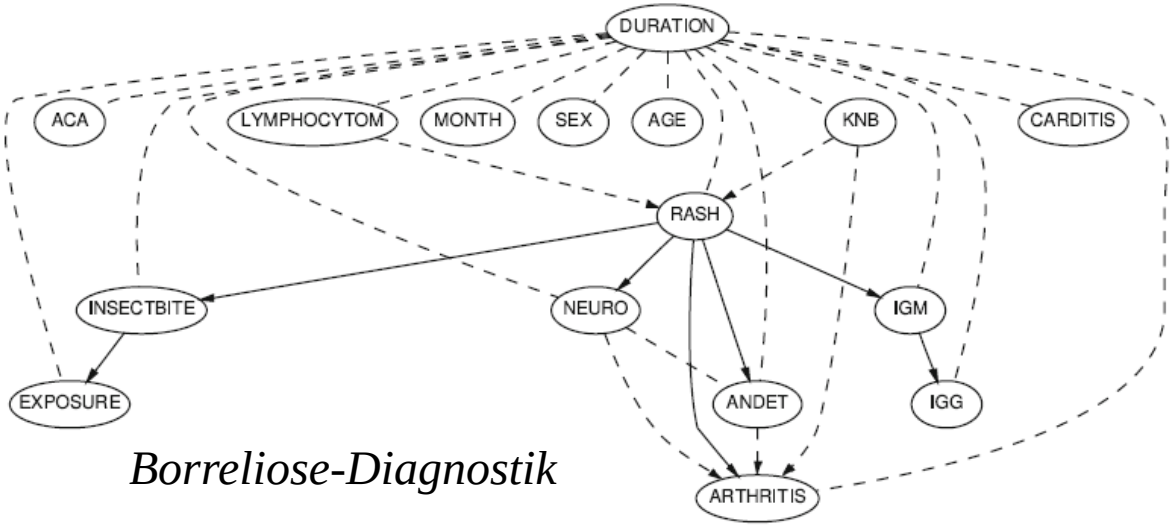
Universität Potsdam
Institut für Informatik
Lehrstuhl Maschinelles Lernen

Graphische Modelle

Niels Landwehr

Überblick: Graphische Modelle

- Graphische Modelle: Werkzeug zur Modellierung einer Domäne mit verschiedenen Zufallsgrössen
- Beispielsweise medizinische Anwendungsbereiche: gemeinsame Verteilung über Attribute von Patienten, Symptome, und Krankheiten
- Beliebige Wahrscheinlichkeitsanfragen



Borreliose-Diagnostik

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

Überblick

- Graphische Modelle: Syntax und Semantik
- Graphische Modelle im Maschinellen Lernen
- Inferenz in Graphischen Modellen (exakt, approximativ)
- Sequenzmodelle

Überblick

- Graphische Modelle: Syntax und Semantik
- Graphische Modelle im Maschinellen Lernen
- Inferenz in Graphischen Modellen (exakt, approximativ)
- Sequenzmodelle

Erinnerung: Zufallsvariablen, Verteilungen

- Zufallsvariablen: $X, Y, Z, \ldots$
  - Diskrete ZV: Verteilungen beschrieben durch Wahrscheinlichkeiten
  - Kontinuierliche ZV: Verteilungen beschrieben durch Dichten
- Gemeinsame Verteilung $p(X, Y)$
- Bedingte Verteilung $p(X \mid Y) = \frac{p(X, Y)}{p(Y)}$
- Produktregel:

$$p(X, Y) = p(X \mid Y)\, p(Y) \quad \text{diskret oder kontinuierlich}$$

- Summenregel:

$$p(x) = \sum_{y} p(x, y) \quad \text{diskrete Verteilungen}$$

$$p(x) = \int_{-\infty}^{\infty} p(x, y)\, dy \quad \text{kontinuierliche Verteilungen}$$

Unabhängigkeit von Zufallsvariablen

- Unabhängigkeit (diskret oder kontinuierlich)

X,Y unabhängig genau dann wenn $p(X,Y) = p(X)p(Y)$

X,Y unabhängig genau dann wenn $p(X \mid Y) = p(X)$

X,Y unabhängig genau dann wenn $p(Y \mid X) = p(Y)$

- Bedingte Unabhängigkeit (diskret oder kontinuierlich)

