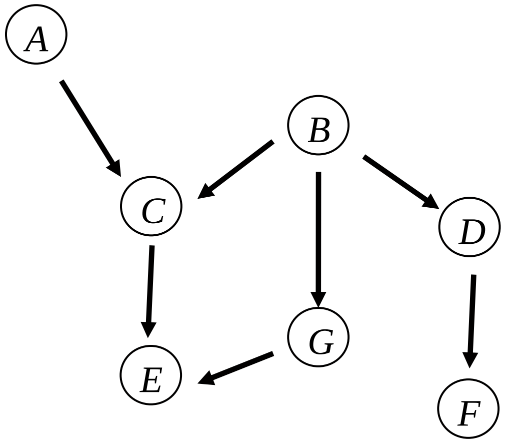
$$p(X_1, \ldots, X_N) = \prod_{i=1}^{N} p(X_i \mid pa(X_i))$$

that depends on the conditional distributions $p(X_i \mid pa(X_i))$.

- **Theorem** (D-separation is correct and complete)
  - If X,Y are d-separated given C in G, then $X \perp Y \mid C$.
  - There are no other independencies that hold irrespective of the choice of the conditional distribution $p(X_i \mid pa(X_i))$.

- Of course, additional independencies can exists because of the choice of particular $p(X_i \mid pa(X_i))$.

D-Separation: Example



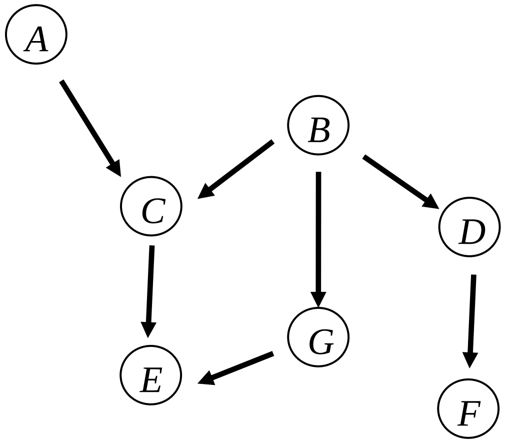
Does $\mathrm{A} \perp \mathrm{F} \mid D$ hold?

Does $\mathrm{B} \perp \mathrm{E} \mid C$ hold?

Does $A \perp \mathrm{E} \mid C$ hold?

- A path …-X-Y-Z-… *is*
  - Blocked at Y, if
    - Diverging connection, and Y is observed, or
    - Serial connection, and Y is observed, or
    - Converging connection, and neither Y nor any of ist descendants $Y' \in Descendants(Y)$ is observed.
  - Otherwise the path is free at Y.

D-Separation: Example



Does $A \perp F \mid D$ hold? Yes

Does $B \perp E \mid C$ hold? No: $B - G - E$

Does $A \perp E \mid C$ hold? No: $A - C - B - G - E$

- A path …-*X-Y-Z-*… *is*
  - Blocked at Y, if
    - Diverging connection, and Y is observed, or
    - Serial connection, and Y is observed, or
    - Converging connection, and neither Y nor any of ist descendants $Y' \in Descendants(Y)$ is observed.
  - Otherwise the path is free at Y.

Bayesian Networks: Causality

- Often Bayesian networks are constructed in such a way that directed edges correspond to causal influences

$G \quad E \longrightarrow A$ „Earthquakes trigger the alarm system“

- However, equivalent model:

$G' \quad E \longleftarrow A$ „The alarm system triggers an earthquake“

- Definition: $\mathrm{I(G)} = \{ (\mathrm{X} \perp Y \mid C) : X$ and Y are d-separated given C in $G\}$

„All independence assumptions encoded in G “

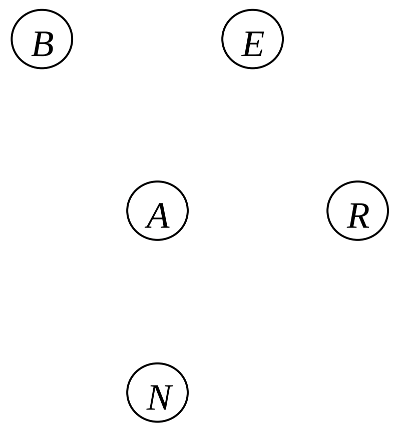
- $I(G) = I(G') = \varnothing$:
  - Not statistical reasons to prefer one of the models.
  - We cannot distinguish between the models based on data.
  - But „causal“ models often more intuitive.

Models of Different Complexity

- Complexity of a model depends on the number (and location) of edges in the graph
  - Many edges: few independence assumptions, many parameters, large class of distributions can be represented.
  - Few edges: many independence assumptions, few parameters, small class of distributions can be represented.

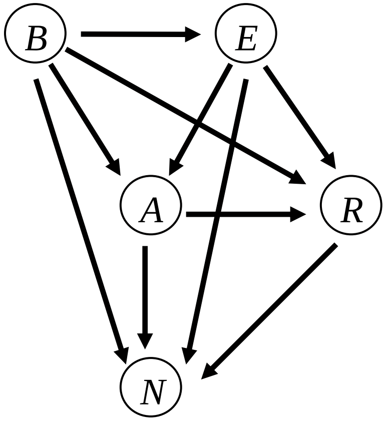
Models of Different Complexity

- Adding edges: family of representable distributions becomes larger, *I(G)* becomes smaller.
- Extreme cases: graph without any edges, graph completely connected (as an undirected graph)



N parameter (for binary variables)

$I(G) = \{(X \perp Y \mid C) : X, Y \text{ RV}, C \text{ set of RV}\}$



2^N -1 parameters (for binary variables)

$I(G) = \varnothing$