X,Y unabhängig gegeben Z genau dann wenn $p(X,Y \mid Z) = p(X \mid Z)p(Y \mid Z)$

X,Y unabhängig gegeben Z genau dann wenn $p(Y \mid X,Z) = p(Y \mid Z)$

X,Y unabhängig gegeben Z genau dann wenn $p(X \mid Y,Z) = p(X \mid Z)$

... einfach Anwendung des Unabhängigkeitsbegriffs auf die bedingte gemeinsame Verteilung p(X,Y|Z)

Graphische Modelle: Idee/Ziel

- Ziel: Modellierung der gemeinsame Verteilung $p(X_1,...,X_N)$ einer Menge von ZV $X_1,...,X_N$
- Aus $p(X_1,...,X_N)$ lassen sich berechnen…
  - Alle Randverteilungen (Summenregel)

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

  - Alle bedingten Verteilungen (aus Randverteilungen)

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

- Damit lassen sich alle probabilistischen Fragestellungen (*Inferenzprobleme*) über $X_1,...,X_N$ beantworten

Graphische Modelle: Idee/Ziel

- Graphische Modelle: Kombination von Wahrscheinlichkeitstheorie und Graphentheorie
- Kompakte, intuitive Modellierung von $p(X_1,...,X_N)$
  - Graphstruktur repräsentiert Struktur der Verteilung (Abhängigkeiten zwischen Variablen $X_1,...,X_N$)
  - Einsicht in Struktur des Modells; einfach, Vorwissen einzubringen
  - Effiziente Algorithmen für Inferenz, die Graphstruktur ausnutzen
- Viele Methoden des maschinellen Lernens lassen sich in Form von GM darstellen
- Fragestellungen wie MAP Lernen, Bayessche Vorhersage lassen sich als Inferenzprobleme in GM formulieren

Graphische Modelle: Beispiel

- Beispiel: „Alarm“ Szenario
  - Unser Haus in LA hat eine Alarmanlage.
  - Wir sind im Urlaub. Unser Nachbar ruft an, falls er den Alarm hört. Wenn eingebrochen wurde, wollen wir zurück kommen.
  - Leider ist der Nachbar nicht immer zu Hause.
  - Leider geht die Alarmanlage auch bei kleinen Erdbeben los.
- 5 binäre Zufallsvariablen

B Burglary – Einbruch hat stattgefunden

E Earthquake – Erdbeben hat stattgefunden

A Alarm – Alarmanlage geht los

N NeighborCalls – Nachbar ruft an

R RadioReport – Bericht über Erdbeben im Radio

Graphische Modelle: Beispiel

- Zufallsvariablen haben eine gemeinsame Verteilung $p(B,E,A,N,R)$. Wie angeben? Welche Abhängigkeiten gelten?
- Beispiel für Inferenzproblem: Nachbar hat angerufen ($N=1$), wie wahrscheinlich ist Einbruch ($B=1$)?
  - Hängt von verschiedenen Faktoren ab
    - Wie wahrscheinlich Einbruch a priori?
    - Wie wahrscheinlich Erdbeben a priori?
    - Wie wahrscheinlich, dass Alarmanlage auslöst?
    - …

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

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

Graphische Modelle: Beispiel

- Wie modellieren wir $p(B,E,A,N,R)$?
  - 1. Versuch: vollständige Tabelle

2^N

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
...	...	...	...	...	...

+Alle Verteilungen $p(B,E,A,N,R)$ können repräsentiert werden

- Anzahl Parameter exponentiell
- Schwierig anzugeben

  - 2. Versuch: alles unabhängig

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

+Anzahl Parameter linear

- Zu restriktiv, Unabhängigkeitsannahme erlaubt keine sinnvolle Inferenz

Graphische Modelle: Beispiel

- Graphisches Modell: Selektive Unabhängigkeitsannahmen, durch Vorwissen motiviert
- Wähle Variablenordnung: z.B. $B<E<A<N<R$
- Produktregel:

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

Faktoren beschreiben die Verteilung einer Zufallsvariablen in Abhängigkeit anderer Zufallsvariablen.

Können wir diese Faktoren vereinfachen?
Welche dieser Abhängigkeiten bestehen wirklich?

Graphische Modelle: Beispiel

- Zerlegung in Faktoren nach Produktregel:

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

- Annahme bedingter Unabhängigkeiten (Entfernen von Variablen aus Bedingungsteil)

$p(E\,\|\,B) = p(E)$	*Erdbeben hängt nicht von Einbruch ab*
$p(A\,\|\,B,E) = p(A\,\|\,B,E)$	*Alarm hängt von Einbruch und Erdbeben ab*
$p(N\,\|\,B,E,A) = p(N\,\|\,A)$	*Anruf von Nachbar hängt nur von Alarm ab*
$p(R\,\|\,B,E,A,N) = p(R\,\|\,E)$	*Nachricht im Radio hängt nur von Erdbeben ab*

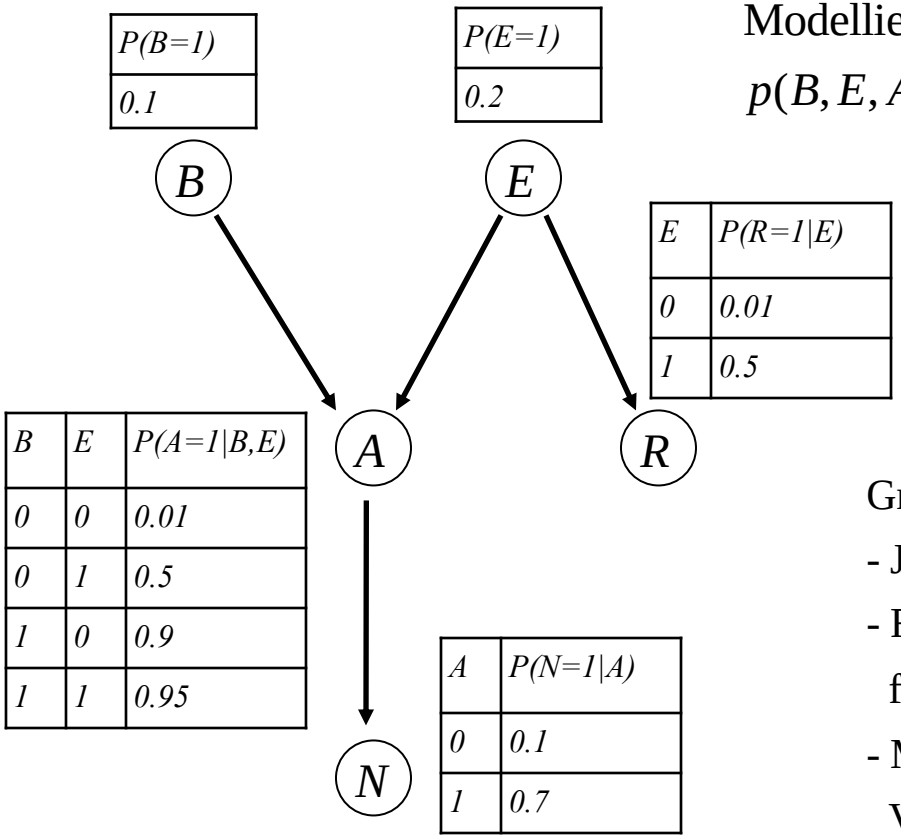
- Vereinfachte Darstellung der gemeinsamen Verteilung:

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

Vereinfachte Faktoren

Graphische Modelle: Beispiel

- Graphisches Modell für „Alarm“ Szenario



Modellierte Verteilung:

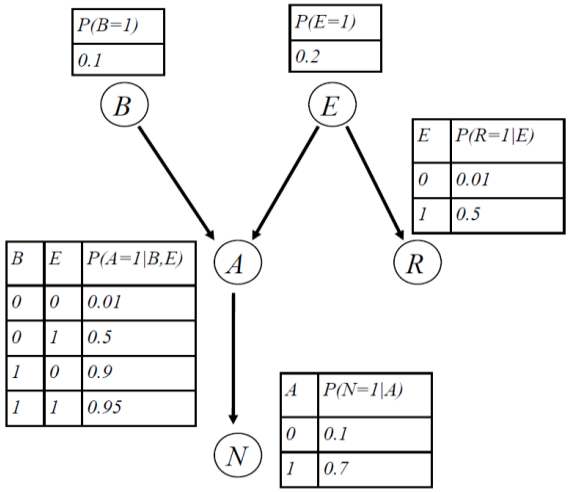
$$p(B,E,A,N,R) = p(B)\,p(E)\,p(A \mid E,B)\,p(N \mid A)\,p(R \mid E)$$

Graphisches Modell:

- Jede ZV ist ein Knoten
- Für jeden Faktor der Form $p(X \mid X_1,...,X_k)$ fügen wir gerichtete Kanten von den X_i zu X ein
- Modell ist parametrisiert mit den bedingten Verteilungen $p(X \mid X_1,...,X_k)$

Graphische Modelle: Beispiel

- Graphisches Modell für „Alarm“ Szenario



  - Anzahl Parameter: $O(N*2^K)$, K= max. Anzahl von Elternknoten
  - Hier 1+1+2+2+4 statt 2^5-1 Parameter
- Gerichtete graphische Modelle heißen auch **Bayessche Netze**

Gerichtete Graphische Modelle: Definition

- Gegeben eine Menge von ZV $\{X_1,...,X_N\}$
- Ein gerichtetes graphisches Modell über den ZV $\{X_1,...,X_N\}$ ist ein gerichteter Graph mit
  - Knotenmenge $X_1,...,X_N$
  - Es gibt keine gerichteten Zyklen $X_{i_1} \to X_{i_2} \to ... \to X_{i_k} \to X_{i_1}$
  - Knoten sind mit parametrisierten bedingten Verteilungen $p(X_i \mid pa(X_i))$ assoziiert, wobei $pa(X_i) = \{X_j \mid X_j \to X_i\}$ die Menge der Elternknoten eines Knoten ist

- Das graphische Modell modelliert eine gemeinsame Verteilung über $X_1,...,X_N$ durch

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

Gerichtete Graphische Modelle: Definition

- Warum muss der Graph azyklisch sein?
  - Satz aus der Graphentheorie:

    G ist azyklisch $\Leftrightarrow$ es gibt Ordnung $\leq_G$ der Knoten, so dass gerichtete Kanten die Ordnung respektieren ($N \rightarrow N' \Rightarrow N \leq_G N'$)

  - Damit ergibt sich

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

Vor X_i in Variablenordnung

    aus Produktregel + bedingten Unabhängigkeitsannahmen (Variablen entsprechend $\leq_G$ umsortieren)

- Gegenbeispiel (kein graphisches Modell):

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

Graphische Modelle: Unabhängigkeit

- Die Graphstruktur eines Graphischen Modells impliziert (bedingte) Unabhängigkeiten zwischen ZV
- Notation: Für Variablen X, Y, Z schreiben wir

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

"X unabhängig von Y gegeben Z"

- Erweiterung auf disjunkte Mengen A, B, C von ZV

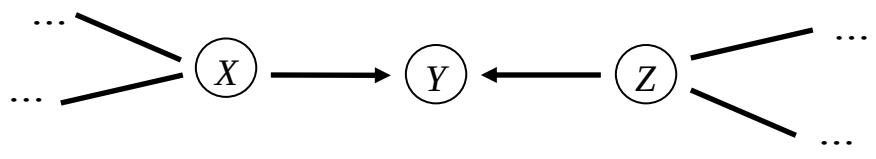
$$\mathrm{A} \perp \mathrm{B} \mid C \Leftrightarrow p(A \mid B, C) = p(A \mid C)$$

Graphische Modelle: Unabhängigkeit

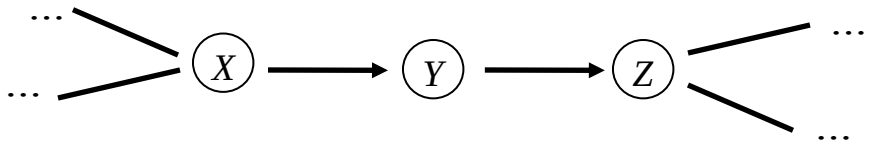
- Welche Unabhängigkeiten der Form $A \perp B \mid C$ werden durch die Graphstruktur modelliert?
  - Im Prinzip auszurechnen durch Summen/Produktregel aus der gemeinsamen Verteilung
  - Bei graphischen Modellen aber direkt aus der Graphstruktur ableitbar→ viel einfacher
  - „D-separation“: Menge einfacher Regeln, nach denen sich alle Unabhängigkeiten ableiten lassen
  - „D“ in „D-separation“ steht für „Directed“, weil sich die Regeln auf Unabhängigkeit in gerichteten graphischen Modellen beziehen (gerichteter Graph).

Graphische Modelle: Unabhängigkeit

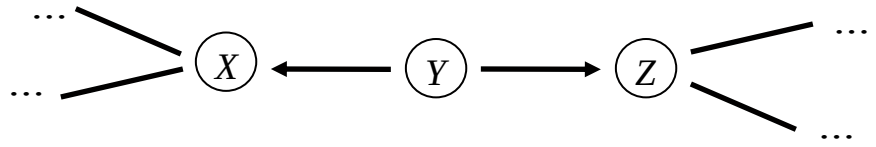
- D-separation: Welche Unabhängigkeiten der Form $A \perp B \mid C$ werden durch die Graphstruktur modelliert?
- Wichtige Rolle spielen Pfade im Graphen, die ZV verbinden
- Notation:



Pfad zwischen X und Z hat eine „**konvergierende**“ Verbindung bei Y („head to head“)

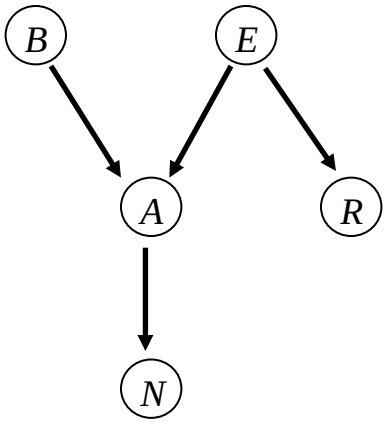


Pfad zwischen X und Z hat eine „**serielle**“ Verbindung bei Y („head to tail“)



Pfad zwischen X und Z hat eine „**divergierende**“ Verbindung bei Y („tail-to-tail“)

Divergierende Verbindungen



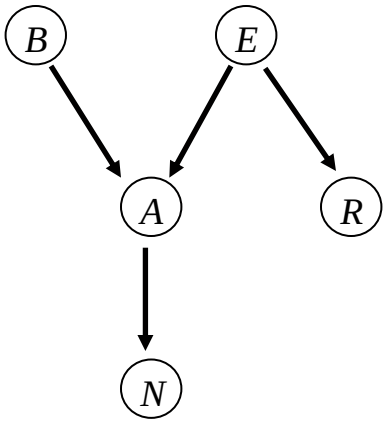
Gemeinsame Verteilung:

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

B=„Einbruch“ *N=„Nachbar ruft an“*
E=„Erdbeben“ *R=„Radio Bericht“*
A=„Alarm“

- Betrachte Pfad $A \leftarrow E \rightarrow R$. Gilt $\mathrm{A} \perp R \,|\, \varnothing$?
  -

Divergierende Verbindungen



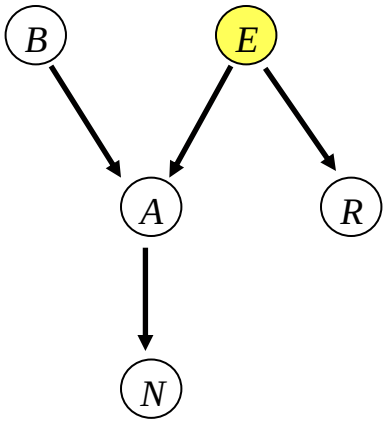
Gemeinsame Verteilung:

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

B=„Einbruch“ *N=„Nachbar ruft an“*
E=„Erdbeben“ *R=„Radio Bericht“*
A=„Alarm“

- Betrachte Pfad $A \leftarrow E \rightarrow R$. Gilt $A \perp R \mid \varnothing$?
  - Nein, $p(A\mid R) \neq p(A)$ [Ausrechnen mit gemeinsamer Verteilung]
  - Intuitiv:
    RadioReport $\Rightarrow$ wahrscheinlich Erdbeben $\Rightarrow$ wahrscheinlich Alarm
    $p(A=1\mid R=1) > p(A=1\mid R=0)$
  - ZV R beeinflusst ZV A über die divergierende Verbindung $R \leftarrow E \rightarrow A$

Divergierende Verbindungen



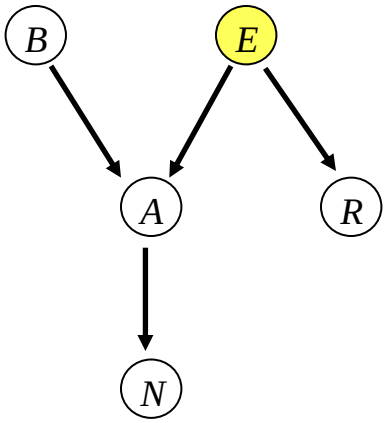
Gemeinsame Verteilung:

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

beobachteter Knoten

- Betrachte Pfad $A \leftarrow E \rightarrow R$. Gilt $A \perp R \,|\, E$?
  -

Divergierende Verbindungen



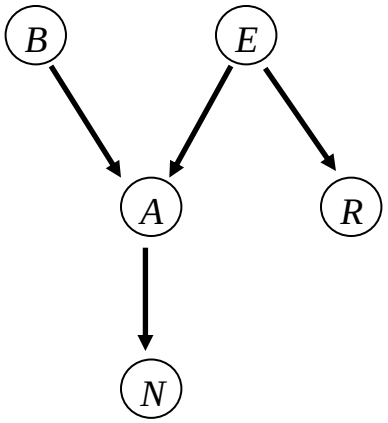
Gemeinsame Verteilung:

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

beobachteter Knoten

- Betrachte Pfad $A \leftarrow E \rightarrow R$. Gilt $A \perp R \,|\, E$?
  - Ja, $p(A\,|\,R,E) = p(A\,|\,E)$ [Ausrechnen mit gemeinsamer Verteilung]
  - Intuitiv:
    Wenn wir schon wissen, dass ein Erdbeben eingetreten ist, wird die Wahrscheinlichkeit für Alarm nicht höher/niedriger durch RadioReport
  - Der divergierende Pfad $A \leftarrow E \rightarrow R$ wird durch Beobachtung von E blockiert

Serielle Verbindungen

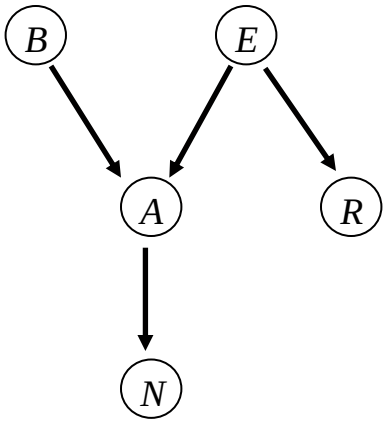


Gemeinsame Verteilung:

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

- Betrachte Pfad $N \leftarrow A \leftarrow B$. Gilt $\mathrm{B} \perp \mathrm{N} \,|\, \varnothing$?
  -

Serielle Verbindungen



Gemeinsame Verteilung:

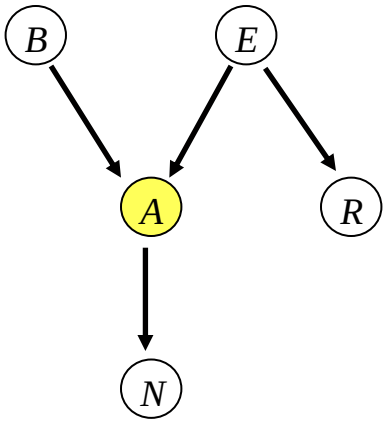
$$p(B,E,A,N,R) =$$
$$p(B)\,p(E)\,p(A\,|\,E,B)\,p(N\,|\,A)\,p(R\,|\,E)$$

- Betrachte Pfad $N \leftarrow A \leftarrow B$. Gilt $\mathrm{B} \perp \mathrm{N} \,|\, \varnothing$?
  - Nein, $p(B\,|\,N) \neq p(B)$ [Ausrechnen mit gemeinsamer Verteilung]
  - Intuitiv:

    NeighborCalls $\Rightarrow$ wahrscheinlich Alarm $\Rightarrow$ wahrscheinlich Burglary

    $p(B=1\,|\,N=1) > p(B=1\,|\,N=0)$
  - ZV N beeinflusst ZV B über den seriellen Pfad $\mathrm{N} \leftarrow A \leftarrow B$

Serielle Verbindungen



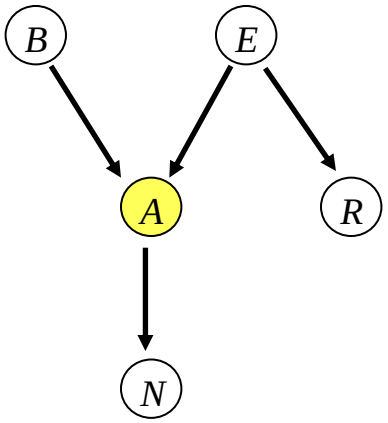
Gemeinsame Verteilung:

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

beobachteter Knoten

- Betrachte Pfad $N \leftarrow A \leftarrow B$. Gilt $\mathrm{B} \perp \mathrm{N}\,|\,A$?
  -

Serielle Verbindungen



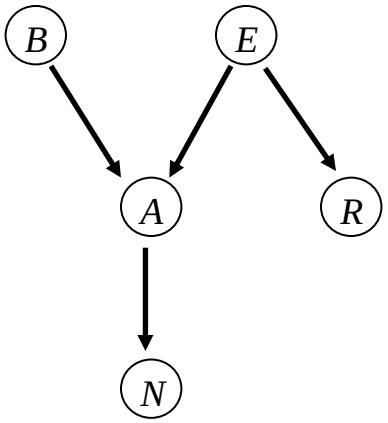
Gemeinsame Verteilung:

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

beobachteter Knoten

- Betrachte Pfad $N \leftarrow A \leftarrow B$. Gilt $\mathrm{B} \perp \mathrm{N}\,|\,A$?
  - Ja, $p(B\,|\,N,A) = p(B\,|\,A)$ [Ausrechnen mit gemeinsamer Verteilung]
  - Intuitiv:
    Wenn wir schon wissen, dass der Alarm ausgelöst wurde, sinkt/steigt die Wahrscheinlichkeit für Einbruch nicht dadurch, dass Nachbar anruft
  - Der serielle Pfad $N \leftarrow A \leftarrow B$ wird durch Beobachtung von A blockiert.

Konvergierende Verbindung

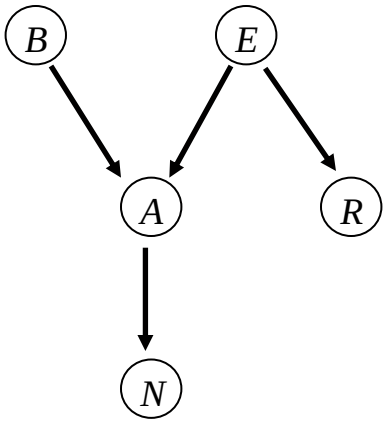


Gemeinsame Verteilung:

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

- Betrachte Pfad $B \rightarrow A \leftarrow E$. Gilt $\mathrm{B} \perp \mathrm{E} \,|\, \varnothing$?
  -

Konvergierende Verbindung

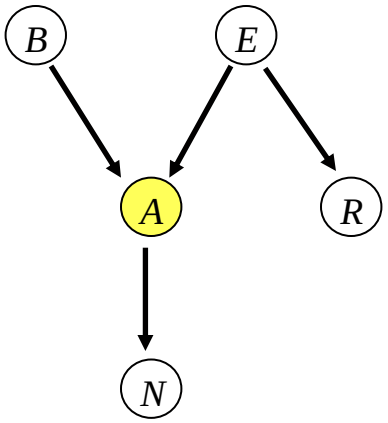


Gemeinsame Verteilung:

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

- Betrachte Pfad $B \rightarrow A \leftarrow E$. Gilt $\mathrm{B} \perp \mathrm{E} \,|\, \varnothing$?
  - Ja, $p(B\,|\,E) = p(B)$ [Ausrechnen mit gemeinsamer Verteilung]
  - Intuitiv:
    Einbrüche treten nicht häufiger/seltener auf an Tagen mit Erdbeben
  - Der konvergierende Pfad $\mathrm{B} \rightarrow A \leftarrow E$ ist blockiert wenn A **nicht** beobachtet ist

Konvergierende Verbindung



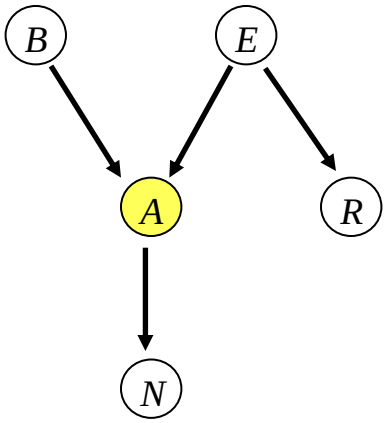
Gemeinsame Verteilung:

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

beobachteter Knoten

- Betrachte Pfad $B \rightarrow A \leftarrow E$. Gilt $\mathrm{B} \perp \mathrm{E}\,|\,A$?
  -

Konvergierende Verbindung



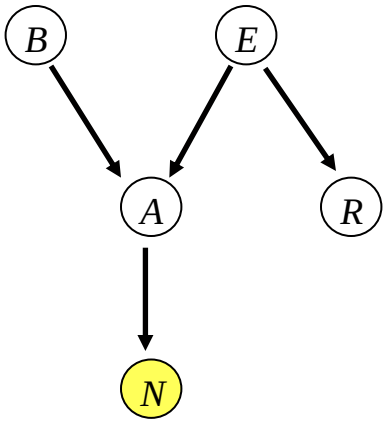
Gemeinsame Verteilung:

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

beobachteter Knoten

- Betrachte Pfad $B \rightarrow A \leftarrow E$. Gilt $\mathrm{B} \perp \mathrm{E}\,|\,A$?
  - Nein, $p(B\,|\,E,A) \neq p(B\,|\,A)$ [Ausrechnen mit gemeinsamer Verteilung]
  - Intuitiv:
    Alarm wurde ausgelöst. Falls wir ein Erdbeben beobachten, erklärt das den Alarm, Wahrscheinlichkeit für Einbruch sinkt ("explaining away").
  - Der konvergierende Pfad $\mathrm{B} \rightarrow A \leftarrow E$ wird **freigegeben** durch Beobachtung von A

Konvergierende Verbindung



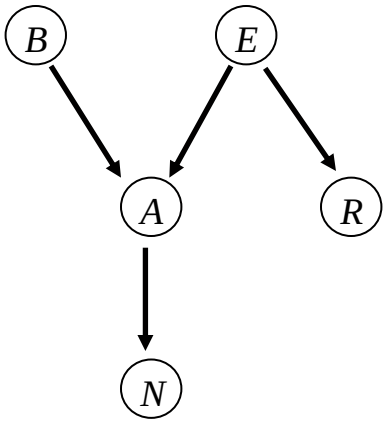
Gemeinsame Verteilung:

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

beobachteter Knoten

- Betrachte Pfad $B \rightarrow A \leftarrow E$. Gilt $\mathrm{B} \perp \mathrm{E}\,|\,N$?
  - Nein, $p(B\,|\,N,A) \neq p(B\,|\,A)$ [Ausrechnen mit gemeinsamer Verteilung]
  - Intuitiv:
    NeighborCalls indirekte Beobachtung von Alarm. Erdbebenbeobachtung erklärt den Alarm, Wahrscheinlichkeit für Einbruch sinkt ("explaining away").
  - Der konvergierende Pfad $\mathrm{B} \rightarrow A \leftarrow E$ wird **freigegeben** durch Beobachtung von N

Zusammenfassung Pfade



Gemeinsame Verteilung:

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

- Zusammenfassung: ein Pfad …-X-Y-Z-... ist
  - Blockiert bei Y, wenn
    - Divergierende Verbindung, und Y beobachtet, oder
    - Serielle Verbindung, und Y beobachtet, oder
    - Konvergierende Verbindung, und weder Y noch einer seiner Nachfahren $Y' \in Descendants(Y)$ beobachtet
    - $Descendants(Y) = \{Y' | \text{es gibt gerichteten Pfad von } Y \text{ zu } Y'\}$
  - Sonst ist der Pfad frei bei Y